Journal of Modern Applied Statistical Methods

Volume 9 | Issue 2 Article 30

11-1-2010

Vol. 9, No. 2 (Full Issue)

JMASM Editors

Journal of Modern Applied Statistical Methods
November, 2010, Vol. 9, No. 2, 332-603

1538 – 9472/10/$95.00

Journal Of Modern Applied Statistical Methods

בס"ד
Journal of Modern Applied Statistical Methods
November, 2010, Vol. 9, No. 2, 332-603

1538 – 9472/10/$95.00

Journal Of Modern Applied Statistical Methods

Journal of Modern Applied Statistical Methods
November, 2010, Vol. 9, No. 2, 332-603

1538 – 9472/10/$95.00

Journal of Modern Applied Statistical Methods
November, 2010, Vol. 9, No. 2, 332-603

1538 – 9472/10/$95.00

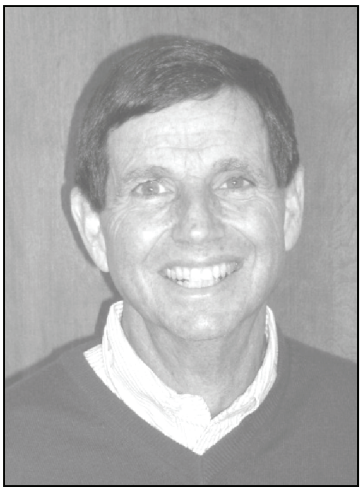
JMASM is an independent print and electronic journal (http://www.jmasm.com/), publishing (1) new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computer-intensive Monte Carlo, bootstrap, jackknife, or resampling methods, (2) the study of nonparametric, robust, permutation, exact, and approximate randomization methods, and (3) applications of computer programming, preferably in Fortran (all other programming environments are welcome), related to statistical algorithms, pseudo-random number generators, simulation techniques, and self-contained executable code to carry out new or interesting statistical methods.

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 332-339

1538 – 9472/10/$95.00

INVITED DEBATE

The Not-So-Quiet Revolution: Cautionary Comments on the Rejection of Hypothesis Testing in Favor of a "Causal" Modeling Alternative



Daniel H. Robinson
University of Texas

Joel R. Levin
University of Arizona

Rodgers (2010) recently applauded a revolution involving the increased use of statistical modeling techniques. It is argued that such use may have a downside, citing empirical evidence in educational psychology that modeling techniques are often applied in cross-sectional, correlational studies to produce unjustified causal conclusions and prescriptive statements.

Key words: Modeling, hypothesis testing, SEM, HLM, causation.

Daniel H. Robinson is Professor of Educational Psychology and editor of *Educational Psychology Review*. His research interests include educational technology innovations that may facilitate learning, team-based approaches to learning, and examining gender trends concerning authoring, reviewing, and editing articles published in various educational journals and societies. Email him at: dan.robinson@mail.utexas.edu.

Joel R. Levin is Professor Emeritus at the University of Arizona and the University of Wisconsin-Madison. He is former editor of the *Journal of Educational Psychology*, and former Chief Editorial Advisor for journal publications of the American Psychological Association. His research interests include the design and statistical analysis of educational research, as well as cognitive-instructional strategies that improve students' learning. Email him at: jrlevin@u.arizona.edu.

Introduction

Over the years, we have found that Joseph Rodgers (e.g., Rodgers, Cleveland, van den Oord, & Rowe, 2000; Rodgers & Nicewander, 1988) has something academically interesting, meaty, and instructive to say. Against that backdrop, Rodgers' most recent essay, provocatively titled "The epistemology of mathematical and statistical modeling: A quiet methodological revolution" (Rodgers, 2010) merits close examination and extensive

commentary. Rodgers appeared to have missed the mark in two critical respects; both reflected in the subtitle "A quiet methodological revolution," because as will become apparent in the following discussion, the revolution is neither quiet nor methodological.

The Null Hypothesis Hullabaloo

Rodgers is correct in stating that serious concerns about null hypothesis significance testing (NHST) have been mounting over the past several decades. Yet, as is well represented in Harlow, Mulaik, & Steiger's (1997) impressive volume, NHST criticisms have hardly been expressed quietly, but rather with full sound and fury. Moreover, in making his case, Rodgers provided a one-sided view of the controversy. Although several sources that indict NHST were cited, short shrift was given to approaches that have defended reasonable and proper applications of statistical hypothesis testing, including, among others, deciding whether a "believed-random" process is truly random (e.g., Abelson, 1997), "intelligent hypothesis testing" (Levin, 1998a), "equivalence testing" (e.g., Serlin & Lapsley, 1993), and hypothesis testing supplemented by effect-size estimation and/or confidence-interval construction (Steiger, 2004).

In addition, numerous authors have defended the use of NHST when mindfully applied (e.g., Frick, 1996; Hagen, 1997; Robinson & Levin, 1997; Wainer & Robinson, 2003). Rodgers cited social-sciences statistical sage Jacob Cohen (1994) as one who dismissed NHST practices in his 1994 seminal article, "The Earth is Round ($p < .05$)." Yet, in the same article, one could easily interpret Cohen's (p. 1001) comment about the "nonexistence of magical alternatives to NHST" as conceding that for whatever "good" NHST does, there are no adequate substitutes.

Rodgers (p. 2) described the fundamental difference between the Fisherian and Neyman-Pearson approaches, with the latter "emphasiz[ing] the importance of the individual decision." However, he characterized NHST as a hybrid and condemned it. Just because a technique is often misused is not a sufficient reason to abandon it. For example, it is argued below that in educational psychology we have observed frequent misapplication of the Rodgers' favored causal modeling techniques. In recognizing that misapplication, however, our goal is not to deter researchers from adopting modeling techniques, but rather to encourage researchers to apply such techniques appropriately and to interpret wisely the results that they pump out. (Back in the Neanderthal age of computers, "grind out" would have been a much more fitting description.)

As researchers who have spent most of our careers conducting randomized experiments, we have sought to apply NHST judiciously, typically adopting or adapting Neyman-Pearson *a priori* Type I, Type II error, effect-size, and sample-size specification principles. Accordingly, we have found that in experiments conducted with rationally (or better, optimally) determined sample sizes - that is, sample sizes associated with enough statistical power to detect nontrivial differences but with not too much power to detect trivial differences (see, for example, Levin, 1998b; and Walster & Cleary, 1970) - NHST provides useful information concerning whether one has an experimental effect worth pursuing. In this context, pursuing means that obtaining a statistically significant effect is followed by a sufficient number of independent replications until the researcher has confidence that the initially observed effect is a statistically reliable one (see, for example, Levin & Robinson, 2003).

In that sense as well, we have regarded NHST primarily as a screening device, similar in function to what Sir Ronald had in mind (e.g., Fisher, 1935). Much of the hullabaloo about NHST is caused by too many researchers focusing on the results of a single study rather than on a series of studies that are part of a program of research (Levin & Robinson, 2000). Fisher was never satisfied with an effect identified in a single study, even if it had a p value of less than 0.05! Instead, he believed that a treatment was only worth writing home about when it had consistently appeared in numerous experiments. As is implied in the following section, whatever purported advantages modeling techniques have over NHST also vanish unless researchers test *a priori* models in multiple experiments.

Rodgers (p. 3) also condemned the

NHST jurisprudence model while aptly referring to Tukey's (1977) "confirmatory data analysis" strategy as being judicial (or quasi-judicial) in nature. Yet, Rodgers mischaracterized Tukey's exploratory data analysis strategy insofar as the detective nature of that hypothesis-generating approach clearly is not jurisprudence. It is this detective role that one emphasizes when using NHST simply as a research-based screening process to determine whether posited effects exist. To us, convincing a jury of one's peers that a prescription for practice should be based on a single research study is rarely, if ever, justified.

Rodgers' (p. 9) assertion that a fundamental problem with NHST is one of testing valueless nil null hypotheses has been advanced by many critics. As researchers who endeavor to use intelligent forms of hypothesis testing with experimental data, we regard the problem of nil nulls not as a statistical issue but as a methodological one. Specifically, it makes little or no conceptual sense to apply NHST when comparing an instructional treatment with a "closet" (Levin, 1994, p. 233) control group (i.e., a condition in which participants sit in a dark room and do nothing), just as it is inane to compute *p*-values for reliability correlations (see, for example, Thompson, 1996). Educational psychology is filled with such examples of comparing new innovations with ridiculous straw-person control conditions that no sane researcher would ever consider using. A more appropriate formulation of a nil null is when an investigator wishes to compare a newly developed and previously untested experimental treatment with the best treatment that is currently available.

According to Rodgers, "the [1999 task force assembled by the American Psychological Association] concluded that NHST was broken in [a] certain respect" (p. 3). Task-force member Wainer and the present first author (Wainer & Robinson, 2003) provided a different view of the task force's brief consideration of the recommendation to issue an outright ban on NHST. As we have argued previously (e.g., Levin & Robinson, 1999) and in our preceding discussion, adopting such an extreme stance would be akin to calling for a ban on hammers because hammerers were hammering their fingers instead of nails (for additional discussion, see Levin & Robinson, 2003). Even the outspoken NHST critic Rozeboom (1997) acknowledged via another "tools" analogy that "the sharpest of scalpels can only create a mess if misdirected by the hand that drives it," (p. 335). Fortunately, in the case of the most recent (6th) edition of the *APA Publication Manual* (American Psychological Association, 2010), the hypothesis-testing baby was not thrown out with the bath water.

"Causal" Modeling Techniques

Contemporary modeling techniques, including structural equations modeling (SEM) and hierarchical linear modeling (HLM), among others, which emerge from a theoretical/conceptual framework, are statistical/data-analytic and not methodological in nature. So, whence Rodgers' "methodological" revolution? Even he noted on p. 8 that "SEM has been built into a powerful analytic method and is a prototype of the first approach [a model-comparison framework] to postrevolutionary modeling" (p. 8).

That a statistical modeling tail often wags the methodological dog may have contributed to what we consider a major misuse of causal modeling: researchers attempting to squeeze causality out of observational or correlational data. Because of the unfortunate "causal" nomenclature, we fear that many researchers may be deluded into believing that the statistical control that such techniques provide for correlational (non-experimental) data is on a par with the genuine experimental control of randomized experiments (Levin & O'Donnell, 2000, p. 211). This in turn results in causal-model appliers issuing causal conclusions that they mistakenly believe are scientifically valid. As Cliff (1983) previously noted, "Literal acceptance of the results of fitting 'causal' models to correlational data can lead to conclusions that are of questionable value" (p. 115).

In addition, because causal-model researchers' conclusions typically flow from revised data-driven models rather than from *a priori* theory-based model specifications, in the absence of independent validations those causal conclusions present even more cause for

concern. As with our previous hammers vs. hammerers distinction, Rodgers is well aware of researchers' potential shoddy application of causal modeling techniques. Yet, he could have sent a stronger cautionary message to the relatively uninitiated model builder than his innocuous pronouncement that "the success of SEM depends on the extent to which it is applied in many research settings" (p. 8).

To illustrate what we mean by prescriptive statements appearing in articles that include statistical modeling techniques, we offer very recent examples that appeared in a reputable educational psychology research journal. To avoid redundancy, we offer only two such unjustified causal excerpts here, from numerous ones that we have encountered in multiple teaching-and-learning research journals that we have recently read or reviewed (see Robinson, Levin, Thomas, Pituch, & Vaughn, 2007, and the following section).

Ciani, Middleton, Summers and Sheldon (2010)'s Study

The following summary appeared in Ciani et al.'s study abstract:

> Multilevel modeling was used to test student perceptions of three contextual buffers: classroom community, teacher's autonomy support, and a mastery classroom goal structure…Results provide practitioners with tools for counteracting potential negative implications of emphasizing performance in the classroom. (p. 88)

There was one predictor variable; one outcome variable, a three-item scale that measured students' motivation to learn; and three moderator variables, a three-item scale that measured student perceptions of classroom community, a four-item scale that measured student perceptions of instructor autonomy support, and a three-item scale that measured student perceptions of the extent to which their teacher emphasizes developing competence in the classroom. All measures were collected at a single point in time and HLM was used to analyze the data. Here are a couple causal conclusions from the discussion section:

> However, it appears that comparing students' achievement publicly, or using the work of the highest achieving students as an example for everyone, may not be so pernicious a practice when students in the classroom perceive a sense of community among their fellow classmates.
>
> [O]ur findings demonstrate that if students feel respected by the teacher, such that their preferences and ways of doing things are acknowledged and accommodated as much as possible, then a strong performance orientation on the part of the teacher is not harmful. Autonomy support enables students to internalize what they are doing, so that they view their activity as important even if it is not enjoyable, or if it creates stress and pressure. Thus, it appears that emphasizing competition between students is not necessarily undermining of student mastery goals, if the teacher can communicate and promote the performance structure in a non-controlling way. These findings are reassuring, showing that performance orientations are not necessarily corrosive – certainly an important message, given the performance necessities that all students face. (p. 95)

As with most of these articles based on correlational data and yet that offer prescriptive recommendations, certain limitations of the research are explicitly acknowledged by the authors:

> The most significant limitation to the current study is that all data reported are correlational.
>
> Gathering data at one point in time also creates a limitation regarding the causal relationships among the variables in this study. (p. 96)

These limitations aside (or ignored?), the authors proceeded to offer the following prescriptive:

> Our findings, along with other goal theorists (e.g., Urdan & Midgley, 2003), suggest that given current prevailing attitudes and policy it may be more fruitful to emphasize adaptive instructional practices in the classroom, as opposed to trying to reduce maladaptive practices. (p. 97)

Thus, the authors made recommendations for practice ("prescriptive statements") in the absence of convincing evidence that such practices are clearly causally related to student outcomes.

Chen, Wu, Kee, Lin & Shui's (2009) Study

Chen et al. used SEM to analyze relations among fear of failure, achievement goals, and self-handicapping. Causal relations among the variables are implied in the Discussion section:

> This finding shows fear of failure as a distal determinant of self-handicapping and achievement goals (MAv and PAv) as proximal determinants of self-handicapping, demonstrating the motivational process of self-handicapping. (p. 302)

The authors revealed the perceived magical quality of SEM allowing researchers to coax causality from correlational data:

Since SEM analysis examines many variables' relationships simultaneously, we rely on its results as the basis for our conclusions and discussion. (p. 303)

The Limitations section is predictable:

> Although we used the SEM approach to estimate the proposed model, the data in the study are cross-sectional in nature and causal relations cannot be drawn. The longitudinal approach is preferred in order to ascertain the causal pattern and to further clarify the chronic effects of mastery-avoidance and performance-approach goals on achievement-related outcomes. (p. 304)

In contrast, what follows are the grand prescriptives that appeared in the Implications and Conclusions:

> We believe that the integrative model can help educators develop effective interventions to reduce students' self-handicapping, especially since we found that the mid-level achievement goals (MAv and PAv) mediate the relationships between fear of failure and self-handicapping… it is suggested that teachers use multiple indices to offer more opportunities for students to attain success. In addition, teachers should encourage students to embrace a multiple goals perspective in which doing one's best and outperforming others are not in conflict with each other. (p. 304)

Rodgers (2010, p. 8) previously proffered caveat aside, in both of the just-presented examples, cross-sectional (one time point), correlational (no variables were manipulated) data were tossed into a statistical modeling analysis and what popped out were causal conclusions.

Correlational Data and Causal Conclusions

Over the past few years, we have examined empirical articles published in widely read teaching- and-learning research journals and have found that:

1. In one journal survey (Hsieh et al., 2005), the proportion of articles based on intervention and experimental (random assignment) methodology had decreased from 47% in 1983 to 23% in 2004.

2. In another journal survey (Robinson et al., 2007), the proportion of articles based on intervention methods had decreased from 45% in 1994 to 33% in 2004. Meanwhile, the proportion of nonintervention articles that contained prescriptive statements increased from 34% in 1994 to 43% in 2004. The proportion of nonintervention (non-experimental and correlational) articles that included prescriptive statements (in the form of causally implied implications for

educational practice) increased from 33% in 1994 to 45% in 2004.

3. In a follow-up to the just-described Robinson et al. (2007) survey (Shaw, Walls, Dacy, Levin & Robinson, 2010), although only 19 nonintervention studies in 1994 included prescriptive statements, these statements were repeated in 30 subsequent articles that had cited the original 19.

For the present article, we examined the first two issues of the 1999 volume of the APA-published journal, the *Journal of Educational Psychology*, and again for the 2009 volume. We looked specifically at the comparative proportions of articles based on correlational methods and those that involved interventions (either randomized experimental or nonrandomized but researcher manipulated), as well as the proportion of correlational methods articles in which prescriptive statements were offered. The results are summarized in Table 1.

Although roughly half of the articles appearing in only one of the five journals that were part of Robinson et al.'s (2007) study were surveyed, the findings support the reported trends. Intervention studies (both randomized and nonrandomized) are becoming increasingly rare and instead researchers are basing their recommendations for practice on weaker evidence. Moreover, it appears that statistical and nonrandomized) are becoming increasingly rare and instead researchers are nonrandomized) modeling techniques are becoming more popular - having increased from only 3% of the correlational research articles in 1999 to 40% in 2009 - which may in turn contribute to the concomitant 10-year increase in prescriptive statements appearing in such articles.

Thus, we have witnessed widespread application of SEM, HLM, and other sophisticated statistical procedures in correlational data contexts, where causality is sought but the critical conditions needed to attribute causality are missing (e.g., Marley & Levin, 2011; Robinson, 2010). Rodgers states that "researchers who are scientists…should be focusing on building a model…embedded within well-developed theory" (p. 4-5). Here we agree with former Institute for Educational Science Director Grover Whitehurst who argued that - at least in the field of education - we have enough theory development studies and need more studies that address practical "what works" questions.

It is our fear that a research approach where the question, "Does the data fit my model?" is far more dangerous than the question, "Is there anything here worth pursuing?" As we have seen, an affirmative answer to the former question seems to entitle a researcher to form a model that indicates a causal relationship between, say, students' self-efficacy and their achievement. The researcher then develops a self-efficacy scale that measures

Table 1: Summary of Selected Results of Surveyed Articles Appearing in the *Journal of Educational Psychology* (1999 and 2009) Based on Either Correlational or Intervention Methods

	1999		2009	
	Type of Study		Type of Study	
	Correlational	Intervention	Correlational	Intervention
Number of Articles	18 (60%)	12 (40%)	23 (66%)	12 (34%)
Prescriptive Statements	9 (50%)	------[a]	13 (57%)	------[a]
Statistical Modeling	1 (3%)	0 (0%)	14 (40%)	2 (6%)
Prescriptive Statements	1	------[a]	7 (50%)	------[a]

Note: This table includes preliminary data from a larger study recently completed by Reinhart, Haring, Levin, Patall, and Robinson (2011). [a] Not assessed in the present survey

students' self-perceptions and also measures achievement. The data may fit the model but in the absence of convincing longitudinal data, ruling out alternative explanations, and independent replications based on the previous nice-fitting model, this practice may lead to dangerous causal conclusions. For the just-presented self-efficacy example, it is just as likely that high achievers feel better about their effectiveness as learners rather than the other way around. Apparently, many researchers believe that it is entirely appropriate to apply such modeling techniques and to interpret the results as support for prescriptive statements founded on causality.

Conclusions About Revolutions

To summarize, Rodgers (2010) has written a cogent essay on the vices of statistical hypothesis testing and the virtues of statistical modeling. We believe, however, that his essay painted a somewhat distorted (and potentially misleading) portrait about those statistical "arts." In particular, we take issue with two aspects of Rodgers' so-called "quiet methodological revolution." For one aspect (rejecting statistical hypothesis testing), we argue that the picture is neither as bleak nor as open and shut as Rodgers portrayed. As supporting evidence, witness the sustained presence of hypothesis testing, along with its more intelligent additions and adaptations, in various academic-research disciplines - including the research-and-publication "bible" of both our very own field of psychology and virtually all social-sciences domains, the most recent edition of the APA *Publication Manual* (American Psychological Association, 2010).

For the other aspect of Rodgers' essay that merits critical commentary (accepting modeling techniques), we argue that causal modeling and other related multivariate and multilevel data-analysis tools frequently cause their users to think - in accord with Rodgers' seductive subtitle - that the procedures are methodological randomization-compensating panaceas rather than techniques that do the best they can to provide some degree of statistical control in a "multiply confounded variable" world. The unfortunate consequence of that methodological understanding, then, is that when combined with researcher misapplication of such modern modeling artillery, instead of being on target with their data analyses and research conclusions, weapons are backfiring and researchers are ending up (whether knowingly or not) with a considerable amount of egg on their faces.

References

Abelson, R. P. (1997). The surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science*, *8*, 12-15.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (*6th Ed.*). Washington, DC: American Psychological Association.

Chen, L. H., Wu, C.-H., Kee, Y. H., Lin, M.-S., & Shui, S.-H. (2009). Fear of failure, 2 x 2 achievement goal and self-handicapping: An examination of the hierarchical model of achievement motivation in physical education. *Contemporary Educational Psychology*, *34*, 298-305.

Ciani, K. D., Middleton, M. J., Summers, J. J., & Sheldon, K. M. (2010). Buffering against performance classroom goal structures: The importance of autonomy support and classroom community. *Contemporary Educational Psychology*, *35*, 88-99.

Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research*, *18*, 115-126.

Cohen, J. (1994). The earth is round ($p <$ .05). *American Psychologist*, *49*, 997-1003.

Fisher, R. A. (1935). *The design of experiments*. (Reprinted in 1960). Edinburgh: Oliver & Boyd.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*, 379-390.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test, *American Psychologist*, *52*, 15-24.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests*? Mahwah, NJ: Erlbaum.

Hsieh, P., Hsieh, Y. P., Chung, W. H., Acee, T., Thomas, G. D., Kim, H. J., You, J., Levin, J. R., & Robinson, D. H. (2005). Is educational intervention research on the decline? *Journal of Educational Psychology*, *97*, 523-529.

Levin, J. R. (1994). Crafting educational intervention research that's both credible and creditable. *Educational Psychology Review*, *6*, 231-243.

Levin, J. R. (1998a). To test or not to test H_0? *Educational and Psychological Measurement*, *58*, 313-333.

Levin, J. R. (1998b). What if there were no more bickering about statistical significance tests? *Research in the Schools*, *5*(*2*), 43-53.

Levin, J. R., & O'Donnell, A. M. (1999). What to do about educational research's credibility gaps? *Issues in Education: Contributions from Educational Psychology*, *5*, 177-229.

Levin, J. R., & Robinson, D. H. (1999). Further reflections on hypothesis testing and editorial policy for primary research journals. *Educational Psychology Review*, *11*, 143-155.

Levin, J. R., & Robinson, D. H. (2000). Statistical hypothesis testing, effect-size estimation, and the conclusion coherence of primary research studies. *Educational Researcher*, *29*(1), 34-36.

Levin, J. R., & Robinson, D. H. (2003). The trouble with interpreting statistically nonsignificant effect sizes in single-study investigations. *Journal of Modern Applied Statistical Methods*, *2*, 231-236.

Marley, S. C., & Levin, J. R. (2011). When are prescriptive statements in educational research justified? *Educational Psychology Review*, *23*, 197-206.

Reinhart, A. L., Haring, S. H., Levin, J. R., Patall, E. A., & Robinson, D. H. (2011). *Models of not-so-good behavior: Yet another way to squeeze causality and recommendations for practice out of correlational data*. Unpublished manuscript, University of Texas, Austin.

Robinson, D. H. (2010, May). *Correlational, causal, and prescriptive claims: Guidelines for articles appearing in* Educational Psychology Review. Paper presented at the annual meeting of the American Educational Research Association, Denver.

Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher*, *26*, 21-26.

Robinson, D. H., Levin, J. R., Thomas, G. D., Pituch, K. A., & Vaughn, S. R. (2007). The incidence of "causal" statements in teaching and learning research journals. *American Educational Research Journal*, *44*, 400-413.

Rodgers, J. L., Cleveland, H. H., van den Oord, E., & Rowe, D. C. (2000). Resolving the debate over birth order, family size, and intelligence. *American Psychologist*, *55*, 599-612.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, *65*, 1-12.

Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *American Statistician*, *42*, 59-66.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests*, 335-391. Mahwah, NJ: Erlbaum.

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Statistical issues*, 199-228. Hillsdale, NJ: Lawrence Erlbaum.

Shaw, S. M., Walls, S. M., Dacy, B. S., Levin, J. R., & Robinson, D. H. (2010). A follow-up note on prescriptive statements in nonintervention research studies. *Journal of Educational Psychology*, *102*, 982-988.

Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, *9*, 164-182.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*(*2*), 26-30.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

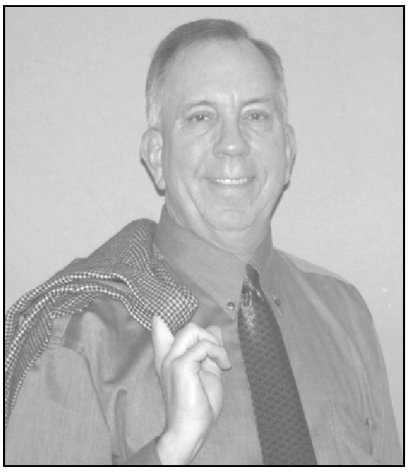
Wainer, H., & Robinson, D. H. (2003). Shaping up the practice of null hypothesis significance testing. *Educational Researcher*, *32*(*7*), 23-31.

Walster, G. W., & Cleary, T. A. (1970). Statistical significance as a decision-making rule. In E. F. Borgatta & G. W. Bohrnstedt (Eds.), *Sociological methodology*, 246-254. San Francisco: Jossey-Bass.

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 340-347

1538 – 9472/10/$95.00

Statistical and Mathematical Modeling versus NHST? There's No Competition!



Joseph Lee Rodgers
University of Oklahoma

Some of Robinson & Levin's critique of Rodgers (2010) is cogent, helpful, and insightful – although limiting. Recent methodology has advanced through the development of structural equation modeling, multi-level modeling, missing data methods, hierarchical linear modeling, categorical data analysis, as well as the development of many dedicated and specific behavioral models. These methodological approaches are based on a revised epistemological system, and have emerged naturally, without the need for task forces, or even much self-conscious discussion. The original goal was neither to develop nor promote a modeling revolution. That has occurred; I documented its development and its status. Two organizing principles are presented that show how both perspectives can be reconciled and accommodated. A program of research that could not have occurred within the standard NHST epistemology, without a modeling perspective, is discussed. An historical and cross-disciplinary analogy suggests their view is similar to Galileo's world view, whereas some branches of social and behavioral science may be ready for something closer to a Newtonian perspective.

Key words: NHST (or Null Hypothesis Significance Testing), Modeling, Mathematical ModelsD.

Joe Rodgers is a Quantitative Psychologist in the University of Oklahoma Department of Psychology, where he is a George Lynn Cross Research Professor and the Robert Glenn Rapp Foundation Presidential Professor. He has held visiting teaching/research positions at Ohio State, University of Hawaii, UNC, Duke, the University of Southern Denmark and the University of Pennsylvania. He and his research team have been continuously funded by NIH since 1987 to develop mathematical models of adolescent development, young adult fertility, and family/friendship interactions. His methodological interests include mathematical modeling, resampling theory, linear statistical models, quasi-experimental design methods, EDA and multidimensional scaling. He has been editor of the applied methods journal *Multivariate Behavioral Research* from 2006 to 2011, and is a past-president of the Society of Multivariate Experimental Psychology, the Society for the Study of Social Biology and the APA's Population and Environmental Psychology Division. Email: jrodgers@ou.edu.

Introduction

Null Hypothesis Significance Testing (NHST) has, for many years, been the primary organizational and epistemological system by which we understand statistical practice in behavioral sciences. NHST has been frequently criticized, and in the late 1990s the criticism was sufficient to create substantial contention, explicit calls for NHST to be outlawed or abandoned, the appointment of an American Psychological Association task force to judge its status and to evaluate proper statistical practice, and a great deal of discussion and argumentation, both informally and in published articles. Before and during this same period, a different epistemological system, what I referred to in Rodgers (2010) as a modeling revolution was in development. With little discussion (and most of what would have naturally occurred has been largely drowned out by the clamor over NHST), mathematical and statistical modeling have become the set of organizing principles that has the potential to completely replace NHST as the primary epistemological system. And modeling should replace NHST, for several reasons.

The first is because it is a more natural way for researchers to frame, think about, and conduct research, whereas NHST was a creation of and for statisticians. Second, modeling has more flexibility to support the maturation of both statistical and methodological practice within psychology and other behavioral sciences. Third, modeling includes NHST as a special case, and so NHST has not been replaced or even very much revised as a set of procedure.

Robinson and Levin presented position statements that, in my career, I have taught to my students, and have applied in my research. These principles emerged from a strong and coherent philosophical background, including caution against over-interpretation of correlations, which emerged from John Stuart Mills' (1843) inductive canons of scientific inquiry. Another principle is to use randomization if possible, which emerged from Fisher's (1935) answer to the problem that Mill left open -- how can researchers equate groups, on average, before a manipulation? Yet another is to emphasize the importance of replication; this underappreciated practice serves the purpose of correcting the bad luck that can befall a researcher in "gaming with the devil" (see Box, 1978, p. 144), and is another of Fisher's edicts that helped create the philosophical basis of social/behavioral science methodology.

I could (almost) leave this reply hanging, and emphasize how correct and well-founded are many of the positions stated in their critique. If so, though, I would necessarily conclude with some comments about how none of these principles has any import in evaluating either the status of NHST, or the development of statistical/mathematical modeling, or as criticism of my article, because these principles stand firm in relation to either NHST or statistical/ mathematical modeling. However, if I left my reply here, that would obfuscate my initial intent, which I believe has been mischaracterized.

Two basic principles (and some potential quibbles with the language, to follow) are paramount, and within those principles their criticisms and my position statements will be simultaneously accommodated. The first principle is that NHST is the type of statistical paradigm that naturally applies to a rather immature science, whereas statistical modeling naturally fits a more mature, or at least maturing, science. The second principle is that NHST is subsumed within the modeling perspective. The two paradigms need not compete, as Robinson and Levin implied. Accept the modeling perspective, and it can be sharpened to the special case of the NHST perspective at any time; insist that NHST is the one, only, and proper epistemological position, and the full range and power of structural equation modeling (SEM), multi-level modeling (MLM), and dozens of specialty models are relegated to virtual impotence.

Statistical Modeling Reflects and Supports the Maturation of Social/Behavioral Sciences

The development of statistical and mathematical modeling as an epistemological system didn't occur through high-level mandate or management; it has been a natural and emergent methodological feature of the maturing of psychology (and has parallels within education, economics, sociology, and other social/behavioral sciences). In this sense, it is a

mischaracterization to claim that I "condemn" NHST or that I "perceive vices of statistical hypothesis testing." Most of my article was not prescriptive, despite their suggestions to the contrary; the part that is prescriptive has little to do with liking or condemning NHST. Rather, I described a developmental process that is well advanced, though relatively unexamined in historical perspective. As science has advanced, stronger statements are possible, ones that even in some cases move toward legitimate causal attribution. Nowhere in that previous sentence is there encouragement to assert unjustified causality. Further, to suggest that such unjustified claims occur – even to illustrate with specific examples – does no damage to the position that our science is maturing in that direction. Nor is science necessarily advanced by successful causal claims; sometimes, rather, it advances by identifying past mis-attributions, a process which Robinson and Levin support and appreciate. Ironically, though, certain versions of that process would not likely emerge from an NHST perspective. I described an example from my own research program.

For many years, nearly an entire community of research psychologists has ignored a certain type of selection bias, resulting in the kind of mis-attributed causal process that Robinson & Levin (and I) decry. Scarr and McCartney (1983) made a stark statement concerning this design flaw, which is inherent in literally hundreds (perhaps thousands) of previous published papers: "passive genotype-environment effects arise in biologically related families and render all of the research literature on parent-child socialization uninterpretable" (p. 427). Using a quasi-experimental design that takes advantage of siblings to partially control for selection bias, along with a powerful sibling dataset, my colleagues and I have published a series of articles during the past decade that have separated and quantified the difference between certain types of inherent selection bias and the remaining correlational links, within which the causal attributions are logically expected to exist.

I review several of these studies based on the sibling design and on the children-of-siblings design. (Besides these, other quasi-experimental design innovations exist that also can also be used to separate family-based selection bias from parental and family influence; see D'Onofrio, 2003, for description of the children-of-twins design and Rodgers, et al, 2008, for description of the mother-daughter-aunt-niece design). Rodgers, et al. (2000) showed how selection bias has improperly influenced the interpretation of birth order-intelligence links; at least most (perhaps all) of what has appeared to be birth order effects on intelligence in past research has actually been between-family differences in parental education and IQ, among others (see Rodgers, 2001 for further explanation of this logic, Wichman, et al., 2006, for a modeling demonstration of this phenomenon, and Wichman, et al., 2007, for further elaboration). D'Onofrio, et al. (2008) showed how the link between smoking during pregnancy and child conduct problems is at least partially caused by the kind of women who smoke during pregnancy, thus challenging much of the direct causal attribution.

D'Onofrio, et al. (2009) used a similar design to investigate the relationship between family income and child conduct problems, with similar conclusions. Mendle, et al. (2009) applied this type of sibling control to study the link between father absence and age at first intercourse, and found that much of the apparent direct link between father absence and age at first intercourse has likely been caused by shared genetic factors in the background. Harden, et al. (2009) studied whether population density has a direct influence on antisocial behavior during adolescence, or whether the apparent link is due to selection bias; the latter was more strongly supported. Finally, Jaffee, et al. (2011) showed how placement of infants and young children in day care as an influence on both achievement and behavioral problem scores in childhood is almost completely attributable to the type of women who put their children in day care, leaving very little remaining variance to attribute to the direct influence of the day care experience in and of itself on these child outcomes.

For the purposes of this reply, these findings make a strong statement about both modeling and NHST. Each result above depended on strong design logic combined with a statistical modeling exercise. Further, each study contained within it a number of NHST

results, but the *organizational principles* emerged from a research perspective that required longitudinal and within-family data, strong research designs, powerful measurement tools, and sophisticated statistical models. They would not have likely emerged from an NHST epistemology. Nor are the conclusions that emerge from this type of work necessarily causal; indeed, most of the conclusions above challenge previous causal attributions.

In the tradition of Cook and Campbell (1979) and Shadish, et al. (2002), the researchers' goal, whether in quasi-experimental or experiment research, is to address as many threats as possible to internal validity, the validity of causal attribution, and to admit freely and to self-evaluate in the face of those that remain. Robinson and Levin admitted to this maturational challenge: "Our field of educational psychology is filled with such examples of comparing new innovations with ridiculous strawperson control conditions that no sane researcher would ever consider using." So are psychology, sociology, etc., of course. And so the proper and defensible approach is exactly where they stated it should be, using an appropriate set of methodological tools to draw cautious but legitimate conclusions, and to avoid wasting time asking superficial and uninteresting questions. Hopefully, those methodological tools expand to accommodate improvements, maturation, in the science that they support. Statistical modeling is an example of such expansion.

Statistical Modeling Subsumes NHST

There exists a way to view both NHST and statistical modeling that accommodates both Rodgers (2010) and Robinson and Levin's critique. That accommodation was stated in my original article, but here I shall present this argument in different words. Robinson and Levin presumed I was prescriptively criticizing NHST; that I favor modeling and oppose NHST: "Rodgers (2010) has written a cogent essay on what he perceives as the vices of statistical hypothesis testing and the virtues of statistical modeling." First, my article was intended to be more of an historical account than a desideratum about what should be. Second, I was a strong opponent of outlawing, abandoning, or otherwise providing any type of institutional control over NHST (or any other methodology). I have used the NHST paradigm often, in most of my published research. I have also used modeling approaches, when they appeared to be useful and appropriate.

Many of my publications incorporated both, which leads to my third comment: I do become prescriptive when I describe in detail how the statistical modeling strategy subsumes NHST, because I'm convinced of the value of *both* approaches. Hence, the crux of my reply: NHST is a proper paradigm, but it is a special case of a broader and thus more flexible paradigm. I do not agree there are two competing approaches. One is broader and one is a special case. The modeling approach uses NHST as a fundamental part of the modeling framework. As Rodgers (2010) explained:

> As the two models ... are evaluated, no chance-level null hypothesis is posited, nor is an alternative constructed, at least not in the sense that those concepts are usually treated. However, traditional statistical concepts are used in this comparison, such as a test statistic (e.g., Chi-square values), a sampling distribution (the theoretical chi-square), and an alpha level (to tune the trade-off between fit and parsimony). Further, the NHST perspective is embedded within this statistical evaluation in the sense that there is a null hypothesis built into the model comparison (i.e., whether the population parameters ... are equal to one another). (p. 7)

NHST is a tool, as a way to answer a certain question. I've never understood why researchers would be satisfied with the conclusion to reject H_o or fail to reject H_o, unless the research question was simple enough to warrant such a conclusion. It seems to me that when the research questions become more complex, modeling has the potential to provide more complex answers, and to move scientific epistemology forward substantially further than what can be obtained via NHST.

Minor Issues

There are some mischaracterizations in their critique that require a response (though the majority are accounted for by the two principles in the previous sections). They suggested that "he then goes on to discuss NHST as a hybrid and condemns it;" I did not, though I cited Gigerenzer (1993), who did. They implied that I supported a ban on NHST, when I actually opposed such a ban. They claimed that "Rodgers also condemns the NHST 'jurisprudence model," whereas in fact I teach and promote this way of thinking of NHST. They suggested that "Rodgers mischaracterizes Tukey's 'exploratory data analysis' strategy insofar as the detective nature of that hypothesis-generating approach clearly is not jurisprudence," but I did not link the detective and jurisprudence components – after describing the role of jurisprudence within research, I stated "The researcher is *also* a detective" (p. 3, italics added for emphasis). They failed to make the connection that my section titled "Criticism and Adjustment of NHST" was historical; their first sentence in their section "The Null Hypothesis Hullabaloo" recognized historical goals, but the remainder of that section was not about the "hullabaloo," but rather about their perception that I promulgated it, although I did not. Finally, they suggested that "he gives short shrift to approaches that have defended reasonable and proper applications of statistical hypothesis testing," and cited four articles that would have provided more balance. In fact, I discussed three of those articles.

NHST is a worthy, valuable, and useful tool. It helps researchers to answer a certain question, framed in a certain way. However, its weaknesses are well-known, and often discussed (see Wainer, 1999, for a balanced and interesting account, among dozens of others). Further, as the field of behavioral science matures, it should not stand as the epistemological basis of research methodology within the field of psychological science, because modeling is more useful, flexible, and better supports the future of behavioral science research.

This methodological practice should not be banned or outlawed for two reasons. First, such practice should not be managed at the institutional level (any more than the workers' union should decide to ban hammers or electric saws). Second, NHST has served its value in thousands of scientific settings. It has also been misused, and Robinson and Levin provided support for its proper and legitimate use, in this and other published articles.

Regarding "the 'revolution' about which Rodgers writes is neither quiet nor methodological," they were correct, as I originally asserted. The NHST hullabaloo was anything but quiet. But the modeling revolution was so quiet that apparently many didn't notice, and now aren't sure that it occurred. Robinson and Levin contend that the revolution was not methodological, that the issues are entirely statistical. SEM contains both a structural and a measurement model. Multi-level modeling accounts for clustering, which is often caused by sampling processes. Multilevel modeling also cannot be separated from the design issues that generated the different levels. Analytic procedures that handle missing data require specification of the generating processes – sampling, measurement, etc. – that produced the missing values. In other words, modern statistical models account for design, sampling, and measurement, as well as the formal statistical properties of statistical models. As one example, MacCallum and Tucker (1991) could not have developed their conceptualization separating sampling and model error if they had used an NHST epistemology.

It is perhaps not surprising that those whose way of thinking about the advancement of behavioral science is embedded in the NHST tradition would not recognize the modeling revolution as bringing about the expansion of statistical practice to include many other features of the methodological arena. But such broadening is one among many features of statistical/mathematical modeling that make the use of SEM, MLM, missing data approaches, and other modeling methods exciting and useful. To expand their analogy, there are new dangers created in using models, and their misuse cannot be supported (Cliff, 1983). The danger is analogous to learning how to use electric saws, when hand saws used to be the state-of-the-art. We can either decry electric saws, or teach their proper and safe use. One of premier psychology quantitative journals is called *Psychological*

Methods, and publishes articles on design, sampling, measurement, and statistics, as well as how these different areas overlap and inform one another.

Conclusion

Consider an analogy from the history of science to illustrate the points made in my response. The analogy draws on two popular science books, Sobel (2000) and Gleick (2003). The late 16th and early 17th century occupied a remarkable period of scientific ascendancy in the field of astronomy. In 1543, Copernicus offered the insightful (yet heretical) view that the earth revolved around the sun, rather than vice versa. Galileo was born shortly after, and as Sobel noted, "All his [Galilio's] observations lent credence to the unpopular sun-centered universe of Nicolas Copernicus, which had been introduced over half a century previously, but foundered on lack of evidence" (p. 7). The observations to which Sobel referred were of course obtained with Galilio's new invention, the telescope, through which he observed the moons of Jupiter, the face of earth's moon, and the sunspots moving across the face of the sun. Such observations were, in modern language, exploratory evidence in support of a previously proposed theory.

Although probabilistic reasoning was still in its infancy (and was being developed by Fermat and Pascal in France during the same historical epoch), the epistemological basis of scientific inquiry in astronomy during that period was similar to that in psychology during the 20th century. The NHST paradigm that Robinson & Levin vigorously defended was similar to the one used by Galileo and others during the period of time in which they were collecting information (using telescopes and otherwise). Ultimately, such information of course inductively coheres into theoretical propositions. Galileo offered multiple sources of astronomical evidence for a heliocentric view of the solar system, including the movement of sunspots, the eclipses of the moons of Jupiter, and the tides on earth. Each might be viewed as a separate astronomical significance test of the null hypothesis that the earth was at the center of the universe, a hypothesis that we have ultimately rejected. But astronomy quickly moved on beyond the question of whether the Copernican system could be rejected or not.

Kepler, in 1609 and 1619, published his three laws of planetary motion, and Newton (who was born in 1642, the year that Galilio died), published in 1687 his *Principia*, stating formal mathematical models of motion and the universal law of gravity. These "laws" stepped up to a new epistemological level, using previous observations as the basis for mathematical models that were designed to subsume many previous disparate and separate astronomical observations. (The development of the double-helix model of DNA is another example in a different discipline in which disparate observations were brought together inductively using mathematical modeling.)

To bring these historical references to the current discussion, Robinson and Levin wrote: "we agree that - in the field of education - we have enough theory development studies and need more studies that address practical 'what works' questions." Fair enough. They argued that in many domains of our immature science more knowledge is needed, that more educational and psychological telescopes need to be brought to bear on current problems. Nothing in my own teaching, thinking, or research practice holds anything but praise and agreement for such a position. Indeed, two of my primary courses over the past 30+ years of teaching have been Exploratory Data Analysis and Quantitative Methods in Evaluation Research, where students learn to engage exactly this kind of goal, to address practical "what works" questions.

Then, they stated, "It is our fear that a research approach where the question 'Does the data fit my model?' is far more dangerous than the question 'Is there anything here worth pursuing?'" Again, fair enough. Without knowledge, both scientists and those who consume the science (policymakers, the public, etc.) can be led to the modern equivalent of the geocentric universe, and there is indeed danger in promulgating positions both pro and con in the absence of adequate knowledge, or even with substantial knowledge when that knowledge is at odds with societal expectations (just ask Galilio!). But does such lurking danger excuse statisticians and methodologists from

developing proper tools, perspectives, and whole epistemological systems to support the development and evaluation of such models?

My answer, strongly implied throughout the original article, is indeed not. Both the NHST epistemology they promoted for relatively immature science, and the one that they view as dangerous, the modeling approach, should exist side-by-side within the arena of quantitative methods in both education and psychology. I promoted the development of the latter, not erasing the former. The former can only be criticized when it purports to serve the function of the latter. What is dangerous is asking NHST to provide methodological support beyond that for which it was designed. NHST can answer the question, "Is the null hypothesis plausible, or not?" It was not designed to answer the question, "Which of these two competing mathematical models is preferable in the way that it handles the trade-off between fit and parsimony?" In areas of behavioral science that are ready for more strongly confirmatory research – including the development of mathematical and statistical models that contain both causal and explanatory components (which are, of course, not entirely the same thing) – NHST is naturally expanded into the broader modeling epistemology. That expansion was the subject of my article. The earlier view of NHST as providing epistemological support for important but often separate and disparate individual findings is the topic of Robinson and Levin's criticism. Both stand effectively before criticism.

References

Box, J. F. (1978). *R. A. Fisher: The life of a scientist.* New York, NY: Wiley.

Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research*, *18*, 115-126.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

D'Onofrio, B. M., Goodnight, J. A., Van Hulle, C. A., Rodgers, J. L., Rathouz, P. J., Waldman, I. D., & Lahey, B. B. (2009). A quasi-experimental analysis of the association between family income and offspring conduct problems. *Journal of Abnormal Child Psychology*, 37, 415-439.

D'Onofrio, B. M., Turkheimer, E. N., Eaves, L. J., Corey, L. A., Berg, K., Solaas, M.H., & Emery, R. E. (2003). The role of the children of twins design in elucidating causal relations between parent characteristics and child outcomes. *Journal of Child Psychology and Psychiatry*, 44, 1130-1144.

D'Onofrio, B. M., Van Hulle, C. A., Waldman, I. D., Rodgers, J. L., Harden, K. P., Rathouz, P. J. & Lahey, B. B. (2008). Smoking during pregnancy and offspring externalizing problems: An exploration of genetic and environmental confounds. *Development and Psychopathology*, *20*, 139-164.

Fisher, R. A. (1935). *The design of experiments*.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Vol 1. Methodological issues*, 311-339. Hillsdale, NJ: Erlbaum.

Gleick, J. (2003). *Isaac Newton*. New York: Pantheon Books.

Harden, K. P., D'Onofrio, B. M., Van Hulle, C., Turkheimer, E., Rodgers, J. L., Waldman, I. D., & Lahey, B. B. (2009). Population density and youth antisocial behavior. *Child Psychology and Psychiatry*, *50*, 999-1008.

Jaffee, S. R., Van Hulle, S., & Rodgers, J L. (2011). Effects of non-maternal care in the first three years on children's academic skills and behavioral functioning in childhood and early adolescence: A sibling comparison study. . *Child Development*, 84, 1076-1081.

MacCallum, R. C., & Tucker, L. R (1991). Representing sources of error in the common factor model: Implications for theory and practice. *Psychological Bulletin*, *109*, 502-511.

Mendle, J., Harden, K. P., Turkheimer, E., Van Hulle, C. A., D 'Onofrio, B. M., Brooks-Bunn, J., Rodgers, J. L., Emery, R. E., & Lahey. B. B. (2009). Associations between father absence and age of first sexual intercourse. *Child Development*, *80*, 1463-1480.

Mill, J. S. (1843). *A system of logic*. London: John W. Parker.

Rodgers, J. L. (2001). What causes birth order-intelligence patterns? The admixture hypothesis, revived. *American Psychologist*, *56*, 505-510.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, *65*, 1-12.

Rodgers, J. L., Bard, D., Johnson, A., D'Onofrio, B., & Miller, W. B. (2008). The Cross-Generational Mother-Daughter-Aunt-Niece Design: Establishing Validity of the MDAN Design with NLSY Fertility Variables. *Behavior Genetics*, *38*, 567-578.

Rodgers, J. L., Cleveland, H. H., van den Oord, E., & Rowe, D. C. (2000). Resolving the debate over birth order, family size, and intelligence. *American Psychologist*, *65*, 1-12.

Scarr, S. & McCartney, K. (1983). How People Make Their Own Environments: A Theory of Genotype → Environment Effects. *Child Development*, *54*, 424-435

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton-Mifflin.

Sobel, D. (2000). *Galileo's daughter*. New York: Penguin Books.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods*, *4*(*2*), 212-213.

Wichman, A., Rodgers, J. L., & MacCallum, R. C. (2006) A multilevel approach to the relationship between birth order and intelligence. *Personality and Social Psychology Bulletin*, *32*, 117-127.

Wichman, A., Rodgers, J. L., & MacCallum, R. C. (2007) Birth order has no effects on intelligence: A reply and extension of previous findings. *Personality and Social Psychology Bulletin*, *33*, 1195-2000.

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 348-358

1538 – 9472/10/$95.00

On Scientific Research: The Role of Statistical Modeling and Hypothesis Testing

Lisa L. Harlow
University of Rhode Island

Comments on Rodgers (2010a, 2010b) and Robinson and Levin (2010) are presented. Rodgers (2010a) initially reported on a growing trend towards more mathematical and statistical modeling; and a move away from null hypothesis significance testing (NHST). He defended and clarified those views in his sequel. Robinson and Levin argued against the perspective espoused by Rodgers and called for more research using experimentally manipulated interventions and less emphasis on correlational research and ill-founded prescriptive statements. In this response, the goal of science and major scientific approaches are discussed as well as their strengths and shortcomings. Consideration is given to how their recent articles intersect or differ on these points. A summary and suggestions are provided regarding how to move forward with scientific inferences.

Key words: Scientific inference, statistical modeling, null hypothesis significance testing.

Introduction

The Focus of Science

The study and practice of science is complex and encompasses various approaches and methods. Central to all of science is the search for basic principles from which phenomena can be explained and predicted. How are the underlying tenets - the golden nuggets of truth - in a scientific field discovered and illuminated? That is one of the main questions of this commentary.

Herbert Simon (1969), a Nobel Laureate in economics and a noted cognitive psychologist, believed that whereas human behavior is inherently simple, the complexity of the environment in which the behavior occurs can prevent or obscure human understanding of the basic processes. Thus, Simon (1969) viewed the main focus of science as finding the simplicity in the complexity of life.

Lisa L. Harlow is a Professor in the Department of Psychology. Dr. Harlow is Past President for the Society of Multivariate Experimental Psychology, Editor of the Multivariate Application Book Series, Associate Editor for *Psychological Methods* Journal and co-Director of Quantitative Training for Underrepresented Groups. Email her at: lharlow@uri.edu.

Four decades later, Michio Kaku (2009), a theoretical physicist and an advocate of making science understandable, reached a conclusion that was not far afield from Simon. Kaku made a comparison with the basic rules of chess and the actual enactment of a multitude of different possible chess games, elaborating that "the rules of nature may also be finite and simple, but the applications of those rules may be inexhaustible. Our goal is to find the rules." (p. 302). Kaku elucidated that the development and testing of basic principles in science "reveals the ultimate simplicity and harmony of nature at the fundamental level" (pp. 302-302), and that testing in science is most often indirect. As a result, it may be more productive to have multiple and varied ways to approach research and inferences in order to arrive at the most salient, underlying, and often latent, truths.

Consistent with the perspective that scientific understanding is not always directly observable, George Lakoff and Rafael Núñez (2000) emphasized the importance of concepts and analogies in what they call "the metaphorizing capacity" (p. 54) for understanding and applying quantitative methods beyond simple arithmetic and counting. These researchers realized the value of considering how a phenomenon is similar to, and different from, other related quantifiable observations. In a comparable view, Brian Hayes (2011) wrote that by breaking down stimuli into small segments and noticing points of contrast and similarity the most salient aspects are revealed. He summarized this process by stating that "the aim is to explore the kinds of patterns that appear frequently in our environment, in the hope of identifying and understanding some characteristic themes or features" (p. 422).

Another perspective was offered by Paul Rozin (2009) who discussed how published and funded research has tended, perhaps mistakenly, to involve results engendered through hypothesis-testing, controlled experiments and building causal evidence. In contrast, Rozin recommended descriptive or other kinds of studies that may have more external validity in varied, real-world settings. Rozin ventured that, "Elegance and clarity are criteria for publication, but there should be a trade-off with novelty and engagement" (2009, p. 437); and further that "a really interesting study with a flaw may be more valuable than a flawless but uninteresting study" (p. 438).

Stefan Hoffmann (2011) suggested that scientific curiosity is fed by having a great deal of background knowledge about a phenomenon, and then noticing anomalies, developing intuitions and finding connections. It is at the intersection of novelty, uncertainty and understanding that brings about scientific curiosity and discovery. Toby Huff (2011) concurred, speaking of how engaging curiosity and overarching synthesis lead to scientific discovery.

Culling together the perceptions of these and other astute thinkers, what appear to be integral for scientific discovery are the inquiring, understanding, seeking, describing, comparing and testing of credible and innovative ideas and relationships that may initially be difficult to discern; and the potential to assess the import and generalizability of findings with rigorous methodological procedures. I would argue that the methods espoused by Robinson and Levin, and Rodgers incorporate much of these elements of scientific discovery, albeit with differing approaches.

Approaches to Scientific Research

A reasonable question to ask is how scientific research should be approached. To accomplish scientific development and discovery, Simonton (2003) argued that it is important to see connections among diverse situations and processes, as well as to have an experimental, problem solving approach. Cronbach (1957) spoke to this seeming duality when discussing the two disciplines of psychology that involved either a correlational or an experimental focus. Each of these researchers is featuring two valuable, although often divergent, aspects of innovative science: naturalistic flexibility and rigorous control. This apparent dichotomy can also be viewed as striving for broad, generalizable external validity, versus strict and controlled internal validity; objectives endorsed in varying degrees by the Rodgers, and Robinson and Levin articles, respectively. Although there are probably as many approaches to scientific investigation as there are

researchers, two major methods - null hypothesis significance testing and (correlational) statistical modeling - are the main focus of this commentary.

Null Hypothesis Significance Testing

The traditional approach to research, null hypothesis significance testing (NHST), is supported by Robinson and Levin, and minimized but recognized by Rogers. Briefly, NHST centers on an attempt to reject a null hypothesis of no notable import (e.g., two means are equal, a correlation is zero) and thereby attempting to build evidence for an alternate hypothesis that claims a significant difference or relationship. A noted benefit of NHST is that researchers can clearly specify null and alternate hypotheses and can calculate the probability of obtaining sample results as extreme or more so than are achieved in a relevant and randomly collected sample. Thus, if the probability, or *p-value*, is less than a designated level (e.g., 0.05), researchers can conclude that there is very little chance of obtaining the sample results found if the null hypothesis is true in the larger population from which the sample was drawn. This is particularly helpful if a decision is needed as to whether a specific treatment or intervention should be pursued as a viable option, after conducting a rigorous experiment that had adequate power to detect a significant finding and involved satisfactory design (e.g., random selection and assignment) to rule out possible rival hypotheses or confounds.

Devlin (1998) agreed, pointing out how probability theory is useful when it is necessary to make crucial decisions about whether to endorse a particular treatment or intervention. NHST would be helpful in this regard when there is a need to come to a decision about rejecting a null hypothesis with a specified probability. Others also attested to the benefits of NHST. Mulaik, Raju and Harshman (1997) stated that "as long as we have a conception of how variation in results may be due to chance and regard it as applicable to our experience; we will have a need for significance tests in some form or another" (p. 81). Chow (1996) and Cortina and Dunlap (1997), among others, also applauded the advantage of using NHST to rule out a chance finding in research.

Nonetheless, NHST has been extensively discussed and debated by Robinson and Levin, as well as Rodgers, and in numerous other forums (e.g., Balluerka, Gómez & Hidalgo, 2005; Denis, 2003; Harlow, Mulaik & Steiger, 1997; Kline, 2004; Nickerson, 2000). The better part of criticism regarding NHST appears to center on the exclusive focus of the *p-value* from a statistical test, and the accompanying dichotomizing decision to reject or retain the null hypothesis. Cumming (2012) has spoken at length on the volubility of *p-values* and the practice of NHST. Rice and Trafimow (2010) would likely agree with Cumming in arguing for less concern over Type I errors (i.e., rejecting a null hypothesis when the null hypothesis should not be rejected), and more attention to Type II errors, which refer to the failure to reject a null hypothesis when the alternate, scientific hypothesis may actually have more merit.

Noteworthy is that most, if not all, of the proponents and critics of NHST would also promote the use of additional substantiation over and above, or instead of, evidence of a significant *p-value*. Robinson and Levin advocated for correct applications of statistical hypothesis testing that involve randomized experiments, attention to Types I and II errors, effect sizes and sample size considerations, as well as the use of confidence intervals. Rodgers in turn played down hypothesis testing in favor of what he claimed is a broader, more subsuming and organic modeling approach that has emerged in an almost imperceptible methodological revolution. Before discussing the statistical modeling endorsed by Rogers and eschewed by Robinson and Levin, it is worthwhile to mention the merits of complementary procedures to help corroborate research findings.

Supplementing NHST

Any acknowledged advantages of NHST notwithstanding, current guidelines and research call for additional evidence when making scientific inferences. The recent 6th edition of the American Psychological Association (APA: 2010) publication manual "stresses that NHST is but a starting point and that additional reporting elements such as effect

sizes, confidence intervals, and extensive description are needed..." (p. 33); this viewpoint is consistent with that from Robinson and Levin as well as Rodgers and others.

Seven years before the APA guidelines, Denis (2003) presented a balanced overview of NHST and several possible alternatives. Denis suggested that the use of model testing among two or more reasonable alternatives, using good-enough hypotheses, calculating effect sizes and confidence intervals, and providing graphical displays of the findings are all effective and viable alternatives or supplements to NHST. Neither Rogers nor Robinson and Levin would be likely to take issue with much of this suggestion.

Others call for establishing or replicating a finding before it is accepted. Sawilowsky (2003) cautioned that effect sizes should not be widely published if they are not statistically significant. Filkin (1997) stated that "science seeks to separate fact from fiction by finding evidence" (p. 16); and that "for an idea or theory to be accepted as scientifically proven, it has to be tested in such a way that it can be tested over and over again and the result must always confirm the theory" (p. 20). Carl Sagan (1997) would have agreed with the need for replication; he wrote that the only way to find answers to "deep and difficult questions … [is] by real, repeatable, verifiable observations" (p. 63). Robinson and Levin aptly encouraged conducting "independent replications" to verify whether a significant finding is reliable, a practice also backed by Rodgers.

Consistent with replication, Wilson (1998) affirmed that "scientific evidence is accretionary, built from blocks of evidence joined artfully by the blueprints and mortar of theory … as evidence piles upon evidence and theories interlock more firmly, certain bodies of knowledge do gain universal acceptance" (p. 64). Wilson further highlighted the need for "improving the piecemeal approach science has taken to its material properties" (p. 66). Here, Wilson argued for a multivariate approach, as well as more attention to strong theory to ground scientific research. In this issue of the *Journal of Modern Applied Statistical Methods*, the value of theory was touted by Rodgers as well as Robinson and Levin; however, the usefulness of multivariate methodology was championed by the former but discouraged by the latter researchers.

It is also of interest that discussion about the need to augment NHST is not limited to the topic of abstract methodology, but rather intersects with the content and substance of practice and research. In a recent issue of the journal *Psychotherapy*, Thompson-Brenner (2011) introduced a special section on the role of significance testing in clinical trials. The set of articles illuminated considerations for providing the most accurate information on how best to create effective interventions in clinical practice. In the leading article, Krause (2011a), discussed the limitations of significance testing with randomized clinical trials (RCTs) and called for the inclusion of whole outcome distributions from participants in an RCT. Similar to what Cumming (2012) and others promote, Krause (2011a, 2011b) maintained that the significance test and *p-value*, alone, are not very informative about how to proceed with clinical treatments. Gottdiener (2011) responded by advocating the use of effect sizes and confidence intervals when presenting RCT results and asked researchers to supplement these data with information from case studies that can more specifically delineate treatment effectiveness and failure.

It is noteworthy that Gottdiener - as Wilson (1998) did earlier - also encouraged the study of multiple outcomes, arguing that multivariate data are more apt to provide bases for reliable and valid conclusions regarding treatment success or failure. Wise (2011) provided a compelling discussion on the need for evidence of clinically significant change and the use of a reliable change index, which is similar to a pre-post-intervention *z*-score for participants in an RCT. Here, the convergence and divergence of these proposals with respect to views put forth by Rogers, and Robinson and Levin, are not as clear-cut, except, again, that the former would favor multivariate approaches more readily than the latter researchers.

To round out this discourse on significance testing and its supplements, it is of note that Hagen (1997, 1998), a strong proponent of NHST, also recognized that effect sizes and confidence intervals are meaningful to report. Further, Hagen - who was reportedly

"struck by the beauty, elegance, and usefulness of NHST" - went on to acknowledge that "other methods of inference may be equally elegant and even more useful depending on the question being asked" (1998, p. 803). Similarly, whereas Burnham and Anderson (2002) admitted that "for classic experiments (control-treatment, with randomization and replication) we generally support the traditional approaches (e.g., analysis of variance)" (p. viii), largely based on NHST; they more strongly endorsed a modeling perspective. Rozin (2009) would probably agree, stating that hypothesis testing may be more appropriate in fields where there is more knowledge and background. Otherwise, Rozin recommended assessing the nature of the phenomenon and its "generality outside of the laboratory and across cultures" (2009, p. 436), a practice that may be more easily accomplished with modeling. In this regard, it is useful to consider an alternative to NHST, namely, statistical and mathematical modeling.

Statistical and Mathematical Modeling

Rodgers (2010a) argued persuasively for adopting statistical and mathematical modeling, which he claims subsumes the predominant standard of NHST. Rodgers convincingly expressed the benefits and extent of statistical modeling, including such procedures as "structural equation modeling, multi-level modeling, missing data methods, hierarchical linear modeling, categorical data analysis, as well as the development of many dedicated and specific behavioral models." Rodgers further decried the emphasis in NHST on the rejection of a null hypothesis, a practice that, in opposition to Rodgers, was embraced by Robinson and Levin. However, these latter researchers clarify that they view NHST mainly as a screening device (Robinson & Levin, 2010) to illuminate findings worthy of further study, and thus would not be expected to place undue attention on the null hypothesis. Still, as Rodgers pointed out, statistical modeling places the focus on a well-constructed model, as opposed to a null hypothesis, and entails a "powerful epistemological system" of "building and evaluating statistical and scientific models." Rodgers (2010a) further advocated that methodological curriculum should be revised to incorporate a modeling approach, with NHST playing an "an important though not expansive role" (p. 1).

Others would agree with the call for wider use of model testing. Burnham and Anderson (2002) discussed a multi-model approach to understanding and approximating a complex process. Their information-theoretic approach includes comparing a scientific model that has a strong theoretical basis to several reasonable alternative models, while also taking into account parameter estimation, uncertainty and parsimony. In this way, a model or reduced set of models can be retained as the "best approximating model" (p. 2). Their approach represents a balance between over-fitting that would be neither replicable nor externally valid, and under-fitting which would be limiting and lack internal validity. It may seem paradoxical that Robinson and Levin would most likely also go along with the practice of testing multiple models, whereas it could easily be expected that Rodgers would approve of Burnham and Anderson's recommended multi-model testing methodology.

In a similar endorsement, Filkin (1997) described how Stephen Hawking, a renowned physicist, used a method called "sum over histories" to select the most likely approaches or models to understand a specific phenomenon and then to eliminate them one by one until arriving at the most probable solution (p. 272). Likewise, Maxwell and Delaney (2004) presented a convincing and integrative approach to science by proposing the examination of multiple models within a given study, ideally with research based on an experimental design. To varying degrees, Robinson and Levin, as well as Rodgers, would support this emphasis on assessing several viable and relevant models, particularly within the context of rigorous, controlled research.

Congruent with Rodgers' (2010a) focus on statistical modeling that recognizes the role of significance testing, Granaas (1998) claimed that "model fitting combines the NHST ability to falsify hypotheses with the parameter estimation characteristic of confidence intervals" and could still recognize that "effect size estimation is central" (p. 800). In an in-depth and convincing collection of model-based methods, Little,

Bovaird and Card (2007) offered a well-articulated treatise on the benefits of statistical modeling, particularly when taking into account various conditions (e.g., mediation, missing data, moderation, multilevel data, multiple time points). I back each of these efforts, which would - at least in part - be supported by Robinson and Levin as to the value of NHST, considering relevant provisos. I would go further to state that statistical modeling may be more effective than NHST in allowing and even encouraging researchers to be more motivated to study, analyze and integrate their findings into encompassing and coherent streams of research. This position would most assuredly be endorsed by Rodgers.

The capabilities aside, it cannot go unnoticed that Robinson and Levin, as well as numerous other researchers (e.g., Baumrind, 1983; Cliff, 1983; Freedman, 1987a, 1987b; Ragosa, 1987) spelled out the possible hazards of statistical modeling, particularly when making unjustifiable causal claims from information that does not stem from longitudinal data or experimental design with adequate controls. Moreover, Kratochwill and Levin (2010), as well as Robinson and Levin, emphasized the importance of randomization, as well as replication and manipulation of the independent variable in order to achieve experimental control and build causal evidence. These authors argued that even single-case intervention designs can be made more rigorous and allow stronger conclusions, particularly by randomizing the assignment, timing and/or replication of interventions.

Shared Variance

Despite the various approaches to conducting scientific research, and the apparently contended methods of NHST and model testing, the articles in this issue by Rodgers, and Robinson and Levin could be said to agree on a number of practices and perspectives, including the merits of randomization and replication, and the cautions against over-interpreting correlations or using causal language when it is not justified. A careful reading of the viewpoints put forth by these authors, who admittedly come from differing epistemological vantages; concur on the importance of each of the following:

- Conducting exploratory / preliminary research that reveals worthwhile avenues to pursue;
- A strong theoretical framework;
- The use of randomization;
- Addressing threats to the validity of research;
- Emphasizing effect sizes and reasonable sample-size considerations;
- Being cautious to not over-interpret correlations;
- Avoiding causal language when not justified;
- Only making meaningful and justified conclusions;
- Encouraging replication;
- Noting the historical importance and development of NHST;
- Recognizing the value of NHST as part of a larger research process;
- Acknowledging the value of both NHST and statistical modeling;
- Realizing that both NHST and statistical modeling can be misused;
- Not disavowing a statistical procedure just because it is sometimes misused; and
- Accruing ongoing knowledge about scientific findings that address relevant problems.

By any yardstick, it would be difficult to deny significant overlap and agreement in the scientific values of Robinson, Levin, and Rodgers.

Just as it would not be accurate to posit hypothesis testing as the exclusive focus on a dichotomous decision between a null hypothesis and a generic alternative hypothesis, there may not be the need for a sharp contrast between the approaches presented by Rogers, and Robinson and Levin. Unlike Schmidt and Hunter (1997) who claimed that "statistical significance testing…never makes a positive contribution" (p. 37), or even McGrath (1998) who ventured that "it is very appropriate to praise the brilliance of NHST, but having done so, perhaps it is time to bury it" (p. 797), a more inclusive

approach to science would allow for much of what was advocated by Robinson and Levin as well as Rodgers.

Rodgers (2010a) and Robinson and Levin, among others (e.g., APA, 2010; Wilkinson, et al, 1999), supported a broad and accurate approach that incorporates rigorous considerations (e.g., effect sizes, confidence intervals), alongside either NHST or statistical modeling. Hagen (1998), consistent with Robinson and Levin, and Rodgers, raised another issue by contending that "absence of evidence does not equal evidence of absence" (p. 803). By this Hagen clarified that research that fails to reject a null hypothesis cannot claim that the null hypothesis is true, a point that is sometimes mistakenly made with proponents of both NHST and modeling. In this regard, researchers conducting NHST cannot assert finding proof for the null hypothesis when it fails to be rejected. Similarly, those carrying out statistical modeling cannot overstate the benefit of a model in which the proposed model was not found to be significantly different from the pattern of variation and covariation in the data. Rogers, Robinson and Levin would undoubtedly agree that reasonable alternatives, confounds and considerations need ample deliberation, regardless of scientific approach.

Significant Differences or Type I Errors?

Given the recognized points of convergence, it is informative to at least mention that in this issue, Robinson and Levin, and Rodgers set forth differing or detracting points of view, as evinced in the following:

Robinson and Levin believed that Rodgers presents "a one-sided view of the controversy," and argue that they "have seen frequent misapplication of Rodgers' favored causal modeling techniques." Robinson and Levin further argued against a statistical modeling approach, based largely on the possible misuses associated with such an approach, for example, making unwarranted causal conclusions and overly prescriptive statements when using cross-sectional and correlational data. It is likely that most researchers, including Rodgers, would agree with their encouragement to use hypothesis testing wisely and to supplement with effect sizes and confidence intervals. Similarly, Rodgers and other researchers are apt to endorse their concern with ascribing causality when the research design did not include the necessary controls (e.g., randomization, manipulation, temporal ordering, isolation of effect, repetition).

Whereas Rodgers' (2010b) claim that statistical modeling could serve as a larger framework that subsumes NHST could be acknowledged, some of the writing may be too dismissive. For example, Rodgers charged that NHST does not have status and involves immature and simple science, compared with an epistemological system such as mathematical and statistical modeling. It may be more accurate to state that NHST can focus on more specific research questions, particularly in areas in which there is sufficient background knowledge to make informed and relevant hypotheses (see Rozin, 2009 for more discussion on this point).

Robinson and Levin occasionally made statements that may be overstated or inaccurate, such as using the qualifier "causal" numerous times when referring to modeling procedures or advocates, even when the term "causal" was not necessarily appropriate or endorsed by what was being described. This misattribution of causal language is evident in the title of their article, when referring to "Rodgers' favored causal modeling techniques," when speaking about "causal modeling techniques" and "unfortunate 'causal' nomenclature, "as well as "causal-model researchers," among other instances. Robinson and Levin also provided what they claimed as examples of "unjustified 'causal' excerpts" that are said to have overstated the use of causal language, when the research they describe does not explicitly appear to have done so and where, in some cases, the researchers have cautioned against making causal conclusions. For example, in an article that is critiqued, researchers claimed that "the data in the study are cross-sectional in nature and causal relations cannot be drawn" (Chen, et al, 2009, p. 304) although Robinson and Levin dismissed the stated limitation as "predictable."

Rodgers could also offer more elaboration and careful language when describing relevant examples that would favor

modeling research, such as when stating how "selection bias has improperly influenced the interpretation of birth order-intelligence links," on illustrating a "type of sibling control," and on how "findings make a strong statement about both modeling and NHST." When describing each of these examples, there did not appear to be enough information provided to come to the conclusions that Rodgers set forth. Additionally, it would be preferred to use the word "parents" instead of "women" when discussing problems that are "almost completely attributable to the type of women who put their children in day care."

Regarding the use of language, Robinson and Levin occasionally used glib or dismissive terms when describing "the perceived magical quality of SEM allowing researchers to coax causality from correlational data," or referring to "grand prescriptives" in published conclusions. Moreover, these authors chided that cross-sectional and correlational data are "tossed into a statistical modeling analysis and what 'popped out' were causal conclusions", and allude to Rodgers' "seductive subtitle" that could purportedly "cause" researchers to see modeling as "methodological randomization compensating panaceas."

Another point worth noting is that Robinson and Levin, as well as Rodgers, expressed concern about the nature of the articles cited and, conversely, omitted from their respective manuscripts, when almost half of the citations in each manuscript involve one or more of the corresponding authors (i.e., 11 of 24 references are self-citations in Rodgers; and 14 of 33 references in Robinson & Levin similarly involve one or both of the authors). Whereas it is not unusual to cite relevant articles with which one is familiar, there may be some degree of selection bias in what is referenced in both manuscripts.

Are these points indicative of significant differences between Rodgers, and Robinson and Levin, or possibly just Type I errors in some cases? The reader may best decide.

Reconciling Different Approaches to Scientific Inference

Is it possible to come to agreement on how to approach scientific research? As Simon (1969) and Kaku (2009) expounded, whereas the world around us appears complex and unknowable, the role of scientists is to use whatever means are available to see through to the essence or set of truths in a field. These efforts will most likely involve thoughtful theoretical frameworks alongside sophisticated quantitative analysis to uncover what is not easily distinguished on the surface, positions that many scientists, including Rodgers, Robinson and Levin would endorse. Without specifying a precise approach, Devlin, a mathematician, writes that "where the real world is concerned, we have to go out and collect data. We enter the world of statistics" (1998, p. 156). Lakoff & Núñez (2000) affirmed that "mathematics is a magnificent example of the beauty, richness, complexity, diversity, and importance of human ideas" (p. 379), and Galton (1889) eloquently spoke of the wonder of statistics when used judiciously, stating:

> Some people hate the very name of statistics, but I find them full of beauty and interest. Whenever they are not brutalised, but delicately handled by the higher methods, and are warily interpreted, their power of dealing with complicated phenomena is extraordinary. They are the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue … Science. (p. 62-63)

Advocates of both NHST and statistical modeling would most likely agree with Galton on the overriding splendor of quantitative methods when used responsibly, regardless of the particular approach to scientific research.

Hayes (2011) maintained that scientists may fare well when using statistical, probabilistic models. He argued that, in contrast to using a strictly deductive process and seeking deterministic principals, it is preferable to actively engage with the data by "forming and evaluating hypotheses, building conceptual models, and applying iterative procedures to refine the models or replace them when necessary" (p. 421). This description aptly depicts what Rodgers advocated with statistical

modeling, and incorporates what Robinson and Levin encourage with testing hypotheses with randomized experiments "followed by a sufficient number of independent replications until the researcher has confidence that the initially observed effect is a statistically reliable one."

When considering the overall value of hypothesis testing and modeling, Rodgers (2010b) acclaimed that "NHST is a worthy, valuable, and useful tool" and "is still a proper paradigm, but it is a special case of a broader and thus more flexible paradigm." Hagen (1998) also acknowledged, along with Granaas (1998), that statistical modeling may well have advantages over NHST, although knowledge and use of modeling may not be as widely available as NHST, a position endorsed by Rodgers, as well. Certainly, the longer history of NHST as adopted in classrooms and research labs, has found its way into books and scholarly articles in larger volume than that of statistical and mathematical modeling procedures. It could only facilitate the progression of scientific knowledge to encourage more attention to well-tempered modeling to complement the pervasive availability and use of significance testing.

Ultimately, creative science depends on the ability to conduct specifically-focused, controlled studies that involve randomization and allow for causal inference. At the same time, there is a need for more broad-based and overarching statistical modeling that allows more flexible hypothesizing, analyzing and synthesizing of relationships among multiple relevant variables. There need not be an artificial dichotomy between these approaches to scientific research. Indeed, Rodgers (2010b) recognized that hypothesis testing and modeling "can be reconciled and accommodated" (p. 340).

As long as researchers keep in mind what can and cannot be claimed on the basis of their particular studies, the adoption of multiple approaches can only enhance and further the realm of science. A new journal is now available, the *Journal of Causal Inference*, edited by Judea Pearl and others, to encourage a rigorous multidisciplinary exchange of ideas regarding causation in scientific research. It is hypothesized that ongoing and open dialogue among foremost scientific researchers will help clarify the value of maintaining controlled and specific NHST, as well as revolutionary and overarching statistical modeling.

References

APA. (2010). *Publication manual of the American Psychological Association* (*6th Ed.*). Washington, DC: American Psychological Association.

Balluerka, N., Gómez, J., & Hidalgo, D. (2005). The controversy over null hypothesis significance testing revisited. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*(*2*), 55-70.

Baumrind, D. (1983). Specious causal attributions in the social sciences: The reformulated stepping-stone theory of heroin use as exemplar. *Journal of Personality and Social Psychology*, *45*, 1289-1298.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (*2nd Ed.*). New York: Springer.

Chen, L. H., Wu, C.-H., Kee, Y. H., Lin, M.-S., & Shui, S.-H. (2009). Fear of failure, 2 x 2 achievement goal and self-handicapping: An examination of the hierarchical model of achievement motivation in physical education. *Contemporary Educational Psychology*, *34*, 298-305.

Chow, S. L. (1996). *Statistical significance: Rationale, validity and utility.* Beverly Hills, CA: Sage.

Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research*, *18*, 115-126.

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, *2*, 161-172.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671-684.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York: Routledge.

Denis, D. (2003). Alternatives to null hypothesis significance testing. *Theory & Science*, *4*, 1-21. Retrieved 9/8/11 from: (http://theoryandscience.icaap.org/content/vol4.1/02_denis.html).

Devlin, K. (1998). *Life by the numbers*. New York: Wiley.

Filkin, D. (1997). *Stephen Hawking's universe.* New York: Basic Books.

Freedman, D. A. (1987a). As others see us: A case study in path analysis. *Journal of Educational and Behavioral Statistics*, *12*, 101-128.

Freedman, D. A. (1987b). A rejoinder on models, metaphors, and fables. *Journal of Educational and Behavioral Statistics*, *12*, 206-223.

Galton, F. (1889). *Natural inheritance.* London: MacMillan.

Gottdiener, W. H. (2011). Improving the relationship between the randomized clinical trial and real-world clinical practice. *Psychotherapy*, *48*, 231-233.

Granaas, M. M. (1998). Model fitting: A better approach. *American Psychologist*, *53*, 800-801.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, *52*, 15-24.

Hagen, R. L. (1998). A further look at wrong reasons to abandon statistical testing. *American Psychologist*, *53*, 801-803.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates.

Hayes, B. (2011). Making sense of the world. *American Scientist*, *99*, 420-422.

Hoffman, R. (2011). That's interesting. *American Scientist*, *99*, 374-377.

Huff, T. E. (2011). *Intellectual curiosity and the scientific revolution: A global perspective.* Cambridge: Cambridge University Press.

Kaku, M. (2009). *Physics of the impossible.* New York: Anchor Books.

Kline, R. B., (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*, *15*, 124-144.

Krause, M. S. (2011). Statistical significance testing and clinical trials. *Psychotherapy*, *48*, 217-222.

Krause, M. S. (2011). What are the fundamental facts of a comparison of two treatments' outcomes? *Psychotherapy*, *48*, 234-236.

Lakoff, G., & Núñez, R. E. (2000). *Where mathematics comes from: How the embodied mind brings mathematics into being.* New York: Basic Books.

Little, T. D., Bovaird, J. A., & Card, N. A. (Eds.) (2007). *Modeling contextual effects in longitudinal studies.* Mahwah, NJ: Erlbaum.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (*2nd Ed.*). Mahwah, NJ: Erlbaum.

McGrath, R. E. (1998). Significance testing: Is there something better? *American Psychologist*, *53*, 796-797.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significant tests?*, 65-115. Hillsdale, NJ: Erlbaum.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241-302.

Ragosa, D. R. (1987). Causal models do not support scientific conclusions: A comment in support of Freedman. *Journal of Educational and Behavioral Statistics*, *12*, 185-195.

Rice, S., & Trafimow, D. (2010). How many people have to die over a type II error? *Theoretical Issues in Ergonomics Science*, *11*(*5*), 387-401.

Robinson, D. H., & Levin, J. R. (2010). The not-so-quiet revolution: Cautionary comments on the rejection of hypothesis testing in favor of a 'causal' modeling alternative. *Journal of Modern Applied Statistical Methods*, *9*(*2*), 332-339.

Rodgers, J. L. (2010a). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, *65*, 1-12.

Rodgers, J. L. (2010b). Statistical and mathematical modeling versus NHST? There's no competition! *Journal of Modern Applied Statistical Methods*, *9*(*2*), 340-347.

Rozin, P. (2009). What kind of empirical research should we publish, fund, and reward? A different perspective. *Perspectives on Psychological Science*, *4*, 435-439.

Sagan, C. (1997). *Billions and billions: Thoughts on life and death at the brink of the millennium*. New York: Ballentine Books.

Sawilowsky, S. S. (2003). You think you've got trivials? *Journal of Modern Applied Statistical Methods*, *2*, 218-225.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significant tests?*, 37-64. Hillsdale, NJ: Erlbaum.

Simon, H. A. (1969). *The sciences of the artificial*. Cambridge, MA: The M.I.T. Press.

Simonton, D. K. (2003). Scientific creativity as constrained stochastic behavior: The integration of product, person, and process perspectives. *Psychological Bulletin*, *129*, 475-494.

Thompson-Brenner, H. (2011). Introduction to the special section: Contextualizing significance testing in clinical trials. *Psychotherapy*, *48*, 215-216.

Wilkinson, L., & the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.

Wilson, E. O. (1998). *Consilience: The unity of knowledge*. New York: Vintage Books.

Wise, E. A. (2011). Statistical significance testing and clinical effectiveness studies. Psychotherapy, *48*, 225-228.

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 359-368

1538 – 9472/10/$95.00

REGULAR ARTICLES

Recommended Sample Size for Conducting Exploratory Factor Analysis on Dichotomous Data

Robert H. Pearson
University of Northern Colorado,
Greeley, CO USA

Daniel J. Mundfrom
New Mexico State University,
Las Cruces, NM USA

Minimum sample sizes are recommended for conducting exploratory factor analysis on dichotomous data. A Monte Carlo simulation was conducted, varying the level of communalities, number of factors, variable-to-factor ratio and dichotomization threshold. Sample sizes were identified based on congruence between rotated population and sample factor loadings.

Key words: Exploratory Factor Analysis, dichotomous data, sample size.

Introduction

Selecting a sample size is one of the most important decisions to be made when planning an empirical study. Often the choice is based on the minimum necessary sample size to obtain reliable results from the statistical procedures to be conducted. For many procedures (e.g., *t*-test, *F*-test) an exact minimum can be found which will allow relationships in the population (if they exist) to be detected with high probability. The issue of sample size for exploratory factor analysis (EFA) is not as straightforward, however, because an exact minimum cannot easily be found analytically and because the procedure's use involves a greater degree of subjectivity.

Although factor analysis has been used in a vast array of scientific fields, it is most frequently used as a tool to investigate the structure of scores obtained via psychometric measures. Such research seeks to identify and possibly measure a small number of unobservable traits that are hypothesized to explain a large portion of the covariation among observed variables. The statistical problem for EFA is the estimation of communalities and - perhaps more importantly - factor loadings. If the results of a factor analysis are to be useful beyond a particular study, then the estimated loadings must be reasonable approximations of true population loadings. Thus, reliable guidelines for selecting a sample size that is likely to produce a factor solution which closely matches a population factor structure would be a boon to researchers planning factor analytic studies.

Until recently, most of the published sample size recommendations were simplified rules based on experts' experience. Several of the most frequently cited guidelines are absolute numbers. Gorsuch (1983) and Kline (1994) suggested sampling at least 100 subjects. Comrey and Lee (1992) provided the following scale of sample size adequacy: 50 – very poor, 100 – poor, 200 – fair, 300 – good, 500 – very good, and 1,000 or more – excellent. Authors have also proposed minimum ratios of sample size to the number of variables (*n*:*p*). Cattell (1978) suggested three to six subjects per variable, Gorsuch (1983) suggested this ratio be at least five and both Everitt (1975) and

Robert H. Pearson is an Assistant Professor of Applied Statistics in the Department of Applied Statistics & Research Methods. His research interests include multivariate statistics, statistical computing and factor analysis. Email: robert.pearson@unco.edu. Daniel J. Mundfrom is an Associate Professor of Applied Statistics in the Department of Economics. His research interests include multivariate statistics, statistical methods and applications of statistics. Email: mundfrom@nmsu.edu.

Nunnally (1978) recommended sampling at least ten times as many subjects as variables.

MacCallum, Widaman, Zhang, and Hong (1999) demonstrated mathematically and empirically that sample size requirements are contingent upon two aspects of the factor structure. Specifically, they showed that both mathematical overdetermination (the extent to which the common factors are sufficiently represented by an adequate number of variables) and the size of communalities have a considerable effect on the agreement between sample and population factor loadings. In a Monte Carlo study they showed that communality had an estimated effect size ($\hat{\omega}^2$) nearly three times greater than sample size and overdetermination had an effect nearly as large as sample size. Mundfrom, Shaw and Ke (2005) subsequently provided sample size recommendations for 180 population conditions on the basis of a Monte Carlo study that varied the number of factors, the ratio of variables to factors (an important aspect of overdetermination) and communalities.

In practice, data are often measured on ordinal or nominal scales, particularly in the social sciences (Hip & Bollen, 2006; Lee & Song, 2003; Schoenberg & Arminger, 1989). Exploratory factor analysis is often applied to ordinal or dichotomous data to examine their relationship with underlying factors (Baños & Franklin, 2002; Mundfrom, Bradley, & Whiteside, 1994; Tomás-Sábado & Gómez-Benito, 2005). Many authors have suggested other approaches for this situation (Bartholomew & Knott, 1999; Bock & Aitkin, 1981; Muthén, 1978), however, a traditional factor analysis can be useful as long as a meaningful and interpretable set of factors can be identified, regardless of the measurement level of the input data. Johnson and Wichern (2002) refer to this as the WOW criterion: "If, while scrutinizing the factor analysis, the investigator can shout 'Wow, I understand these factors,' the application is deemed successful" (p. 524).

Darlington (1997) described this use of factor analysis as heuristic rather than absolute. It is understood that any factor solution is only one among many that are possible. If the retained factor structure can be cross-validated or together with other evidence supports a broader theory, then the analysis is successful. Mulaik (1989) discussed how this approach fits with theory development throughout science:

> Theoretical physics, for example, is continuously occupied with differing speculations designed to synthesize the same sets of diverse experimental data. All of these differing theoretical speculations may yield models that fit equally well the data already at hand, but in time some or all of these speculative models may be eliminated from further consideration by their inconsistency with new data obtained to test certain predictions derived from them. (p. 54)

For a factor solution to be replicable across studies it must represent a structure that truly exists in the population.

The primary purpose of this study was to provide sample size recommendations for researchers who are planning factor analytic studies that will involve dichotomous variables. It was also of interest to compare the results of this study to requirements for continuous data (Mundfrom, et al., 2005). From a methodological standpoint, the extent to which these results differ from those found by Mundfrom, et al. (2005) lends insight into the effect that scale of measurement has on this statistical procedure. Because the case of dichotomous data is the most extreme departure from continuity, these recommendations represent an upper bound for minimum necessary sample size. Therefore, these recommendations were also intended to serve as conservative guidelines for EFA of ordinal data.

Methodology

Monte Carlo simulation was used for this study. Population data were generated using the SAS System v9.1.3 (SAS Institute Inc., 2007). One-hundred matrices of dichotomous data, each conceptually representing a unique population of 100,000 observations on p variables, were generated for each condition determined by four manipulated variables: the number of common factors (m), the variable-to-factor ratio (p:m), the

variable communalities and the dichotomization threshold. Populations were randomly generated using the following two-stage process.

In the first stage, the procedure described by Tucker, Koopman, and Linn (1969) was used to randomly generate population correlation matrices with specified factor structures. A total of 180 factor structures were investigated by crossing the number of factors ($1 \leq m \leq 6$), the variable-to-factor ratio ($3 \leq p{:}m \leq 12$), and the variable communalities. Three levels of variable communalities were examined: high, in which communalities were randomly assigned values of 0.6, 0.7 or 0.8; wide, in which they could have values from 0.2 to 0.8 in increments of 0.1; and low, in which they could have values of 0.2, 0.3, or 0.4 (Tucker, Koopman & Linn, 1969). Ten correlation matrices were generated for each factor structure.

In the second stage, ten matrices of binary data were generated from each population correlation matrix (R). Each data matrix consisted of 100,000 rows of values on p dichotomous variables. First, a matrix X was created by taking the product of the Cholesky root of R and a matrix of multivariate-normal deviates. Elements of each column of X were then dichotomized according to three conditions. In the first condition, all variables were dichotomized to have a 50/50 split. This condition results in the smallest amount of information loss due to dichotomization (Cohen, 1983) and can be considered the best case. In the second condition, all variables were dichotomized to have an 80/20 split. This condition was used in simulation studies by Parry and McArdle (1991) and Weng and Cheng (2005), and is similar to the 84/16 split used by Bernstein and Teng (1989) which they likened to item distributions found in symptom description scales such as in the MMPI or a difficult ability test. In the remaining condition, half of the variables were dichotomized using an 80/20 split and half using a 50/50 split.

Because differences in item means limit the maximum possible value of the product-moment correlation it was important to investigate the resulting effect on factor loading estimates. As a result, one-hundred population data matrices (hereafter referred to as populations) were generated for each combination of communality level, number of factors, variable-to-factor ratio and dichotomization threshold.

Each population was factor analyzed using maximum likelihood estimation and varimax rotation. One-hundred simple random samples of a specific size were then selected from each population. If a sample correlation matrix was non-positive-definite, another was generated and used instead. Each sample was factor analyzed and the rotated factor loadings were compared to those in the population using a coefficient of congruence.

Sample sizes were chosen by first starting with a sample size that was too small based on the recommendations of Mundfrom, et al. (2005). Sample sizes were then increased systematically according to the following algorithm:

- while $n < 30$, it was increased by 1;
- while $30 \leq n < 100$, it was increased by 5;
- while $100 \leq n < 300$, it was increased by 10;
- while $300 \leq n < 500$, it was increased by 20;
- while $500 \leq n < 1,000$, it was increased by 50;
- while $n \geq 1,000$, it was increased by 200.

This system of increments is nearly identical to that used by Mundfrom, et al. (2005). The procedure was stopped when the sample and population correlation matrices met criteria based on a coefficient of congruence. These criteria are defined below. The procedure was also stopped if a sample size greater than 5,000 was necessary.

In summary, a $3 \times 6 \times 10 \times 3$ factorial design was implemented, corresponding to the experimental variables communality level, number of factors, variable-to-factor ratio, and dichotomization threshold, resulting in a total of 540 population conditions. One-hundred populations were randomly generated for each population condition and 100 samples were taken from each population for every sample size considered. Thus, a total of 10,000 samples

were taken for each population condition and sample size combination.

Coefficient of Congruence

A coefficient of congruence was calculated to assess the degree of correspondence between the sample and population solutions (MacCallum, et al., 1999; Tucker, et al., 1969). The coefficient for the k^{th} factor was calculated using the formula:

$$\phi_k = \frac{\sum_{j=1}^{p} \lambda_{jk(s)} \lambda_{jk(t)}}{\sqrt{\left(\sum_{j=1}^{p} \lambda_{jk(s)}^2\right)\left(\sum_{j=1}^{p} \lambda_{jk(t)}^2\right)}},$$

where $\lambda_{jk(t)}$ is the true population factor loading for variable j on factor k, and $\lambda_{jk(s)}$ is the corresponding sample loading. To assess the degree of congruence for a given solution, the mean value of ϕ_k across the m factors was computed and denoted K. For any solution with m factors there were $m!$ possible arrangements of the factors and therefore $m!$ possible values of K. The maximum value of K was used for each solution, thus representing the sample solution that was most similar to the targeted population solution.

For each population, 100 samples were taken and factor analyzed, resulting in 100 values of K. The fifth percentile of these coefficients, denoted K_{95}, was used to represent the lower bound of a 95% confidence interval for a particular population. Subsequently, 100 values of K_{95} were obtained for each population condition, corresponding to the 100 generated populations.

MacCallum, et al. (1999) provided the following guidelines for interpreting values of the coefficient of congruence: 0.98 to 1.00 = excellent, 0.92 to 0.98 = good, 0.82 to 0.92 = borderline, 0.68 to 0.82 = poor, and below 0.68 = terrible. Because the purpose of this study was to determine minimum recommended sample sizes, only those that provided good and excellent levels of agreement were retained. For a given population condition and sample size, the proportions of K_{95}s that were greater than 0.92 and 0.98 were respectively denoted P_{92} and P_{98}.

For a particular condition, a sample size was determined to meet the good criterion if either of the following occurred (Mundfrom, et al., 2005):

- The P_{92} from three successive sample sizes was at least 0.95.
- The P_{92} from two successive sample sizes was at least 0.95, the P_{92} from the next sample size was less than 0.95 and the P_{92} from the next two successive sample sizes was at least 0.95.

The same system was used to select a sample size to meet the excellent criterion. Thus, for every population condition, two sample sizes were chosen as recommendable according to the two criteria.

Results

Minimum necessary sample sizes were identified using a Monte Carlo simulation that manipulated four population characteristics. Factor structures were determined by crossing three levels of communality (high, wide and low), six numbers of factors (1 to 6), and ten variable-to-factor ratios (3 to 12). The three variable distributions considered were 50/50, 80/20 and a third distribution, hereafter referred to as mix, for which half the variables had a 50/50 split and half had an 80/20 split. The minimum necessary sample sizes for each of the 540 population conditions and two agreement criteria are presented in Tables 1, 2, and 3 for the high, wide and low levels of communality respectively.

A few cautions should be observed when interpreting these results. First, the methodology employed did not consider sample sizes beyond 5,000, so this was an artificial ceiling in this study. Secondly, frequent computational errors occurred for conditions when the p:m ratio was three: all results for these conditions should be interpreted cautiously. In addition, the three conditions involving one-factor models with p:m = 3 could not be run by SAS PROC FACTOR with maximum likelihood estimation. Thirdly, the observed results for the mix condition were unstable for models with four to six factors. This instability may be an

Table 1: Minimum Sample Size for Two Agreement Criteria - High Level of Communality

p:m	*Excellent (0.98) Criterion*						*Good (0.92) Criterion*					
	F1	*F2*	*F3*	*F4*	*F5*	*F6*	*F1*	*F2*	*F3*	*F4*	*F5*	*F6*
50/50 Variable Distribution												
3	.	1,200	3,000	5,000	5,000	5,000	.	400	1,400	3,800	5,000	5,000
4	120	270	750	1,600	5,000	5,000	40	90	380	800	3,600	5,000
5	80	280	460	1,800	5,000	5,000	35	85	180	550	2,600	5,000
6	75	250	500	650	1,800	2,600	28	85	200	250	650	700
7	70	250	340	750	1,000	1,200	26	85	120	360	340	400
8	60	270	260	500	1,800	1,000	23	100	90	170	340	460
9	55	320	200	400	1,200	1,400	22	95	65	150	300	700
10	65	260	200	290	480	1,400	25	75	70	110	140	420
11	55	200	220	440	380	800	22	85	75	150	130	250
12	50	160	250	400	550	900	20	60	100	150	170	280
80/20 Variable Distribution												
3	.	2,000	5,000	5,000	5,000	5,000	.	420	5,000	5,000	5,000	5,000
4	230	750	1,600	5,000	5,000	5,000	75	320	900	3,200	3,800	5,000
5	170	900	1,200	2,400	4,400	5,000	65	340	400	900	1,400	4,600
6	150	360	800	2,400	3,800	5,000	55	120	250	500	1,400	2,000
7	130	340	1,200	1,600	3,200	2,200	55	120	420	950	1,200	1,600
8	120	270	650	1,600	2,000	2,000	50	110	230	300	650	900
9	120	240	700	800	1,600	1,800	50	75	190	420	500	650
10	100	320	400	600	950	1,400	45	100	180	200	360	380
11	100	240	440	800	1,400	1,000	45	75	150	290	460	380
12	95	400	700	1,200	850	1,400	45	120	180	320	250	460
Half 50/50 and Half 80/20												
3	.	5,000	2,200	5,000	5,000	5,000	.	4,200	800	5,000	5,000	5,000
4	180	2,000	5,000	5,000	5,000	5,000	55	600	4,000	5,000	5,000	5,000
5	130	480	1,400	2,400	5,000	5,000	40	300	550	1,400	1,400	5,000
6	120	480	1,000	3,200	4,200	5,000	45	190	380	2,200	1,800	3,400
7	95	480	950	1,400	1,600	3,200	40	160	320	460	600	850
8	95	260	500	2,600	1,800	3,000	40	85	180	1,200	600	1,200
9	85	200	340	600	1,200	3,200	35	65	140	240	360	650
10	85	180	340	480	1,800	3,800	35	60	120	160	550	1,200
11	75	140	320	380	1,800	3,600	27	50	100	140	900	750
12	80	190	240	440	650	1,800	30	55	80	150	220	550

Note: F1 denotes one-factor models, F2 two-factor models, etc.

Table 2: Minimum Sample Size for Two Agreement Criteria - Wide Level of Communality

p:m	*Excellent (0.98) Criterion*						*Good (0.92) Criterion*					
	F1	*F2*	*F3*	*F4*	*F5*	*F6*	*F1*	*F2*	*F3*	*F4*	*F5*	*F6*
50/50 Variable Distribution												
3	.	4,000	5,000	5,000	5,000	5,000	.	1,800	5,000	5,000	5,000	5,000
4	700	1,400	5,000	5,000	5,000	5,000	200	480	2,400	5,000	5,000	5,000
5	320	1,400	5,000	5,000	5,000	5,000	95	480	1,400	5,000	5,000	4,600
6	250	950	1,600	2,800	4,000	3,600	75	380	550	1,000	2,200	1,400
7	280	360	1,000	1,600	5,000	5,000	90	180	360	550	1,600	1,600
8	150	460	600	1,400	3,600	3,800	50	190	210	380	1,800	1,400
9	210	650	600	1,800	1,200	2,200	65	170	230	460	420	850
10	150	420	600	1,600	1,400	1,600	55	150	220	550	420	550
11	140	320	700	1,200	1,600	1,600	45	110	210	320	460	550
12	170	440	500	700	950	1,600	55	140	170	180	320	550
80/20 Variable Distribution												
3	.	5,000	5,000	5,000	5,000	5,000	.	5,000	5,000	5,000	5,000	5,000
4	650	5,000	5,000	5,000	5,000	5,000	180	2,000	5,000	5,000	5,000	5,000
5	500	2,000	3,800	5,000	5,000	5,000	160	850	1,400	5,000	5,000	5,000
6	440	1,200	2,000	5,000	5,000	5,000	140	460	500	3,000	5,000	5,000
7	340	1,800	1,800	2,800	4,400	5,000	110	550	600	800	1,600	3,200
8	340	950	1,200	3,000	2,800	4,400	110	270	420	700	1,400	1,600
9	320	550	1,000	1,400	2,600	5,000	100	230	300	550	750	2,000
10	240	550	1,000	1,600	2,200	3,600	85	200	360	550	750	1,400
11	220	400	850	1,200	1,600	2,200	75	130	270	360	480	650
12	210	420	650	950	1,600	1,800	70	140	180	320	460	600
Half 50/50 and Half 80/20												
3	.	4,200	5,000	5,000	5,000	5,000	.	2,200	4,200	5,000	5,000	5,000
4	600	1,800	5,000	5,000	5,000	5,000	200	1,200	5,000	5,000	5,000	5,000
5	290	900	5,000	5,000	5,000	5,000	90	460	3,800	5,000	5,000	5,000
6	300	750	3,600	5,000	5,000	5,000	85	300	1,400	1,200	1,800	5,000
7	210	700	900	5,000	5,000	5,000	70	200	420	2,000	2,800	2,200
8	210	850	1,600	5,000	2,800	5,000	70	300	360	1,200	1,200	2,400
9	210	1,200	650	2,600	2,600	3,000	70	380	220	900	1,200	1,600
10	180	750	800	1,200	1,400	3,000	55	260	250	460	550	850
11	190	500	750	1,600	2,000	5,000	65	180	280	420	600	1,600
12	280	700	1,000	1,200	3,600	3,600	85	240	240	340	1,200	1,400

Note: F1 denotes one-factor models, F2 two-factor models, etc.

Table 3: Minimum Sample Size for Two Agreement Criteria - Low Level of Communality

p:m	*Excellent (0.98) Criterion*						*Good (0.92) Criterion*					
	F1	*F2*	*F3*	*F4*	*F5*	*F6*	*F1*	*F2*	*F3*	*F4*	*F5*	*F6*
50/50 Variable Distribution												
3	.	5,000	5,000	5,000	5,000	5,000	.	5,000	5,000	5,000	5,000	5,000
4	950	3,000	5,000	5,000	5,000	5,000	280	1,200	2,000	5,000	5,000	5,000
5	900	5,000	3,800	5,000	5,000	5,000	270	1,800	1,600	5,000	5,000	5,000
6	650	2,600	3,600	5,000	5,000	5,000	200	1,200	1,400	3,600	5,000	5,000
7	460	2,400	1,600	3,000	5,000	5,000	140	750	600	1,200	5,000	2,800
8	400	950	2,200	5,000	5,000	5,000	120	340	700	1,800	5,000	5,000
9	380	1,400	2,600	2,800	5,000	3,400	120	480	900	1,000	1,600	1,400
10	380	600	1,800	2,200	3,200	4,200	110	180	750	1,000	1,200	1,600
11	340	850	1,400	1,800	5,000	3,200	95	260	400	500	5,000	1,200
12	290	1,000	1,600	2,000	5,000	5,000	85	320	700	700	2,400	2,600
80/20 Variable Distribution												
3	.	5,000	5,000	5,000	5,000	5,000	.	5,000	5,000	5,000	5,000	5,000
4	1,800	5,000	5,000	5,000	5,000	5,000	550	2,600	3,800	5,000	5,000	5,000
5	2,000	5,000	5,000	5,000	5,000	5,000	550	2,600	5,000	5,000	5,000	5,000
6	1,200	2,200	5,000	5,000	5,000	5,000	320	750	2,600	5,000	5,000	5,000
7	800	2,600	2,800	5,000	5,000	5,000	230	650	1,200	2,000	2,800	5,000
8	700	1,800	5,000	5,000	5,000	5,000	200	480	5,000	5,000	3,000	5,000
9	700	1,600	3,400	4,400	4,600	5,000	200	600	1,000	1,800	2,000	4,600
10	600	1,800	3,400	2,400	5,000	5,000	180	650	1,200	800	2,400	2,600
11	550	1,400	2,800	2,800	4,400	5,000	160	420	650	950	1,600	3,200
12	550	1,000	1,200	2,400	4,400	4,400	160	360	1,000	850	1,600	1,600
Half 50/50 and Half 80/20												
3	.	5,000	5,000	5,000	5,000	5,000	.	5,000	5,000	5,000	5,000	5,000
4	2,000	5,000	5,000	5,000	5,000	5,000	600	5,000	3,000	5,000	5,000	5,000
5	950	5,000	5,000	5,000	5,000	5,000	260	2,600	5,000	5,000	5,000	5,000
6	700	1,800	5,000	5,000	5,000	5,000	220	700	5,000	5,000	5,000	5,000
7	550	1,800	5,000	5,000	5,000	5,000	170	500	5,000	3,000	3,800	2,800
8	550	1,600	2,600	5,000	5,000	5,000	170	600	1,200	1,600	2,600	2,800
9	420	1,400	2,400	5,000	5,000	5,000	130	460	950	2,000	2,800	3,800
10	460	1,200	5,000	2,400	5,000	5,000	140	400	1,800	850	4,400	2,000
11	400	1,000	2,800	2,600	4,000	5,000	120	260	1,000	950	1,600	2,800
12	360	2,800	1,800	4,000	2,800	5,000	110	700	650	2,600	950	1,800

Note: F1 denotes one-factor models, F2 two-factor models, etc.

artifact of the methodology used to generate the data.

Overall, the sample sizes needed to analyze dichotomous data are higher than those needed for continuous data as presented by Mundfrom, et al. (2005). For many models with high communalities, three or fewer factors, and high *p*:*m* ratios, sample sizes below 100 are likely to achieve good agreement. Conversely, sample sizes in the thousands are necessary to meet that criterion for most cases when all variables have low communalities or the factors are weakly determined.

Some relationships are apparent from Tables 1 and 2. For a given distribution, level of communality and number of factors, the necessary sample size tends to decrease sharply as the *p*:*m* ratio increases until some elbow after which changes in sample size are very small. This elbow tends to occur at *p*:*m* ratios between seven and ten. For a fixed *p*:*m* ratio, the minimum sample size tends to increase as the number of factors increases. These relations mimic those reported by Mundfrom, et al. (2005) for continuous data, but with more extreme patterns.

Among the three dichotomization conditions, the 50/50 distribution generally requires the lowest sample size. No generalizations are evident as to which of the 80/20 and mix conditions require a lower sample size. The disparity between continuous and binary conditions is smallest for the most well-defined factor structures, especially those with high *p*:*m* ratios. Differences among the binary distribution conditions tend to be small relative to their differences from the continuous data requirements.

Conclusion

One purpose of this study was to provide sample size recommendations to be used by researchers planning studies involving factor analysis of dichotomous data; these are provided in Tables 1, 2 and 3. Although the requirements for analyzing binary data are uniformly higher than those for continuous data across varied aspects of factor model design, they are still reasonable for well-defined factor models. A sample size of 100, which Gorsuch (1983) called the absolute minimum and Comrey and Lee (1992) labeled as poor, is enough to achieve a good level of agreement for models having one or two factors, as well as for three-factor models with at least 24 variables when communalities are high and variables have a symmetric distribution. When the *p*:*m* ratio is high, a sample size of 300 results in good agreement for many models in the wide communality condition and all three examined variable distribution conditions. This sample size is also enough to achieve excellent loading agreement for small models (one or two factors) when variables have high communalities.

The necessary sample size to achieve good agreement between sample and population loadings is grossly inflated for poorly-defined factor models. When communalities are all in the low range, sample sizes in the thousands are necessary for most of the examined conditions. The same is true for most models having four or more factors and *p*:*m* ratios of five or lower.

Another goal of this study was to investigate how dichotomization affects the necessary sample size for EFA. Cohen (1983) showed that when two continuous variables with a joint correlation of *r* are dichotomized at their means, the correlation between the resulting variables is attenuated to a value of .637*r*. One effect of the reduced correlations is that the communalities estimates are concordantly reduced. As described by Schiel and Shaw (1992), 36% of the information is lost when a perfectly reliable continuous variable is dichotomized at the mean. Hence, the communalities are deterministically reduced and additional error is present in the correlation estimates themselves.

MacCallum, et al. (1999) illustrated the role that sampling error has in the formula for the sample factor model. In the presence of sampling error the unique factors will neither have zero correlations with each other nor with the common factors. The terms that are affected by this error are weighted by the size of the unique factor loadings, which are inversely related to communalities.

In summary, dichotomization results in increased sampling error in correlation estimates and attenuated correlation coefficients, which in turn results in decreased communalities. The latter outcome produces larger unique variances which places more weight on the lack of fit

terms in the sample factor model. Thus, there is more sampling error and more weight placed on its detrimental effects.

Dichotomization has the greatest deleterious impact on necessary sample size when communalities are low, the ratio of variables to factors is low or the number of factors is high. The direct and interaction effects of communality follow directly from the previous argument. The other two characteristics affect the overdetermination of common factors. Although the variable-to-factor ratio is not the sole basis of overdetermination, it is an important aspect of it. Many authors have suggested the importance of having a high *p*:*m* ratio (Comrey & Lee, 1992; Tucker, Koopman, and Linn, 1969).

Mundfrom, et al. (2005) demonstrated that the *p*:*m* ratio both has a strong direct relationship with sample size for a fixed *m* as well as a moderating effect on the relationships between sample size, communality, and the number of factors. Moreover, the results of the present study show that the ratio also moderates the effects of dichotomization and variable distribution. At high *p*:*m* ratios, the sample size requirements between the 50/50, 80/20, and mix distributions are fairly similar and in some cases (high communalities, one or two factors) are not that discrepant from those for continuous data. On the contrary, when the ratio is low and the common factors have a low degree of overdetermination, then other changes to the factor model have dramatic consequences on the necessary sample size.

Unless extremely large samples are tenable, some general strategies are recommended when binary data will be factor analyzed. Using variables with high communalities substantially reduces sample size requirements. However, this aspect of the study may be the most difficult to control in practice, especially in survey development. A more manageable design aspect is the *p*:*m* ratio. Having at least eight variables per factor is advised, and a ratio of ten or more should be preferred. This practical step may ameliorate unexpected problems of skewed variables and occasional low communalities.

Results of this study provide direct guidelines to applied researchers who are selecting a sample size for research that will involve exploratory factor analysis of dichotomous data. It is also intended for these results to serve as conservative guidelines for research involving ordinal data. Although the use of dichotomous measures does necessitate larger samples, if many high-quality indicators are used to measure a small number of factors, then applied researchers can be confident that a small to moderate sample size will be adequate to produce a reliable factor solution.

References

Baños, J. H., & Franklin, L. M. (2002). Factor structure of the mini-mental state examination in adult psychiatric inpatients. *Psychological Assessment*, *14*(*4*), 397-400.

Bartholomew, D. J., & Knott, M. (1999). *Latent Variable Models and Factor Analysis*. London: Arnold.

Bock, R. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443-459.

Browne, M. (1968). A comparison of factor analytic techniques. *Psychometrika*, *33*, 267-334.

Cattell, R. (1978). *The Scientific Use Of Factor Analysis*. New York: Plenum.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*(*3*), 249-253.

Comrey, A., & Lee, H. (1992). *A first course in factor analysis*. Hillsdale, NJ: Erlbaum.

Darlington, R. (1997). *Factor Analysis*. Retrieved June 2, 2008, from http:// www.psych.cornell.edu/darlington/factor.htm.

Everitt, B. (1975). Multivariate analysis: The need for data, and other problems. *British Journal of Psychiatry*, *126*, 237-240.

Gorsuch, R. L. (1983). *Factor Analysis* (*2nd Ed*.). Hillsdale, NJ: Erlbaum.

Gorsuch, R. L. (1997). Exploratory factor analysis: its role in item analysis. *Journal of Personality Assessment*, *68*(*3*), 532-560.

Hip, J., & Bollen, K. (2006). Model Fit in Structural Equation Models with Censored, Ordinal, and Dichotomous Variables: Testing Vanishing Tetrads. *Sociological Methodology*, *33*, 267-305.

Johnson, R., & Wichern, D. (2002). *Applied Multivariate Statistical Analysis* (*5th Ed.*). Upper Saddle River, NJ: Prentice Hall.

Kline, P. (1994). *An Easy Guide To Factor Analysis*. New York: Routledge.

Lee, S., & Song, X. (2003). Bayesian analysis of structural equation models with dichotomous variables. *Statistics in Medicine*, *22*(*19*), 3073-3088.

MacCallum, R., Widaman, K., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(*1*), 84-99.

Mulaik, S. (1989). Blurring the distinctions between component analysis and common factor analysis. *Multivariate Behavioral Research*, *25*(*1*), 53-59.

Mundfrom, D., Bradley, R., & Whiteside, L. (1993). A factor analytic study of the infant/toddler and early childhood versions of the HOME Inventory. *Educational and Psychological Measurement*, *53*, 479-489.

Mundfrom, D., Shaw, D., & Ke, T. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, *5*(*2*), 159-168.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*, 551-560.

Nunnally, J. (1978). *Psychometric Theory* (*2nd Ed.)*. New York : McGraw-Hill.

Parry, C., & McArdle, J. (1991). An applied comparison of methods for least-squares factor analysis of dichotomous variables. *Applied Psychological Measurement, 15*(*1*), 35-46.

Schiel, J., & Shaw, D. (1992). Information retention as a function of the number of intervals and the reliability of continuous variables. *Applied Measurement in Education, 5*(*3*), 213-223.

Schoenberg, R., & Arminger, G. (1989). Latent variable models of dichotomous data. *Multivariate Behavioral Research*, *18*(*1*), 164-182.

Tomás-Sábado, J., & Gómez-Benito, J. (2005). Construction and Validation of the Death Anxiety Inventory (DAI). *European Journal of Psychological Assessment*, *21*(*2*), 108-114.

Tucker, R., Koopman, R., & Linn, R. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika*, *34*(*4*), 421-459.

Weng, L., & Cheng, C. (2005). Parallel analysis with unidimensional binary data. *Educational and Psychological Measurement, 65*(*5*), 697-716.

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 369-378

1538 – 9472/10/$95.00

Generalized Variances Ratio Test for Comparing k Covariance Matrices from Dependent Normal Populations

Marcelo Angelo Cirillo Daniel Furtado Ferreira Thelma Sáfadi
Federal University of Lavras,
Lavras Brazil
Eric Batista Ferreira
Federal University of Alfenas,
Brazil

New tests based on the ratio of generalized variances are presented to compare covariance matrices from dependent normal populations. Monte Carlo simulation concluded that the tests considered controlled the Type I error, providing empirical probabilities that were consistent with the nominal level stipulated.

Key words: Dependent normal, bootstrap, variances generalized, power, type I error.

Introduction

Most statistical techniques assume that samples must be independent; however, practical situations where the samples come from dependent populations cannot be ignored. For example, a typical situation is a bioequivalence assay, the objective of which is to verify if a new drug presents effectiveness similar to a brand-name drug. Thus, both drugs are applied to the same sample units, which are classified in two distinct groups and differentiated by the receiving order. The responses of such experiments are correlated and associated to a specific correlation structure.

A naturally appearing hypothesis in this type of experiment regards the equality of covariance matrices between a new drug and a brand-name drug (Wang, et al., 1999). The Bartlett test mentioned by O'Brien (1992) could not be used in this case, because its construction assumes independence of the samples. Due to the restriction of the current tests, the main goal of this study is to propose multivariate tests to verify the equality of covariance matrices considering dependence among multivariate observations along populations.

Another motivation justifying the need for a general test for the equality of covariance matrices of correlated data in time or space are the suppositions of analysis of variance and the Hotelling T^2 test. It is required that the data submitted to multivariate analysis of variance have p-variate normal residues, with null mean vector and constant covariance matrices. To check the assumption of constant covariances for k populations or treatments, a more general test is required. As noted, such tests do not exist or have limited properties for dependence structure situations.

Finney (1938) studied this problem considering the univariate case (p = 1) and two populations (k = 2) under a known correlation coefficient between the same variable in both populations. Pitman (1939) and Morgan (1939) proposed a likelihood ratio test for the case of k = 2 populations, however with an unknown correlation matrix. Since that time, many authors have explored these results, all have considered

Marcelo Cirillo is an Adjunct Lecturer III in the Exact Sciences Department. E-mail: macufla@gmail.com. Daniel Ferreira is an Associated Lecturer I in the Exact Sciences Department. Email: danielff@ufla.br. Thelma Sáfadi is an Adjunct Lecturer III in the Exact Sciences Department. Email: safadi@ufla.br. Eric Ferreira is an Adjunct Lecturer I. Email: eric.ferreira@unifal-mg.edu.br.

only the univariate case (p = 1), although with different numbers of populations.

Roy and Potthoff (1958) concentrated on the bidimensional case, that is, k = 2 and p ≥ 2 variables. However, they did not succeed in test construction. Jiang, et al. (1999) evidenced that the test considered by Roy and Potthoff (1958) presented deficiencies in the imposed presuppositions. Smith and Kshirsagar (1985) presented a likelihood ratio test to compare covariance matrices, coming from two dependent normal populations. However, the authors had not obtained the analytical expression of the maximum likelihood estimator under the null hypothesis. Due to some numerical problems in the maximization of the likelihood functions, the authors surrounded the problem using initial values such that the estimate of the covariance matrix was positive definite.

In a more general situation, represented by a number of populations k ≥ 2 and by a number of variables p ≥ 2, Krishnaiah (1975) considered a test to compare two or more covariance matrices coming from dependent normal populations. This test was formalized under the assumption that the diagonals of the covariance matrices were equal; however, the main criticism to this test was that any restriction or assumption was made for the dependence structure between those matrices.

Jiang, et al. (1999) used Monte Carlo simulation to evaluate some tests based on a likelihood ratio used in the comparison of covariance matrices of dependent normal populations. The differentiation between each test was made under different corrections in the degrees of freedom as proposed by several authors. It was such that - for each correction - new statistics had arisen. Results were restricted to the bidimensional case, and the extension of these tests for p dimensions became impracticable in the face of the numerical problem in the likelihood maximization. Because finding a general test based upon the likelihood ratio to compare k dependent population covariances is a difficult task, the bootstrap method can be used (Manly, 1997). Bootstrapping is typically used to round problems for which an analytical solution is not straightforward. Due to the dependency between populations, Hall, et al., (1995) recommend the use of implicit resample in bootstrap, which must be done in blocks. This article proposes multivariate tests for comparing covariance matrices from k dependent multivariate normal populations, as well as studying their power and type I error probability.

Methodology

The multivariate tests considered in this article have been constructed considering the multivariate observation represented by the vector of random variables $\underset{\sim}{X}$, where each component $\underset{\sim}{X}_1^t,...,\underset{\sim}{X}_k^t$ is composed of p-dimensional vectors of random variables $\underset{\sim}{X}_j = (X_{j1},...,X_{jp})^t$, j = 1..., k, where k refers to the total number of populations and p to the number of variables. The vector $\underset{\sim}{X}$ is then a pk-dimensional random variable from a multivariate normal distribution, $N_{pk}\left(\underset{\sim}{\mu}, \Sigma\right)$, whose parameters are defined as:

$$\underset{\sim}{\mu}_{pk\times 1} = \begin{bmatrix} \underset{\sim}{\mu}_1 \\ \underset{\sim}{\mu}_2 \\ \vdots \\ \underset{\sim}{\mu}_k \end{bmatrix} \tag{1a}$$

and

$$\Sigma_{pk\times pk} = \begin{pmatrix} \Sigma_{11} & \cdots & \Sigma_{1k} \\ \vdots & \ddots & \vdots \\ \Sigma_{k1} & \cdots & \Sigma_{kk} \end{pmatrix} \tag{1b}$$

The off diagonal elements indicate non-null covariances between populations because independence was not assumed. Each element in the diagonal of Σ represents the covariance matrix of the j^{th} population. The hypothesis of interest is: H_0: $\Sigma_{11} = \Sigma_{22} = ... = \Sigma_{kk}$ versus H_1: At least one covariance matrix Σ_{jj} differs from the others.

Statistics of the proposed tests were specified by the function of the ratio of generalized variances, as follows:

$$\lambda_{1(b)} = \frac{\max_j \left(\left| S_{jj} \right| \right)}{\min_j \left(\left| S_{jj} \right| \right)}; \qquad \lambda_{2(b)} = \frac{\max_j \left(\text{Trace} \left[S_{jj} \right] \right)}{\min_j \left(\text{Trace} \left[S_{jj} \right] \right)} \tag{2}$$

where S_{jj} are estimators of the sum of squares and products matrices. Each test was differentiated by the criterion used in the composition of the ratio, namely determinant or trace. Estimators of the sum of squares and products matrices of the j^{th} population (j=1, 2, ..., k) were only considered after the imposition of H_0 through the bootstrap method (Figure 1).

After defining the test statistics, the multivariate samples considering equicorrelation structure were generated in order to evaluate the performance of the new tests. Thus, specifying the matrix Σ, proceeded as follows. A global (population) correlation matrix R_b, where each block element in the diagonal represents a correlation structure referring to the j^{th} population (the area delimited by hatched lines) is given by:

$$R_b = \begin{bmatrix} 1 & \rho & \cdots & \rho & \rho & \rho & \rho & \cdots & \cdots & \rho \\ \rho & 1 & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \rho \\ \vdots & \vdots & \ddots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \rho \\ \rho & \cdots & \cdots & 1 & \ddots & \ddots & \ddots & \ddots & \ddots & \rho \\ \rho & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \rho \\ \rho & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \rho & \ddots & \ddots & \ddots & \ddots & \ddots & 1 & \rho & \cdots & \rho \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \rho & 1 & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \rho & \rho & \cdots & \rho & \cdots & \cdots & 1 \end{bmatrix}. \tag{3}$$

The global covariance matrix is obtained from the following relation:

$$\Sigma^* = V^{\frac{1}{2}} R_b V^{\frac{1}{2}}, \tag{4}$$

where $V^{\frac{1}{2}}$ is a diagonal matrix of the population standard deviations which are all equal to 1, without loss of generality.

After defining, samples were generated using the Monte Carlo method; an algorithm was developed using R software version 2.6.2 assuming multivariate normal distribution $N_{pk}(\underset{\sim}{0}, \Sigma^*)$. The algorithm first evaluated the Type I error rates of the related tests when applied to samples simulated under the null hypothesis H_0. Power was not measured at this stage because all diagonal block elements of Σ^* were considered equal.

Power rates were evaluated for those tests applied to samples simulated under the alternative hypothesis. The global population covariance matrix should be defined in such a way that each population matrix (diagonal blocks) would have to obey the heterogeneity settled in an intended value δ. In both situations, under null and alternative hypotheses, those matrices were evaluated in situations of low and high correlation, originated from structures represented by parametric values ρ fixed in 0.2 and 0.8.

Under H_1, the Σ_g matrix was defined as:

$$\Sigma_g = V^{\frac{1}{2}} R_b V^{\frac{1}{2}} \tag{5}$$

where

$$V = \text{diag}\left[1\,1 \ldots 1 \vdots \sqrt[2p]{d_2}\ \sqrt[2p]{d_2} \cdots \sqrt[2p]{d_2} \vdots \ldots \vdots \sqrt[2p]{d_k} \ldots \sqrt[2p]{d_k} \right] \tag{6}$$

Each block (6) was p-dimensional and

$$d_j = \sqrt[p]{1 + \frac{(j-1) \times (\delta - 1)}{k - 1}} \tag{7}$$

for j = 1, 2, …, k, and δ = 2, 4, 8, 16.

After defining the covariance matrix parameters Σ^* (δ=1) and Σ_g (δ>1), multivariate sample observations used in the evaluation of the considered tests were simulated. The N vector set generated formed the matrix of sample data:

$$X=\begin{bmatrix} \underset{\sim}{X}_1^t \\ \vdots \\ \underset{\sim}{X}_N^t \end{bmatrix}, \qquad (8)$$

where $\underset{\sim}{X}_l$ is a pk×1 vector and N is the sample size.

The construction the matrix was carried through using the vector of observations coming from the joint distribution of the k populations, generated according to the multivariate normal distribution,

$$\underset{\sim}{X}_\ell = \underset{\sim}{\underset{\sim}{F}}\,\underset{\sim}{Z}_\ell + \underset{\sim}{\mu} \quad (\ell = 1, 2, ..., N),$$

where F is the Cholesky factor (Bock, 1975) of the population covariance matrix Σ_g or Σ^*; and $\underset{\sim}{Z}_\ell$ is a kp × 1 vector of independent standard normal variables, generated by the inversion of the distribution function of the standard univariate normal in a random point U, U~U[0,1].

After obtaining the multivariate normal samples, the vector of sample means of kp variables was estimated by $\overline{\underset{\sim}{X}} = \left(\overline{\underset{\sim}{X}}_1^t, \overline{\underset{\sim}{X}}_2^t, \cdots, \overline{\underset{\sim}{X}}_k^t\right)^t$, where (j = 1, ..., k). The deviations of the vector of means were then computed in order to allow no influence of possible different averages between the k populations on the estimators of the covariance matrices. Thus, the inference was made considering the matrix of deviations X_d, defined as:

$$X_d = X - \underset{\sim}{1}\overline{\underset{\sim}{X}}^t, \qquad (9)$$

where $\underset{\sim}{1}$ is a vector N×1. The sum of squares and products matrix was estimated by

$$S = \left(X_d\right)^t Q X_d, \qquad (10)$$

where the projection matrix is given by

$$Q = I - \frac{\underset{\sim}{1}\underset{\sim}{1}^t}{n}, \qquad (11)$$

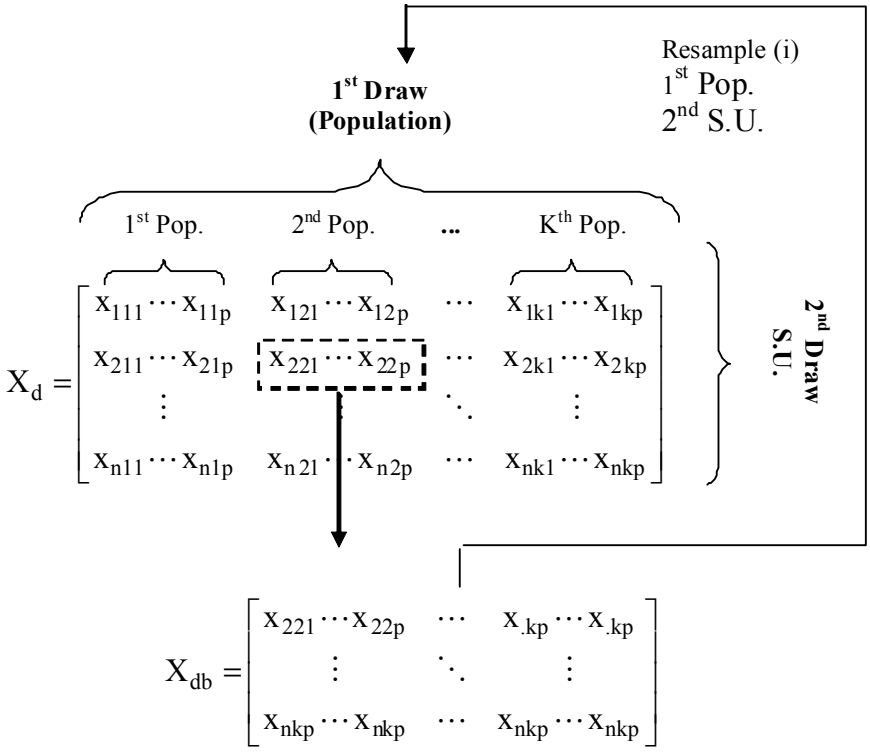
where $\underset{\sim}{1}$ is a vector of 1's (N × 1) and I (N × N) is an identity matrix.

Given the random sample X_d, 1,000 resamples were drawn. In each resample a new bootstrap sample X_{db}, was obtained, of which the matrix of sum of squares and products was estimated and named S_b^*, (b = 1, ..., 1,000). The elements of the diagonal blocks $\hat{S}_{b(jj)}^*$ (j = 1, 2, ..., k) of dimensionality (p × p) represent the estimators of the population sum of squares and products matrices, used to determine the statistics based on the generalized variances ratio. In each resample, values $\lambda_{1(b)}$ and $\lambda_{2(b)}$ were computed and compared with λ_1 and λ_2, obtained in the original sample of the Monte Carlo simulation. The critical region for the considered tests was constructed on the empirical distribution of the values of the statistics $\lambda_{1(b)}$ and $\lambda_{2(b)}$.

The critical stage of this procedure was setting the null hypothesis of equality of the population covariance matrices, surrounding all restrictions of the numerical methods of likelihood function maximization. The bootstrap method (Figure 1) considers as randomization unit the multivariate sample unit (SU) of each population considering p variables, thus characterizing H_0, which was set considering the dependence between the variables of all k-populations.

For each situation designed by the combination of the number of variables (p = 2, 3, 8), number of populations (k = 2, 8, 12), sample size (N = 20, 50, 100) and nominal values 1% and 5%, the empirical probabilities were computing by the times that the values of the statistics $\lambda_{1(b)}$ and $\lambda_{2(b)}$ were greater than or equal to the values λ_1 e λ_2 respectively. These values were obtained in the original sample in relation to the total number of bootstrap. The empirical type I error rates and power had been computed considering the proportion of times that H_0 was rejected by the nominal levels of 1% and considered 5% under H_0 and H_1, respectively.

Figure 1: Bootstrap Process Used to Estimate the Matrices of Sum of Squares and Products Coming from Dependent Multivariate Normal Populations



Results

Probabilities of Type I Error

Using a 95% confidence interval for the adopted nominal level, it can be inferred that the test was conservative if the value of the probability was less than the inferior limit. However, probability values contained between the interval limits demonstrated that the tests had provided effective control of type I error rates, that is, they have exact size. Table 2 contains the empirical Type I error rates, where was used the generalized variances obtained from the ratio of determinants.

Results in Table 2 show that the test based on the ratio of determinants submitted to low covariances ($\rho = 0.20$) controlled the Type I error with probabilities equal to or less than the nominal level set at 5% in almost all the evaluated situations. The exception occurred when the test was submitted to a high number of populations and had a small sample (N = 20). Increasing the value of the global correlation for $\rho = 0.80$ implied greater averages in the covariances and it was verified that the test was conservative in general. Results in Table 3 show the probabilities of the test considering the ratio of total variances (trace) evaluated with the same situations as that of the previous test.

In general, when using the ratio of total variances statistic, conservative results were obtained for those samples submitted with low global correlation ($\rho = 0.20$). In the high correlation cases ($\rho = 0.80$), the results of the test remain conservative, despite using determinant or trace.

Comparing the results of the generalized variances ratio tests presented in this article with the likelihood ratio tests considered by Bartlett (1937), Box (1949) and Krishnaiah (1975), it can be affirmed that the likelihood ratio tests are not adequate to compare dependent multivariate populations. Such affirmation is based on the fact that these tests have been compared with results presented by Jiang, et al. (1999) who

Table 2: Type I Error Rates for the Situations of Low and High Correlation, Evaluated In the Combinations of Number of Populations (k), Number of Variables (p) Considering the Test Defined by the Determinant Ratio

	k = 2			k = 8		k = 12
	p = 2	p = 3	p = 8	p = 2	p = 3	p = 8
N	ρ = 0.20					
20	0.039	0.04	0.029*	0.045	0.016*	0.036*
50	0.045	0.040	0.038	0.039	0.043	0.037
100	0.037	0.039	0.047	0.046	0.050	0.048
N	ρ = 0.80					
20	0.016*	0.015*	0.011*	0.011*	0.014*	0.018*
50	0.006*	0.014*	0.004*	0.005*	0.036*	0.039
100	0.004*	0.020*	0.006*	0.006*	0.052	0.036*

*empirical probabilities under the lower limit of the 95% confidence interval (0.037; 0.065)

Table 3: Type I Error Rates for the Situations of Low and High Correlation Evaluated In the Combinations of Number of Populations (k), Number of Variables (p) Considering the Test Defined By the Ratio of Traces

	k = 2			k = 8		k = 12
	p = 2	p = 3	p = 8	p = 2	p = 3	p = 8
N	ρ = 0.20					
20	0.042	0.044	0.032*	0.037*	0.031*	0.035*
50	0.039	0.035*	0.032*	0.046*	0.030	0.031*
100	0.040	0.036*	0.031*	0.037	0.034*	0.043
N	ρ = 0.80					
20	0.000*	0.000*	0.000*	0.001*	0.006*	0.001*
50	0.000*	0.000*	0.000*	0.001*	0.000*	0.001*
100	0.000*	0.000*	0.000*	0.001*	0.002*	0.000*

*empirical probabilities under the lower limit of the 95% confidence interval (0.037; 0.065)

used Monte Carlo simulations to verify that, in general, the likelihood ratio tests did not control type I error when N = 10, 15, 20, 25, 50, 75 and 100 under several correlation structures. However, asymptotic tests considered by the authors did control type I error for samples greater than 50 (N > 50) with probabilities close to the nominal level.

For the tests evaluated herein, it was observed that, for samples sizes smaller than 50 (N < 50), the tests were conservative under correlation ρ = 0.80. It is noteworthy that results obtained by other authors were related to bivariate populations only. This limitation was due to the maximization of the likelihood functions problem. Thus, for larger numbers of

populations and variables no results exist in the literature, regarding means of the likelihood theory that could be compared with the results of this present work. Results shown in Table 4 were obtained under the same configurations previously evaluated, but with the nominal level set to 1%. However, k = 8 populations on p = 12 variables were evaluated in particular, because this represents an extreme case and because cases considering k > 2 could not be found in the literature.

Similarly, by estimating a 95% confidence interval for this nominal level it can be verified whether or not the test was conservative. It was observed that the results for a 1% level of significance had the same pattern as results at the 5% level. Due to the similarity in results of the type I error rates, it is expected that the power function would be similar and coherent for both nominal levels 1% and 5%. It is worth noting that this similarity to the pattern of type I error rates between 1% and 5% also was observed in other configurations evaluated in k variables and p populations, thus, not all results are shown.

Power of the Multivariate Tests for Comparing Covariance Matrices of k Dependent Normal Populations

Power results corresponded to the empirical probabilities, which were obtained under the same configurations evaluated in the control of type I error rate discussed previously using the bootstrap method (see Figure 1). Results shown in Table 5 consider low global correlation ($\rho = 0.20$).

Analyzing the results in Table 5, it is observed that by increasing the degree of heterogeneity (δ) in all evaluated situations the power of the test suffers incrementally. However, for sample sizes N = 50 and greater, cases of $\delta = 8$ were similar to situations where $\delta = 16$. This suggests that - for any degree of heterogeneity ($\delta > 8$) between covariance matrices - the considered test was powerful when the population covariances had relatively low correlation.

An interesting result can be observed in power evaluation as the number of populations (k) rises. The power of the test presents few oscillations under a degree of heterogeneity (δ)

Table 4: Probabilities of Type I Error Considering the Generalized Variance Given By the Ratio of Determinants and the Ratio of Traces In the Two Evaluated Global Correlations with Nominal Significance Level 1%, k = 8 and p = 12

Ratio of Determinants		
N	$\rho = 0.20$	$\rho = 0.80$
20	0.0116	0.0033*
50	0.0050	0.0066
100	0.0150	0.0133
Ratio of Traces		
N	$\rho = 0.20$	$\rho = 0.80$
20	0.0100	0.0000*
50	0.0016*	0.0000*
100	0.0083	0.0000*

greater than 8 and sample sizes greater than 50, but does not hold for the case where a high number of variables are considered (p = 12). Regarding performance, when the number of variables (p) increases for a settled number of populations (k) and when bivariate populations (k = 2) are considered, the test becomes more sensitive, thus decreasing its power. Under low heterogeneity (δ), the test showed discrepant results for small samples (N = 20). Clearly, for a great number of variables (p = 8), the reduction of power was even more drastic. With respect to k = 8 populations, the number of variables caused less reduction of power, considering a maximum degree of heterogeneity of this study (δ = 16).

Regarding the effect of increasing the sample size (N), the power for cases with small samples was small, what agrees with empirical Type I error rate probabilities (Table 2). In such situations the test was revealed to be conservative. Note such deficiency of power, caused by the conservative property of the test (Table 2), does not invalidate it. Tests comparing k dependent population covariance matrices for many populations do not exist in the literature. Results shown in Table 6 emphasize the performance of the generalized variances test as represented by the determinants ratio under a global correlation (ρ = 0.8) and considering the same situations evaluated previously.

Table 5: Power Empirical Values for the Bootstrap Generalized Likelihood Ratio Test for Different Sample Sizes (n), Numbers of Populations (k), Variables (p), Degrees of Heterogeneity (δ) Under Low Global Correlation (ρ = 0.20)

N	k = 2 p = 2	k = 2 p = 3	k = 2 p = 8	k = 8 p = 2	k = 8 p = 3	k = 8 p = 12
δ = 2						
20	0.150	0.080	0.050	0.090	0.090	0.070
50	0.370	0.230	0.220	0.200	0.180	0.080
100	0.610	0.650	0.470	0.430	0.270	0.120
δ = 4						
20	0.490	0.320	0.120	0.300	0.180	0.100
50	0.900	0.790	0.820	0.820	0.550	0.190
100	0.970	0.980	0.950	0.980	0.930	0.520
δ = 8						
20	0.810	0.600	0.150	0.710	0.430	0.090
50	0.970	1.000	0.950	0.980	0.910	0.470
100	0.980	1.000	0.950	0.980	1.000	0.900
δ = 16						
20	0.95	0.890	0.220	0.950	0.760	0.180
50	0.98	1.000	0.950	0.980	1.000	0.730
100	0.980	1.000	0.950	0.980	1.000	0.990

Table 6: Power Values for the Bootstrap Generalized Likelihood Ratio Test for Different Sample Sizes (n), Numbers of Populations (k), Variables (p), Degrees of Heterogeneity (δ) Under Low Global Correlation (ρ = 0.80)

N	k = 2 p = 2	k = 2 p = 3	k = 2 p = 8	k = 8 p = 2	k = 8 p = 3	k = 8 p = 12
δ = 2						
20	0.110	0.080	0.030	0.020	0.020	0.050
50	0.380	0.230	0.100	0.180	0.030	0.060
100	0.730	0.590	0.330	0.410	0.110	0.100
δ = 4						
20	0.500	0.370	0.070	0.170	0.040	0.070
50	0.930	0.780	0.720	0.760	0.200	0.220
100	0.980	0.950	0.700	0.990	0.500	0.580
δ = 8						
20	0.890	0.600	0.090	0.580	0.090	0.110
50	0.990	0.980	0.950	0.960	0.450	0.480
100	0.990	0.980	0.980	1.000	0.900	0.930
δ = 16						
20	1.000	0.870	0.210	0.920	0.180	0.200
50	1.000	1.000	0.950	1.000	0.800	0.840
100	1.000	1.000	0.980	1.000	0.980	0.990

Comparing results in Tables 5 and 6, observe that an increment of the degree of heterogeneity (δ) yields an increment of the power values. However, this increment was small, since for N = 20,50 and degree of heterogeneity δ = 4, the test remains not so powerful.

In a general manner, the number of populations (k) is related to a reduction of power, retaining the same highlighted properties of when the population covariance matrices presented, in average, low correlation (Table 5). However, in comparison to results shown in Table 6, it is suggested that increasing the global correlation yields an even greater reduction in power. Therefore, it may be concluded that increasing the number of populations (k) where population covariances present high correlations results in a great loss of power. In turn, when the number of variables (p) is increased with a set number of populations (k), the test retained the same properties. The increment of the global correlation from ρ = 0.20 for ρ = 0.80 also did not affect the power of the test when the number of variables (p) was increased for a set number of populations (k).

Regarding the sample size, results shown in Table 6 agree with previous results. It is advisable to use the considered test to deal with small samples when comparing bivariate populations (k = 2) with a degree of heterogeneity greater than 8. Such a test can be used for other cases; however, exploratory studies of the populations must be done.

Conclusion

Generalized variances tests controlled type I error according to a set nominal level. Tests based on the ratio of traces, in general, provided more conservative results. The simulation results clearly demonstrated that the procedure based on the determinant could more effectively control

the type I error rate to α, particularly when the off-diagonal elements of R_j, the correlation matrix corresponding to Σ_j, are small. Power of the generalized variances tests was reduced by increasing the number of variables and populations in both global correlations evaluated.

References

Bock, R. D. (1975). Multivariate Statistical Methods in Behavioral Research, New York: McGraw-Hill.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Statistical Society*, *160*, 268-282.

Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika*, *36*, 317-346.

Finney, D. J. (1938). The distribution of the ratio of estimates of the two variances in a sample from a normal bivariate population. *Biometrika*, *30*, 190-192.

Hall, P., Horowitz,J. L., & Jiang, B. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika*, *82*, 561-574.

Jiang, G., Sarkar, K. S., & Hsuan, F. (1999). A likelihood ratio test and its modifications for the homogeneity of the covariance matrices of dependent multivariate normals. *Journal of Statistical Planning and Inference*, *81*, 95-111.

Krishnaiah, P. R. (1975). Tests for the equality of the matrices covariance of correlated multivariate normal populations. In J. N. Srivastava (Ed.), *A survey statistical design and linear models*, 355-366. Amsterdam: North-Holland.

Manly, B. F. J. (1997). Randomization Bootstrap and Monte Carlo Methods in Biology, New York: Chapman & Hall.

Morgan, W. A. (1939). A test for the significance of the difference between the two variances in a sample from a normal bivariate population. *Biometrika*, *31*, 9-12.

O'Brien, C. O. P. (1992). Robust Procedures for Testing Equality of Covariance Matrices. *Biometrics*, *48*, 819-827.

Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika*, *31*, 9-12.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Roy, S. N., & Potthoff, R. F. (1958). Confidence bounds on vector analogues of the ratio of means and the ratio of variances for two correlated normal variates and some associated tests. *Annals of Mathematical Statistics*, *29*, 829-841.

Smith, P. L., & Kshirsagar, A. M. (1985). Testing for the equality of the variance-covariance matrices of two jointly normal vector variables. *Biometrika*, *27*, 581-589.

Wang, W., Hwang, J. T. G., & Dasgupta, A. (1999). Statistical tests for multivariate bioequivalence, *Biometrika*, *86*, 395-402.

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 379-387

1538 – 9472/10/$95.00

Notes on Hypothesis Testing under a Single-Stage Design in Phase II Trial

Kung-Jong Lui
San Diego State University,
San Diego, CA USA

A primary objective of a phase II trial is to determine future development is warranted for a new treatment based on whether it has sufficient activity against a specified type of tumor. Limitations exist in the commonly-used hypothesis setting and the standard test procedure for a phase II trial. This study reformats the hypothesis setting to mirror the clinical decision process in practice. Under the proposed hypothesis setting, the critical points and the minimum required sample size for a desired power of finding a superior treatment at a given α-level are presented. An example is provided to illustrate how the power of finding a superior treatment by accounting for a secondary endpoint may be improved without inflating the given Type I error.

Key words: Phase II trial, Type I error, power, union-intersection test, sample size, equivalence.

Introduction

One of the primary objectives in a phase II trial for a new anti-cancer treatment is to make a preliminary determination on whether the treatment has sufficient activity or benefits against a specified type of tumor to warrant its further development. Based on subjective knowledge, researchers commonly choose two response rates in advance p_0 and p_1 (where $0 < p_0 < p_1 < 1$) for the uninteresting and desirable levels, respectively. Test hypotheses: $H_0: p \le p_0$ versus $H_a: p \ge p_1$ (Simon, 1989; Lin, Allred & Andrews, 2008; Lu, Jin & Lamborn, 2005) are considered using p_1 to determine the minimum required sample size for a desired power $1-\beta$ of rejecting $H_0: p \le p_0$ at a nominal α-level when $p = p_1$. This hypothesis setting can cause clinicians to misinterpret their findings that rejecting the null hypothesis $H_0: p \le p_0$ is equivalent to supporting the alternative hypothesis $H_a: p \ge p_1$ and vice versa (Storer, 1992).

Kung-Jong Lui is a Professor in the Department of Mathematics and Statistics. Email him at: kjl@rohan.sdsu.edu.

Despite employing a large sample size to meet a desired power, the probability of excluding a potentially interesting treatment from further consideration can still be large. To illustrate the above points, for example, consider testing $H_0: p \le 0.30$ versus $H_a: p \ge 0.50$ at the 0.05 level. When using the common sample size calculation formula for a desired 90% power of rejecting $H_0: p \le 0.30$ for p = 0.50 at the 0.05-level, the minimum required sample size is determined to be 49 patients.

Suppose that ($x =$) 20 patients respond among these ($n =$) 49 patients (i.e., the sample proportion response $\hat{p} = 20/49$ = 0.408). Using these data, the p-value for testing $H_0: p \le 0.30$ is 0.049 (on the basis of normal approximation) and thereby, H_0 is rejected at the 0.05 level. Note that because $\hat{p} = 20/49$ (= 0.41) is less than 0.50, there is no evidence that the underlying response rate p is larger than 0.50. Conversely, there is statistically significant evidence, given $\hat{p} = 20/49$, to indicate that the underlying response rate p is less than the desirable level 0.50 at the 10% level for testing

$$H_0: p \ge 0.50$$

versus

$$H_a: p < 0.50.$$

Furthermore, when a treatment has the response rate p = 0.35 (which is larger than the uninteresting level p = 0.30) and is of potential interest, it can be shown that the probability of terminating this treatment for further consideration by not rejecting $H_0: p \le 0.30$ is approximately 80%.

The above concerns and criticisms are partially due to the fact that the complement of $\{p|p \le p_0\}$ is not the set $\{p|p \ge p_1\}$ and there is no explicit instruction about what should be done when the underlying response rate p falls in the borderline region $\{p|p_0 < p < p_1\}$. This motivates the recent development of a design with three outcomes, including an outcome allowable to account for other factors, including toxicity, cost or convenience, when making a decision (Storer, 1992; Sargent, Chan & Goldberg, 2001; Hong & Wang, 2007). One intuitive and logical justification of this practice is that - if the response rate of a new treatment was not much different from that of the standard treatment - it would be reasonable to recommend the new treatment for further study if the new treatment was less toxic, cheaper and/or easier to administer.

Treating both $H_0: \{p|p \le p_0\}$ and $H_a: \{p|p \ge p_1\}$ as two separate competing null hypotheses, Storer (1992) proposed a three-outcome design to accommodate the situation in which one might reject neither H_0 nor H_a and he suggested sample size calculation based on $P(X \ge r_u|H_0) \le \alpha$, $P(X \le r_l|H_a) \le \beta$, and $P($ rejecting $H_i|p_m) \le \gamma$ for $i = 0, a$, where r_u and r_l are minimum and the maximum critical points satisfying the above probability constrains and where $p_m \approx (p_0 + p_1)/2$.

On the basis of Simon's setting (1989) and the normal approximation for the binomial distribution, Sargent, Chan & Goldberg (2001) proposed a three-outcome test procedure with an inconclusive region in which neither H_0 nor H_a were rejected and they discussed sample size calculation for given errors of α and β, and the minimum probabilities of concluding correctly. Hong & Wang (2007) further extended sample size calculation to accommodate a two-sample randomized comparative trial. In fact, the design suggested by Sargent, Chan & Goldberg (2001) can be expressed in terms of Storer's setting (1989) by treating H_0 and H_a as two competing null hypotheses in the following:

(1) testing $H_0: p \le p_0$ (versus $p > p_0$) at α-level, and rejecting H_0 when $X \ge r_u$ where r_u is the minimum point satisfying $P(X \ge r_u|p=p_0) \le \alpha$;
(2) testing $H_a: p \ge p_1$ (versus $p < p_1$) at β-level, and rejecting H_a when $X \le r_l$, where r_l is the maximum point satisfying $P(X \le r_l|p=p_0) \le \beta$.

The inconclusive region then simply corresponds to the set of sample points $\{r_l < X < r_u\}$. Based on the normal approximation, it can be shown that the inconclusive region consists of

$$\{\mathrm{X}|\, np_1 - Z_\beta\sqrt{np_1(1-p_1)}\text{ -0.5} < X < np_0 + Z_\alpha\sqrt{np_0(1-p_0)} + 0.5\},$$

where Z_α is the upper $100(\alpha)^{\text{th}}$ percentile of the standard normal distribution. Note that this inconclusive region is a function of errors α, β, and the sample size, which are all operating parameters of the statistical test procedure rather than the biological characteristics of patient response to treatments. Various choices of α, β, or the sample size can lead to obtain different inconclusive regions despite that the underlying p_1 and p_0 are fixed. This is not appealing because the inconclusive region should represent the values falling in the borderline between the uninteresting and desirable levels and should be related to the biological aspects. Furthermore, it is possible that both $H_0: p \le p_0$ and $H_a: p \ge p_1$ may be rejected in the design proposed by Sargent, Chan & Goldberg (2001); in this case, the above inconclusive region will no longer exist. This

can occur even when the sample size is moderate and both α and β errors are controlled.

To clarify this point, consider the above example of testing $H_0: p \le 0.30$ versus $H_a: p \ge 0.50$. Given ($x =$) 20 patients with response among ($n =$) 49 patients, $H_0: p \le 0.30$ can be rejected at $\alpha =$ 0.05 level and $H_a: p \ge 0.50$ would be rejected at $\beta =$ 0.10 level. When choosing $\alpha =$ 0.05 and $\beta =$ 0.10, by definition $r_u < r_l$ in this case and the inconclusive region does not exist. There is no discussion on what action to take when both $H_0: p \le p_0$ and $H_a: p \ge p_1$ are rejected in the three-outcome design as proposed previously (Storer, 1992; Sargent, et al., 2001; Hong & Wang, 2007).

When determining in practice whether a new treatment warrants further study at the end of a phase II trial the decision is almost always based on multiple risk/benefit considerations rather than the testing result of a single primary endpoint, especially when no clear decision can be derived from the testing result. In other words, unless the response rate of the new treatment can be shown to be different from that of the standard treatment by a magnitude of clinical importance, relevant factors are incorporated into the determination of whether the new treatment should be studied further. Thus, it is desirable to design a test procedure that can mirror the clinical decision process in reality.

To avoid distracting readers' attention from the main focus of this article, discussion is restricted to a single-stage design. Under the proposed setting, the critical points and the minimum required sample size for a desired power of finding a superior treatment in a variety of situations are presented. Furthermore, using an idea suggested by Lin, Allred and Andrews (2008) and Lu, Jin and Lamborn (2005), an example is included to illustrate how the power of detecting a superior treatment may be improved by considering a secondary endpoint without inflating the given Type I error. Finally, another alternative procedure is considered and its difference, advantage, and disadvantage are noted and compared with the proposed procedure.

Notation and Hypothesis Testing

Consider a phase II trial in which a random sample of size *n* patients is taken from a studied population and assigned to receive a new treatment under study. Suppose that *x* out of *n* patients are obtained with objective (or primary) response. Let p_0 denote the objective response rate determined from the historical data for the standard treatment. Let δ denote the level of difference such that, if the objective response rate p is larger than $p_u = p_0 + \delta$, the new treatment is regarded as superior to the standard treatment and hence is warranted for further study.

Similarly, if the objective response rate p is less than $p_l = p_0 - \delta$, the new treatment is regarded as inferior to the standard treatment and is terminated from further investigation. Recall that in the standard setting, statistical significance against $H_0: p \le p_0$ does not provide information on how large the difference p - p_0 is between the new and standard treatments. By contrast, statistical significance evidence to support that $p > p_0 + \delta$ (i.e., the new treatment is larger than the standard treatment by a magnitude δ of clinical significance) will provide better evidence. Conversely, when statistically significant evidence exists that the new treatment is inferior to the standard treatment (i.e., $p < p_0 - \delta$), the new treatment may be excluded from further consideration for ethical reasons. This occurrence will not be known unless the data against the hypothesis $p \ge p_0 - \delta$ is examined. Thus, despite the fact that the main interest in a phase II trial is to find a potentially promising treatment, the critical region may also include the sample points to test the hypothesis $p \ge p_0 - \delta$. However, the calculation of sample size required for power of detecting a given $p (< p_0 - \delta)$ is of no practical interest. Defining $p_l = p_0 - \delta$ and $p_u = p_0 + \delta$, therefore, the hypotheses considered in testing are:

$$H_0: p_l \le p \le p_u$$

versus (1)

$$H_a: p > p_u \text{ or } p < p_l .$$

$H_0: p_l \le p \le p_u$ will be rejected at the α-level if $x \ge x_u(\alpha_1)$ or $x \le x_l(\alpha_2)$, where $\alpha = \alpha_1 + \alpha_2$, $x_u(\alpha_1)$ is the minimum point such that

$$P(X \ge x_u(\alpha_1) \mid p_u) = \sum_{x=x_u(\alpha_1)}^{n} \binom{n}{x} p_u^x (1-p_u)^{n-x} \le \alpha_1, \quad (2)$$

and $x_l(\alpha_2)$ is the maximum point such that

$$P(X \le x_l(\alpha_2) \mid p_l) = \sum_{x=0}^{x_l(\alpha_2)} \binom{n}{x} p_l^x (1-p_l)^{n-x} \le \alpha_2. \quad (3)$$

Note that the hypothesis setting (1) is simply a switch between the null and alternative hypotheses when testing equivalence (Dunnett & Gent, 1997; Westlake, 1979; Liu & Weng, 1995; Liu & Chow, 1992; Hauck & Anderson, 1984; Lui, 1997a, 1997b; Lui & Cumberland, 2001a, 2001b). Note also that the above test procedure for (1) is a union-intersection test (Casella & Berger, 1990). When making an error in recommending an ineffective or harmful treatment for phase III trial is considered more serious than making an error of missing a potentially interesting treatment, an investigator may wish to choose $\alpha_1 \le \alpha_2$.

For a given true value $p \in \{p \mid p > p_u\}$, the power is equal to

$$\Phi(n, p, \alpha_1, \alpha_2, \delta) = P(X \le x_l(\alpha_2) \mid p) + P(X \ge x_u(\alpha_1) \mid p). \quad (4)$$

Thus, given p, α_1, α_2, and δ, a trial-and-error procedure can be applied to determine the critical points: $x_l(\alpha_2)$ and $x_u(\alpha_1)$, as well as the minimum required sample size n for a desired power $1-\beta$ based on (4) such that

$$\Phi(n, p, \alpha_1, \alpha_2, \delta) \ge 1-\beta. \quad (5)$$

Sample Size Determination and Critical Points

Programs were written in SAS (1990) to find the minimum required sample size n satisfying equation (5). For illustration purposes, $\delta =$ 2.5% was arbitrarily chosen for the following discussion. Table 1 summarizes the critical points $x_u(\alpha_1)$, $x_l(\alpha_2)$, and the minimum required sample size n for $\alpha_1 = \alpha_2 = 0.10$ calculated from $\Phi(n, p, \alpha_1, \alpha_2) \ge 1-\beta$ (5) for a desired power $1-\beta = 0.80, 0.90$ in testing

$$H_0: p_l \le p \le p_u$$

versus

$$H_a: p > p_u \text{ or } p < p_l ,$$

where $p_l = p_0 - \delta$, $p_u = p_0 + \delta$, $\delta =$ 2.5%; $p_0 = 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75$; and $p = p_0 + 0.15, p_0 + 0.20$.

For example, consider testing

$$H_0: 0.325 \le p \le 0.375 \text{ (i.e., } p_0 = 0.35)$$

versus

$$H_a: p > 0.375 \text{ or } p < 0.325$$

at levels of $\alpha_1 = \alpha_2 = 0.10$. If the desired power for rejecting H_0 when the underlying objective response rate p equals 0.50 is 80%, for example, based on equation (5), 77 patients would be required. Furthermore, Table 1 shows that if $(x_u(\alpha_1)) = 35$ or more patients are obtained with an objective response out of the 77 patients, then the new treatment would be recommended for further study.

On the other hand, if 19 or less patients are obtained with objective responses, the new treatment would be terminated from further consideration. Finally, if the number of patients with objective responses falls between 20 and 34, other factors would be considered to determine whether the experimental treatment warrants further study. Table 2 summarizes the corresponding critical points $x_l(\alpha_2)$, $x_u(\alpha_1)$

and the minimum required sample size *n* for α_1 = 0.05 and α_2=0.15 in the same configurations as those considered in Table 1.

Discussion

Multiple factors are almost always accounted for at the end of a phase II trial to determine whether a new treatment warrants further study unless there is a clear cut decision in the testing results. The test procedure proposed herein has the advantage of resembling the actual clinical decision process more closely than the standard test procedure. By contrast, in Simon's setting, the determination of a new treatment for further study may completely depend on the testing result of a single primary point, but this may not be the case in practice. Furthermore, in the three-outcome design, the inconclusive region depends on the operating characteristics, such as errors α, β, and the sample size, of a test procedure. Thus, the inconclusive region can change or may not even exist for different given values of these parameters even when the underlying objective response rate is fixed. For this reason the inconclusive region is defined here in terms of biological equivalence. Based on the proposed hypothesis setting (1), it is possible to control both the errors of recommending a non-superior treatment and of terminating a non-inferior treatment to be less than a given error-level.

When there is no statistical evidence against the hypothesis $H_0: p \in [p_l, p_u]$ based on the primary endpoint, a reasonable and appealing action can be to consider a secondary endpoint to improve power. For example, in traditional phase II trials, the total response (TR) rate, the sum of the complete response (CR) rate and the partial response (PR) rate, is often used as the objective (or primary) response rate *p*. Because CR is generally rare for many tumors, even a small increase in the number of CRs can be important in evaluation of the efficacy of a treatment. Thus, clinicians will welcome a decision rule that accepts a new treatment for further study based on an improved CR rate even when the treatment does not achieve the desirable objective response rate of TR (Lin, Allred & Andrews, 2008; Lu, Jin & Lamborn, 2005).

Table 1: The critical points $x_l(\alpha_2)$, $x_u(\alpha_1)$ and the minimum required sample size *n* calculated from $\Phi(n, p, \alpha_1, \alpha_2) \geq 1-\beta$ in equation (5) for a desired power $1-\beta = 0.80$, 0.90 in testing $H_0: p_l \leq p \leq p_u$ versus $H_a: p > p_u$ or $p < p_l$ where $p_l = p_0 - \delta$, $p_u = p_0 + \delta$, $\delta =$ 2.5%; $p_0 = 0.15$, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75; $p = p_0 + 0.15$, $p_0 + 0.20$; $\alpha_1 = 0.10$ and α_2 = 0.10.

p_0	p	n	$x_l(\alpha_2)$	$x_u(\alpha_1)$
$1-\beta = 0.80$				
0.15	0.30	51	2	13
	0.35	31	1	9
0.25	0.40	68	10	24
	0.45	36	4	14
0.35	0.50	77	19	35
	0.55	41	9	20
0.45	0.60	77	26	43
	0.65	37	11	22
0.55	0.70	73	32	48
	0.75	36	14	25
0.65	0.80	59	31	45
	0.85	30	14	24
0.75	0.90	39	24	34
	0.95	16	8	15
$1-\beta = 0.90$				
0.15	0.30	79	5	19
	0.35	45	2	12
0.25	0.40	94	15	32
	0.45	52	7	19
0.35	0.50	109	28	48
	0.55	53	12	25
0.45	0.60	105	37	57
	0.65	54	17	31
0.55	0.70	101	46	65
	0.75	50	21	34
0.65	0.80	83	45	62
	0.85	41	21	32
0.75	0.90	61	39	52
	0.95	22	12	20

Table 2: The critical points $x_l(\alpha_2)$, $x_u(\alpha_1)$ and the minimum required sample size n calculated from $\Phi(n, p, \alpha_1, \alpha_2) \geq 1-\beta$ in equation (5) for a desired power $1-\beta =$ 0.80, 0.90 in testing $H_0: p_l \leq p \leq p_u$ versus $H_a: p > p_u$ or $p < p_l$, where $p_l = p_0 - \delta$, $p_u = p_0 + \delta$, $\delta =$ 2.5%; $p_0 =$ 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75; $p = p_0 + 0.15, p_0 + 0.20$; $\alpha_1 = 0.05$ and α_2=0.15.

p_0	p	n	$x_l(\alpha_2)$	$x_u(\alpha_1)$
$1-\beta = 0.80$				
0.15	0.30	73	5	19
	0.35	41	2	12
0.25	0.40	92	16	33
	0.45	48	7	19
0.35	0.50	102	27	47
	0.55	50	12	25
0.45	0.60	103	38	58
	0.65	53	18	32
0.55	0.70	95	44	63
	0.75	48	21	34
0.65	0.80	81	45	62
	0.85	41	21	33
0.75	0.90	56	36	49
	0.95	26	15	24
$1-\beta = 0.90$				
0.15	0.30	102	8	25
	0.35	55	3	15
0.25	0.40	121	21	42
	0.45	66	10	25
0.35	0.50	136	38	61
	0.55	71	18	34
0.45	0.60	140	52	77
	0.65	72	25	42
0.55	0.70	129	61	84
	0.75	64	28	44
0.65	0.80	110	62	83
	0.85	53	28	42
0.75	0.90	78	51	67
	0.95	32	20	29

When studying the efficacy of a treatment for brain tumors the TR rate can be small as well. In this case, the objective response can be stabilization disease (SD) progression for six months after post-treatment initiation, while the secondary endpoint can be either CR or PR. For both of the above examples, a critical region may be found based on the objective and secondary responses such that if the objective response rate cannot be used to decide whether a new treatment warrants further study, an opportunity may still exist to justify the acceptance of the new treatment based on its secondary response rate subject to the originally given α_1 error. To illustrate this point, consider the example for patients with glioblastomas. On the basis of the standard for the North American Brain Tumor Consortium (NABTC), interest lies in determining whether the objective response rate of SD increases from $p_0 = 0.15$ to $p = 0.35$ (Lu, Jin & Lamborn, 2005). Thus, testing

$$H_0: 0.125 \leq p \leq 0.175 \text{ (with } \delta = 2.5\%)$$

versus

$$H_a: p > 0.175 \text{ or } p < 0.125$$

is considered. From equation (5), the minimum required number of patients is determined to be 31 patients for a desired power of 80% when p = 0.35 at ($\alpha_1 = \alpha_2 =$) 0.10-level and the corresponding critical points $x_l(\alpha_2)$ and $x_u(\alpha_1)$ are 1 and 9, respectively (Table 1).

When no evidence exists to claim the experimental treatment to be superior (i.e., $p > 0.175$) to the standard treatment based on the objective response rate of SD, for example, the experimental treatment may be still determined to warrant further study. This could occur if the secondary response rate, p_S, that the tumor shrinkage is sufficient to be regarded as either CR or PR for a 6-month interval is larger than 0.05.

Let x_s denote the number of patients with the secondary response among 31 patients. While keeping the above critical point $x_u(\alpha_1)$ for the objective response of SD, SAS programs are written to search for the secondary endpoint

for the critical point x_{CS}, which is the minimum point x_S such that the probability $P(X \geq 9 \text{ or } X_s \geq x_S | p_u = 0.175, p_S = 0.05) \leq 0.10$. The critical point, x_{CS}, is 5 if an observation (x, x_S) = (8, 6) is obtained. Although the number (x = 8) of patients with the objective response of SD is not ≥ 9, the experimental treatment may be recommended for further development because the number of (x_S = 6) patients with the secondary response is above the critical point (x_{CS}=5). In fact, the joint power for given values p and p_S based on the trinomial distribution can also be calculated:

$$P(X \geq 9 \text{ or } X_s \geq 5 | p, p_S) = \sum_i \sum_j 1_{\{i+j \geq 9 \text{ or } i \geq 5\}} \frac{31!}{i!\, j!(n-i-j)!} \times p_S^i (p - p_S)^j (1-p)^{(31-i-j)} \quad (6)$$

where the indicator function, $1_{\{condition\}}$, equals 1 if the condition in braces is true, and equals 0 otherwise.

For example, when $p = 0.35$ and p_S = 0.20, the joint power obtained from (6) ≈ 0.88, which is larger than the original desired actual power $P(X \geq 9 | p = 0.35) \approx 0.81$ exclusively based on the objective response by approximately 7%. Note that because the binomial distribution is discrete, the true Type I error $P(X \geq 9 | p_u = 0.175)$ based on the objective response is actually equal to 0.079, which is less than the nominal ($\alpha_1 = 0.10$) level. This is the reason why the critical region can be expanded from $\{X \geq 9\}$ to $\{X \geq 9 \text{ or } X_S \geq 5\}$ to increase power without the necessity of inflating the given α_1 error. Conaway & Petroni (1995) proposed methods for designing group sequential phase II trials with two binary endpoints.

Conaway & Petroni (1995) also focused discussion on the situation in which a new treatment is recommended for further study when the new treatment has both a high response and lower toxicity. By contrast, consider the situation in which the new treatment is recommended for further study if the new treatment has either a high objective response rate or a high secondary response rate. Thus, Conaway & Petroni's results cannot be applicable to the situations discussed here.

It may be shown that

$$P(X \geq x | p) \left(= \sum_{X=x}^{n} \binom{n}{X} p^x (1-p)^{n-x}\right) \leq \alpha^*$$

if and only if the 100(1-α^*)% lower confidence limit (LCL) (one-sided), given by $x/(x+(n-x+1)F_{2(n-x+1),2x,\alpha^*})$, falls above the underlying response rate p, where $F_{2(n-x+1),2x,\alpha^*}$ is the upper 100(α^*)th percentile of the central F-distribution with degrees of freedom $2(n-x+1)$ and $2x$, respectively (Casella & Berger, 1990; Lui, 2004). Similarly, it can be shown that $P(X \leq x | p) \leq \alpha^*$ if and only if the 100(1- α^*)% upper confidence limit (UCL) (one-sided), given by

$$\frac{\{(x+1)F_{2(x+1),2(n-x),\alpha^*}\}}{\{(n-x)+(x+1)F_{2(x+1),2(n-x),\alpha^*}\}}$$

falls below p. Thus, the hypothesis setting and test procedure defined in (1-3) is equivalent to the decision procedure defined as follows: when the UCL with $\alpha^* = \alpha_2$ falls below $p_l (= p_0 - \delta)$, the new treatment is terminated; when the LCL with $\alpha^* = \alpha_1$ falls above $p_u (= p_0 + \delta)$, the new treatment warrants further consideration; when neither of the above conditions hold relevant factors are accounted for in the final decision. Compared with hypothesis testing, the use of confidence intervals to present the testing results may shed light on the magnitude of the difference between the two treatments under comparison.

Rather than excluding a new treatment from further consideration when it is shown to be inferior (i.e., $p < p_0 - \delta$) to the standard treatment in the procedure proposed, an alternative procedure can be considered by including a new treatment into further

consideration only when it is shown to be non-inferior to the latter (i.e., $p>p_0-\delta$). That is, the following design may be employed: (1) for the LCL with a given $\alpha^*=\alpha_1$ falling below $p_0-\delta$, the new treatment is excluded from further consideration; (2) for the LCL with $\alpha^*=\alpha_1$ falling into [$p_0-\delta$, $p_0+\delta$], accounting for other factors; and (3) for the LCL falling above $p_0+\delta$, the new treatment is recommended for further study. To avoid missing a potentially useful treatment when a new treatment for a specified type of cancer is hard to find, the hypothesis setting and test procedure (1-3) described herein may be employed to terminate a new treatment only when it is shown to be inferior to the standard treatment. To alleviate the concern of including an inferior treatment for phase III trials, a large value for α_2 may be chosen (e.g., 0.15) in (3); on the other hand, when new experimental treatments are easier to find, the alternative decision procedure, including only those treatments shown to be non-inferior to the standard treatment for further consideration, can be of potential use.

In summary, limitations in the commonly-used hypothesis setting and the recently proposed three-outcome design have been described. The hypothesis testing has been reformatted and a test procedure proposed to more closely resemble the clinical decision process. The minimum required sample size for a desired power of finding a superior treatment at a given α-level has been presented and the corresponding critical points in a variety of situations provided. Discussion and an example were used to illustrate how power may be improved by accounting for the secondary endpoint without inflating the given Type I error in the proposed test procedure. Also included was a discussion on an alternative procedure and for which situations in which this procedure can be of use. The findings and the discussion should be helpful for clinicians when exploring a new treatment in a phase II trial.

References

Casella, G., & Berger, R. L. (1990). *Statistical Inference*. Belmont, CA: Duxbury.

Conaway, M. R., & Petroni, G. R. (1995). Bivariate sequential designs for phase II trials. *Biometrics*, *51*, 656-664.

Dunnett, C. W., & Gent, N. (1997). Significance testing to establish equivalence between treatments, with special reference to data in the form of 2x2 tables. *Biometrics*, *33*, 593-602.

Hauck, W. W., & Anderson, S. (1984). A new statistical procedure for testing equivalence in two group comparative bioavailability trials. *Journal of Pharmacokinetics and Biopharmaceutics*, *12*, 83-91.

Hong, S., & Wang, Y. (2007). A three-outcome design for randomized comparative phase II clinical trials. *Statistics in Medicine*, *26*, 3525-3534.

Lin, X., Allred, R., & Andrews, G. (2008). A two-stage phase II trial design utilizing both primary and secondary endpoints. *Pharmaceutical Statistics*, *7*, 88-92.

Liu, J.-P., & Chow, S.-C. (1992). Sample size determination for the two one-sided tests procedure in bioequivalence. *Journal of Pharmacokinetics and Biopharmaceutics*, *20*, 101-104.

Liu, J.-P., & Weng, W.-S. (1995). Bias of two one-sided tests procedures in assessment of bioequivalence. *Statistics in Medicine*, *14*, 853-861.

Lu, Y., Jin, H., & Lamborn, K. R. (2005). A design of phase II cancer trials using total and complete response endpoints. *Statistics in Medicine*, *24*, 3155-3170.

Lui, K.-J. (1997a). Sample size determination for repeated measurements in bioequivalence test. *Journal of Pharmacokinetics and Biopharmaceutics*, *25*, 507-513.

Lui, K.-J. (1997b). Exact equivalence test for risk ratio and its sample size determination under inverse sampling. *Statistics in Medicine*, *16*, 1777-1786.

Lui, K.-J. (2004). *Statistical Estimation of Epidemiological Risk*. New York: Wiley.

Lui, K.-J., & Cumberland, W. G. (2001a). A test procedure of equivalence in ordinal data with matched-pairs. *Biometrical Journal*, *43*, 977-983.

Lui, K-J., & Cumberland, W. G. (2001b). Sample size determination for equivalence test using rate ratio of sensitivity and specificity in paired-sample data. *Controlled Clinical Trials*, *22*, 373-389.

Sargent, D. J., Chan, V., & Goldberg, R. M. (2001). A three-outcome design for phase II clinical trials. *Controlled Clinical Trials*, *22*, 117-125.

SAS Institute Inc. (1990). *SAS Language Version* 6, 1st edition. Cary, North Carolina SAS Institute.

Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*, *10*, 1-10.

Storer, B. E. (1992). A class of phase II designs with three possible outcomes. *Biometrics*, *48*, 55-60.

Westlake, W. J. (1979). Statistical aspects of comparative bioavailability trials. *Biometrics*, *35*, 273-280.

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 388-402

1538 – 9472/10/$95.00

Effect of Measurement Errors on the Separate and Combined Ratio and Product Estimators in Stratified Random Sampling

Housila P. Singh Namrata Karpe
Vikram University,
Ujjain India

Separate and combined ratio, product and difference estimators are introduced for population mean μ_Y of a study variable Y using auxiliary variable X in stratified sampling when the observations are contaminated with measurement errors. The bias and mean squared error of the proposed estimators have been derived under large sample approximation and their properties are analyzed. Generalized versions of these estimators are given along with their properties.

Key words: Auxiliary variate, bias, mean squared error, measurement error, study variate.

Introduction

Statistical procedures for the analysis of data presume that observations are correct measurements for the characteristics being studied. When applied to a real world data set, it is assumed it is possible to take measurements without error on the theoretical construct of the variables. This is untenable in many applied situations when observation errors are a rule rather than an exception.

Hence, an auxiliary variable is commonly used in survey sampling to improve the precision of estimates. When auxiliary variable information is available researchers are able to utilize it in methods of estimation to obtain the most efficient estimator. Examples are ratio, product and regression estimation methods. Using auxiliary information at the estimation stage, a large number of estimation procedures for approximating the population mean μ_Y of a study variable Y have been proposed and their properties studied based on data originating under various kinds of sampling schemes and under the supposition that observations have been recorded without error. Such an assumption may not be tenable in actual practice and data may contain observational or measurement errors due to various reasons (Cochran, 1968; Sukhatme, 1984).

Chandhok and Han (1990) have studied the properties of a ratio estimator under two sampling schemes; simple random sampling without replacement and the Mizuno scheme when measurement errors are present. Shalabh (1997) studied the properties of the classical ratio estimator in simple random sampling when the data on both the characteristics Y (study variable) and X (auxiliary variable) are subject to measurement errors. Manisha and Singh (2001), Maneesha and Singh (2002) and Singh and Karpe (2008a) have also considered the problem of estimating the population mean using auxiliary information in the presence of measurement errors. Later Singh and Karpe (2008b, 2009a, 2009c) studied the effect of measurement errors on the classes of estimators proposed for population variance and coefficient of variation. This article discusses the properties of separate and combined ratio and product estimators in stratified random sampling when the data are subject to measurement errors on both the characteristics Y and X.

Housila P. Singh is a Professor in the School of Studies in Statistics. Email: hpsujn@rediffmail.com. Namrata Karpe is a Research Scholar in the School of Studies in Statistics. Email: namratakarpe@yahoo.com.

Suggested Estimators

Separate Ratio Estimator in Stratified Random Sampling in the Presence of Measurement Errors

Consider a finite population $U=(u_1,u_2,...,u_N)$ of size N and let Y and X respectively be the study and auxiliary variables associated with each unit $u_j=(j=1,2,...,N)$ of the population. Let the population of size N be stratified into L strata with the h^{th} stratum containing N_h units, where $h=1,2,...,L$ such that $\sum_{h=1}^{L}N_h=N$. A simple random sample size n_h is drawn without replacement from the h^{th} stratum such that $\sum_{h=1}^{L}n_h=n$. Let (y_{hi},x_{hi}) be the observed pair values instead of true pair values (Y_{hi},X_{hi}) of two characteristics (Y, X) on i^{th} unit of the h^{th} stratum, where $i=1,2,...,N_h$ and $h=1,2,...,L$. In addition, let:

$$\left(\bar{y}_h=\frac{1}{n_h}\sum_{i=1}^{n_h}y_{hi},\bar{x}_h=\frac{1}{n_h}\sum_{i=1}^{n_h}x_{hi}\right),$$

$$\left(\bar{y}_{st}=\sum_{h=1}^{L}W_h\bar{y}_h,\bar{x}_{st}=\sum_{h=1}^{L}W_h\bar{x}_h\right),$$

$$\left(\mu_{Yh}=\frac{1}{N_h}\sum_{i=1}^{N_h}y_{hi},\mu_{Xh}=\frac{1}{N_h}\sum_{i=1}^{N_h}x_{hi}\right),$$

and

$$\left(\mu_Y=\sum_{h=1}^{L}W_h\mu_{Yh},\mu_X=\sum_{h=1}^{L}W_h\mu_{Xh}\right)$$

be the samples means and population means of (Y, X) respectively, where $W_h=\frac{N_h}{N}$ is the stratum weight. Let the observational or measurement errors be

$$u_{hi}=y_{hi}-Y_{hi} \qquad (2.1)$$

and

$$v_{hi}=x_{hi}-X_{hi} \qquad (2.2)$$

which are stochastic in nature and are uncorrelated with mean zero and variances σ_{Uh}^2 and σ_{Vh}^2 respectively. Further let ρ_h be the population correlation coefficient between Y and X in the h^{th} stratum.

For simplicity in exposition, assume that $u_{hi}'s$ and $v_{hi}'s$ are uncorrelated although (Y_{hi},X_{hi}) are correlated; such a specification can be relaxed at the cost of some algebraic complexity. It is also assumed that the finite population correction terms $(1-f_h)$ and $(1-f)$ can be ignored where $f_h=\frac{n_h}{N_h}$ and $f=\frac{n}{N}$.

To estimate the population mean μ_Y, the traditional unbiased estimator (i.e., stratified sample mean) $\bar{y}_{st}$ is used, but it does not utilize the sample information on auxiliary characteristic X. Assuming that μ_{Xh} is known and is different from zero, this method yields a separate ratio estimator of the population mean μ_Y:

$$t_{RS}=\sum_{h=1}^{L}W_h\bar{y}_h\frac{\mu_{Xh}}{\bar{x}_h}. \qquad (2.3)$$

To obtain the bias and mean squared error of t_{RS}: $\bar{y}_h=\mu_{Yh}(1+\delta_{\bar{y}_h})$, and $\bar{x}_h=\mu_{Xh}(1+\delta_{\bar{x}_h})$, such that

$$E(\delta_{\bar{y}_h})=E(\delta_{\bar{x}_h})=0$$

$$E(\delta_{\bar{y}_h}^2)=\frac{C_{Yh}^2}{n_h}\left(1+\frac{\sigma_{Uh}^2}{\sigma_{Yh}^2}\right)=\frac{C_{Yh}^2}{n_h\theta_{Yh}},$$

$$E(\delta_{\bar{x}_h}^2)=\frac{C_{Xh}^2}{n_h}\left(1+\frac{\sigma_{Vh}^2}{\sigma_{Xh}^2}\right)=\frac{C_{Xh}^2}{n_h\theta_{Xh}},$$

$$E\left(\delta_{\bar{y}_h}\delta_{\bar{x}_h}\right)=\frac{1}{n_h}\rho_h C_{Yh}C_{Xh},$$

where

$$C_{Yh}=\frac{\sigma_{Yh}}{\mu_{Yh}},$$

$$C_{Xh}=\frac{\sigma_{Xh}}{\mu_{Xh}},$$

$$\theta_{Yh}=\frac{\sigma_{Uh}^2}{\sigma_{Uh}^2+\sigma_{Yh}^2},$$

and

$$\theta_{Xh}=\frac{\sigma_{Vh}^2}{\sigma_{Vh}^2+\sigma_{Xh}^2}.$$

Expressing (2.3) in terms of $\delta' s$ as

$$t_{RS}=\sum_{h=1}^{L}W_h\mu_{Yh}\left(1+\delta_{\bar{y}_h}\right)\left(1+\delta_{\bar{x}_h}\right)^{-1} \quad (2.4)$$

Assuming $\left|\delta_{\bar{x}_h}\right|<1$, the right hand side of (2.4) is expanded as

$$\mathrm{t_{RS}}=\sum_{\mathrm{h=1}}^{\mathrm{L}}\mathrm{W_h\mu_{Yh}}\left(1+\delta_{\bar{\mathrm{y}}_\mathrm{h}}\right)\begin{pmatrix}1-\delta_{\bar{\mathrm{x}}_\mathrm{h}}+\delta_{\bar{\mathrm{x}}_\mathrm{h}}^2\\-\delta_{\bar{\mathrm{x}}_\mathrm{h}}^3+\ldots\end{pmatrix}$$

$$=\sum_{\mathrm{h=1}}^{\mathrm{L}}\mathrm{W_h\mu_{Yh}}\begin{Bmatrix}1+\delta_{\bar{\mathrm{y}}_\mathrm{h}}-\delta_{\bar{\mathrm{x}}_\mathrm{h}}-\delta_{\bar{\mathrm{y}}_\mathrm{h}}\delta_{\bar{\mathrm{x}}_\mathrm{h}}\\+\delta_{\bar{\mathrm{x}}_\mathrm{h}}^2+\delta_{\bar{\mathrm{y}}_\mathrm{h}}\delta_{\bar{\mathrm{x}}_\mathrm{h}}^2--\delta_{\bar{\mathrm{x}}_\mathrm{h}}^3+\ldots\end{Bmatrix}$$

Neglecting terms of $\delta' s$ having power greater than two, results in

$$\mathrm{t_{RS}}=\sum_{\mathrm{h=1}}^{\mathrm{L}}\mathrm{W_h\mu_{Yh}}\begin{Bmatrix}1+\delta_{\bar{\mathrm{y}}_\mathrm{h}}-\delta_{\bar{\mathrm{x}}_\mathrm{h}}-\delta_{\bar{\mathrm{y}}_\mathrm{h}}\delta_{\bar{\mathrm{x}}_\mathrm{h}}\\+\delta_{\bar{\mathrm{x}}_\mathrm{h}}^2+\delta_{\bar{\mathrm{y}}_\mathrm{h}}\delta_{\bar{\mathrm{x}}_\mathrm{h}}^2\end{Bmatrix},$$

$$\mathrm{t_{RS}}=\mu_\mathrm{Y}+\sum_{\mathrm{h=1}}^{\mathrm{L}}\mathrm{W_h\mu_{Yh}}\begin{Bmatrix}\delta_{\bar{\mathrm{y}}_\mathrm{h}}-\delta_{\bar{\mathrm{x}}_\mathrm{h}}-\delta_{\bar{\mathrm{y}}_\mathrm{h}}\delta_{\bar{\mathrm{x}}_\mathrm{h}}\\+\delta_{\bar{\mathrm{x}}_\mathrm{h}}^2+\delta_{\bar{\mathrm{y}}_\mathrm{h}}\delta_{\bar{\mathrm{x}}_\mathrm{h}}^2\end{Bmatrix},$$

$$\left(\mathrm{t_{RS}}-\mu_\mathrm{Y}\right)=\sum_{\mathrm{h=1}}^{\mathrm{L}}\mathrm{W_h\mu_{Yh}}\begin{Bmatrix}\delta_{\bar{\mathrm{y}}_\mathrm{h}}-\delta_{\bar{\mathrm{x}}_\mathrm{h}}-\delta_{\bar{\mathrm{y}}_\mathrm{h}}\delta_{\bar{\mathrm{x}}_\mathrm{h}}\\+\delta_{\bar{\mathrm{x}}_\mathrm{h}}^2+\delta_{\bar{\mathrm{y}}_\mathrm{h}}\delta_{\bar{\mathrm{x}}_\mathrm{h}}^2\end{Bmatrix}, \quad (2.5)$$

Taking the expectation of both sides of (2.5) results in the bias of t_{RS} to the first degree of approximation,

$$B\left(t_{RS}\right)=\sum_{h=1}^{L}W_h\mu_{Yh}\left(\frac{C_{Xh}^2}{n_h\theta_{Xh}}\right)\left(1-\theta_{Xh}K_h\right) \quad (2.6)$$

where

$$K_h=\rho_h\left(\frac{C_{Yh}}{C_{Xh}}\right).$$

Squaring both sides of (2.5), neglecting terms of $\delta' s$ having power greater than two and then taking the expectation of both sides gives the mean squared error of t_{RS} to the first degree of approximation as

$$\mathrm{MSE\left(t_{RS}\right)}=$$
$$\sum_{\mathrm{h=1}}^{\mathrm{L}}\mathrm{W_h^2}\left(\frac{\mu_{\mathrm{Yh}}^2}{\mathrm{n_h}}\right)\left[\frac{\mathrm{C_{Yh}^2}}{\theta_{\mathrm{Yh}}}+\frac{\mathrm{C_{Xh}^2}}{\theta_{\mathrm{Xh}}}\left(1-2\mathrm{K_h}\theta_{\mathrm{Xh}}\right)\right] \quad (2.7)$$

The variance of $\bar{y}_{st}$ is:

$$Var\left(\bar{y}_{st}\right)=\sum_{h=1}^{L}W_h^2\frac{\mu_{Yh}^2C_{Yh}^2}{n_h\theta_{Yh}} \quad (2.8)$$

and, from (2.7) and (2.8),

$$\mathrm{MSE\left(t_{RS}\right)-Var\left(\bar{y}_{st}\right)}=$$
$$\sum_{\mathrm{h=1}}^{\mathrm{L}}\mathrm{W_h^2}\left(\frac{\mu_{\mathrm{Yh}}^2}{\mathrm{n_h}}\right)\frac{\mathrm{C_{Xh}^2}}{\theta_{\mathrm{Xh}}}\left(1-2\mathrm{K_h}\theta_{\mathrm{Xh}}\right)$$

which is less than zero if

$$\sum_{h=1}^{L} W_h^2 \left(\frac{\mu_{Yh}^2}{n_h}\right)\frac{C_{Xh}^2}{\theta_{Xh}}\left(1-2K_h\theta_{Xh}\right)<0$$
$$\text{or if } \left(1-2K_h\theta_{Xh}\right)<0$$
$$\text{or if } \frac{1}{2\theta_{Xh}}<K_h \quad (2.9)$$
$$\text{or if } K_h>\frac{1}{2\theta_{Xh}}$$

Thus, the proposed separate ratio estimator t_{RS} is more efficient than the usual unbiased estimator $\bar{y}_{st}$ if condition (2.9) holds.

If the observations on both the variables X and Y are recorded without error, then the MSE of t_{RS} at (2.7) reduces to:

$$MSE(t_{RS})_t = \sum_{h=1}^{L} W_h^2\left(\frac{\mu_{Yh}^2}{n_h}\right)\left[C_{Yh}^2 + C_{Xh}^2(1-2K_h)\right] \quad (2.10)$$

Expression (2.10) can be obtained from (2.7) by setting $\theta_{Xh}=\theta_{Yh}=1$. From (2.7) and (2.10):

$$\mathrm{MSE}(t_{RS})-\mathrm{MSE}(t_{RS})_t = \sum_{h=1}^{L} W_h^2\left(\frac{\mu_{Yh}^2}{n_h}\right)\left[\left(\frac{1-\theta_{Yh}}{\theta_{Yh}}\right)C_{Yh}^2+\left(\frac{1-\theta_{Xh}}{\theta_{Xh}}\right)C_{Xh}^2\right] \quad (2.11)$$

which is always positive. Thus, it follows from (2.11) that the presence of measurement errors associated with both variables are accountable for increasing the mean squared error of the separate ratio-estimator t_{RS}.

Separate Product Estimator in Stratified Random Sampling in the presence of Measurement Errors

Next, define the product estimators in stratified random sampling in the presence of measurement errors of the population mean μ_Y as

$$t_{PS} = \sum_{h=1}^{L} W_h \bar{y}_h \frac{\bar{x}_h}{\mu_{Xh}} \quad (2.12)$$

Express (2.12) in terms of $\delta' s$ as

$$t_{PS} = \sum_{h=1}^{L} W_h \mu_{Yh}\left(1+\delta_{\bar{y}_h}\right)\left(1+\delta_{\bar{x}_h}\right) \quad (2.13)$$

Assuming that $\left|\delta_{\bar{x}_h}\right|<1$, the right hand side of (2.13) is expanded as

$$t_{PS} = \sum_{h=1}^{L} W_h \mu_{Yh}\left\{1+\delta_{\bar{y}_h}+\delta_{\bar{x}_h}+\delta_{\bar{y}_h}\delta_{\bar{x}_h}\right\},$$

$$t_{PS} = \mu_Y + \sum_{h=1}^{L} W_h \mu_{Yh}\left\{\delta_{\bar{y}_h}+\delta_{\bar{x}_h}+\delta_{\bar{y}_h}\delta_{\bar{x}_h}\right\},$$

or

$$\left(t_{PS}-\mu_Y\right) = \sum_{h=1}^{L} W_h \mu_{Yh}\left\{\delta_{\bar{y}_h}+\delta_{\bar{x}_h}+\delta_{\bar{y}_h}\delta_{\bar{x}_h}\right\}, \quad (2.14)$$

and taking the expectation of both sides of (2.14) results in the bias of t_{PS} to the first degree of approximation,

$$B(t_{PS}) = \sum_{h=1}^{L} W_h \mu_{Yh} C_{Xh}^2 K_h \quad (2.15)$$

Squaring both sides of (2.14) and neglecting terms of $\delta' s$ having power greater than two and taking expectations of both sides, provides the mean squared error of t_{PS} to the first degree of approximation as

$$MSE(t_{PS}) = \sum_{h=1}^{L} W_h^2\left(\frac{\mu_{Yh}^2}{n_h}\right)\left[\frac{C_{Yh}^2}{\theta_{Yh}}+\frac{C_{Xh}^2}{\theta_{Xh}}\left(1+2K_h\theta_{Xh}\right)\right] \quad (2.16)$$

From (2.16) and (2.8)

$$\mathrm{MSEP}(t_{RS})-\mathrm{Var}(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2\left(\frac{\mu_{Yh}^2}{n_h}\right)\frac{C_{Xh}^2}{\theta_{Xh}}\left(1+2K_h\theta_{Xh}\right)$$

which is less than zero if

$$\sum_{h=1}^{L} W_h^2 \left(\frac{\mu_{Yh}^2}{n_h} \right) \frac{C_{Xh}^2}{\theta_{Xh}} \left(1 + 2K_h \theta_{Xh}\right) < 0$$

$$\text{or if } \left(1 + 2K_h \theta_{Xh}\right) < 0 \qquad (2.17)$$

$$\text{or if } K_h < -\frac{1}{2\theta_{Xh}}$$

Thus, the proposed separate product estimator t_{PS} is more efficient than the usual unbiased estimator $\bar{y}_{st}$ if condition (2.17) holds.

If the observations on both the variables X and Y are recorded without error, then the MSE of the separate product estimator t_{PS} is given by

$$MSE(t_{RS})_t = \sum_{h=1}^{L} W_h^2 \left(\frac{\mu_{Yh}^2}{n_h} \right) \left[C_{Yh}^2 + C_{Xh}^2 (1 + 2K_h) \right] \qquad (2.18)$$

Expression (2.18) can be obtained from (2.16) by setting $\theta_{Xh} = \theta_{Yh} = 1$. From (2.16) and (2.18):

$$MSE(t_{PS}) - MSE(t_{PS})_t =$$

$$\sum_{h=1}^{L} W_h^2 \left(\frac{\mu_{Yh}^2}{n_h} \right) \left[\left(\frac{1 - \theta_{Yh}}{\theta_{Yh}} \right) C_{Yh}^2 + \left(\frac{1 - \theta_{Xh}}{\theta_{Xh}} \right) C_{Xh}^2 \right]. \qquad (2.19)$$

which is always positive. Thus, it follows from (2.19) that the presence of measurement errors associated with both variables are accountable for increasing the mean squared error of the separate product-estimator t_{PS}.

Separate Difference Estimator in Stratified Random Sampling in the presence of Measurement Errors

A separate difference estimator is defined in stratified random sampling in the presence of measurement errors for population mean μ_Y, as

$$t_{dS} = \sum_{h=1}^{L} W_h \{ \bar{y}_h + d_h (\mu_{Xh} - \bar{x}_h) \} \qquad (2.20)$$

where $d_h's$ are suitably chosen constants.

It can be observed that the estimator t_{dS} is an unbiased estimator for the population mean μ_Y, and the variance of t_{dS} is given by

$$Var(t_{dS}) = \sum_{h=1}^{L} W_h^2 \left\{ \begin{matrix} Var(\bar{y}_h) + d_h^2 Var(\bar{x}_h) \\ -2Cov(\bar{y}_h, \bar{x}_h) \end{matrix} \right\} \qquad (2.21)$$

where

$$Var(\bar{y}_h) = \frac{\sigma_{Yh}^2 + \sigma_{Uh}^2}{n_h} = \frac{\sigma_{Yh}^2}{n_h \theta_{Yh}},$$

$$Var(\bar{x}_h) = \frac{\sigma_{Xh}^2 + \sigma_{Vh}^2}{n_h} = \frac{\sigma_{Xh}^2}{n_h \theta_{Xh}},$$

$$Cov(\bar{y}_h, \bar{x}_h) = \frac{\sigma_{YXh}}{n_h} = \frac{\sigma_{Xh}^2}{n_h} \beta_{YXh},$$

$$\sigma_{YXh} = Cov(y_{hi}, x_{hi})$$

$$= E\{(y_{hi} - E(y_{hi}))(x_{hi} - E(x_{hi}))\}$$

and

$$\beta_{YXh} = \frac{\sigma_{YXh}}{\sigma_{Xh}^2}.$$

Thus,

$$Var(t_{dS}) = \sum_{h=1}^{L} \frac{W_h^2}{n_h} \left\{ \frac{\sigma_{Yh}^2}{\theta_{Yh}} + d_h^2 \frac{\sigma_{Xh}^2}{\theta_{Xh}} - 2d_h \sigma_{XY}^2 \beta_{YXh} \right\}$$

$$= \sum_{h=1}^{L} \frac{W_h^2}{n_h} \left\{ \frac{\sigma_{Yh}^2}{\theta_{Yh}} + d_h \frac{\sigma_{Xh}^2}{\theta_{Xh}} (d_h - 2\beta_{YXh} \theta_{Xh}) \right\} \qquad (2.22)$$

which is minimized for

$$d_h = \beta_{YXh} \theta_{Xh} \qquad (2.23)$$

where $\beta_{YXh} = \frac{\sigma_{YXh}}{\sigma_{Xh}^2}$ is the population regression coefficient of Y on X in the h^{th} stratum. Thus, the resulting (minimum) variance of t_{dS} is given by

$$min.Var(t_{dS}) = \sum_{h=1}^{L} \frac{W_h^2}{n_h} \left\{ \frac{\sigma_{Yh}^2}{\theta_{Yh}} - \beta_{YXh}^2 \theta_{Xh} \sigma_{Xh}^2 \right\}$$

or

$$min.Var(t_{dS}) = \sum_{h=1}^{L} \frac{W_h^2 \sigma_{Yh}^2}{n_h \theta_{Yh}} \left(1 - \rho_h^2 \theta_{Xh} \theta_{Yh}\right) \tag{2.24}$$

When data are recorded without error for the variables X and Y, the variance of t_{dS} at (2.22) reduces to:

$$Var(t_{dS})_t = \sum_{h=1}^{L} \frac{W_h^2}{n_h} \left\{ \sigma_{Yh}^2 + d_h \sigma_{Xh}^2 (d_h - 2\beta_{YXh}) \right\} \tag{2.25}$$

From (2.22) and (2.25):

$$\mathrm{Var}(t_{dS}) - \mathrm{Var}(t_{dS})_t = \sum_{h=1}^{L} \frac{W_h^2}{n_h} \left\{ \left(\frac{1-\theta_{Yh}}{\theta_{Yh}} \right) \sigma_{Yh}^2 + d_h^2 \left(\frac{1-\theta_{Xh}}{\theta_{Xh}} \right) \right\} \geq 0. \tag{2.26}$$

Observe from (2.26) that the presence of measurement error in both the variables X an Y inflates the variance of t_{dS}. The expression (2.25) is minimized for:

$$d_h = \beta_{YXh} \tag{2.27}$$

and the resulting (minimum) variance of t_{dS} in the absence of measurement errors is given by

$$min.Var(t_{dS})_t = \sum_{h=1}^{L} \frac{W_h^2}{n_h} \sigma_{Yh}^2 \left(1 - \rho_h^2\right) \tag{2.28}$$

From (2.24) and (2.28):

$$\min.\mathrm{Var}(t_{dS}) - \min.\mathrm{Var}(t_{dS})_t = \sum_{h=1}^{L} \frac{W_h^2 \sigma_{Yh}^2}{n_h} \left\{ \left(\frac{1-\theta_{Yh}}{\theta_{Yh}} \right) + \rho_h^2 (1-\theta_{Xh}) \right\} \tag{2.29}$$

It is observed from (2.29) that the difference [$min.Var(t_{dS}) - min.Var(t_{dS})_t$], is always positive. Thus, the presence of measurement errors in both variables X and Y inflates the variance of t_{dS} at optimum condition, which disturbs the optimal properties of the difference estimator t_{dS}.

A Separate Class of Estimators in Stratified Random Sampling in the Presence of Measurement Errors

Whatever the sample chosen, let $(\bar{y}_h, \bar{x}_h)$ assume values in a bounded, closed subset, P_h, of the two-dimensional real space containing the point (μ_{Yh}, μ_{Xh}). Following an approach similar to that adopted by Srivatava (1971, 1980) for defining a class of estimators of the population μ_{Yh}, consider the class of estimators of the population μ_Y, defined by

$$t_S = \sum_{h=1}^{L} W_h t_h (\bar{y}_h, \bar{x}_h), \tag{2.30}$$

where $t_h(\bar{y}_h, \bar{x}_h)$ is a function of $(\bar{y}_h, \bar{x}_h)$

$$t_h(\mu_{Yh}, \mu_{Xh}) = \mu_{Yh} \tag{2.31}$$

$$\Rightarrow \mathrm{t_{h1}}(\mu_{Yh}, \mu_{Xh}) = \left. \frac{\partial \mathrm{t_h}(\mu_{Yh}, \mu_{Xh})}{\partial \bar{y}_h} \right|_{(\bar{y}_h, \bar{x}_h)}$$
$$= (\mu_{Yh}, \mu_{Xh})$$
$$= 1$$

such that it satisfies the following conditions:

i. The function $t_h(\bar{y}_h, \bar{x}_h)$ is continuous and bounded in P_h; and
ii. The first, second and third order partial derivatives of $t_h(\bar{y}_h, \bar{x}_h)$ exist and are continuous and bounded in P_h.

Expanding the function $t_h(\bar{y}_h, \bar{x}_h)$ about the point $(\bar{y}_h, \bar{x}_h) = (\mu_{Yh}, \mu_{Xh})$ in a third-order partial derivative, results in

$$t_S = \sum_{h=1}^{L} W_h \left[\begin{array}{l} \mu_{Yh} + (\bar{y}_h - \mu_{Yh})\, t_{h1}(\mu_{Yh}, \mu_{Xh}) \\ + (\bar{x}_h - \mu_{Xh})\, t_{h2}(\mu_{Yh}, \mu_{Xh}) \\ + \frac{1}{2}\left\{ \begin{array}{l} (\bar{y}_h - \mu_{Yh})^2\, t_{h11}(\mu_{Yh}, \mu_{Xh}) \\ + 2(\bar{y}_h - \mu_{Yh})(\bar{x}_h - \mu_{Xh})\, t_{h12}(\mu_{Yh}, \mu_{Xh}) \\ + (\bar{x}_h - \mu_{Xh})^2\, t_{h22}(\mu_{Yh}, \mu_{Xh}) \end{array} \right\} \\ + \frac{1}{6}\left\{ \begin{array}{l} (\bar{y}_h - \mu_{Yh})^3\, t_{h111}(\bar{y}_h^*, \bar{x}_h^*) \\ + 3(\bar{y}_h - \mu_{Yh})(\bar{x}_h - \mu_{Xh})^2\, t_{h122}(\bar{y}_h^*, \bar{x}_h^*) \\ + 3(\bar{y}_h - \mu_{Yh})^2 (\bar{x}_h - \mu_{Xh})\, t_{h112}(\bar{y}_h^*, \bar{x}_h^*) \\ + (\bar{x}_h - \mu_{Xh})^3\, t_{h222}(\bar{y}_h^*, \bar{x}_h^*) \end{array} \right\} \end{array} \right] \tag{2.32}$$

where $\bar{y}_h^* = \{\mu_{Yh} + \theta(\bar{y}_h - \mu_{Yh})\}$, $\bar{x}_h^* = \{\mu_{Xh} + \theta(\bar{x}_h - \mu_{Xh})\}$, $0 < \theta < 1$ and θ may depend on $(\bar{y}_h^*, \bar{x}_h^*)$, and $t_{hijk}(\bar{y}_h^*, \bar{x}_h^*)$ denotes the third order partial derivative of $t_h(\bar{y}_h^*, \bar{x}_h^*)$ with respect to $(\bar{y}_h, \bar{x}_h)$ at the point $(\bar{y}_h, \bar{x}_h) = (\bar{y}_h^*, \bar{x}_h^*)$.

Taking the expectation of (2.32) the bias of the estimator t_S up to the terms of the order n^{-1} is obtained

$$B(t_S) = \frac{1}{2}\sum_{h=1}^{L} W_h \left\{ \begin{array}{l} \mathrm{Var}(\bar{x}_h)\, t_{h22}(\mu_{Yh}, \mu_{Xh}) + \\ 2\mathrm{Cov}(\bar{y}_h, \bar{x}_h)\, t_{h12}(\mu_{Yh}, \mu_{Xh}) \end{array} \right\},$$

or

$$B(t_S) = \frac{1}{2}\sum_{h=1}^{L} \frac{W_h^2}{n_h} \left\{ \begin{array}{l} \frac{\sigma_{Xh}^2}{\theta_{Xh}}\, t_{h22}(\mu_{Yh}, \mu_{Xh}) \\ + 2\sigma_{YXh}\, t_{h12}(\mu_{Yh}, \mu_{Xh}) \end{array} \right\}$$

$$= \frac{1}{2}\sum_{h=1}^{L} \frac{W_h^2 \sigma_{Yh}^2}{n_h \theta_{Xh}} \left\{ \begin{array}{l} t_{h22}(\mu_{Yh}, \mu_{Xh}) \\ + 2\beta_{YXh}\, t_{h12}(\mu_{Yh}, \mu_{Xh}) \end{array} \right\}. \tag{2.33}$$

Up to the terms of order n^{-1}, the MSE of t_S is :

$$MSE(t_S) = \frac{1}{2}\sum_{h=1}^{L} W_h^2 \left\{ \begin{array}{l} \mathrm{Var}(\bar{y}_h) \\ + \mathrm{Var}(\bar{x}_h)\, t_{h2}^2(\mu_{Yh}, \mu_{Xh}) \\ + 2\mathrm{Cov}(\bar{y}_h, \bar{x}_h)\, t_{h12}(\mu_{Yh}, \mu_{Xh}) \end{array} \right\}$$

$$= \sum_{h=1}^{L} \frac{W_h^2}{n_h} \left\{ \begin{array}{l} \frac{\sigma_{Yh}^2}{\theta_{Yh}} + \frac{\sigma_{Xh}^2}{\theta_{Xh}}\, t_{h2}^2(\mu_{Yh}, \mu_{Xh}) \\ + 2\sigma_{YXh}\, t_{h2}(\mu_{Yh}, \mu_{Xh}) \end{array} \right\}$$

$$= \sum_{h=1}^{L} \frac{W_h^2}{n_h} \left[\begin{array}{l} \frac{\sigma_{Yh}^2}{\theta_{Yh}} + \frac{\sigma_{Xh}^2}{\theta_{Xh}}\, t_{h2}(\mu_{Yh}, \mu_{Xh}) \\ \left\{ t_{h2}(\mu_{Yh}, \mu_{Xh}) + 2\beta_{YXh}\theta_{Xh} \right\} \end{array} \right] \tag{2.34}$$

The MSE(t_S) is minimized for

$$t_{h12}(\mu_{Yh}, \mu_{Xh}) = -\left(\frac{\sigma_{YXh}}{\sigma_{Xh}^2}\right)\theta_{Xh} = -\beta_{YXh}\theta_{Xh} \tag{2.35}$$

Thus, the resulting minimum MSE of t_S is given by

$$min.MSE(t_S) = \sum_{h=1}^{L} \frac{W_h^2 \sigma_{Yh}^2}{n_h \theta_{Yh}} \left(1 - \rho_h^2 \theta_{Xh} \theta_{Yh}\right) \tag{2.36}$$

Theorem 2.1

Based on the previous discussion, the following theorem is put forth. To the first degree of approximation,

$$MSE(t_S) \geq \sum_{h=1}^{L} \frac{W_h^2 \sigma_{Yh}^2}{n_h \theta_{Yh}} \left(1 - \rho_h^2 \theta_{Xh} \theta_{Yh}\right),$$

with equality holding if

$$t_{h12}(\mu_{Yh}, \mu_{Xh}) = -\beta_{YXh} \theta_{Xh}.$$

Note the lower bound of the MSEs of the separate class of estimators t_S at (2.30) in the variance of the optimum separate difference estimator (OSDE)

$$t_{d0S} = \sum_{h=1}^{L} W_h \{\bar{y}_h + d_{0h}(\mu_{Yh} - \bar{x}_h)\}$$

with

$$d_{0h} = \beta_{YXh} \theta_{Xh},$$

which shows that the estimators belonging to the class of separate estimators t_S at (2.30) are asymptotically no more efficient than the optimum difference estimator (ODE) t_{d0S}.

Any parametric function $t_h(\bar{y}_h, \bar{x}_h)$ satisfying conditions (i) and (ii) can define an μ_{Yh}. The class of such estimators is very large. For example, the following estimators:

$$t_{S1} = \sum_{h=1}^{L} W_h \bar{y}_h \left(\frac{\mu_{Yh}}{\bar{x}_h}\right)^{\alpha_h},$$

$$t_{S2} = \sum_{h=1}^{L} W_h \bar{y}_h \left\{2 - \left(\frac{\bar{x}_h}{\mu_{Yh}}\right)^{\alpha_h}\right\},$$

$$t_{S3} = \sum_{h=1}^{L} W_h \bar{y}_h \frac{\mu_{Xh}}{\{\mu_{Xh} + \alpha_h(\bar{x}_h - \mu_{Xh})\}},$$

$$t_{S4} = \sum_{h=1}^{L} W_h \{\bar{y}_h + \alpha_h(\bar{x}_h - \mu_{Xh})\},$$

are particular members of a proposed class of estimator where α_h is a suitably chosen constant. The optimum value of constant α_h in $t_{Sj}, j = 1 to 4$, which minimizes the mean squared error of the resulting estimator are obtained from (2.36).

It follows from (2.7). (2.8), (2.16) and (2.34) that the proposed separate class of estimators t_S is more efficient than:

i. the usual unbiased estimator $\bar{y}_{st}$ if

$$\min.\{0, -\beta_{YXh}\theta_{Xh}\} < t_{h2}(\mu_{Yh}.\mu_{Xh}) < \max.\{0, -\beta_{YXh}\theta_{Xh}\} \quad (2.37)$$

ii. the separate ratio estimator $\bar{y}_{RS}$ if

$$\min.\{-R_h, (R_h - 2\beta_{YXh}\theta_{Xh})\} < t_{h2}(\mu_{Yh}.\mu_{Xh}) < \max.\{-R_h, (R_h - 2\beta_{YXh}\theta_{Xh})\} \quad (2.38)$$

iii. the separate product estimator $\bar{y}_{PS}$ if

$$\min.\{R_h, -(R_h + 2\beta_{YXh}\theta_{Xh})\} < t_{h2}(\mu_{Yh}.\mu_{Xh}) < \max.\{R_h, -(R_h + 2\beta_{YXh}\theta_{Xh})\} \quad (2.39)$$

A Combined Ratio Estimator in Stratified Random Sampling in the Presence of Measurement Errors

For the estimation of population mean, μ_Y, the following combined ratio estimator is defined in the presence of measurement errors:

$$t_{RC} = \bar{y}_{st} \frac{\mu_X}{\bar{x}_{st}} \quad (2.40)$$

To the first degree of approximation, the bias and mean squared error of the combined ratio estimator t_{RC} are respectively given by

$$B(t_{RC}) = \mu_Y \sum_{h=1}^{L} \frac{W_h^2}{n_h}\left[\frac{(\sigma_{Xh}^2 + \sigma_{Vh}^2)}{\mu_X^2} - \frac{\rho_h \sigma_{Xh}\sigma_{Yh}}{\mu_X \mu_Y}\right]$$
$$= \mu_Y \sum_{h=1}^{L} \frac{W_h^2}{n_h}\left[\frac{\sigma_{Xh}^2}{\mu_X^2 \theta_{Xh}} - \frac{\beta_{YXh}\sigma_{Xh}^2}{\mu_X \mu_Y}\right]$$
$$= \mu_Y \sum_{h=1}^{L} \frac{W_h^2 \sigma_{Xh}^2}{n_h}\left[\frac{R}{\mu_X^2 \theta_{Xh}} - \frac{\beta_{YXh}}{\mu_X}\right]$$
$$= \left(\frac{1}{\mu_X}\right)\sum_{h=1}^{L} \frac{W_h^2 \sigma_{Xh}^2}{n_h \theta_{Xh}}(R - \beta_{YXh}\theta_{Xh}) \quad (2.41)$$

and

$$MSE(t_{RC}) = \sum_{h=1}^{L} \frac{W_h^2}{n_h}\left[\frac{\sigma_{Yh}^2}{\theta_{Yh}} + R\left(\frac{\sigma_{Xh}^2}{\theta_{Xh}}\right)(R - 2\beta_{YXh}\theta_{Xh})\right] \quad (2.42)$$

From (2.8) and (2.42):

$$Var(\bar{y}_{st}) - MSE(t_{RC}) = -\sum_{h=1}^{L} \frac{W_h^2}{n_h}\left(\frac{\sigma_{Xh}^2}{\theta_{Xh}}\right) R(R - 2\beta_{YXh}\theta_{Xh}),$$

which is positive if

$$\frac{\sum_{h=1}^{L} \frac{W_h^2}{n_h}\beta_{YXh}\left(\frac{\sigma_{Xh}^2}{\theta_{Xh}}\right)}{R\sum_{h=1}^{L} \frac{W_h^2}{n_h}\sigma_{Xh}^2} > \frac{1}{2} \quad (2.43)$$

When the data are recorded without error then the expression (2.42) reduces to:

$$MSE(t_{RC})_t = \sum_{h=1}^{L} \frac{W_h^2}{n_h}\left[\sigma_{Yh}^2 + R\sigma_{Xh}^2(R - 2\beta_{YXh})\right] \quad (2.44)$$

From (2.42) and (2.44):

$$MSE(\bar{y}_{RC}) - MSE(t_{RC})_t = \sum_{h=1}^{L} \frac{W_h^2}{n_h}\left[\left(\frac{1-\theta_{Yh}}{\theta_{Yh}}\right)\sigma_{Yh}^2 + R^2\left(\frac{1-\theta_{Xh}}{\theta_{Xh}}\right)\sigma_{Xh}^2\right] \quad (2.45)$$

which is always positive. It follows that the presence of measurement errors in both the variables *X* and *Y* inflates the $MSE(t_{RC})$.

A Combined Product Estimator in Stratified Random Sampling in the Presence of Measurement Errors

The following combined product estimator is defined for the population mean μ_Y in the presence of measurement errors as:

$$t_{PC} = \bar{y}_{st}\frac{\bar{x}_{st}}{\mu_X}. \quad (2.46)$$

The exact bias and mean squared error to the first degree of approximation of the combined ratio estimator t_{PC} are respectively given by

$$B(t_{PC}) = \mu_Y \sum_{h=1}^{L}\left(\frac{W_h^2}{n_h}\right)\frac{\beta_{YXh}\sigma_{Xh}^2}{\mu_X} \quad (2.47)$$

and

$$MSE(t_{PC}) = \sum_{h=1}^{L} \frac{W_h^2}{n_h}\left[\frac{\sigma_{Yh}^2}{\theta_{Yh}} + R\left(\frac{\sigma_{Xh}^2}{\theta_{Xh}}\right)(R + 2\beta_{YXh}\theta_{Xh})\right] \quad (2.48)$$

From (2.8) and (2.48):

$$Var(\bar{y}_{st}) - MSE(t_{PC}) = -\sum_{h=1}^{L} \frac{W_h^2}{n_h}\left(\frac{\sigma_{Xh}^2}{\theta_{Xh}}\right) R(R + 2\beta_{YXh}\theta_{Xh}) \quad (2.49)$$

which is positive if

$$\frac{\sum_{h=1}^{L}\frac{W_h^2}{n_h}\beta_{YXh}\left(\frac{\sigma_{Xh}^2}{\theta_{Xh}}\right)}{R\sum_{h=1}^{L}\frac{W_h^2}{n_h}\sigma_{Xh}^2} < -\frac{1}{2} \qquad (2.50)$$

Assuming that the observations for X and Y be recorded without error, then expression (2.48) reduces to:

$$MSE\left(t_{PC}\right)_t = \sum_{h=1}^{L}\frac{W_h^2}{n_h}\left[\sigma_{Yh}^2 + R\sigma_{Xh}^2\left(R+2\beta_{YXh}\right)\right] \qquad (2.51)$$

which can be obtained from (2.48) by setting $\theta_{Yh} = \theta_{Xh} = 1$. From (2.48) and (2.51):

$$\mathrm{MSE}\left(\mathrm{t_{PC}}\right) - \mathrm{MSE}\left(\mathrm{t_{PC}}\right)_{\mathrm{t}} = \sum_{\mathrm{h=1}}^{\mathrm{L}}\frac{\mathrm{W_h^2}}{\mathrm{n_h}}\left[\left(\frac{1-\theta_{\mathrm{Yh}}}{\theta_{\mathrm{Yh}}}\right)\sigma_{\mathrm{Yh}}^2 + \mathrm{R}^2\left(\frac{1-\theta_{\mathrm{Xh}}}{\theta_{\mathrm{Xh}}}\right)\sigma_{\mathrm{Xh}}^2\right] \qquad (2.52)$$

which is always positive. Expression (2.52) is the same as that obtained in (2.45). Thus, the presence of measurement errors in both variables X and Y are responsible for increasing the MSE of the combined product estimator t_{PC}.

A Combined Difference Estimator in Stratified Random Sampling in the presence of Measurement Errors

A combined difference estimator in stratified random sampling is defined in the presence of measurement errors for a population mean, μ_Y, as

$$t_{dC} = \bar{y}_{st} + d\left(\mu_X - \bar{x}_{st}\right) \qquad (2.53)$$

where d is a suitably chosen constant. It can be seen that the combined difference estimator t_{dC} is unbiased. The variance of t_{dC} is given by

$$\begin{aligned}
&\mathrm{Var}\left(\mathrm{t_{dC}}\right) = \\
&= \mathrm{Var}\left(\bar{\mathrm{y}}_{\mathrm{st}}\right) + \mathrm{d}^2\left(\bar{\mathrm{x}}_{\mathrm{st}}\right) - 2\mathrm{dCov}\left(\bar{\mathrm{y}}_{\mathrm{st}}, \bar{\mathrm{x}}_{\mathrm{st}}\right) \\
&= \sum_{\mathrm{h=1}}^{\mathrm{L}}\frac{\mathrm{W_h^2}}{\mathrm{n_h}}\left[\frac{\sigma_{\mathrm{Yh}}^2}{\theta_{\mathrm{Yh}}} + \mathrm{d}^2\left(\frac{\sigma_{\mathrm{Xh}}^2}{\theta_{\mathrm{Xh}}}\right) - 2\mathrm{d}\sigma_{\mathrm{YXh}}\right] \qquad (2.54) \\
&= \sum_{\mathrm{h=1}}^{\mathrm{L}}\frac{\mathrm{W_h^2}}{\mathrm{n_h}}\left(\frac{\sigma_{\mathrm{Yh}}^2}{\theta_{\mathrm{Yh}}}\right) + \mathrm{d}^2\sum_{\mathrm{h=1}}^{\mathrm{L}}\frac{\mathrm{W_h^2}}{\mathrm{n_h}}\left(\frac{\sigma_{\mathrm{Xh}}^2}{\theta_{\mathrm{Xh}}}\right) \\
&\quad - 2\mathrm{d}\sum_{\mathrm{h=1}}^{\mathrm{L}}\frac{\mathrm{W_h^2}}{\mathrm{n_h}}\beta_{\mathrm{YXh}}\sigma_{\mathrm{Xh}}^2
\end{aligned}$$

which is minimized for

$$d = \frac{\sum_{h=1}^{L}\frac{W_h^2}{n_h}\beta_{YXh}\sigma_{Xh}^2}{\sum_{h=1}^{L}\frac{W_h^2}{n_h}\left(\frac{\sigma_{Xh}^2}{\theta_{Xh}}\right)} \qquad (2.55)$$

Thus, the resulting minimum variance of t_{dC} is given by

$$min.Var\left(t_{dC}\right) = \sum_{h=1}^{L}\frac{W_h^2}{n_h}\left(\frac{\sigma_{Yh}^2}{\theta_{Yh}}\right) - \frac{\left\{\sum_{h=1}^{L}\frac{W_h^2}{n_h}\beta_{YXh}\sigma_{Xh}^2\right\}^2}{\sum_{h=1}^{L}\frac{W_h^2}{n_h}\left(\frac{\sigma_{Xh}^2}{\theta_{Xh}}\right)} \qquad (2.56)$$

Assume the data associated with variables X and Y are recorded without error; in such a case, the expressions (2.54) reduce to:

$$min.Var\left(t_{dC}\right) = \sum_{h=1}^{L}\frac{W_h^2}{n_h}\left(\frac{\sigma_{Yh}^2}{\theta_{Yh}}\right) - \frac{\left\{\sum_{h=1}^{L}\frac{W_h^2}{n_h}\beta_{YXh}\sigma_{Xh}^2\right\}^2}{\sum_{h=1}^{L}\frac{W_h^2}{n_h}\left(\frac{\sigma_{Xh}^2}{\theta_{Xh}}\right)} \qquad (2.57)$$

From (2.54) and (2.57):

$$\mathrm{Var}\left(\mathrm{t_{dC}}\right)-\mathrm{Var}\left(\mathrm{t_{dC}}\right)_{\mathrm{t}}=$$
$$\sum_{\mathrm{h=1}}^{\mathrm{L}}\frac{\mathrm{W_h^2}}{\mathrm{n_h}}\sigma_{\mathrm{Yh}}^2\left(\frac{1-\theta_{\mathrm{Yh}}}{\theta_{\mathrm{Yh}}}\right)+\mathrm{d}^2\sum_{\mathrm{h=1}}^{\mathrm{L}}\frac{\mathrm{W_h^2}}{\mathrm{n_h}}\sigma_{\mathrm{Xh}}^2\left(\frac{1-\theta_{\mathrm{Xh}}}{\theta_{\mathrm{Xh}}}\right) \tag{2.58}$$

which is always positive. It follows from (2.56) that the presence of measurement errors in both variables X and Y enhances the variance of t_{dC} regardless of the value of d.

The $Var(t_{dC})$ at (2.57) is minimized for

$$d=\frac{\sum_{h=1}^{L}\frac{W_h^2}{n_h}\beta_{YXh}\sigma_{Xh}^2}{\sum_{h=1}^{L}\frac{W_h^2}{n_h}\sigma_{Xh}^2} \tag{2.59}$$

Combining (2.59) with (2.57) results in the minimum value of $Var(\bar{y}_{st})$ as

$$min.Var(t_{dC})_t=\sum_{h=1}^{L}\frac{W_h^2}{n_h}\sigma_{Yh}^2-\frac{\left(\sum_{h=1}^{L}\frac{W_h^2}{n_h}\beta_{YXh}\sigma_{Xh}^2\right)^2}{\sum_{h=1}^{L}\frac{W_h^2}{n_h}\sigma_{Xh}^2} \tag{2.60}$$

From (2.56) and (2.58):

$$\min.\mathrm{Var}\left(\mathrm{t_{dC}}\right)-\min.\mathrm{Var}\left(\mathrm{t_{dC}}\right)_{\mathrm{t}}=$$
$$\sum_{\mathrm{h=1}}^{\mathrm{L}}\frac{\mathrm{W_h^2}}{\mathrm{n_h}}\sigma_{\mathrm{Yh}}^2\left(\frac{1-\theta_{\mathrm{Yh}}}{\theta_{\mathrm{Yh}}}\right)$$
$$+\left\{\sum_{\mathrm{h=1}}^{\mathrm{L}}\frac{\mathrm{W_h^2}}{\mathrm{n_h}}\beta_{\mathrm{YXh}}\sigma_{\mathrm{Xh}}^2\right\}^2\frac{\sum_{\mathrm{h=1}}^{\mathrm{L}}\frac{\mathrm{W_h^2}}{\mathrm{n_h}}\sigma_{\mathrm{Xh}}^2\left(\frac{1-\theta_{\mathrm{Xh}}}{\theta_{\mathrm{Xh}}}\right)}{\left\{\sum_{\mathrm{h=1}}^{\mathrm{L}}\frac{\mathrm{W_h^2}}{\mathrm{n_h}}\sigma_{\mathrm{Xh}}^2\right\}\left\{\sum_{\mathrm{h=1}}^{\mathrm{L}}\frac{\mathrm{W_h^2}}{\mathrm{n_h}}\left(\frac{\sigma_{\mathrm{Xh}}^2}{\theta_{\mathrm{Xh}}}\right)\right\}} \tag{2.61}$$

It can be observed from (2.61) that the difference $\left[min.Var(t_{dC})-min.Var(t_{dC})_t\right]_t$ is always positive. Thus the presence of measurement error in both variables X and Y inflates the variance of t_{dC} at their optimum conditions, which disturbs the optimal properties of the difference estimator t_{dC}.

A Combined Class of Estimators in Stratified Random Sampling in the presence of Measurement Errors

Following the same procedure as adopted by Srivastava (1971, 1980) a class of combined estimators of the population mean μ_Y in the presence of measurement errors is suggested, such as

$$t_C=t(\bar{y}_{st},\bar{x}_{st}), \tag{2.62}$$

where $t(\bar{y}_{st},\bar{x}_{st})$ is a function of $(\bar{y}_{st},\bar{x}_{st})$ such that

$$t\ (\mu_Y,\mu_X)=\mu_Y \tag{2.63}$$

$$\Rightarrow t_1(\mu_Y,\mu_X)=\left.\frac{\partial t_1(\mu_Y,\mu_X)}{\partial\,\bar{y}_{st}}\right|_{(\mu_Y,\mu_X)}=1,$$

and satisfies the same conditions as given in 2.4 for t_S.

Expanding the function $t(\bar{y}_{st},\bar{x}_{st})$ about the point $(\bar{y}_{st},\bar{x}_{st})=(\mu_Y,\mu_X)$ in a third-order Taylor's series results in

$$\mathrm{t_C}=$$
$$\mu_{\mathrm{Y}}+\left[\begin{array}{l}(\bar{\mathrm{y}}_{\mathrm{st}}-\mu_{\mathrm{Y}})\,\mathrm{t_1}(\mu_{\mathrm{Y}},\mu_{\mathrm{X}})\\+(\bar{\mathrm{x}}_{\mathrm{st}}-\mu_{\mathrm{X}})\,\mathrm{t_2}(\mu_{\mathrm{Y}},\mu_{\mathrm{X}})\\+\frac{1}{2}\left\{\begin{array}{l}(\bar{\mathrm{y}}_{\mathrm{st}}-\mu_{\mathrm{Y}})^2\,\mathrm{t_{11}}(\mu_{\mathrm{Y}},\mu_{\mathrm{X}})\\+2(\bar{\mathrm{y}}_{\mathrm{st}}-\mu_{\mathrm{Y}})(\bar{\mathrm{x}}_{\mathrm{st}}-\mu_{\mathrm{X}})\,\mathrm{t_{12}}(\mu_{\mathrm{Y}},\mu_{\mathrm{X}})\\+(\bar{\mathrm{x}}_{\mathrm{st}}-\mu_{\mathrm{X}})^2\,\mathrm{t_{22}}(\mu_{\mathrm{Y}},\mu_{\mathrm{X}})\end{array}\right\}\\+\frac{1}{6}\left\{\begin{array}{l}(\bar{\mathrm{y}}_{\mathrm{st}}-\mu_{\mathrm{Y}})^2\,\mathrm{t_{111}}(\bar{\mathrm{y}}^*_{\mathrm{st}},\bar{\mathrm{x}}^*_{\mathrm{st}})\\+3(\bar{\mathrm{y}}_{\mathrm{st}}-\mu_{\mathrm{Y}})(\bar{\mathrm{x}}_{\mathrm{st}}-\mu_{\mathrm{X}})^2\,\mathrm{t_{122}}(\bar{\mathrm{y}}^*_{\mathrm{st}},\bar{\mathrm{x}}^*_{\mathrm{st}})\\+3(\bar{\mathrm{y}}_{\mathrm{st}}-\mu_{\mathrm{Y}})^2(\bar{\mathrm{x}}_{\mathrm{st}}-\mu_{\mathrm{X}})\,\mathrm{t_{112}}(\bar{\mathrm{y}}^*_{\mathrm{st}},\bar{\mathrm{x}}^*_{\mathrm{st}})\\+(\bar{\mathrm{x}}_{\mathrm{st}}-\mu_{\mathrm{X}})^3\,\mathrm{t_{222}}(\bar{\mathrm{y}}^*_{\mathrm{st}},\bar{\mathrm{x}}^*_{\mathrm{st}})\end{array}\right\}\end{array}\right]$$

or

$$
\begin{aligned}
&(t_C-\mu_Y)= \\
&\left[\begin{array}{l}
(\bar{y}_{st}-\mu_Y)\,t_1(\mu_Y,\mu_X) \\
+(\bar{x}_{st}-\mu_X)\,t_2(\mu_Y,\mu_X) \\
+\frac{1}{2}\left\{\begin{array}{l}(\bar{y}_{st}-\mu_Y)^2\,t_{11}(\mu_Y,\mu_X) \\ +2(\bar{y}_{st}-\mu_Y)(\bar{x}_{st}-\mu_X)\,t_{12}(\mu_Y,\mu_X) \\ +(\bar{x}_{st}-\mu_X)^2\,t_{22}(\mu_Y,\mu_X)\end{array}\right\} \\
+\frac{1}{6}\left\{\begin{array}{l}(\bar{y}_{st}-\mu_Y)^3\,t_{111}(\bar{y}^*_{st},\bar{x}^*_{st}) \\ +3(\bar{y}_{st}-\mu_Y)(\bar{x}_{st}-\mu_X)^2\,t_{122}(\bar{y}^*_{st},\bar{x}^*_{st}) \\ +3(\bar{y}_{st}-\mu_Y)^2(\bar{x}_{st}-\mu_X)\,t_{112}(\bar{y}^*_{st},\bar{x}^*_{st}) \\ +(\bar{x}_{st}-\mu_X)^3\,t_{222}(\bar{y}^*_{st},\bar{x}^*_{st})\end{array}\right\}
\end{array}\right]
\end{aligned}
\tag{2.64}
$$

where $\bar{y}^*_{st}=\{\mu_Y+\xi(\bar{y}_{st}-\mu_Y)\}$, $\bar{x}^*_{st}=\{\mu_X+\xi(\bar{x}_{st}-\mu_X)\}$, and $0<\xi<1$. Also, ξ may depend on $(\bar{y}^*_{st},\bar{x}^*_{st})$ and $t_{ijk}(\bar{y}^*_{st},\bar{x}^*_{st})$ denotes the third order partial derivative of $t(\bar{y}_{st},\bar{x}_{st})$ with respect to $(\bar{y}_{st},\bar{x}_{st})$ at the point $(\bar{y}_{st},\bar{x}_{st})=(\bar{y}^*_{st},\bar{x}^*_{st})$.

Taking the expectation of (2.64) provides the bias of the estimator t_C up to the terms of the order n^{-1},

$$
B(t_C)=\frac{1}{2}\left\{\begin{array}{l}\mathrm{Var}(\bar{x}_{st})\,t_{22}(\mu_Y,\mu_X) \\ +2\mathrm{Cov}(\bar{y}_{st},\bar{x}_{st})\,t_{12}(\mu_Y,\mu_X)\end{array}\right\},
$$

$$
\begin{aligned}
B(t_C)&=\frac{1}{2}\sum_{h=1}^{L}\frac{W_h^2}{n_h}\left\{\begin{array}{l}\frac{\sigma^2_{Xh}}{\theta_{Xh}}\,t_{22}(\mu_Y,\mu_X) \\ +2\sigma_{YXh}\,t_{12}(\mu_Y,\mu_X)\end{array}\right\} \\
&=\frac{1}{2}\left[\begin{array}{l}\sum_{h=1}^{L}\frac{W_h^2\sigma^2_{Yh}}{n_h\theta_{Xh}}t_{22}(\mu_Y,\mu_X) \\ +2\sum_{h=1}^{L}\frac{W_h^2}{n_h}\beta_{YXh}\sigma^2_{Xh}\,t_{12}(\mu_Y,\mu_X)\end{array}\right]
\end{aligned}
\tag{2.65}
$$

Squaring both sides of (2.64) and neglecting terms $(\bar{y}_{st}-\mu_Y)$ *and* $(\bar{x}_{st}-\mu_X)$ having power greater than two results in

$$
(t_C-\mu_Y)^2=\left[\begin{array}{l}(\bar{y}_{st}-\mu_Y)^2\,t_1^2(\mu_Y,\mu_X) \\ +(\bar{x}_{st}-\mu_X)\,t_2^2(\mu_Y,\mu_X) \\ +2\mathrm{Cov}(\bar{y}_{st},\bar{x}_{st})\,t_{12}(\mu_Y,\mu_X)\end{array}\right]
$$

Noting that $t_1(\mu_Y,\mu_X)$=1 and taking the expectation of both sides of the above expression, provides the mean squared error of the class of combined estimators t_C as

$$
\mathrm{MSE}(t_C)=\left[\begin{array}{l}\mathrm{Var}(\bar{y}_{st})+\mathrm{Var}(\bar{x}_{st})\,t_{22}(\mu_Y,\mu_X) \\ +2\mathrm{Cov}(\bar{y}_{st},\bar{x}_{st})\,t_2(\mu_Y,\mu_X)\end{array}\right],
$$

or

$$
\begin{aligned}
\mathrm{MSE}(t_C)&=\sum_{h=1}^{L}\frac{W_h^2}{n_h}\left\{\begin{array}{l}\frac{\sigma^2_{Yh}}{\theta_{Yh}}+\frac{\sigma^2_{Xh}}{\theta_{Xh}}t_2^2(\mu_Y,\mu_X) \\ +2\sigma_{YXh}\,t_2(\mu_Y,\mu_X)\end{array}\right\} \\
&=\left[\begin{array}{l}\sum_{h=1}^{L}\frac{W_h^2}{n_h}\frac{\sigma^2_{Yh}}{\theta_{Yh}}+\sum_{h=1}^{L}\frac{W_h^2}{n_h}\frac{\sigma^2_{Xh}}{\theta_{Xh}}t_2^2(\mu_Y,\mu_X) \\ +2\sum_{h=1}^{L}\frac{W_h^2}{n_h}\beta_{YXh}\sigma^2_{Xh}\,t_2(\mu_Y,\mu_X)\end{array}\right]
\end{aligned}
\tag{2.66}
$$

which is minimized for

$$
t_2(\mu_Y,\mu_X)=-\frac{\sum_{h=1}^{L}\frac{W_h^2}{n_h}\beta_{YXh}\sigma^2_{Xh}}{\sum_{h=1}^{L}\frac{W_h^2}{n_h}\left(\frac{\sigma^2_{Xh}}{\theta_{Xh}}\right)}
\tag{2.67}
$$

Thus the resulting minimum MSE of t_C is given by

$$\min.\mathrm{MSE}(t_C) = \sum_{h=1}^{L} \frac{W_h^2}{n_h}\left(\frac{\sigma_{Yh}^2}{\theta_{Yh}}\right) - \frac{\left\{\sum_{h=1}^{L} \frac{W_h^2}{n_h}\beta_{YXh}\sigma_{Xh}^2\right\}^2}{\sum_{h=1}^{L} \frac{W_h^2}{n_h}\left(\frac{\sigma_{Xh}^2}{\theta_{Xh}}\right)} \qquad (2.68)$$

Theorem 2.2

Based on the above, the following theorem is put forth. To the first degree of approximation,

$$MSE(t_C) \geq \left[\sum_{h=1}^{L} \frac{W_h^2}{n_h}\left(\frac{\sigma_{Yh}^2}{\theta_{Yh}}\right) - \frac{\left\{\sum_{h=1}^{L} \frac{W_h^2}{n_h}\beta_{YXh}\sigma_{Xh}^2\right\}^2}{\sum_{h=1}^{L} \frac{W_h^2}{n_h}\left(\frac{\sigma_{Xh}^2}{\theta_{Xh}}\right)}\right]$$

with equality holding if

$$t_2(\mu_Y, \mu_X) = -\frac{\sum_{h=1}^{L} \frac{W_h^2}{n_h}\beta_{YXh}\sigma_{Xh}^2}{\sum_{h=1}^{L} \frac{W_h^2}{n_h}\left(\frac{\sigma_{Xh}^2}{\theta_{Xh}}\right)}.$$

Note that the lower bound of the MSE of the combined class of estimators t_C at (2.62) is the variance of the optimum combined difference estimator (OCDE)

$$t_{d0C} = \bar{y}_{st} + d_0(\mu_X - \bar{x}_{st}) \qquad (2.69)$$

with

$$d_0 = \frac{\sum_{h=1}^{L} \frac{W_h^2}{n_h}\beta_{YXh}\sigma_{Xh}^2}{\sum_{h=1}^{L} \frac{W_h^2}{n_h}\left(\frac{\sigma_{Xh}^2}{\theta_{Xh}}\right)},$$

which demonstrates that the estimators belonging to the class of combined estimators t_C at (2.62) are asymptotically no more efficient than the optimum difference estimator (ODE) t_{d0C} at (2.69).

Any parametric function $t(\bar{y}_{st}, \bar{x}_{st})$ satisfying the regularity conditions as described for t_C, can define a μ_Y. The class of such estimators is very large. For example, the following estimators:

$$t_{C1} = \bar{y}_{st}\left(\frac{\mu_Y}{\bar{x}_{st}}\right)^{\psi},$$

$$t_{C2} = \bar{y}_{st}\left\{2 - \left(\frac{\bar{x}_{st}}{\mu_Y}\right)^{\psi}\right\},$$

$$t_{C3} = \bar{y}_{st}\frac{\mu_X}{\{\mu_X + \psi(\bar{x}_{st} - \mu_X)\}},$$

$$t_{C4} = \{\bar{y}_{st} + \psi(\bar{x}_{st} - \mu_X)\},$$

are particular members of the proposed class of estimator, where ψ is a suitably chosen constant. The optimum value of constant ψ in $t_{Cj}, j = 1 to 4$ which minimizes the mean squared error of the resulting estimator are obtained from (2.68).

It follows from (2.42), (2.8), (2.48) and (2.66) that the proposed separate class of estimators t_C is more efficient than:

i. the usual unbiased estimator $\bar{y}_{st}$ if

$$\min.\left\{0, \frac{-2\sum_{h=1}^{L} p_h\beta_{YXh}\theta_{Xh}}{\sum_{h=1}^{L} p_h}\right\} < t_2(\mu_Y.\mu_X) < \max.\left\{0, \frac{-2\sum_{h=1}^{L} p_h\beta_{YXh}\theta_{Xh}}{\sum_{h=1}^{L} p_h}\right\} \qquad (2.70)$$

ii. the combined ratio estimator $\bar{y}_{CS}$ if

$$\min.\left\{-R, R-2\left(\frac{\sum_{h=1}^{L} p_h \beta_{YXh}\theta_{Xh}}{\sum_{h=1}^{L} p_h}\right)\right\} < t_2\left(\mu_{Yh}.\mu_{Xh}\right) < \max.\left\{-R, R-2\left(\frac{\sum_{h=1}^{L} p_h \beta_{YXh}\theta_{Xh}}{\sum_{h=1}^{L} p_h}\right)\right\} \quad (2.71)$$

iii. the combined product estimator $\bar{y}_{PC}$ if

$$\min.\left\{-R, -R-2\left(\frac{\sum_{h=1}^{L} p_h \beta_{YXh}\theta_{Xh}}{\sum_{h=1}^{L} p_h}\right)\right\} < t_2\left(\mu_{Yh}.\mu_{Xh}\right) < \max.\left\{-R, -R-2\left(\frac{\sum_{h=1}^{L} p_h \beta_{YXh}\theta_{Xh}}{\sum_{h=1}^{L} p_h}\right)\right\} \quad (2.72)$$

where

$$p_h = \sum_{h=1}^{L} \frac{W_h^2}{n_h}\left(\frac{\sigma_{Xh}^2}{\theta_{Xh}}\right).$$

Theoretical Comparisons

From (2.7) (or (2.36)) and (2.42) (or (2.68)):

$$MSE\left(t_{RS} \text{ or } t_S\right) - MSE\left(t_{RC}\right) = \sum_{h=1}^{L} \frac{W_h^2}{n_h}\left(\frac{\sigma_{Xh}^2}{\theta_{Xh}}\right)\left[\left(R-\beta_{YXh}\theta_{Xh}\right)^2 - \left(R_h-\beta_{YXh}\theta_{Xh}\right)^2\right]$$

which is positive if

$$\left(R-\beta_{YXh}\theta_{Xh}\right)^2 > \left(R_h-\beta_{YXh}\theta_{Xh}\right)^2 \quad (3.1)$$

It follows that t_{RS} will be more efficient than t_{RC} if and only if $(\beta_{YXh}\theta_{Xh})$ is nearer to R_h than to R.

From (2.16) and (2.48):

$$MSE\left(t_{PS}\right) - MSE\left(t_{PC}\right) = \sum_{h=1}^{L} \frac{W_h^2}{n_h}\left(\frac{\sigma_{Xh}^2}{\theta_{Xh}}\right)\left[\left(R+\beta_{YXh}\theta_{Xh}\right)^2 - \left(R_h+\beta_{YXh}\theta_{Xh}\right)^2\right]$$

which is positive if

$$\left(R+\beta_{YXh}\theta_{Xh}\right)^2 > \left(R_h+\beta_{YXh}\theta_{Xh}\right)^2 \quad (3.2)$$

Thus, the separate product estimator t_{PC} is more efficient than the combined product estimator t_{PC} if the inequality (3.2) holds.

From (2.24) and (2.56):

$$\min.MSE\left(t_{dC}\right) - \min.MSE\left(t_{dS}\right) = \frac{\left\{\sum_{h=1}^{L} \frac{W_h^2}{n_h}\beta_{YXh}\sigma_{Xh}^2\right\}^2}{\sum_{h=1}^{L} \frac{W_h^2}{n_h}\left(\frac{\sigma_{Xh}^2}{\theta_{Xh}}\right)} - \left\{\sum_{h=1}^{L} \frac{W_h^2}{n_h}\beta_{YXh}^2\theta_{Xh}\sigma_{Xh}^2\right\}$$

$$= \frac{1}{\sum_{h=1}^{L} \frac{W_h^2}{n_h}\left(\frac{\sigma_{Xh}^2}{\theta_{Xh}}\right)}\left[\left\{\sum_{h=1}^{L} \frac{W_h^2}{n_h}\beta_{YXh}\sigma_{Xh}^2\right\}^2 - \left\{\sum_{h=1}^{L} \frac{W_h^2}{n_h}\beta_{YXh}^2\theta_{Xh}\sigma_{Xh}^2\right\}\sum_{h=1}^{L} \frac{W_h^2}{n_h}\left(\frac{\sigma_{Xh}^2}{\theta_{Xh}}\right)\right]$$

$$= \frac{\left\{\sum_{h=1}^{L} p_h \beta_{YXh}\theta_{Xh}\right\}^2}{\sum_{h=1}^{L} p_h} - \sum_{h=1}^{L} p_h \beta_{YXh}^2\theta_{Xh}^2$$

$$= \left\{\frac{\sum_{h=1}^{L} p_h \beta_{YXh}\theta_{Xh}}{\sum_{h=1}^{L} p_h}\right\}^2 \sum_{h=1}^{L} p_h - \sum_{h=1}^{L} p_h \beta_{YXh}^2\theta_{Xh}^2$$

$$= \sum_{h=1}^{L} p_h \left[\left\{ \frac{\sum_{h=1}^{L} p_h \beta_{YXh} \theta_{Xh}}{\sum_{h=1}^{L} p_h} \right\}^2 - \beta_{YXh}^2 \theta_{Xh}^2 \right]$$

$$= \sum_{h=1}^{L} p_h \left(\beta_{YXh} \theta_{Xh} - a \right)^2 \qquad (3.3)$$

where

$$a = \frac{\sum_{h=1}^{L} p_h \beta_{YXh} \theta_{Xh}}{\sum_{h=1}^{L} p_h}$$

and

$$p_h = \frac{W_h^2}{n_h} \frac{\sigma_{Xh}^2}{\theta_{Xh}}.$$

Observe the expression (3.3) is always positive. Thus, unless the term $(\beta_{YXh}\theta_{Xh})$ is the same from stratum to stratum, the separate difference estimator t_{dS} (or the separate class of the estimators t_S) at its optimum condition, that is, OSDE $t_{d0S} = \sum_{h=1}^{L} W_h \{\bar{y}_h + d_{0h}(\mu_{Yh} - \bar{x}_h)\}$ with $d_{0h} = \beta_{YXh}\theta_{Xh}$) is more efficient than the combined difference estimator t_{dC} (or the combined class of the estimators t_C) at optimum (i.e., the OCDE $t_{d0C} = \bar{y}_{st} + d_0(\mu_X - \bar{x}_{st})$ with

$$d_0 = \frac{\sum_{h=1}^{L} \frac{W_h^2}{n_h} \beta_{YXh} \sigma_{Xh}^2}{\sum_{h=1}^{L} \frac{W_h^2}{n_h} \left(\frac{\sigma_{Xh}^2}{\theta_{Xh}} \right)}.$$

References

Chandhok, P. K., & Han, C. P. (1990). On the efficiency of the ratio estimator under Mizuno scheme with measurement errors. *Journal of the Indian Statistical Association*, *28*, 31-39.

Cochran. W. G. (1968). Errors of measurement in Statistics. *Technometrics*, *10*, 637-666.

Manisha, & Singh, R. K. (2001). An estimation of population mean in the presence of measurement errors. *Journal of the Indian Society of Agricultural Statistics*, *54*(*1*), 13-18.

Maneesha, & Singh, R. K. (2002). Role of regression estimator involving measurement errors. *Brazilian Journal of Probability and Statistics*, *16*, 39-46.

Shalabh (1997). Ratio method of estimation in the presence of measurement errors, *Journal of the Indian Society of Agricultural Statistics*, 50(*2*), 150-155.

Singh, H. P., & Karpe, N. (2008a). Ratio-Product estimator for population mean in presence of measurement errors. *Journal of Applied Statistical Sciences*, *16*(*4*), 49-64.

Singh, H. P., & Karpe, N. (2008b). Estimation of population variance using auxiliary information in the presence of Measurement Errors. *Statistics in Transition-New Series, 9*(*3*), 443-470.

Singh, H. P., & Karpe, N. (2009a). A class of estimators using auxiliary information for estimating finite population variance in presence of Measurements Errors. *Communications in Statistical Theory and Methods*, *38*(*5*), 734-741.

Singh, H. P., & Karpe, N. (2009b). A General Procedure for Estimating the General parameter using Auxiliary Information in Presence of Measurement Errors**.** *Communications in Korean Journal of Applied Statistics*, (*Accepted for publication*).

Singh, H. P., & Karpe, N. (2009c). On the estimation of ratio and product of two population means using supplementary information in presence of measurement errors. *Statistica*, (*Accepted for publication*).

Srivastava, A. K., & Shalabh. (2001). Effect of measurement errors on the regression method of estimation in survey sampling. *Journal of Statistical Research*, *35*(*2*), 35-44.

Srivastava, S. K. (1971). A generalized estimator for the mean of a finite population using multi-auxiliary information. *Journal of the American Statistical Association*, *66*, 404-407.

Srivastava, S. K. (1980). A class of estimators using auxiliary information in sample surveys. *Canadian Journal of Statistics*, *8*, 253-254.

Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S., & Asok, C. (1984). *Sampling Theory of Surveys With Applications*. Des Moines: Iowa State University Press.

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 403-413

1538 – 9472/10/$95.00

Reducing Selection Bias in Analyzing Longitudinal Health Data with High Mortality Rates

Xian Liu Charles C. Engel
Uniformed Services University of the Health Sciences, Bethesda MD and Walter Reed National Military Medical Center, Bethesda MD

Han Kang
Department of Veterans Affairs, Washington, DC

Kristie L. Gore
Walter Reed National Military Medical Center, Bethesda MD and Uniformed Services University of the Health Sciences, Bethesda MD

Two longitudinal regression models, one parametric and one nonparametric, are developed to reduce selection bias when analyzing longitudinal health data with high mortality rates. The parametric mixed model is a two-step linear regression approach, whereas the nonparametric mixed-effects regression model uses a retransformation method to handle random errors across time.

Key words: Longitudinal data, mortality rates, nonrandom dropouts, selection bias.

Introduction

Analyzing large-scale longitudinal health data poses special challenges to statisticians, demographers and other quantitative methodologists. Most longitudinal surveys collect random and unbiased samples at baseline. Among older persons, however, a considerable proportion of the baseline respondents will not survive to the ensuing phases of investigation. As a result, longitudinal health outcomes are based on several follow-up samples selected by values of the dependent health variable because physically frailer, functionally disabled and environmentally disadvantaged persons are more likely to die. Thus, follow-up data of a longitudinal health survey on these populations often bear little resemblance to the initial sample, making dropouts non-ignorable. Consequently, currently existing longitudinal regression models, such as the random-effects linear regression model, can be highly sensitive to untestable assumptions and inestimable parameters (Hedeker & Gibbons, 2006; Hogan, Roy, & Korkontzelou 2004; Little & Rubin, 2003; Schafer & Graham, 2002).

There is abundant literature devoted to modeling non-ignorable longitudinal missing data in biostatistics (Demirtas, 2004; Hedeker & Gibbons, 2006; Hogan, Roy & Korkontzelou, 2004; Little, 1995; Little & Rubin, 2003; Robins, Rotnitzky & Zhao, 1995; Yao, Wei & Hogan, 1998). The primary focus of this literature, however, is dropout in clinical trials. Here the missingness is primarily due to reasons other than death and is closely related to outcomes being measured (Schafer & Graham, 2002). In large-scale longitudinal health data for older persons, high death rates are usually the primary reason for dropouts in follow-up waves; in a strict sense, this cannot be simply viewed as missing because the deceased no longer

Xian Liu is Associate Research Professor in the Department of Psychiatry at the F. Edward Hebert School of Medicine, Uniformed Services University of the Health Sciences. Email him at: Xian.Liu@usuhs.edu. Charles C. Engel, Jr. is Associate Chair of the Department of Psychiatry and DoD Director of the Deployment Health Clinical Center at Walter Reed in Bethesda MD. Email him at: cengel@usuhs.mil. Han Kang is Director of Environmental Epidemiology Service, Veterans Health Administration of Department of Veterans Affairs. E-mail him at: han.kang@mail.va.gov. Kristie L. Gore is Director of Research and Program Evaluation at the DoD Deployment Health Clinical Center and Assistant Research Professor in the Department of Psychiatry at the Uniformed Services University of the Health Sciences. E-mail her at: Kristie.gore@med.navy.mil.

possesses any values or characteristics to estimate (Hogan, Roy & Korkontzelou, 2004; Pauler, McCoy &Moinpour, 2003). On the other hand, although assumptions on measurability of the deceased's health outcomes are imperceptible and inappropriate, the influence of high mortality on the distribution of survivors' health data cannot be ignored. When creating a longitudinal model with high death rates, researchers should establish the statistical structure needed to account for the potential lack of independence that often exists among those who have been selected from the survival of the fittest process.

Some researchers have proposed the use of joint modeling, originally developed by Heckman (1979), for longitudinal and survival data that link the health outcomes by means of a common selection factor (Egleston, Scharfstein, Freeman & West, 2006; Fu, Winship & Mare, 2004; Kurland & Heagerty, 2005; Leigh, Ward & Fries, 1993; Pauler, McCoy & Moinpour, 2003; Ratcliffe, Guo & Ten Have, 2004). Given specification of the selection factor, the two responses, survival and longitudinal health outcomes, are thought to be conditionally independent, hence more efficient and less-biased parameter estimates can be obtained from this type of statistical modeling. However, the two-step parametric joint modeling has been criticized because of its considerable dependence on distributional assumptions for the non-ignorable missing data that are impossible to verify (Demirtas, 2004; Hedeker & Gibbons, 2006; Hogan, Roy & Korkontzelou, 2004; Little & Rubin, 2003; Winship & Mare, 1992). Due to the unique characteristics involved in health transitions among older persons, the restrictive assumptions of this method on the parametric disturbance function can be readily violated, thereby degrading the quality of parameter estimates and model-based prediction.

This research develops two longitudinal regression models to account for the selection bias from high mortality rates, one parametric and one nonparametric. The parametric model is a two-step statistical technique developed as a joint model combining longitudinal and survival data. By contrast, the nonparametric longitudinal model uses a retransformation approach, taking into account the missing data mechanism by assuming a skewed distribution of disturbances. Empirical examples are employed to illustrate the new methods developed herein and to discuss the merits and weaknesses in each of the two-step estimators.

Impact of Selection Bias from Mortality

For a baseline sample of I individuals and J follow-up time points, for convenience of analysis, a disability severity score, Y_{it}, is defined to indicate health status for individual i (i = 1, 2, …., I) at time t (t = 0, 1, …., J). It is then assumed that a hypothetical disability severity score exists instantaneously before dying for those who have been deceased between time (t - 1) and time t (t = 1, …., J). It is further assumed that the hypothetical disability severity score for the deceased, denoted by Y_{it}^{d}, is greater than or equal to a constant C_t, and the disability severity scores among survivors, Y_{it}^{s}, are all smaller than this constant.

Heckman's (1979) perspective serves to exhibit the impact of selection bias from mortality. Beginning with two longitudinal random-effects linear regression models, the complete model that includes all members of the baseline sample and a truncated model that consists of survivors only, given by

$$Y = X_1'\beta_1 + Z_1'\gamma_1 + \varepsilon_1 \qquad (1a)$$

$$Y \,|\, Y < C = X_2'\beta_2 + Z_2'\gamma_2 + \varepsilon_2, \qquad (1b)$$

where **Y** represents the (n × 1) vector of observed outcome data within the framework of a block design (n = I × [J + 1]). The matrix **X** is an (n × p) matrix for p − 1 independent variables and **Z** is a (n × r) design matrix for the random effects. The matrices **β** and **γ** are parameters for **X** and **Z** respectively. The random effects are assumed to be normally distributed with mean 0 and variance matrix **G**. The joint distribution of $\boldsymbol{\varepsilon_1}$ $\boldsymbol{\varepsilon_2}$ is assumed to be a singular distribution with covariance matrix $\boldsymbol{\sigma_{12}}$. While the residual term $\boldsymbol{\varepsilon_1}$ is assumed to be normally distributed with mean 0 and variance matrix $\boldsymbol{\sigma}_1^2$, it is implausible to assume that $\boldsymbol{\varepsilon_2}$ be normally distributed with zero

expectation, because the error term in (1b) may not be independent of the covariates.

Because $\mathbf{Y^d}$ is not observable, a dichotomous factor δ_{it} is defined to indicate the survival status for individual i between time (t – 1) and time t (t = 1, 2, …, J) and is used as a proxy for C, such that

$$\begin{cases} \delta_{it} = 0 \text{ if individual i dies between} \\ \quad \text{time (t-1) and t } \left(Y_{it} \geq C_t\right) \\ \delta_{it} = 1 \text{ if individual i survives from} \\ \quad \text{time } \left(\text{t-1}\right) \text{ and time t } \left(Y_{it} < C_t\right) \end{cases}.$$

Specifically, the disability severity score is viewed at time t as a joint distribution of two sequential events: the likelihood of survival between time (t - 1) and time t (S_t; t = 1,2, …., J) and the conditional density function on the disability severity score (Y_t) among those who have survived to t. Given the aforementioned assumptions, the expected disability severity score for individual i at time t can be estimated by the following equation

$$\mathrm{E}\left(Y_{it} \middle| X_{2i}, Z_{2i}, \delta_{it} = 1\right) = Pr\left(\delta_{it} = 1 \middle| X_{1i}\right) \left\{ \begin{array}{l} X'_{2i}\beta_2 + Z'_{2i}\gamma_2 \\ +E\left[\varepsilon_{2i} \middle| \varepsilon_{2i} < C_t - \left(X'_{1i}\beta_1 + Z'_{1i}\gamma_1\right)\right] \end{array} \right\}. \quad (2)$$

As demonstrated by (2), the conditional mean of the disturbance in the survivors sample is a function of $\mathbf{X}_{1i}$ and $\mathbf{Z}_{1i}$. The estimation of equation (2) without considering this correlation will lead to inconsistent parameter estimates and prediction biases. Therefore, modeling longitudinal processes of this disability severity score can be much beyond what a conventional single-equation linear regression can handle. Next, two refined longitudinal models are developed for reducing the selection bias in the analysis of longitudinal health data for older persons, one parametric and one nonparametric.

Parametric Joint Model

The parametric joint mixed model begins by constructing a selection model using survival rates as the dependent variable. Specifically, a Probit survival model is developed using the rationale of Heckman's (1979) two-step perspective to estimate the proportion surviving between time (t – 1) and time t (t =1, 2, …., J). Some empirical studies with joint modeling of longitudinal and survival data have used other statistical functions to estimate survival rates such as the Cox proportional hazard rate model and logistic regression (Egleston, Scharfstein, Freeman & West, 2006; Kurland & Heagerty, 2005; Leigh, Ward & Fries 1993; Pauler, McCoy & Moinpour, 2003; Ratcliffe, Guo & Ten Have, 2004). The Probit function is used here for convenience of illustration assuming survival probabilities are normally distributed. Specification of other functions would lead to the same results (Greene, 2003; Kalbfleisch & Prentice, 2002).

For individual i at time (t – 1), the probability of his or her survival to time t is given by

$$Pr\left(Y_{it} \middle| \delta_{it} = 1\right) = \Phi\left(X'_{i(t-1)}\beta_p + Z'_{i(t-1)}\gamma_p\right) \quad (3)$$
$$\text{t} = 1, 2, 3, ..., \text{J}$$

where Φ(.) represents the cumulative normal distribution function (Probit). From this equation, estimated survival rates can be obtained for each individual at J – 1 observation intervals. The estimates of $\Phi(\mathbf{X'\beta} + \mathbf{Z'\gamma})$ are then saved for each individual at each follow-up time point as an unbiased estimate of the survival rate.

Given the assumption that the hypothetical disability severity score for those who have been deceased between time (t – 1) and time t (t = 1, 2, …, J), the distribution of survivors' disability severity scores at time t is truncated on the right. Accordingly, the inverse Mills ratio for individual i at time t can be given by

$$\lambda_{it} = -\frac{\varphi\left(X'_{i(t-1)}\beta_p + Z'_{i(t-1)}\gamma_p\right)}{\Phi\left(X'_{i(t-1)}\beta_p + Z'_{i(t-1)}\gamma_p\right)} \text{ if } \delta_{it} = 1 \ (\text{Y} < \text{C}), \quad (4a)$$

$$\lambda_{it} = \frac{\varphi\left(X'_{i(t-1)}\beta_p + Z'_{i(t-1)}\gamma_p\right)}{1-\Phi\left(X'_{i(t-1)}\beta_p + Z'_{i(t-1)}\gamma_p\right)} \quad \text{if } \delta_{it} = 0 \ (Y \geq C), \tag{4b}$$

where φ(.) represents the standard normal density function. Values of λ's at time 0 (first wave) are all zero because no selection bias is present from deaths at the outset of the longitudinal investigation. As defined, the inverse Mills ratio for the deceased is the hazard rate of surviving between two adjacent time points; for those who have survived, it represents the risk of not surviving within an observational interval (Greene, 2003).

With the vector **λ** created, a conditionally unbiased truncated random-effects model is developed on the disability severity score at J time points, given by

$$Y\left(Y \middle| \delta = 1\right) = X'_2\beta_3 + Z'_2\gamma_3 + \sigma'_{e_2 v}\lambda + \varepsilon_3, \tag{5}$$

where $\boldsymbol{\sigma}_{ev}$ is a vector of covariance between $\boldsymbol{\varepsilon}_1$ and **v**, the latent error vector from (3), specified in the estimation process as a vector of the regression coefficients of **λ**, with elements assumed to be normally distributed. Because the survival rate and the disability severity score are inversely correlated, elements in $\boldsymbol{\sigma}_{ev}$ – with the exception of the first – are expected to take negative signs. With **λ** included in the estimation process, the error term $\boldsymbol{\varepsilon}_3$ is assumed to have mean 0 and variance σ_3^2, and to be uncorrelated with $\mathbf{X}_2$, $\mathbf{Z}_2$, and **λ**. When all assumptions on error distributions are satisfied, equation (5) generates unbiased and consistent parameter estimates because observations are presumably conditionally independent of each other.

Note that in equation (5), the inclusion of **λ** and **σ** accounts for the covariance between two error terms, $\boldsymbol{\varepsilon}_1$ and **v**, thereby indicating that the joint distribution of two sequential equations, represented by equation (2), is empirically embedded in (5).

Nonparametric Joint Random-Effects Model

The traditional two-step linear regression estimator and the joint longitudinal models depend on several strong assumptions regarding error distributional functions. When the assumption of multivariate normality for **ε** cannot be satisfied, as is often the case in health transitions (Liu, 2000; Manning, Duan & Rogers, 1987), Equation (5) cannot derive correct estimates for the underlying disability severity score. In these circumstances, Duan's (1983) and Liu's (2000) retransformation methods are extended into the context of repeated measures, assuming a nonparametric distribution of disturbances. One of the advantages of this approach is that researchers do not need to specify a parametric selection model to consider the missing data mechanisms. Rather, the selection bias is handled indirectly through estimating a smearing effect in the estimation process (Duan, 1983; Liu, 2000).

The log transformed nonzero value of the underlying disability severity score is used to address the possible non-linearity of its distribution among those with any disability. For this reason, a two-step procedure is proposed with the first equation meant to estimate the likelihood of having a nonzero disability score. The two-stage nonparametric mixed model is given by

$$\Pr\left(Y > 0\right) = \Phi\left(X'_2\beta_4 + Z'_2\gamma_4\right) \tag{6a}$$

$$log\left(Y \middle| Y > 0\right) = \left(X'_2\beta_5 + Z'_2\gamma_5 + \varepsilon_5\right)\xi, \tag{6b}$$

where ξ serves as a nonparametric adjustment factor for selection bias from high mortality. The expected disability severity score at various points in time can be expressed by the following joint distribution:

$$E\left(\hat{Y} \middle| S = 1\right) = \Phi\left(X'_2\hat{\beta}_4 + Z'_2\hat{\gamma}_4\right) exp\left(X'_2\hat{\beta}_5 + Z'_2\hat{\gamma}_5\right)\hat{\xi}. \tag{7}$$

As previously indicated, the distribution of the error term in health transition data is often skewed without following an identifiable pattern (Duan, 1983; Liu, 2000; Manning, Duan & Rogers, 1987). However, empirical data can be used to estimate values of **ξ** when the error distributional function is uncertain. First, assuming **X** to have full rank:

$$E(Y|Y>0) = E\left[log\left(X_2'\beta_5 + Z_2'\gamma_5 + \varepsilon_5\right)\right] = \int\left[log\left(X_2'\beta_5 + Z_2'\gamma_5 + \varepsilon_5\right)\right]dF(\varepsilon_5). \quad (8)$$

When the error distributional function F is unknown, this cumulative density function, F, is replaced by its empirical estimate $\hat{\mathbf{F}}_j$ at time-point t; this is referred to as the smearing estimate and is given by

$$E\left(\hat{Y}_t|Y_t>0\right) = E\int\left[log\left(X_{2t}'\beta_5 + Z_{2t}'\gamma_5 + \varepsilon_{5t}\right)d\hat{F}_{n_j}\left(\varepsilon_{5t}\right)\right] = \frac{1}{n_t}\sum_{i=1}^{n_t} log\left(X_{2it}'\beta_5 + Z_{2it}'\gamma_5 + \hat{\varepsilon}_{5it}\right) = log\left(X_{2t}'\hat{\beta}_5 + Z_{2t}'\hat{\gamma}_5\right)n_t^{-1}\sum_{i=1}^{n_{tj}} exp\left(\hat{\varepsilon}_{5it}\right), \quad (9)$$

where n_t is the number of observations at time t with nonzero disability severity scores and $\hat{\boldsymbol{\beta}}_5$ and $\hat{\boldsymbol{\gamma}}_5$ can be estimated by employing the maximum likelihood procedure without specifying a disturbance distributional function (Liu, 2000). When the sample size for a longitudinal study is large enough to derive a reliable expected value of errors, such a smearing estimate for the retransformation in log-linear equations is consistent, robust and efficient (Duan, 1983; Liu, 2000; Manning, Duan & Rogers, 1987).

The estimate of ξ at time t can be calculated by the equation

$$\xi_t = \frac{\sum_{i=1}^{n_t} exp\left[log\left(Y_{it}|Y_{it}>0\right) - \left(X_{2it}'\hat{\beta}_5 + Z_{2it}'\hat{\gamma}_5\right)\right]}{n_t}. \quad (10)$$

As presented, the nonparametric random-effects model does not depend on the specification of a given selection process; rather, it estimates an unknown error distribution by the empirical cumulative density function of the estimated regression residuals, and then takes the desired expectation with respect to the expected error distribution. If skeptical whether observations are conditionally independent, researchers might use the inverse Mills ratio as a covariate to account for the potential clustering among survivors thereby deriving more reliable parameter estimates. The complete dependence of this nonparametric approach on empirical data is obvious: If the longitudinal attrition due to reasons other than death is not random making the missingness non-ignorable, then the model-based predicted values of the disability severity score can be still severely biased.

Methodology

Illustrations

Data used for empirical demonstrations are from the Survey of Asset and Health Dynamics among the Oldest Old (AHEAD), a nationally representative investigation of older Americans. This survey, conducted by Institute of Social Research (ISR), University of Michigan, is funded by National Institute on Aging as a supplement to the Health and Retirement Study (HRS). At present, the survey consists of six waves of investigation; the Wave I survey was conducted between October 1993 and April 1994. Specifically, a sample of individuals aged 70 or older (born in 1923 or earlier) was identified throughout the HRS screening of an area probability sample of households in the nation. This procedure identified 9,473 households and 11,965 individuals in the target area range. AHEAD obtains detailed information on a number of domains, including demographics, health status, health care use, housing structure, disability, retirement plans and health and life insurance. Survival information throughout the six waves has been obtained by a link to the data of National Death Index (NDI). The present research uses data of all six waves: 1993, 1995, 1998, 2000, 2002 and 2004.

Disability severity, standing for an individual's health status in this study, is measured by a score of activities of daily living (ADL), instrumental activities of daily living (IADL), and other types of functional limitations (Liu, Engel, Kang & Cowan, 2005). A score of one is given to an individual who has any difficulty with a specific physical or social activity and the number of items for which difficulties are reported is then summed. As a

result, the score ranges from 0 (functional independence) to 15 (maximum disability). When predicting the survival rate (for the parametric joint model) or the probability of having any functional limitation (for the nonparametric joint model), such covariates as: veterans status (1 = veteran, 0 = non-veteran), age, gender (1 = female), education (years in school), ethnicity (1 = white, 0 = others), marital status (1 = currently married, 0 = other), smoking cigarettes and drinking alcohol, the number of serious health conditions, and self-rated health (5 scales: 1 = poor, 5 = excellent) are considered. The first four of these covariates (veteran status, age, gender and education) are used as the control variables in estimating the random-effects models and are rescaled to be centered about their means for analytic convenience. Specification of different sets of covariates at two different estimation stages helps reduce the occurrence of collinearity (Winship & Mare, 1992).

Three sets of the predicted number of functional limitations are compared at six time points; these are derived, respectively, from the conventional single-equation random-effects model, the parametric two-step joint model, and the nonparametric joint model. This provides the basis for examining how well each of these three random-effects longitudinal models fits the observed data for the following two reasons. First, if longitudinal dropouts due to reasons other than death are missing at random (MAR), the trajectory of the observed mean number of functional limitations is approximately unbiased. Here, the accurate description of empirical data serves as a criterion for the quality of a statistical model. Second, even if dropouts due to other reasons are missing not at random (MNAR), useful theoretical implications can be obtained by deviations of model-based predicted values from the empirical data.

The SAS PROC MIXED procedure with repeated measures is used to compute both fixed and random effects and to derive the predicted number of functional limitations at each time point (Littell, Milliken, Stroup, Wolfinger & Schabenberger 2006). Because intervals between two adjacent time points are unequally spaced in the AHEAD longitudinal data the REPEATED/TYPE = SP option was used in executing the SAS PROC.MIXED procedure to represent the autoregressive error structure of the data (Littell, et al., 2006). For analytic simplicity without loss of generality, between-individuals random effects are not further specified with the presence of a specific residual variance/covariance structure. Statistically, a combination of both error types is often found to fit the data about the same as does a model of either type (Hedeker & Gibbons, 2006). Hence, in the estimation process the variable time is treated as a series of dichotomous variables with the last time point, time 5 (time = 0, 1, 2, 3, 4, and 5), used as the reference.

Results

Table 1 presents the results of three random-effects models, the conventional, the parametric two-step and the nonparametric two-stage. In terms of the fixed effects, the intercept suggests the population estimate of the dependent variable at time 5 (year 2004); this time point is used as the reference in specification of five time dichotomous variables and all other covariates are centered about their sample means. The combined regression coefficients of the five time variables demonstrate an inverse-U shaped nonlinear function for the trajectory of transitions in the number of functional limitations, revealing the strong impact of the survival-of-the-fittest selection process among older Americans.

Of the control variables, veterans, older persons and women are expected to have a higher number of functional limitations than do their non-veteran, younger and male counterparts, other variable being equal. All regression coefficients, except those of veteran status, are statistically significant. The regression coefficient of lambda, the inverse Mills ratio, estimated for the parametric second-step random-effects model is sizable (-4.8184), statistically significant and takes a negative sign as expected. This suggests the importance of accounting for clustering effects when analyzing the longitudinal health data of older persons.

All estimates of the random effects are statistically significant. The SP variance/ covariance structure covers a relatively small but statistically significant portion of total variance for the conventional and the parametric two-step

random-effects longitudinal models. The relative size of this variance component increases considerably for the nonparametric random-effects model in which the dependent variable is the natural logarithm of the number of functional limitations among those with any functional limitation. The values of ξ's at the six time points, the adjustment factors in the means for the retransformation in the nonparametric random-effects model (not presented in Table 1) are, respectively, 1.3678 at time 0, 1.2448 at time 1, 1.1371 at time 2, 1.1491 at time 3, 1.1408 at time 4, and 1.2616 at time 5, all are statistically significant. The model Chi-square for each mixed model, reported in the last row of the table, is calculated as the difference in the value of -2 × (log likelihood) between the model with covariates and the model without any covariates.

Table 2 shows four sets of mean numbers of functional limitations in older Americans at six time points - 1993, 1995, 1998, 2000, 2002 and 2004 - derived from observed data and the three types of longitudinal random-effects models, respectively. Compared to the observed data, the conventional single-equation linear random-effects model systematically overestimates the number of functional limitations at every time point except the baseline and this overestimation increases as the survey progresses. The parametric two-step longitudinal joint model somewhat reduces such overestimation, but the adjustment appears very limited and deviations from the observed data are still considerable and systematic. By contrast, the nonparametric longitudinal joint model derives the closest set of the estimates to describe transitions in the number of functional limitations in older Americans.

Table 1: Results of Three Random-Effects Models on Number of Functional Limitations in Older Americans: AHEAD Longitudinal Survey (n = 8,443)

Explanatory Variables and Other Statistics	Conventional Mixed Model	Parametric 2-Step Model[a]	Nonparametric 2-Step Model[b]
Fixed Effects:			
Intercept	5.5045**	5.3967**	1.4515**
Time 0 (1993)	-3.0158**	-2.9079**	-0.4582**
Time 1 (1995)	-0.2583**	-0.1320	0.0028
Time 2 (1998)	0.8780**	0.9613**	0.2348**
Time 3 (2000)	0.9984**	1.0416**	0.2287**
Time 4 (2002)	1.2367**	1.2575**	0.2569**
Veteran status	0.1613	0.1023	0.0292
Age	0.1742**	0.1320**	0.0274**
Female	0.7360**	0.8773**	0.0849**
Education	-0.1665	-0.1519**	-0.0269**
Lambda (λ)		-4.8184**	
Random Effects:			
Spatial power (POW)	0.5651**	0.5295**	0.4571**
Residual	12.3156**	11.5321**	0.4939**
Model Chi-Square	13367.1**	16715.9**	6100.3**

*0.01 < P < 0.05; **P < 0.01; [a] Results of the second-step mixed model; [b] Results of the second-step mixed model for those with at least one functional limitation, with the dependent variable being the natural logarithm of the number of functional limitations

Table 2: Predicted Number of Functional Limitations in Older Americans Derived From Three Random-Effects Models (n = 8,443)

Time Point	Observed and Predicted Number of Functional Limitations			
	Observed	Conventional	Parametric	Nonparametric
1993	2.4887	2.4996	2.4759	2.6918
1995	5.1514	5.2571	5.2518	5.1184
1998	6.1378	6.3934	6.3451	6.1197
2000	6.1602	6.5138	6.4254	6.1598
2002	6.3348	6.7521	6.6413	6.3056
2004	4.9608	5.5154	5.3838	4.9088

Note: All predicted values derived from the three mixed models are statistically significant relative to value zero.

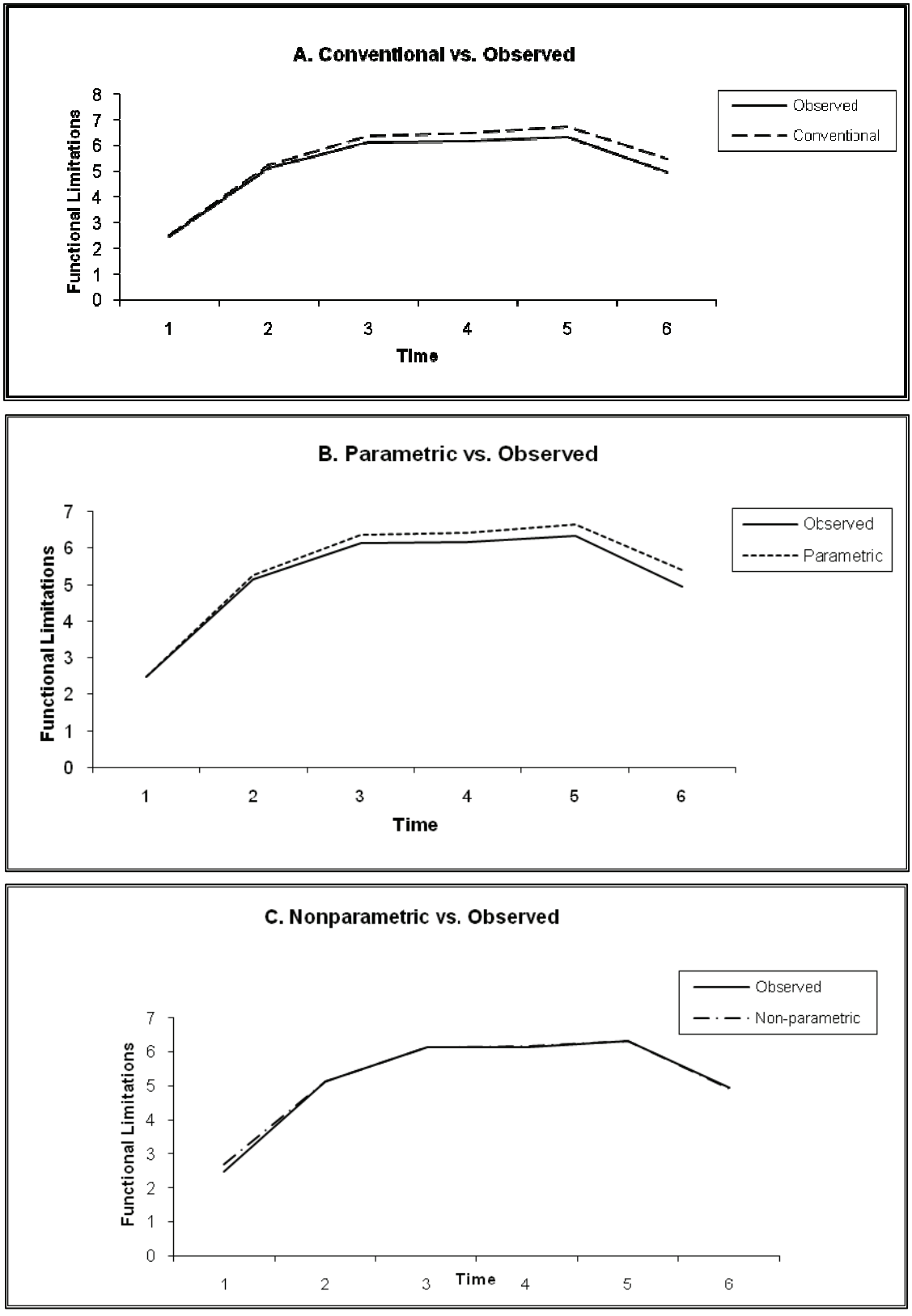
Figure 1 illustrates deviations in the predicted number of functional limitations derived from the three types of mixed models. Panel A compares the observed curve with the predicted values derived from the conventional single-equation random-effects model and shows distinct and systematic separations between the two growth curves. At each time point following the baseline survey, the predicted number of functional limitations obtained from the conventional single-equation random-effects model is considerably higher than the corresponding observed number. The predicted growth curve in Panel B, derived from the parametric longitudinal joint model, displays mitigated separation from the observed curve; however, the deviations remain sizable and systematic thereby reflecting the restriction of using parametric approach to correct for selection bias. In Panel C, the two curves almost coincide, demonstrating the accurate description of the empirical data by applying the nonparametric longitudinal joint modeling, which builds upon observed pattern of health transitions rather than impose strong assumptions on error distributions.

Conclusion

Non-ignorable missing data are important issues in longitudinal data analysis. Despite an abundant literature on this subject, none of the currently existing statistical models has the capacity to handle all types of non-ignorable dropouts (Hogan, Roy & Korkontzelou, 2004). Most models of this type are created for the analysis of longitudinal missing data in clinical experimental studies where repeated measures are often narrowly spaced and mortality is almost nonexistent. With respect to large-scale longitudinal data of older persons, currently available models are not specifically developed to reflect the unique influence of high mortality on estimating and predicting health outcomes at older ages. Because those who have been deceased between assessment periods no longer exist, various assumptions on the measurability of health status for dropouts are not plausible and meaningful.

When mortality rates are high, the direct application of conventional random-effects linear models on longitudinal health data can be associated with serious selection bias. As previously noted, mechanisms leading to biases on parameter estimates have been well documented (Egleston, Scharfstein, Freeman & West, 2006; Hogan, Roy & Korkontzelou, 2004; Kurland, & Heagerty, 2005; Leigh, Ward & Fries, 1993; Liu, 2000; Manning, Duan & Rogers, 1987; Pauler, McCoy & Moinpour, 2003; Ratcliffe, Guo & Ten Have, 2004). This study introduced two refined random-effects joint models and sought to substantially reduce

Figure 1: Transitions in Functional Limitations in Older Americans:
Growth Curves Derived from Three Approaches



bias incurred from changes in the distribution of health outcome data at multiple time points. The parametric longitudinal model is an extension of Heckman's (1979) traditional two-step estimator which, like other parametric joint models, is based on several restrictive assumptions on the joint modeling and error distributional functions. Researchers have questioned and discussed the validity and reliability of this type of two-step estimator. Much of the literature about this estimator focuses on the ill effects of violations against assumptions regarding λ, X and the error distributions (Demirtas, 2004; Fu, Winship & Mare, 2004; Hedeker & Gibbons, 2006; Hogan, Roy & Korkontzelou, 2004; Little & Rubin 2003; Manning, Duan & Rogers, 1987; Winship & Mare, 1992).

This study shows that - as an extended case of the Heckman's perspective - the parametric two-step random-effects joint model has the capacity to reduce some of the deviations from the observed data; however, the degree of this adjustment is limited and deviations remain considerable and systematic. The limited effects of this approach are further evidenced by the similarity between the growth curve derived from this two-step estimator and the curve from the single-equation random-effects model (see Table 1 and Figure 1). In view of the difficulty in verifying assumptions on parametric distributional functions at multiple time points, the use of a nonparametric approach seems a more promising way of modeling longitudinal health data for older persons.

In reality, it is not possible to verify or contradict whether missingness is random by examination of the observed data (Demirtas, 2004; Little & Rubin, 2002). However, if non-death dropouts are missing at random, the selection bias from high mortality rates can be identified by examining the model fitness with observed health transition data. In many empirical applications in which mortality is low, the true cause of the missingness is often thought to be an unmeasured variable that is only moderately correlated with the response, not the response itself. Failure to account for the cause seems to introduce only minor bias (Schafer & Graham, 2002). If this phenomenon can be viewed as a general rule, the agreement of the model-based longitudinal trajectory with the observed curve can be used to measure the sensitivity of predicted health scores in older persons. The nonparametric longitudinal joint model presented herein is created particularly to correct for the selection bias from high mortality rates when the observed data are trustworthy and the non-death longitudinal dropouts are missing at random and thereby ignorable. This nonparametric regression model has the added advantage that the selection information (survival in the present study) does not need to be accounted for directly in the estimation process.

Because the nonparametric approach presented is meant to correct for the selection bias using empirical adjustments, its application must be based on researchers' confidence that biases from ignoring missing data from other causes are minor (Little, 1995). Therefore, its practicality is limited within the circumstances that non-response due to mortality is the only source of non-ignorable dropouts.

If non-death dropouts are missing not at random (MNAR), which is thought to be exceptional by some researchers (Schafer & Graham, 2002), investigators need to compare results generated from various statistical models handling non-ignorable dropouts, such as selection, semi-parametric, pattern-mixture models (Demirtas, 2004; Hedeker & Gibbons, 2006; Hogan, Roy & Korkontzelou, 2004; Little, 1995; Pauler, McCoy & Moinpour, 2003; Robins, Rotnitzky & Zhao, 1995), and the present nonparametric joint approach. However, the effects of dropouts from different reasons on the longitudinal selection bias should be dealt with separately before a unified statistical model handling multi-cause dropouts can be eventually developed (Demirtas, 2004; Hedeker & Gibbons, 2006; Hogan, Roy & Korkontzelou, 2004). For example, dropouts due to mortality, sickness, migration or difficulty in answering sensitive questions may each involve a unique missing data mechanism. To fulfill this task, researchers must collect as much information as possible about various reasons for dropouts and incorporate this information into model development (Little, 1995).

Acknowledgement

This research was supported partly by the National Institute on Aging (NIH/NIA Grant No.: R03AG20140-01). Address correspondence to Dr. Xian Liu, Department of Psychiatry, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814. Email: Xian.Liu@usuhs.edu.

References

Hakan. D. (2004). Modeling incomplete longitudinal data. *Journal of Modern Applied Statistical Methods*, *3*, 305-321.

Duan, N. (1983). Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association*, *78*, 605-610.

Egleston, B. L., Scharfstein, D. O., Freeman, E. E., & West, S. K. (2006). Causal inference for non-mortality outcomes in the presence of death. *Biostatistics*, *8*, 526-545.

Fu, V. K., Winship, C., & Mare, R. D. (2004). Sample selection bias models. In *Handbook of data analysis*, M. Hardy & A. Bryman (Eds.), 409-430. London: Sage.

Greene, W. H. (2003). *Econometric analysis* (*5th Ed.*). New Jersey: Prentice.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Annals of Econometrica*, *47*, 153-161.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Hoboken, NJ: Wiley.

Hogan, J. W., Roy, J., & Korkontzelou, C. (2004). Tutorial in biostatistics: handling drop-out in longitudinal studies. *Statistics in Medicine*, *23*, 1455-1497.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data* (*2nd Ed.*). New York: Wiley.

Kurland, B. F., & Heagerty, P. J. (2005). Directly parameterized regression conditioning on being alive: analysis of longitudinal data truncated by death. *Biostatistics*, *6*, 241-258.

Leigh, J. P., Ward, M. M., & Fries. J. F. (1993). Reducing attrition bias with an instrumental variable in a regression model: results from a panel of rheumatoid arthritis patients. *Statistics in Medicine*, *12*, 1005-1018.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models* (*2nd Ed.*). Gary, NC: SAS Institute, Inc.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, *90*, 1112-1121.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (*2nd Ed.*). New York: Wiley.

Liu, X. (2000). Development of a structural hazard rate model in sociological research. *Sociological Methods & Research*, *29*, 77-117.

Liu, X., Engel, C. C., Kang, H., & Cowan, D. (2005). The Effect of Veteran Status on Mortality among Older Americans and its Pathways. *Population Research and Policy Review*, *24*, 573-592.

Manning, W., Duan, N., & Rogers, W. (1987). Monte Carlo evidence on the choice between sample selection and two-part models. *Journal of Econometrics*, *35*, 59-82.

Pauler, D. K., McCoy, S., & Moinpour, C. (2003). Pattern mixture models for longitudinal quality of life studies in advanced stage disease. *Statistics in Medicine*, *22*, 795-809.

Ratcliffe, S. J., Wensheng G., & Ten Have, T. R. (2004). Joint modeling of longitudinal and survival data via a common frailty. *Biometrics*, *60*, 892-899.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, *90*, 106-121.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*, 147-177.

Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual Reviews of Sociology*, *18*, 327-350.

Yao, Q., Wei, L. J., & Hogan, J. W. (1998). Analysis of incomplete repeated measurements with dependent censoring times. *Biometrika*, *85*, 139-149.

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 414-442

1538 – 9472/10/$95.00

Use of Two Variables Having Common Mean to Improve the Bar-Lev, Bobovitch and Boukai Randomized Response Model

Oluseun Odumade
Educational Testing Service,
Princeton, New Jersey

Sarjinder Singh
Texas A&M University,
Kingsville

A new method to improve the randomized response model due to Bar-Lev, Bobovitch and Boukai (2004) is suggested. It has been observed that if two sensitive (or non sensitive) variables exist that are related to the main study sensitive variable, then those variables could be used to construct ratio type adjustments to the usual estimator of the population mean of a sensitive variable due to Bar-Lev, Bobovitch and Boukai (2004).The relative efficiency of the proposed estimators is studied with respect to the Bar-Lev, Bobovitch and Boukai (2004) models under different situations.

Key words: Randomized response sampling, estimation of population mean, sensitive quantitative variable.

Introduction

The problem of estimating the population total of a sensitive quantitative variable is well known in survey sampling. Warner (1965) was the first to suggest a method to estimate the proportion of sensitive characters (e.g., induced abortions, drugs used) via use of a randomization device such as a deck of cards or a spinner such that respondents' privacy would be protected (Tracy and Mangat (1996) presented a rich description of the literature). Mangat and Singh (1990) proposed a two-stage randomized response model. Leysieffer and Warner (1976) and Lanke (1975, 1976) studied different randomized response procedures at equal levels of protection of the respondents; later Nayak (1994), Bhargava (1996), Zou (1997), Bhargava and Singh (2001, 2002) and Moors (1997) found that the Mangat and Singh (1990) and Warner (1965) models remain equally efficient at equivalent protection; however, this result is not true for all randomized response models (Bhargava, 1996; Bhargava & Singh, 2002). Singh (2003) shows that the Mangat (1994) model remains more efficient than the Warner (1965) model at equal protection: note that the Mangat (1994) model is a special case of the Kuk (1990) model, which is further improved and studied by Gjestvang and Singh (2006). A two stage model developed by Mangat and Singh (1990) was studied by both Kim and Elam (2005) and Kim and Warde (2005). Eichorn and Hayre (1983) suggested a multiplicative model to collect information on sensitive quantitative variables such as, income, tax evasion or amounts of drugs used; this model was further studied by Arnab (1995, 1996). According to Eichorn and Hayre (1983), each respondent in the sample is requested to report a scrambled response $Z_i = SY_i$, where Y_i is the real value of the sensitive quantitative variable, and S is the scrambling variable whose distribution is assumed to be known. Thus, $E_R(S) = \theta$ and $V_R(S) = \gamma^2$ are assumed to be known and positive, therefore, an estimator of the population mean $\overline{Y} = N^{-1}\sum_{i\in\Omega} Y_i$ under simple random with replacement (SRSWR) sampling is given by:

$$\overline{y}_{EH} = \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i}{\theta} \qquad (1.1)$$

Oluseun Odumade is an Associate Psychometric Analyst. Email: oluseunodumade@yahoo.com. Sarjinder Singh is an Assistant Professor in the Department of Mathematics. Email: sarjinder@yahoo.com.

with variance

$$V(\bar{y}_{\text{EH}}) = \frac{1}{n}\sigma_y^2 + \frac{1}{n}C_\gamma^2\bar{Y}^2(1+C_y^2) \quad (1.2)$$

where

$$C_\gamma^2 = \gamma^2/\theta^2,\ \bar{Y} = Y/N$$

and

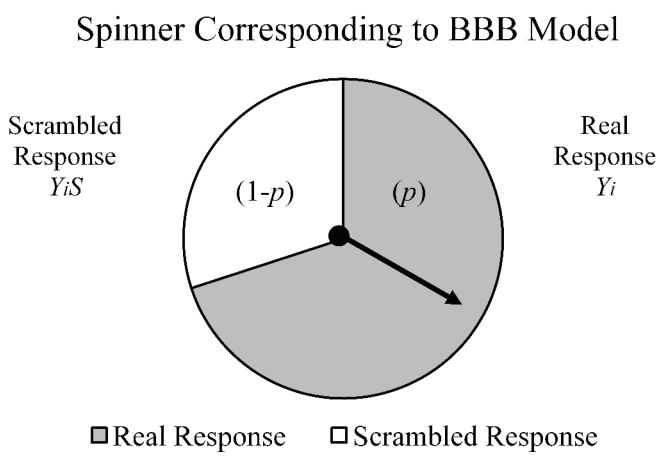
$$C_y = \sigma_y/\bar{Y}.$$

In a randomized response model recently developed by Bar-Lev, Bobovitch and Boukai (2004) (hereafter referred to as the BBB model), the distribution of the responses is given by:

$$Z_i = \begin{cases} Y_i S & \textit{with probability (1-p)} \\ Y_i & \textit{with probability } p \end{cases} \quad (1.3)$$

Thus, each respondent is requested to rotate a spinner unobserved by the interviewer, if the spinner stops in the shaded area, the respondent is requested to report the real response on the sensitive variable, Y_i; if the spinner stops in the non-shaded area, the respondent is requested to report the scrambled response, $Y_i S$, where S is any scrambling variable with a known distribution. Assume that $E(S) = \theta$ and $V(S) = \gamma^2$ are known. Let p be the proportion of the shaded area of the spinner and $(1-p)$ be the non-shaded area of the spinner as shown in Figure 1.

Figure 1: BBB Randomized Response Device



An unbiased estimator of population mean $\bar{Y}$ is given by:

$$\bar{y}_{\text{BBB}} = \frac{1}{n\{(1-p)\theta + p\}}\sum_{i=1}^{n} Z_i \quad (1.4)$$

with variance under SRSWR sampling given by:

$$V[\bar{y}_{\text{BBB}}] = \frac{\bar{Y}^2}{n}[C_y^2 + (1+C_y^2)C_p^2] \quad (1.5)$$

where

$$C_p^2 = \frac{(1-p)\theta^2(1+C_\gamma^2)+p}{[(1-p)\theta+p]^2} - 1. \quad (1.6)$$

Notations

Let $\bar{X}_{1i} = \bar{X}_{2i} = \bar{X}$ be two auxiliary sensitive variables that have a common mean (Tripathi & Chaubey, 1992), and let Y_i be the sensitive variable under study whose mean is to be estimated. Consider a simple random sample of n respondents selected with replacement (SRSWR), where each respondent selected in the sample is requested to rotate three spinners (see Figure 2).

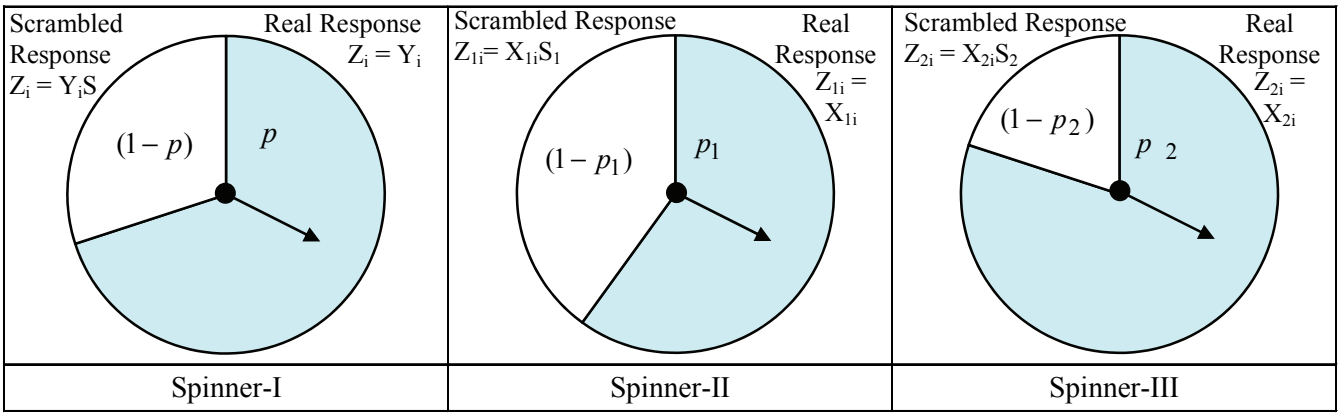
The first spinner is used to collect scrambled response Z_i on the real study variable Y_i with the distribution of responses as:

$$Z_i = \begin{cases} Y_i & \textit{with probability } p \\ Y_i S & \textit{with probability } (1-p) \end{cases} \quad (2.1)$$

where the value of p is assumed to be known. The second spinner is used to collect scrambled response Z_{1i} on the first auxiliary sensitive variable X_{1i} with the distribution of responses as:

$$Z_{1i} = \begin{cases} X_{1i} & \textit{with probability } p_1 \\ S_1 X_{1i} & \textit{with probability } (1-p_1) \end{cases} \quad (2.2)$$

Figure 2: Three Spinners



where the value of p_1 is assumed to be known. The third spinner is used to collect scrambled response Z_{2i} on the second auxiliary sensitive variable X_{2i} with the distribution of responses as:

$$Z_{2i} = \begin{cases} X_{2i} & \text{with probability } p_2 \\ S_2 X_{2i} & \text{with probability } (1-p_2) \end{cases} \quad (2.3)$$

where the value of p_2 is assumed to be known.

Assume that the sample means of the scrambled responses obtained from the respondents in the sample Z_i, Z_{1i} and Z_{2i} are given by:

$$\bar{y}^* = \frac{1}{p+(1-p)\theta}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right),$$

$$\bar{x}_1^* = \frac{1}{p_1+(1-p_1)\theta_1}\left(\frac{1}{n}\sum_{i=1}^{n} Z_{1i}\right),$$

and

$$\bar{x}_2^* = \frac{1}{p_2+(1-p_2)\theta_2}\left(\frac{1}{n}\sum_{i=1}^{n} Z_{2i}\right).$$

Defining $\in = \frac{\bar{y}^*}{\bar{Y}} - 1$, $\delta = \frac{\bar{x}_1^*}{\bar{X}} - 1$, $\eta = \frac{\bar{x}_2^*}{\bar{X}} - 1$, such that $E(\in) = E(\delta) = E(\eta) = 0$, it can be shown that

$$E(\in^2) = \frac{1}{\bar{Y}^2} V(\bar{y}^*) = \frac{1}{n}\left[C_y^2 + (1+C_y^2)C_p^2\right],$$

$$E(\delta^2) = \frac{1}{n}\left[C_{x_1}^2 + (1+C_{x_1}^2)C_{p_1}^2\right],$$

$$E(\eta^2) = \frac{1}{n}\left[C_{x_2}^2 + (1+C_{x_2}^2)C_{p_2}^2\right],$$

$$E(\in \delta) = \frac{1}{n[p+(1-p)\theta][p_1+(1-p_1)\theta_1]}\rho_{YX_1}C_yC_{x_1},$$

$$E(\in \eta) = \frac{1}{n[p+(1-p)\theta][p_2+(1-p_2)\theta_2]}\rho_{YX_2}C_yC_{x_2}$$

and

$$E(\delta\eta) = \frac{1}{n[p_1+(1-p_1)\theta_1][p_2+(1-p_2)\theta_2]}\rho_{X_1X_2}C_{x_1}C_{x_2}$$

where

$$E(Z_i) = pY_i + (1-p)Y_i\theta = [p+(1-p)\theta]Y_i,$$

$$V(X_{1i}) = \sigma_{X_1}^2, V(X_{2i}) = \sigma_{X_2}^2,$$

$$C_p^2 = \frac{(1-p)\theta^2(1+C_\gamma^2)+p}{((1-p)\theta+p)^2} - 1,$$

$$C_{p_1}^2 = \frac{(1-p_1)\theta_1^2\left(1+C_{\gamma_1}^2\right)+p_1}{\left((1-p_1)\theta_1+p_1\right)^2}-1,$$

$$C_{p_2}^2 = \frac{(1-p_2)\theta_2^2\left(1+C_{\gamma_2}^2\right)+p_2}{\left((1-p_2)\theta_2+p_2\right)^2}-1,$$

and

$$\begin{aligned}
&Cov\left(\bar{y}^*,\bar{x}_1^*\right)\\
&= E_1\left[C_2\left(\bar{y}^*,\bar{x}_1\right)\right]+C_1\left[E_2\left(\bar{y}^*\right),E_1\left(\bar{x}_1^*\right)\right]\\
&= E_1\left[E_2\left(\bar{y}^*\cdot\bar{x}_1\right)-E_2\left(\bar{y}^*\right)E_2\left(\bar{x}_1\right)\right]\\
&\quad +C_1\left[E_2\left(\bar{y}^*\right),E_1\left(\bar{x}\right)\right]\\
&= E_1\begin{bmatrix}E_2\left\{\left(\frac{1}{n^2}\sum_{i=1}^{n}Z_i\right)\left(\sum_{i=1}^{n}Z_{1i}\right)\right\}\\ -E_2\left(\frac{1}{n}\sum_{i=1}^{n}Z_i\right)E_2\left(\frac{1}{n}\sum_{i=1}^{n}Z_{1i}\right)\end{bmatrix}\\
&= 0+C_1\begin{bmatrix}\frac{1}{p+(1-p)\theta}\frac{1}{n}\sum_{i=1}^{n}y_i,\\ \frac{1}{p_1+(1-p_1)\theta_1}\frac{1}{n}\sum_{i=1}^{n}x_{1i}\end{bmatrix}\\
&= \frac{\bar{Y}\,\bar{X}_1}{n\left[p+(1-p)\theta\right]\left[p_1+(1-p_1)\theta_1\right]}\rho_{X_1Y_1}C_yC_{x_1}
\end{aligned}$$

Proposed Ratio-Type Estimator

A ratio estimator is defined as:

$$\bar{y}_{Ratio}^* = \bar{y}^*\left(\frac{\bar{x}_1^*}{\bar{x}_2^*}\right). \qquad (3.1)$$

Note that:

$$\bar{y}^* = \bar{Y}(1+\in),$$

$$\bar{x}_1^* = \bar{X}(1+\delta)$$

and

$$\bar{x}_2^* = \bar{X}(1+\eta),$$

thus, the ratio estimator (3.1) can be written in terms of $\in$, δ and η as:

$$\begin{aligned}
\bar{y}_{Ratio}^* &= \bar{Y}(1+\in)\frac{\bar{X}(1+\delta)}{\bar{X}(1+\eta)}\\
&= \bar{Y}(1+\in)(1+\delta)(1+\eta)^{-1}\\
&= \bar{Y}\left[1+\in+\delta+\in\delta\right]\left[1-\eta+\eta^2+....\right]\\
&= \bar{Y}\left[1+\in+\delta-\eta+\eta^2+\in\delta-\in\eta-\delta\eta+....\right]
\end{aligned} \qquad (3.2)$$

From this, the following theorems result.

Theorem 3.1

The bias in the proposed ratio estimator $\bar{y}_{Ratio}^*$ is given by

$$\begin{aligned}
B(\bar{y}_{Ratio}^*) = &\frac{\bar{Y}}{n}\Big[C_{x_2}^2+\left(1+C_{x_2}^2\right)C_{p_2}^2\\
&+\frac{1}{\left[p+(1-p)\theta\right]\left[p_1+(1-p_1)\theta_1\right]}\rho_{YX_1}C_yC_{x_1}\\
&-\frac{1}{\left[p+(1-p)\theta\right]\left[p_2+(1-p_2)\theta_2\right]}\rho_{YX_2}C_yC_{x_2}\\
&-\frac{1}{\left[p_1+(1-p_1)\theta_1\right]\left[p_2+(1-p_2)\theta_2\right]}\rho_{X_1X_2}C_{x_1}C_{x_2}\Big]
\end{aligned} \qquad (3.3)$$

Theorem 3.1: Proof

Taking the expected value on both sides of (3.2) results in:

$$E\left(\bar{y}_{Ratio}^*\right) = \bar{Y}\begin{bmatrix}1+0+0-0+E\left(\eta^2\right)+E\left(\in\delta\right)\\ -E\left(\in\eta\right)-E\left(\delta\eta\right)+....\end{bmatrix}$$

Thus the bias in the proposed ratio estimator $\bar{y}_{Ratio}^*$ is given by:

$$
\begin{aligned}
&B\left(\bar{y}^{*}_{Ratio}\right)\\
&=E\left(\bar{y}^{*}_{Ratio}\right)-\bar{Y}\\
&=\bar{Y}\left[E\left(\eta^{2}\right)+E\left(\in\delta\right)-E\left(\in\eta\right)-E\left(\delta\eta\right)\right]\\
&=\frac{\bar{Y}}{n}\Big[C_{x_2}^{2}+\left(1+C_{x_2}^{2}\right)C_{p_2}^{2}\\
&\quad+\frac{1}{\left[p+(1-p)\theta\right]\left[p_1+(1-p_1)\theta_1\right]}\rho_{YX_1}C_yC_{x_1}\\
&\quad-\frac{1}{\left[p+(1-p)\theta\right]\left[p_2+(1-p_2)\theta_2\right]}\rho_{YX_2}C_yC_{x_2}\\
&\quad-\frac{1}{\left[p_1+(1-p_1)\theta_1\right]\left[p_2+(1-p_2)\theta_2\right]}\rho_{X_1X_2}C_{x_1}C_{x_2}\Big]
\end{aligned}
$$

Thus, Theorem 3.1 is proved.

Theorem 3.2

The mean squared error of the proposed ratio estimator $\bar{y}^{*}_{Ratio}$ is given by

$$
\begin{aligned}
&MSE\left(\bar{y}^{*}_{Ratio}\right)=\\
&\quad\frac{\bar{Y}^{2}}{n}\Big[C_y^{2}+\left(1+C_y^{2}\right)C_p^{2}+C_{x_1}^{2}\\
&\qquad+\left(1+C_{x_1}^{2}\right)C_{p_1}^{2}+C_{x_2}^{2}+\left(1+C_{x_2}^{2}\right)C_{p_2}^{2}\\
&\qquad+\frac{2\rho_{YX_1}C_yC_{x_1}}{\left[p+(1-p)\theta\right]\left[p_1+(1-p_1)\theta_1\right]}\\
&\qquad-\frac{2\rho_{YX_2}C_yC_{x_2}}{\left[p+(1-p)\theta\right]\left[p_2+(1-p_2)\theta_2\right]}\\
&\qquad-\frac{2\rho_{X_1X_2}C_{x_1}C_{x_2}}{\left[p_1+(1-p_1)\theta_1\right]\left[p_2+(1-p_2)\theta_2\right]}\Big]
\end{aligned}
$$

Theorem 3.2: Proof

The mean squared error of the ratio estimator $\bar{y}^{*}_{Ratio}$ is given by

$$
\begin{aligned}
&MSE\left(\bar{y}^{*}_{Ratio}\right)\\
&=E\left[\bar{y}^{*}_{Ratio}-\bar{Y}\right]^{2}\\
&=\bar{Y}^{2}E\left[\in+\delta-\eta\right]^{2}\\
&=\bar{Y}^{2}E\begin{bmatrix}\in^{2}+\delta^{2}+\eta^{2}\\+2\in\delta-2\in\eta-2\delta\eta\end{bmatrix}\\
&=\frac{\bar{Y}^{2}}{n}\Big[C_y^{2}+\left(1+C_y^{2}\right)C_p^{2}+C_{x_1}^{2}\\
&\qquad+\left(1+C_{x_1}^{2}\right)C_{p_1}^{2}+C_{x_2}^{2}+\left(1+C_{x_2}^{2}\right)C_{p_2}^{2}\\
&\qquad+\frac{2\rho_{YX_1}C_yC_{x_1}}{\left[p+(1-p)\theta\right]\left[p_1+(1-p_1)\theta_1\right]}\\
&\qquad-\frac{2\rho_{YX_2}C_yC_{x_2}}{\left[p+(1-p)\theta\right]\left[p_2+(1-p_2)\theta_2\right]}\\
&\qquad-\frac{2\rho_{X_1X_2}C_{x_1}C_{x_2}}{\left[p_1+(1-p_1)\theta_1\right]\left[p_2+(1-p_2)\theta_2\right]}\Big]
\end{aligned}
$$

thus proving Theorem 3.2.

Efficiency of the Proposed Ratio Estimator

The proposed ratio estimator $\bar{y}^{*}_{Ratio}$ will be more efficient than the BBB model if

$$V\left(\bar{y}^{*}_{Ratio}\right)<V\left(\bar{y}_{BBB}\right) \qquad (3.1.1)$$

Using (1.5) and (3.4), results in:

$$C_{x_1}^{2}+\left(1+C_{x_1}^{2}\right)C_{p_1}^{2}+C_{x_2}^{2}+\left(1+C_{x_2}^{2}\right)C_{p_2}^{2}$$

being less than (<)

$$
\begin{aligned}
&\frac{2\rho_{YX_2}C_yC_{x_2}}{\left[p+(1-p)\theta\right]\left[p_2+(1-p_2)\theta_2\right]}\\
&+\frac{2\rho_{X_1X_2}C_{x_1}C_{x_2}}{\left[p_1+(1-p_1)\theta_1\right]\left[p_2+(1-p_2)\theta_2\right]}\\
&-\frac{2\rho_{YX_1}C_yC_{x_1}}{\left[p+(1-p)\theta\right]\left[p_1+(1-p_1)\theta_1\right]}
\end{aligned}
\qquad (3.1.2)
$$

In order to see the magnitude of the proposed ratio estimator $\bar{y}^*_{Ratio}$ with respect to the BBB model the percent relative efficiency is computed as:

$$\mathrm{RE}(\bar{y}_{\mathrm{BBB}}, \bar{y}^*_{\mathrm{Ratio}}) = \frac{V(\bar{y}_{\mathrm{BBB}})}{\mathrm{MSE}(\bar{y}^*_{\mathrm{Ratio}})} \times 100\% \tag{3.1.3}$$

The relative efficiency of the ratio estimator depends on a few parameters such as P, P_1, P_2, C_y, C_{x_1}, C_{x_2}, C_γ, C_{γ_1}, C_{γ_2}, θ, θ_1 and θ_2. The percent relative efficiency (RE) is free from the sample size n and main parameter of interest $\bar{Y}$. Fortran programs were developed in order to find the values of the parameters C_y, C_{x_1}, C_{x_2}, C_γ, C_{γ_1}, C_{γ_2}, ρ_{yx_1}, ρ_{yx_2}, $\rho_{x_1x_2}$, θ, θ_1 and θ_2 by holding P, P_1, and P_2 equal to 0.7 such that the percent RE remains greater than 200%. (Detailed results are shown in Table 3.1 in the Appendix.)

Values of C_y, C_{x_1}, C_{x_2}, C_γ, C_{γ_1}, and C_{γ_2} were changed between 0.1 and 0.5 with a step of 0.2, and the values θ, θ_1 and θ_2 were changed between 0 and 1 with a step of 0.5. It was observed that selecting larger values for θ, θ_1 and θ_2 may lead to inefficient results, thus the choice of these values is critical when using the proposed ratio method in practice.

Table 3.2: Descriptive Statistics of the Percent Relative Efficiency

Mean	289.9
Standard Error	2.9
Median	270.0
Standard Deviation	77.4
Sample Variance	5994.6
Kurtosis	-0.1
Skewness	0.9
Range	298.9
Minimum	200.9
Maximum	499.8
Count	724

The values of ρ_{yx_2} and $\rho_{x_1x_2}$ were changed between 0.1 and 0.9 with a step of 0.2, and ρ_{yx_1} was changed between –0.9 to +0.9 with a step of 0.2. The average percent relative efficiency was 289.9% with a standard deviation of 77.4, median of 270.0%, minimum of 200.9% and maximum of 499.8% (see Table 3.2). It was observed that 724 cases exist in which the RE of the proposed ratio estimator remained between 200.9% and 499.8%.

Proposed Power Transformation Type Estimator

By following the repeated substitution method developed by Garcia and Cebrian (1996), consider a new power transformation ratio type estimator $\bar{y}^*_{Power}$ as:

$$\bar{y}^*_{Power} = \bar{y}^* \left(\frac{\bar{x}_1^*}{\bar{x}_2^*} \right)^\alpha \tag{4.1}$$

where α is a suitably chosen real constant. For example if $\alpha = 0$ then the proposed power transformation ratio type estimator $\bar{y}^*_{Power}$ reduces to the BBB estimator $\bar{y}_{BBB}$. If $\alpha = 1$ then the proposed power transformation ratio type estimator $\bar{y}^*_{Power}$ reduces to the ratio estimator $\bar{y}^*_{Ratio}$.

Note that the proposed transformation type estimator $\bar{y}^*_{Power}$ in terms of $\in$, δ and η can be written as:

$$\begin{aligned}
\bar{y}^*_{Power} &= \bar{Y}(1+\in)\left[\frac{\bar{X}(1+\delta)}{\bar{X}(1+\eta)}\right]^\alpha \\
&= \bar{Y}(1+\in)\left[(1+\delta)(1+\eta)^{-1}\right]^\alpha \\
&= \bar{Y}(1+\in)\left[(1+\delta)(1-\eta+\eta^2+\ldots)\right]^\alpha \\
&= \bar{Y}(1+\in)\left[1+\delta-\eta+\delta^2-\delta\eta+\ldots\right]^\alpha
\end{aligned}$$

{step 4 assumes that $\left|\delta-\eta+\delta^2-\delta\eta+\ldots\right| < 1$}

$$
\begin{aligned}
&= \bar{Y}(1+\in)\begin{bmatrix}1+\alpha(\delta-\eta)+\alpha\delta^2\\ -\alpha\delta\eta+\ldots\end{bmatrix}\\
&= \bar{Y}\begin{bmatrix}1+\in+\alpha(\delta-\eta)+\alpha^2\delta^2\\ -\alpha\delta\eta+\alpha(\in\delta-\in\eta)+\ldots\end{bmatrix}\\
&= \bar{Y}\begin{bmatrix}1+\in+\alpha(\delta-\eta)+\alpha^2\delta^2\\ -\alpha\delta\eta+\alpha\in\delta-\alpha\in\eta+\ldots\end{bmatrix}
\end{aligned}
\qquad (4.2)
$$

This leads to two additional theorems.

Theorem 4.1

The bias in the proposed power transformation ratio type estimator $\bar{y}^*_{Power}$ is given by:

$$
\begin{aligned}
&B\left(\bar{y}^*_{Power}\right)=\\
&\frac{\bar{Y}}{n}\Bigg[\alpha^2\left\{C_{x_1}^2+\left(1+C_{x_1}^2\right)C_{p_1}^2\right\}\\
&\quad-\alpha\Bigg\{\frac{\rho_{X_1X_2}C_{x_1}C_{x_2}}{\left[p_1+(1-p_1)\theta_1\right]\left[p_2+(1-p_2)\theta_2\right]}\\
&\quad-\frac{\rho_{YX_1}C_yC_{x_1}}{\left[p+(1-p)\theta\right]\left[p_1+(1-p_1)\theta_1\right]}\\
&\quad+\frac{\rho_{YX_2}C_yC_{x_2}}{\left[p+(1-p)\theta\right]\left[p_2+(1-p_2)\theta_2\right]}\Bigg\}\Bigg]
\end{aligned}
\qquad (4.3)
$$

Theorem 4.1: Proof

Taking expected value on both sides of (4.2), and using

$$B\left(\bar{y}^*_{Power}\right)=E\left(\bar{y}^*_{Power}\right)-\bar{Y}$$

results in

$$
\begin{aligned}
&B\left(\bar{y}^*_{Power}\right)=\\
&\frac{\bar{Y}}{n}\Bigg[\alpha^2\left\{C_{x_1}^2+\left(1+C_{x_1}^2\right)C_{p_1}^2\right\}\\
&\quad-\alpha\Bigg\{\frac{\rho_{X_1X_2}C_{x_1}C_{x_2}}{\left[p_1+(1-p_1)\theta_1\right]\left[p_2+(1-p_2)\theta_2\right]}\\
&\quad-\frac{\rho_{YX_1}C_yC_{x_1}}{\left[p+(1-p)\theta\right]\left[p_1+(1-p_1)\theta_1\right]}\\
&\quad+\frac{\rho_{YX_2}C_yC_{x_2}}{\left[p+(1-p)\theta\right]\left[p_2+(1-p_2)\theta_2\right]}\Bigg\}\Bigg]
\end{aligned}
$$

Theorem 4.2

The minimum mean squared error of the proposed power transformation ratio type estimator $\bar{y}^*_{Power}$ is given by formula (4.4) as shown below.

Theorem 4.2: Proof

By the definition of mean squared error,

$$
\begin{aligned}
&Min.MSE\left(\bar{y}^*_{Power}\right)=\\
&\frac{\bar{Y}^2}{n}\Bigg[C_y^2+\left(1+C_y^2\right)C_p^2-\frac{\dfrac{1}{\left[p+(1-p)\theta\right]^2}\left\{\dfrac{\rho_{YX_1}C_yC_{x_1}}{\left[p_1+(1-p_1)\theta_1\right]}-\dfrac{\rho_{YX_2}C_yC_{x_2}}{\left[p_2+(1-p_2)\theta_2\right]}\right\}^2}{C_{x_1}^2+\left(1+C_{x_1}^2\right)C_{p_1}^2+C_{x_2}^2+\left(1+C_{x_2}^2\right)C_{p_2}^2-\dfrac{2\rho_{X_1X_2}C_{x_1}C_{x_2}}{\left[p_1+(1-p_1)\theta_1\right]\left[p_2+(1-p_2)\theta_2\right]}}\Bigg]
\end{aligned}
\qquad (4.4)
$$

$$
\begin{aligned}
MSE\left(\bar{y}^{*}_{Power}\right) &= E\left[\bar{y}^{*}_{Power}-\bar{Y}\right]^2 \\
&= \bar{Y}^2 E\left[\in +\alpha\left(\delta-\eta\right)\right]^2 \\
&= \bar{Y}^2 E\begin{bmatrix}\in^2 +\alpha^2\left(\delta-\eta\right)^2 \\ +2\alpha\in\left(\delta-\eta\right)\end{bmatrix} \\
&= \bar{Y}^2 E\begin{bmatrix}\in^2 +\alpha^2\left(\delta^2+\eta^2-2\delta\eta\right) \\ +2\alpha\left(\in\delta-\in\eta\right)\end{bmatrix} \\
&= \frac{\bar{Y}^2}{n}\Big[C_y^2+\left(1+C_y^2\right)C_p^2 \\
&\quad +\alpha^2\left\{\begin{matrix}C_{x_1}^2+\left(1+C_{x_1}^2\right)C_{p_1}^2+C_{x_2}^2+\left(1+C_{x_2}^2\right)C_{p_2}^2 \\ -2\dfrac{\rho_{X_1X_2}C_{x_1}C_{x_2}}{\left[p_1+\left(1-p_1\right)\theta_1\right]\left[p_2+\left(1-p_2\right)\theta_2\right]}\end{matrix}\right\} \\
&\quad +2\alpha\left\{\begin{matrix}\dfrac{\rho_{YX_1}C_yC_{x_1}}{\left[p+\left(1-p\right)\theta\right]\left[p_1+\left(1-p_1\right)\theta_1\right]} \\ -\dfrac{\rho_{YX_2}C_yC_{x_2}}{\left[p+\left(1-p\right)\theta\right]\left[p_2+\left(1-p_2\right)\theta_2\right]}\end{matrix}\right\}\Big]
\end{aligned}
$$

(4.5)

Differentiating (4.5) with respect to α and setting it equal to zero the optimum value of α as shown in Formula 4.6, results in the minimum MSE of $\bar{y}^{*}_{Power}$ as given by (4.4).

Based on these, it is clear that the proposed $\bar{y}^{*}_{Power}$ estimator remains more efficient than $\bar{y}_{BBB}$ for any choice of parameters in the proposed spinners or the design based parameters.

Methodology

Relative Efficiency of the Power Transformation Type Estimator with Respect to the BBB Model

In order to determine the magnitude of the proposed power transformation type estimator $\bar{y}^{*}_{Power}$ with respect to the BBB model the percent RE was computed as:

$$
\mathrm{RE}(\bar{y}_{\mathrm{BBB}},\bar{y}^{*}_{\mathrm{Power}})=\frac{V(\bar{y}_{\mathrm{BBB}})}{\mathrm{MSE}(\bar{y}^{*}_{\mathrm{Power}})}\times 100\% \qquad (4.1.1)
$$

Again the RE of the power transformation estimator depends on parameters such as P, P_1, P_2, C_y, C_{x_1}, C_{x_2}, C_γ, C_{γ_1}, C_{γ_2}, θ, θ_1 and θ_2; the percent RE is free from the sample size n and main parameter of interest $\bar{Y}$. FORTRAN programs were developed in order to determine the values of the parameters C_y, C_{x_1}, C_{x_2}, C_γ, C_{γ_1}, C_{γ_2}, ρ_{yx_1}, ρ_{yx_2}, $\rho_{x_1x_2}$, θ, θ_1 and θ_2 by holding P, P_1, and P_2 equal to 0.7 such that the percent RE remains higher than 200% (see Table 4.1 in the Appendix for results).

The values of C_y, C_{x_1}, C_{x_2}, C_γ, C_{γ_1}, and C_{γ_2} were changed between 0.1 and 0.5 with a step of 0.2, and the values of θ, θ_1 and θ_2 were changed between 0 and 1 with a step of 0.5. It was observed that larger values of θ, θ_1 and θ_2 may lead to slightly less efficient results, thus the choice of these values is critical when using the proposed power method in practice.

Table 4.2: Descriptive Statistics of the Percent Relative Efficiency

Mean	233.05
Standard Error	7.33
Median	215.47
Standard Deviation	47.53
Sample Variance	2258.76
Kurtosis	4.65
Skewness	2.34
Range	178.60
Minimum	200.61
Maximum	379.21
Count	42

The values of ρ_{yx_2} and $\rho_{x_1x_2}$ were changed between 0.1 and 0.9 with a step of 0.2, and ρ_{yx_1}

$$\alpha = \frac{-\frac{1}{[p+(1-p)\theta]}\left\{\frac{\rho_{YX_1}C_yC_{x_1}}{[p_1+(1-p_1)\theta_1]}-\frac{\rho_{YX_2}C_yC_{x_2}}{[p_2+(1-p_2)\theta_2]}\right\}}{C_{x_1}^2+\left(1+C_{x_1}^2\right)C_{p_1}^2+C_{x_2}^2+\left(1+C_{x_2}^2\right)C_{p_2}^2-\frac{2\rho_{X_1X_2}C_{x_1}C_{x_2}}{[p_1+(1-p_1)\theta_1][p_2+(1-p_2)\theta_2]}}$$

(4.6)

$$MSE\left(\bar{y}^*_{Power}\right) =$$

$$\frac{\bar{Y}^2}{n}\left[C_y^2+\left(1+C_y^2\right)C_p^2-\frac{\frac{1}{[p+(1-p)\theta]^2}\left\{\frac{\rho_{YX_1}C_yC_{x_1}}{[p_1+(1-p_1)\theta_1]}-\frac{\rho_{YX_2}C_yC_{x_2}}{[p_2+(1-p_2)\theta_2]}\right\}^2}{C_{x_1}^2+\left(1+C_{x_1}^2\right)C_{p_1}^2+C_{x_2}^2+\left(1+C_{x_2}^2\right)C_{p_2}^2-\frac{2\rho_{X_1X_2}C_{x_1}C_{x_2}}{[p_1+(1-p_1)\theta_1][p_2+(1-p_2)\theta_2]}}\right]$$

was changed between 0.1 to +0.9 with a step of 0.2. The average percent relative efficiency was 233.5% with a standard deviation of 47.53, a median of 215.47%, minimum of 200.16% and maximum of 379.21% (see Table 4.2). It was observed that 42 cases exist where the RE of the proposed ratio estimator remained between 200.16% and 379.21%. As shown in Table 4.1, the optimum values of α remained between –1.56 and +1.56 with a mean equal to zero, standard deviation of 0.93 and mode of 0.49.

Conclusion

In this study new ratio and power transformation type estimators were proposed and compared to the recently described BBB randomized response model. It was observed that the overall magnitude of the relative efficiency of the ratio estimator - unlike the repeated substitution method due to Garcia and Cebrian (1996) - was better than that of the power transformation estimator in the case of scrambled responses.

References

Arnab, R. (1995). On admissibility and optimality of sampling strategies in randomized response surveys. *Sankhy ā* , *B57*, 385-390.

Arnab, R. (1996). Randomized response trials: A unified approach for qualitative data. *Communications in Statistics-Theory and Methods*, *25*(*6*), 1173-1183.

Bar-Lev, S. K., Bobovitch, E., & Boukai, B. (2004). A note on randomized response models for quantitative data. *Metrika*, 255-260.

Bhargava, M. (1996). *An investigation into the efficiencies of certain randomized response strategies*. Unpublished Ph.D. thesis submitted to Punjab Agricultural University, Ludhiana, India.

Bhargava, M., & Singh, R. (2001). Efficiency comparison of certain randomized response schemes with U-model. *Journal of the Indian Society of Agricultural Statistics*, *54*(*1*), 19-28.

Bhargava, M., & Singh, R. (2002). On the efficiency comparison of certain randomized response strategies. *Metrika*, *55*(*3*), 191-197.

Eichhorn, B. H., & Hayre, L. S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*, *7*, 307-316.

Garcia, M. R., & Cebrian, A. A. (1996). Repeated substitution method: The ratio estimator for the population variance. *Metrika*, *43*, 101-105.

Gjestvang, C. R., & Singh, S. (2006). A new randomized response model. *Journal of the Royal Statistical Society*, *B68*, 523-530.

Kim, J. M., & Elam, M. E. (2005). A two-stage stratified Warner's randomized response model using optimal allocation. *Metrika*, 61, 1-7.

Kim, J.-M. and Warde, W. D. (2004). A stratified Warner's randomized response model. *Journal of Statistical Planning and Inference*, *120*(*1-2*), 155-165.

Kim, J. M., & Warde, W. D. (2005). A mixed randomized response model. *Journal of Statistical Planning and Inference*, *133*, 211-221.

Kuk, A. Y. C. (1990). Asking sensitive questions indirectly. *Biometrika*, *77*, 436-438.

Lanke, J. (1975). On the choice of the unrelated question in Simons version of randomized response. *Journal of the American Statistical Association*, *70*, 80-83.

Lanke, J. (1976). On the degree of protection in randomized interviews. *International Statistical Review*, *44*, 197-203.

Leysieffer, F. W., & Warner, S. L. (1976). Respondent jeopardy and optimal designs in randomized response models. *Journal of the American Statistical Association*, *71*, 649-656.

Mangat, N. S. (1994). An improved randomized response strategy. *Journal of the Royal Statistical Society*, *B56*, 93-95.

Mangat, N. S., & Singh, R. (1990). An alternative randomized response procedure. *Biometrika*, *77*(*2*), 439-442.

Moors, J. J. A. (1997). *A critical evaluation of Mangat's two-step procedure in randomized response*. Discussion paper at Center for Economic Research, Tilburg University, The Netherlands.

Nayak, T. K. (1994). On randomized response surveys for estimating a proportion. *Communications in Statistics-Theory and Methods*, *23*(*1*), 3303-3321.

Singh, H. P., & Mathur, N. (2005). Estimation of population mean when coefficient of variation is known using scrambled response technique. *Journal of Statistical Planning and Inference*, *131*, 135-144.

Singh, S. (2003). *Advanced Sampling Theory with Applications: How Michael "Selected" Amy.* The Netherlands: Kluwer Academic Publishers.

Tracy, D. S., & Mangat, N. S. (1996). Some developments in randomized response sampling during the last decade: A follow up of review by Chaudhuri and Mukerjee. *Journal of Applied Statistical Science*, *4*(*2/3*), 147-158.

Tripathi, T. P., & Chaubey, Y. P. (1992). Improved estimation of a finite population mean based on paired observations. *Communications in Statistics-Theory and Methods*, *21*, 3327-3333.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*, 63-69.

Appendix

Table 3.1: Relative efficiency of the proposed ratio estimator with respect to the BBB model for different choice of parameters with $P = P_1 = P_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	Cx_1	Cx_2	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	θ	θ_1	θ_2	RE
0.1	0.1	0.1	0.5	0.5	0.3	-0.9	0.7	0.7	0.0	0.0	0.0	222.54
0.1	0.1	0.1	0.5	0.5	0.5	-0.9	0.5	0.7	0.0	0.0	0.0	366.67
0.1	0.1	0.1	0.5	0.5	0.5	-0.9	0.7	0.5	0.0	0.0	0.0	366.67
0.1	0.1	0.1	0.5	0.5	0.5	-0.7	0.7	0.7	0.0	0.0	0.0	366.67
0.3	0.3	0.3	0.5	0.5	0.3	-0.9	0.7	0.7	0.0	0.0	0.0	222.54
0.3	0.3	0.3	0.5	0.5	0.5	-0.9	0.5	0.7	0.0	0.0	0.0	366.67
0.3	0.3	0.3	0.5	0.5	0.5	-0.9	0.7	0.5	0.0	0.0	0.0	366.67
0.3	0.3	0.3	0.5	0.5	0.5	-0.7	0.7	0.7	0.0	0.0	0.0	366.67
0.5	0.5	0.5	0.5	0.5	0.3	-0.9	0.7	0.7	0.0	0.0	0.0	222.54
0.5	0.5	0.5	0.5	0.5	0.5	-0.9	0.5	0.7	0.0	0.0	0.0	366.67
0.5	0.5	0.5	0.5	0.5	0.5	-0.9	0.7	0.5	0.0	0.0	0.0	366.67
0.5	0.5	0.5	0.5	0.5	0.5	-0.7	0.7	0.7	0.0	0.0	0.0	366.67
0.1	0.1	0.1	0.3	0.3	0.1	-0.9	0.7	0.7	0.5	0.5	0.5	201.28
0.1	0.1	0.1	0.3	0.3	0.3	-0.9	0.3	0.7	0.5	0.5	0.5	452.41
0.1	0.1	0.1	0.3	0.3	0.3	-0.9	0.5	0.5	0.5	0.5	0.5	452.41
0.1	0.1	0.1	0.3	0.3	0.3	-0.9	0.7	0.3	0.5	0.5	0.5	452.41
0.1	0.1	0.1	0.3	0.3	0.3	-0.7	0.5	0.7	0.5	0.5	0.5	452.41
0.1	0.1	0.1	0.3	0.3	0.3	-0.7	0.7	0.5	0.5	0.5	0.5	452.41
0.1	0.1	0.1	0.3	0.3	0.3	-0.5	0.7	0.7	0.5	0.5	0.5	452.41
0.1	0.1	0.1	0.3	0.3	0.5	-0.9	0.3	0.7	0.5	0.5	0.5	392.90
0.1	0.1	0.1	0.3	0.3	0.5	-0.9	0.5	0.5	0.5	0.5	0.5	392.90
0.1	0.1	0.1	0.3	0.3	0.5	-0.9	0.7	0.3	0.5	0.5	0.5	392.90
0.1	0.1	0.1	0.3	0.3	0.5	-0.5	0.5	0.7	0.5	0.5	0.5	284.07
0.1	0.1	0.1	0.3	0.3	0.5	-0.5	0.7	0.5	0.5	0.5	0.5	284.07
0.1	0.1	0.1	0.3	0.3	0.5	-0.1	0.7	0.7	0.5	0.5	0.5	222.45
0.1	0.1	0.1	0.3	0.5	0.1	-0.9	0.5	0.7	0.5	0.5	0.5	200.89
0.1	0.1	0.1	0.3	0.5	0.1	-0.9	0.7	0.7	0.5	0.5	0.5	249.82
0.1	0.1	0.1	0.3	0.5	0.3	-0.9	0.1	0.5	0.5	0.5	0.5	222.45
0.1	0.1	0.1	0.3	0.5	0.3	-0.9	0.5	0.3	0.5	0.5	0.5	284.07
0.1	0.1	0.1	0.3	0.5	0.3	-0.7	0.1	0.7	0.5	0.5	0.5	222.45
0.1	0.1	0.1	0.3	0.5	0.3	-0.7	0.5	0.5	0.5	0.5	0.5	284.07
0.1	0.1	0.1	0.3	0.5	0.3	-0.5	0.5	0.7	0.5	0.5	0.5	284.07
0.1	0.1	0.1	0.3	0.5	0.5	-0.9	0.5	0.3	0.5	0.5	0.5	259.40
0.1	0.1	0.1	0.3	0.5	0.5	-0.7	0.1	0.7	0.5	0.5	0.5	448.46
0.1	0.1	0.1	0.3	0.5	0.5	-0.7	0.7	0.3	0.5	0.5	0.5	259.40
0.1	0.1	0.1	0.3	0.5	0.5	-0.5	0.3	0.7	0.5	0.5	0.5	448.46
0.1	0.1	0.1	0.3	0.5	0.5	-0.3	0.5	0.7	0.5	0.5	0.5	448.46
0.1	0.1	0.1	0.3	0.5	0.5	-0.1	0.7	0.7	0.5	0.5	0.5	448.46
0.1	0.1	0.1	0.5	0.1	0.3	-0.9	0.7	0.3	0.5	0.5	0.5	218.24
0.1	0.1	0.1	0.5	0.1	0.3	-0.9	0.7	0.5	0.5	0.5	0.5	244.11
0.1	0.1	0.1	0.5	0.1	0.3	-0.9	0.7	0.7	0.5	0.5	0.5	276.92

Appendix

Table 3.1 (continued): Relative efficiency of the proposed ratio estimator with respect to the BBB model for different choice of parameters with $P = P_1 = P_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	Cx_1	Cx_2	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	θ	θ_1	θ_2	RE
0.1	0.1	0.1	0.5	0.1	0.3	-0.7	0.7	0.5	0.5	0.5	0.5	203.84
0.1	0.1	0.1	0.5	0.1	0.3	-0.7	0.7	0.7	0.5	0.5	0.5	226.23
0.1	0.1	0.1	0.5	0.1	0.5	-0.9	0.7	0.1	0.5	0.5	0.5	234.56
0.1	0.1	0.1	0.5	0.1	0.5	-0.9	0.7	0.3	0.5	0.5	0.5	289.51
0.1	0.1	0.1	0.5	0.1	0.5	-0.9	0.7	0.5	0.5	0.5	0.5	378.06
0.1	0.1	0.1	0.5	0.1	0.5	-0.7	0.7	0.3	0.5	0.5	0.5	234.56
0.1	0.1	0.1	0.5	0.1	0.5	-0.7	0.7	0.5	0.5	0.5	0.5	289.51
0.1	0.1	0.1	0.5	0.1	0.5	-0.7	0.7	0.7	0.5	0.5	0.5	378.06
0.1	0.1	0.1	0.5	0.1	0.5	-0.5	0.7	0.5	0.5	0.5	0.5	234.56
0.1	0.1	0.1	0.5	0.1	0.5	-0.5	0.7	0.7	0.5	0.5	0.5	289.51
0.1	0.1	0.1	0.5	0.1	0.5	-0.3	0.7	0.7	0.5	0.5	0.5	234.56
0.1	0.1	0.1	0.5	0.3	0.1	-0.9	0.1	0.5	0.5	0.5	0.5	203.84
0.1	0.1	0.1	0.5	0.3	0.1	-0.9	0.1	0.7	0.5	0.5	0.5	226.23
0.1	0.1	0.1	0.5	0.3	0.1	-0.9	0.3	0.3	0.5	0.5	0.5	218.24
0.1	0.1	0.1	0.5	0.3	0.1	-0.9	0.3	0.5	0.5	0.5	0.5	244.11
0.1	0.1	0.1	0.5	0.3	0.1	-0.9	0.3	0.7	0.5	0.5	0.5	276.92
0.1	0.1	0.1	0.5	0.3	0.1	-0.9	0.5	0.1	0.5	0.5	0.5	234.83
0.1	0.1	0.1	0.5	0.3	0.1	-0.9	0.5	0.3	0.5	0.5	0.5	265.04
0.1	0.1	0.1	0.5	0.3	0.1	-0.9	0.5	0.5	0.5	0.5	0.5	304.18
0.1	0.1	0.1	0.5	0.3	0.1	-0.9	0.5	0.7	0.5	0.5	0.5	356.89
0.1	0.1	0.1	0.5	0.3	0.1	-0.9	0.7	0.1	0.5	0.5	0.5	289.91
0.1	0.1	0.1	0.5	0.3	0.1	-0.9	0.7	0.3	0.5	0.5	0.5	337.40
0.1	0.1	0.1	0.5	0.3	0.1	-0.9	0.7	0.5	0.5	0.5	0.5	403.49
0.1	0.1	0.1	0.5	0.3	0.1	-0.7	0.7	0.5	0.5	0.5	0.5	203.84
0.1	0.1	0.1	0.5	0.3	0.1	-0.7	0.7	0.7	0.5	0.5	0.5	226.23
0.1	0.1	0.1	0.5	0.3	0.3	-0.9	0.1	0.5	0.5	0.5	0.5	239.24
0.1	0.1	0.1	0.5	0.3	0.3	-0.9	0.1	0.7	0.5	0.5	0.5	367.17
0.1	0.1	0.1	0.5	0.3	0.3	-0.9	0.3	0.1	0.5	0.5	0.5	214.34
0.1	0.1	0.1	0.5	0.3	0.3	-0.9	0.3	0.3	0.5	0.5	0.5	311.62
0.1	0.1	0.1	0.5	0.3	0.3	-0.9	0.5	0.1	0.5	0.5	0.5	446.81
0.1	0.1	0.1	0.5	0.3	0.3	-0.7	0.3	0.5	0.5	0.5	0.5	239.24
0.1	0.1	0.1	0.5	0.3	0.3	-0.7	0.3	0.7	0.5	0.5	0.5	367.17
0.1	0.1	0.1	0.5	0.3	0.3	-0.7	0.5	0.1	0.5	0.5	0.5	214.34
0.1	0.1	0.1	0.5	0.3	0.3	-0.7	0.5	0.3	0.5	0.5	0.5	311.62
0.1	0.1	0.1	0.5	0.3	0.3	-0.7	0.7	0.1	0.5	0.5	0.5	446.81
0.1	0.1	0.1	0.5	0.3	0.3	-0.5	0.5	0.5	0.5	0.5	0.5	239.24
0.1	0.1	0.1	0.5	0.3	0.3	-0.5	0.5	0.7	0.5	0.5	0.5	367.17
0.1	0.1	0.1	0.5	0.3	0.3	-0.5	0.7	0.1	0.5	0.5	0.5	214.34
0.1	0.1	0.1	0.5	0.3	0.3	-0.5	0.7	0.3	0.5	0.5	0.5	311.62
0.1	0.1	0.1	0.5	0.3	0.3	-0.3	0.7	0.5	0.5	0.5	0.5	239.24
0.1	0.1	0.1	0.5	0.3	0.3	-0.3	0.7	0.7	0.5	0.5	0.5	367.17

Appendix

Table 3.1 (continued): Relative efficiency of the proposed ratio estimator with respect to the BBB model for different choice of parameters with $P = P_1 = P_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	Cx_1	Cx_2	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	θ	θ_1	θ_2	RE
0.1	0.1	0.1	0.5	0.3	0.5	-0.9	0.1	0.7	0.5	0.5	0.5	282.69
0.1	0.1	0.1	0.5	0.3	0.5	-0.9	0.3	0.3	0.5	0.5	0.5	230.06
0.1	0.1	0.1	0.5	0.3	0.5	-0.9	0.5	0.1	0.5	0.5	0.5	366.52
0.1	0.1	0.1	0.5	0.3	0.5	-0.7	0.3	0.5	0.5	0.5	0.5	230.06
0.1	0.1	0.1	0.5	0.3	0.5	-0.7	0.5	0.3	0.5	0.5	0.5	366.52
0.1	0.1	0.1	0.5	0.3	0.5	-0.5	0.3	0.7	0.5	0.5	0.5	230.06
0.1	0.1	0.1	0.5	0.3	0.5	-0.5	0.5	0.5	0.5	0.5	0.5	366.52
0.1	0.1	0.1	0.5	0.3	0.5	-0.5	0.7	0.1	0.5	0.5	0.5	282.69
0.1	0.1	0.1	0.5	0.3	0.5	-0.3	0.5	0.7	0.5	0.5	0.5	366.52
0.1	0.1	0.1	0.5	0.3	0.5	-0.3	0.7	0.3	0.5	0.5	0.5	282.69
0.1	0.1	0.1	0.5	0.3	0.5	-0.1	0.7	0.5	0.5	0.5	0.5	282.69
0.1	0.1	0.1	0.5	0.3	0.5	0.1	0.7	0.7	0.5	0.5	0.5	282.69
0.1	0.1	0.1	0.5	0.5	0.1	-0.9	0.1	0.1	0.5	0.5	0.5	289.51
0.1	0.1	0.1	0.5	0.5	0.1	-0.9	0.1	0.3	0.5	0.5	0.5	378.06
0.1	0.1	0.1	0.5	0.5	0.1	-0.9	0.3	0.1	0.5	0.5	0.5	378.06
0.1	0.1	0.1	0.5	0.5	0.1	-0.7	0.3	0.7	0.5	0.5	0.5	234.56
0.1	0.1	0.1	0.5	0.5	0.1	-0.7	0.5	0.5	0.5	0.5	0.5	234.56
0.1	0.1	0.1	0.5	0.5	0.1	-0.7	0.5	0.7	0.5	0.5	0.5	289.51
0.1	0.1	0.1	0.5	0.5	0.1	-0.7	0.7	0.3	0.5	0.5	0.5	234.56
0.1	0.1	0.1	0.5	0.5	0.1	-0.7	0.7	0.5	0.5	0.5	0.5	289.51
0.1	0.1	0.1	0.5	0.5	0.1	-0.7	0.7	0.7	0.5	0.5	0.5	378.06
0.1	0.1	0.1	0.5	0.5	0.3	-0.9	0.1	0.1	0.5	0.5	0.5	230.06
0.1	0.1	0.1	0.5	0.5	0.3	-0.7	0.1	0.5	0.5	0.5	0.5	282.69
0.1	0.1	0.1	0.5	0.5	0.3	-0.7	0.3	0.3	0.5	0.5	0.5	282.69
0.1	0.1	0.1	0.5	0.5	0.3	-0.7	0.5	0.1	0.5	0.5	0.5	282.69
0.1	0.1	0.1	0.5	0.5	0.3	-0.5	0.3	0.7	0.5	0.5	0.5	366.52
0.1	0.1	0.1	0.5	0.5	0.3	-0.5	0.5	0.5	0.5	0.5	0.5	366.52
0.1	0.1	0.1	0.5	0.5	0.3	-0.5	0.7	0.3	0.5	0.5	0.5	366.52
0.1	0.1	0.1	0.5	0.5	0.3	-0.3	0.5	0.7	0.5	0.5	0.5	230.06
0.1	0.1	0.1	0.5	0.5	0.3	-0.3	0.7	0.5	0.5	0.5	0.5	230.06
0.1	0.1	0.1	0.5	0.5	0.5	-0.9	0.1	0.3	0.5	0.5	0.5	269.97
0.1	0.1	0.1	0.5	0.5	0.5	-0.9	0.3	0.1	0.5	0.5	0.5	269.97
0.1	0.1	0.1	0.5	0.5	0.5	-0.7	0.1	0.5	0.5	0.5	0.5	269.97
0.1	0.1	0.1	0.5	0.5	0.5	-0.7	0.3	0.3	0.5	0.5	0.5	269.97
0.1	0.1	0.1	0.5	0.5	0.5	-0.7	0.5	0.1	0.5	0.5	0.5	269.97
0.1	0.1	0.1	0.5	0.5	0.5	-0.5	0.1	0.7	0.5	0.5	0.5	269.97
0.1	0.1	0.1	0.5	0.5	0.5	-0.5	0.3	0.5	0.5	0.5	0.5	269.97
0.1	0.1	0.1	0.5	0.5	0.5	-0.5	0.5	0.3	0.5	0.5	0.5	269.97
0.1	0.1	0.1	0.5	0.5	0.5	-0.5	0.7	0.1	0.5	0.5	0.5	269.97
0.1	0.1	0.1	0.5	0.5	0.5	-0.3	0.3	0.7	0.5	0.5	0.5	269.97
0.1	0.1	0.1	0.5	0.5	0.5	-0.3	0.5	0.5	0.5	0.5	0.5	269.97

Appendix

Table 3.1 (continued): Relative efficiency of the proposed ratio estimator with respect to the BBB model for different choice of parameters with $P = P_1 = P_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	Cx_1	Cx_2	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	θ	θ_1	θ_2	RE
0.1	0.1	0.1	0.5	0.5	0.5	-0.3	0.7	0.3	0.5	0.5	0.5	269.97
0.1	0.1	0.1	0.5	0.5	0.5	-0.1	0.5	0.7	0.5	0.5	0.5	269.97
0.1	0.1	0.1	0.5	0.5	0.5	-0.1	0.7	0.5	0.5	0.5	0.5	269.97
0.1	0.1	0.1	0.5	0.5	0.5	0.1	0.7	0.7	0.5	0.5	0.5	269.97
0.3	0.3	0.3	0.3	0.3	0.3	-0.9	0.3	0.7	0.5	0.5	0.5	276.81
0.3	0.3	0.3	0.3	0.3	0.3	-0.9	0.5	0.5	0.5	0.5	0.5	276.81
0.3	0.3	0.3	0.3	0.3	0.3	-0.9	0.7	0.3	0.5	0.5	0.5	276.81
0.3	0.3	0.3	0.3	0.3	0.3	-0.7	0.5	0.7	0.5	0.5	0.5	276.81
0.3	0.3	0.3	0.3	0.3	0.3	-0.7	0.7	0.5	0.5	0.5	0.5	276.81
0.3	0.3	0.3	0.3	0.3	0.3	-0.5	0.7	0.7	0.5	0.5	0.5	276.81
0.3	0.3	0.3	0.3	0.3	0.5	-0.9	0.3	0.7	0.5	0.5	0.5	249.71
0.3	0.3	0.3	0.3	0.3	0.5	-0.9	0.5	0.5	0.5	0.5	0.5	249.71
0.3	0.3	0.3	0.3	0.3	0.5	-0.9	0.7	0.3	0.5	0.5	0.5	249.71
0.3	0.3	0.3	0.3	0.3	0.5	-0.7	0.5	0.7	0.5	0.5	0.5	464.50
0.3	0.3	0.3	0.3	0.3	0.5	-0.7	0.7	0.5	0.5	0.5	0.5	464.50
0.3	0.3	0.3	0.3	0.3	0.5	-0.5	0.5	0.7	0.5	0.5	0.5	202.82
0.3	0.3	0.3	0.3	0.3	0.5	-0.5	0.7	0.5	0.5	0.5	0.5	202.82
0.3	0.3	0.3	0.3	0.3	0.5	-0.3	0.7	0.7	0.5	0.5	0.5	324.81
0.3	0.3	0.3	0.3	0.5	0.3	-0.9	0.3	0.5	0.5	0.5	0.5	324.81
0.3	0.3	0.3	0.3	0.5	0.3	-0.9	0.5	0.3	0.5	0.5	0.5	202.82
0.3	0.3	0.3	0.3	0.5	0.3	-0.9	0.7	0.3	0.5	0.5	0.5	464.50
0.3	0.3	0.3	0.3	0.5	0.3	-0.7	0.3	0.7	0.5	0.5	0.5	324.81
0.3	0.3	0.3	0.3	0.5	0.3	-0.7	0.5	0.5	0.5	0.5	0.5	202.82
0.3	0.3	0.3	0.3	0.5	0.3	-0.7	0.7	0.5	0.5	0.5	0.5	464.50
0.3	0.3	0.3	0.3	0.5	0.3	-0.5	0.5	0.7	0.5	0.5	0.5	202.82
0.3	0.3	0.3	0.3	0.5	0.3	-0.5	0.7	0.7	0.5	0.5	0.5	464.50
0.3	0.3	0.3	0.3	0.5	0.5	-0.9	0.3	0.5	0.5	0.5	0.5	447.19
0.3	0.3	0.3	0.3	0.5	0.5	-0.7	0.1	0.7	0.5	0.5	0.5	264.60
0.3	0.3	0.3	0.3	0.5	0.5	-0.7	0.5	0.5	0.5	0.5	0.5	447.19
0.3	0.3	0.3	0.3	0.5	0.5	-0.5	0.3	0.7	0.5	0.5	0.5	264.60
0.3	0.3	0.3	0.3	0.5	0.5	-0.5	0.7	0.5	0.5	0.5	0.5	447.19
0.3	0.3	0.3	0.3	0.5	0.5	-0.3	0.5	0.7	0.5	0.5	0.5	264.60
0.3	0.3	0.3	0.3	0.5	0.5	-0.1	0.7	0.7	0.5	0.5	0.5	264.60
0.3	0.3	0.3	0.5	0.1	0.3	-0.9	0.7	0.5	0.5	0.5	0.5	209.86
0.3	0.3	0.3	0.5	0.1	0.3	-0.9	0.7	0.7	0.5	0.5	0.5	232.88
0.3	0.3	0.3	0.5	0.1	0.5	-0.9	0.7	0.1	0.5	0.5	0.5	201.43
0.3	0.3	0.3	0.5	0.1	0.5	-0.9	0.7	0.3	0.5	0.5	0.5	239.27
0.3	0.3	0.3	0.5	0.1	0.5	-0.9	0.7	0.5	0.5	0.5	0.5	294.63
0.3	0.3	0.3	0.5	0.1	0.5	-0.9	0.7	0.7	0.5	0.5	0.5	383.32
0.3	0.3	0.3	0.5	0.1	0.5	-0.7	0.7	0.3	0.5	0.5	0.5	201.43

Appendix

Table 3.1 (continued): Relative efficiency of the proposed ratio estimator with respect to the BBB model for different choice of parameters with $P = P_1 = P_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	Cx_1	Cx_2	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	θ	θ_1	θ_2	RE
0.3	0.3	0.3	0.5	0.1	0.5	-0.7	0.7	0.5	0.5	0.5	0.5	239.27
0.3	0.3	0.3	0.5	0.1	0.5	-0.7	0.7	0.7	0.5	0.5	0.5	294.63
0.3	0.3	0.3	0.5	0.1	0.5	-0.5	0.7	0.5	0.5	0.5	0.5	201.43
0.3	0.3	0.3	0.5	0.1	0.5	-0.5	0.7	0.7	0.5	0.5	0.5	239.27
0.3	0.3	0.3	0.5	0.1	0.5	-0.3	0.7	0.7	0.5	0.5	0.5	201.43
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.3	0.5	0.5	0.5	0.5	209.86
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.3	0.7	0.5	0.5	0.5	232.88
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.5	0.1	0.5	0.5	0.5	203.16
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.5	0.3	0.5	0.5	0.5	224.67
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.5	0.5	0.5	0.5	0.5	251.26
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.5	0.7	0.5	0.5	0.5	285.00
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.7	0.1	0.5	0.5	0.5	241.72
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.7	0.3	0.5	0.5	0.5	272.79
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.7	0.5	0.5	0.5	0.5	313.03
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.7	0.7	0.5	0.5	0.5	367.18
0.3	0.3	0.3	0.5	0.3	0.3	-0.9	0.1	0.5	0.5	0.5	0.5	205.55
0.3	0.3	0.3	0.5	0.3	0.3	-0.9	0.1	0.7	0.5	0.5	0.5	289.74
0.3	0.3	0.3	0.5	0.3	0.3	-0.9	0.3	0.3	0.5	0.5	0.5	254.94
0.3	0.3	0.3	0.5	0.3	0.3	-0.9	0.3	0.5	0.5	0.5	0.5	398.56
0.3	0.3	0.3	0.5	0.3	0.3	-0.9	0.5	0.1	0.5	0.5	0.5	335.55
0.3	0.3	0.3	0.5	0.3	0.3	-0.7	0.3	0.5	0.5	0.5	0.5	205.55
0.3	0.3	0.3	0.5	0.3	0.3	-0.7	0.3	0.7	0.5	0.5	0.5	289.74
0.3	0.3	0.3	0.5	0.3	0.3	-0.7	0.5	0.3	0.5	0.5	0.5	254.94
0.3	0.3	0.3	0.5	0.3	0.3	-0.7	0.5	0.5	0.5	0.5	0.5	398.56
0.3	0.3	0.3	0.5	0.3	0.3	-0.7	0.7	0.1	0.5	0.5	0.5	335.55
0.3	0.3	0.3	0.5	0.3	0.3	-0.5	0.5	0.5	0.5	0.5	0.5	205.55
0.3	0.3	0.3	0.5	0.3	0.3	-0.5	0.5	0.7	0.5	0.5	0.5	289.74
0.3	0.3	0.3	0.5	0.3	0.3	-0.5	0.7	0.3	0.5	0.5	0.5	254.94
0.3	0.3	0.3	0.5	0.3	0.3	-0.5	0.7	0.5	0.5	0.5	0.5	398.56
0.3	0.3	0.3	0.5	0.3	0.3	-0.3	0.7	0.5	0.5	0.5	0.5	205.55
0.3	0.3	0.3	0.5	0.3	0.3	-0.3	0.7	0.7	0.5	0.5	0.5	289.74
0.3	0.3	0.3	0.5	0.3	0.5	-0.9	0.1	0.7	0.5	0.5	0.5	233.70
0.3	0.3	0.3	0.5	0.3	0.5	-0.9	0.3	0.5	0.5	0.5	0.5	369.21
0.3	0.3	0.3	0.5	0.3	0.5	-0.9	0.5	0.1	0.5	0.5	0.5	286.22
0.3	0.3	0.3	0.5	0.3	0.5	-0.7	0.3	0.7	0.5	0.5	0.5	369.21
0.3	0.3	0.3	0.5	0.3	0.5	-0.7	0.5	0.3	0.5	0.5	0.5	286.22
0.3	0.3	0.3	0.5	0.3	0.5	-0.5	0.5	0.5	0.5	0.5	0.5	286.22
0.3	0.3	0.3	0.5	0.3	0.5	-0.5	0.7	0.1	0.5	0.5	0.5	233.70
0.3	0.3	0.3	0.5	0.3	0.5	-0.3	0.5	0.7	0.5	0.5	0.5	286.22
0.3	0.3	0.3	0.5	0.3	0.5	-0.3	0.7	0.3	0.5	0.5	0.5	233.70

Appendix

Table 3.1 (continued): Relative efficiency of the proposed ratio estimator with respect to the BBB model for different choice of parameters with $P = P_1 = P_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	Cx_1	Cx_2	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	θ	θ_1	θ_2	RE
0.3	0.3	0.3	0.5	0.3	0.5	-0.1	0.7	0.5	0.5	0.5	0.5	233.70
0.3	0.3	0.3	0.5	0.3	0.5	0.1	0.7	0.7	0.5	0.5	0.5	233.70
0.3	0.3	0.3	0.5	0.5	0.1	-0.9	0.1	0.1	0.5	0.5	0.5	239.27
0.3	0.3	0.3	0.5	0.5	0.1	-0.9	0.1	0.3	0.5	0.5	0.5	294.63
0.3	0.3	0.3	0.5	0.5	0.1	-0.9	0.1	0.5	0.5	0.5	0.5	383.32
0.3	0.3	0.3	0.5	0.5	0.1	-0.9	0.3	0.1	0.5	0.5	0.5	294.63
0.3	0.3	0.3	0.5	0.5	0.1	-0.9	0.3	0.3	0.5	0.5	0.5	383.32
0.3	0.3	0.3	0.5	0.5	0.1	-0.9	0.5	0.1	0.5	0.5	0.5	383.32
0.3	0.3	0.3	0.5	0.5	0.1	-0.7	0.3	0.7	0.5	0.5	0.5	201.43
0.3	0.3	0.3	0.5	0.5	0.1	-0.7	0.5	0.5	0.5	0.5	0.5	201.43
0.3	0.3	0.3	0.5	0.5	0.1	-0.7	0.5	0.7	0.5	0.5	0.5	239.27
0.3	0.3	0.3	0.5	0.5	0.1	-0.7	0.7	0.3	0.5	0.5	0.5	201.43
0.3	0.3	0.3	0.5	0.5	0.1	-0.7	0.7	0.5	0.5	0.5	0.5	239.27
0.3	0.3	0.3	0.5	0.5	0.1	-0.7	0.7	0.7	0.5	0.5	0.5	294.63
0.3	0.3	0.3	0.5	0.5	0.3	-0.9	0.1	0.3	0.5	0.5	0.5	369.21
0.3	0.3	0.3	0.5	0.5	0.3	-0.9	0.3	0.1	0.5	0.5	0.5	369.21
0.3	0.3	0.3	0.5	0.5	0.3	-0.7	0.1	0.5	0.5	0.5	0.5	233.70
0.3	0.3	0.3	0.5	0.5	0.3	-0.7	0.3	0.3	0.5	0.5	0.5	233.70
0.3	0.3	0.3	0.5	0.5	0.3	-0.7	0.5	0.1	0.5	0.5	0.5	233.70
0.3	0.3	0.3	0.5	0.5	0.3	-0.5	0.3	0.7	0.5	0.5	0.5	286.22
0.3	0.3	0.3	0.5	0.5	0.3	-0.5	0.5	0.5	0.5	0.5	0.5	286.22
0.3	0.3	0.3	0.5	0.5	0.3	-0.5	0.7	0.3	0.5	0.5	0.5	286.22
0.3	0.3	0.3	0.5	0.5	0.3	-0.3	0.7	0.7	0.5	0.5	0.5	369.21
0.3	0.3	0.3	0.5	0.5	0.5	-0.9	0.1	0.3	0.5	0.5	0.5	223.29
0.3	0.3	0.3	0.5	0.5	0.5	-0.9	0.3	0.1	0.5	0.5	0.5	223.29
0.3	0.3	0.3	0.5	0.5	0.5	-0.7	0.1	0.5	0.5	0.5	0.5	223.29
0.3	0.3	0.3	0.5	0.5	0.5	-0.7	0.3	0.3	0.5	0.5	0.5	223.29
0.3	0.3	0.3	0.5	0.5	0.5	-0.7	0.5	0.1	0.5	0.5	0.5	223.29
0.3	0.3	0.3	0.5	0.5	0.5	-0.5	0.1	0.7	0.5	0.5	0.5	223.29
0.3	0.3	0.3	0.5	0.5	0.5	-0.5	0.3	0.5	0.5	0.5	0.5	223.29
0.3	0.3	0.3	0.5	0.5	0.5	-0.5	0.5	0.3	0.5	0.5	0.5	223.29
0.3	0.3	0.3	0.5	0.5	0.5	-0.5	0.7	0.1	0.5	0.5	0.5	223.29
0.3	0.3	0.3	0.5	0.5	0.5	-0.3	0.3	0.7	0.5	0.5	0.5	223.29
0.3	0.3	0.3	0.5	0.5	0.5	-0.3	0.5	0.5	0.5	0.5	0.5	223.29
0.3	0.3	0.3	0.5	0.5	0.5	-0.3	0.7	0.3	0.5	0.5	0.5	223.29
0.3	0.3	0.3	0.5	0.5	0.5	-0.1	0.5	0.7	0.5	0.5	0.5	223.29
0.3	0.3	0.3	0.5	0.5	0.5	-0.1	0.7	0.5	0.5	0.5	0.5	223.29
0.3	0.3	0.3	0.5	0.5	0.5	0.1	0.7	0.7	0.5	0.5	0.5	223.29
0.5	0.5	0.5	0.3	0.3	0.3	-0.9	0.5	0.7	0.5	0.5	0.5	285.02
0.5	0.5	0.5	0.3	0.3	0.3	-0.9	0.7	0.5	0.5	0.5	0.5	285.02

Appendix

Table 3.1 (continued): Relative efficiency of the proposed ratio estimator with respect to the BBB model for different choice of parameters with $P = P_1 = P_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	Cx_1	Cx_2	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	θ	θ_1	θ_2	RE
0.5	0.5	0.5	0.3	0.3	0.3	-0.7	0.7	0.7	0.5	0.5	0.5	285.02
0.5	0.5	0.5	0.3	0.3	0.5	-0.9	0.5	0.7	0.5	0.5	0.5	431.57
0.5	0.5	0.5	0.3	0.3	0.5	-0.9	0.7	0.5	0.5	0.5	0.5	431.57
0.5	0.5	0.5	0.3	0.3	0.5	-0.7	0.5	0.7	0.5	0.5	0.5	206.61
0.5	0.5	0.5	0.3	0.3	0.5	-0.7	0.7	0.5	0.5	0.5	0.5	206.61
0.5	0.5	0.5	0.3	0.3	0.5	-0.5	0.7	0.7	0.5	0.5	0.5	316.64
0.5	0.5	0.5	0.3	0.5	0.3	-0.9	0.1	0.7	0.5	0.5	0.5	250.05
0.5	0.5	0.5	0.3	0.5	0.3	-0.9	0.5	0.5	0.5	0.5	0.5	316.64
0.5	0.5	0.5	0.3	0.5	0.3	-0.9	0.7	0.3	0.5	0.5	0.5	206.61
0.5	0.5	0.5	0.3	0.5	0.3	-0.7	0.5	0.7	0.5	0.5	0.5	316.64
0.5	0.5	0.5	0.3	0.5	0.3	-0.7	0.7	0.5	0.5	0.5	0.5	206.61
0.5	0.5	0.5	0.3	0.5	0.3	-0.5	0.7	0.7	0.5	0.5	0.5	206.61
0.5	0.5	0.5	0.3	0.5	0.5	-0.9	0.1	0.7	0.5	0.5	0.5	445.03
0.5	0.5	0.5	0.3	0.5	0.5	-0.9	0.7	0.3	0.5	0.5	0.5	274.07
0.5	0.5	0.5	0.3	0.5	0.5	-0.7	0.3	0.7	0.5	0.5	0.5	445.03
0.5	0.5	0.5	0.3	0.5	0.5	-0.5	0.5	0.7	0.5	0.5	0.5	445.03
0.5	0.5	0.5	0.3	0.5	0.5	-0.3	0.7	0.7	0.5	0.5	0.5	445.03
0.5	0.5	0.5	0.5	0.1	0.5	-0.9	0.7	0.5	0.5	0.5	0.5	209.77
0.5	0.5	0.5	0.5	0.1	0.5	-0.9	0.7	0.7	0.5	0.5	0.5	248.42
0.5	0.5	0.5	0.5	0.1	0.5	-0.7	0.7	0.7	0.5	0.5	0.5	209.77
0.5	0.5	0.5	0.5	0.3	0.1	-0.9	0.5	0.7	0.5	0.5	0.5	208.15
0.5	0.5	0.5	0.5	0.3	0.1	-0.9	0.7	0.3	0.5	0.5	0.5	201.92
0.5	0.5	0.5	0.5	0.3	0.1	-0.9	0.7	0.5	0.5	0.5	0.5	221.85
0.5	0.5	0.5	0.5	0.3	0.1	-0.9	0.7	0.7	0.5	0.5	0.5	246.15
0.5	0.5	0.5	0.5	0.3	0.3	-0.9	0.1	0.7	0.5	0.5	0.5	208.96
0.5	0.5	0.5	0.5	0.3	0.3	-0.9	0.3	0.5	0.5	0.5	0.5	256.69
0.5	0.5	0.5	0.5	0.3	0.3	-0.9	0.3	0.7	0.5	0.5	0.5	390.50
0.5	0.5	0.5	0.5	0.3	0.3	-0.9	0.5	0.1	0.5	0.5	0.5	230.38
0.5	0.5	0.5	0.5	0.3	0.3	-0.9	0.5	0.3	0.5	0.5	0.5	332.69
0.5	0.5	0.5	0.5	0.3	0.3	-0.9	0.7	0.1	0.5	0.5	0.5	472.62
0.5	0.5	0.5	0.5	0.3	0.3	-0.7	0.3	0.7	0.5	0.5	0.5	208.96
0.5	0.5	0.5	0.5	0.3	0.3	-0.7	0.5	0.5	0.5	0.5	0.5	256.69
0.5	0.5	0.5	0.5	0.3	0.3	-0.7	0.5	0.7	0.5	0.5	0.5	390.50
0.5	0.5	0.5	0.5	0.3	0.3	-0.7	0.7	0.1	0.5	0.5	0.5	230.38
0.5	0.5	0.5	0.5	0.3	0.3	-0.7	0.7	0.3	0.5	0.5	0.5	332.69
0.5	0.5	0.5	0.5	0.3	0.3	-0.5	0.5	0.7	0.5	0.5	0.5	208.96
0.5	0.5	0.5	0.5	0.3	0.3	-0.5	0.7	0.5	0.5	0.5	0.5	256.69
0.5	0.5	0.5	0.5	0.3	0.3	-0.5	0.7	0.7	0.5	0.5	0.5	390.50
0.5	0.5	0.5	0.5	0.3	0.3	-0.3	0.7	0.7	0.5	0.5	0.5	208.96
0.5	0.5	0.5	0.5	0.3	0.5	-0.9	0.3	0.5	0.5	0.5	0.5	240.66

Appendix

Table 3.1 (continued): Relative efficiency of the proposed ratio estimator with respect to the BBB model for different choice of parameters with $P = P_1 = P_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	Cx_1	Cx_2	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	θ	θ_1	θ_2	RE
0.5	0.5	0.5	0.5	0.3	0.5	-0.9	0.5	0.1	0.5	0.5	0.5	204.21
0.5	0.5	0.5	0.5	0.3	0.5	-0.9	0.5	0.3	0.5	0.5	0.5	374.24
0.5	0.5	0.5	0.5	0.3	0.5	-0.7	0.3	0.7	0.5	0.5	0.5	240.66
0.5	0.5	0.5	0.5	0.3	0.5	-0.7	0.5	0.3	0.5	0.5	0.5	204.21
0.5	0.5	0.5	0.5	0.3	0.5	-0.7	0.5	0.5	0.5	0.5	0.5	374.24
0.5	0.5	0.5	0.5	0.3	0.5	-0.7	0.7	0.1	0.5	0.5	0.5	292.94
0.5	0.5	0.5	0.5	0.3	0.5	-0.5	0.5	0.5	0.5	0.5	0.5	204.21
0.5	0.5	0.5	0.5	0.3	0.5	-0.5	0.5	0.7	0.5	0.5	0.5	374.24
0.5	0.5	0.5	0.5	0.3	0.5	-0.5	0.7	0.3	0.5	0.5	0.5	292.94
0.5	0.5	0.5	0.5	0.3	0.5	-0.3	0.5	0.7	0.5	0.5	0.5	204.21
0.5	0.5	0.5	0.5	0.3	0.5	-0.3	0.7	0.5	0.5	0.5	0.5	292.94
0.5	0.5	0.5	0.5	0.3	0.5	-0.1	0.7	0.7	0.5	0.5	0.5	292.94
0.5	0.5	0.5	0.5	0.5	0.1	-0.9	0.1	0.3	0.5	0.5	0.5	209.77
0.5	0.5	0.5	0.5	0.5	0.1	-0.9	0.1	0.5	0.5	0.5	0.5	248.42
0.5	0.5	0.5	0.5	0.5	0.1	-0.9	0.1	0.7	0.5	0.5	0.5	304.52
0.5	0.5	0.5	0.5	0.5	0.1	-0.9	0.3	0.1	0.5	0.5	0.5	209.77
0.5	0.5	0.5	0.5	0.5	0.1	-0.9	0.3	0.3	0.5	0.5	0.5	248.42
0.5	0.5	0.5	0.5	0.5	0.1	-0.9	0.3	0.5	0.5	0.5	0.5	304.52
0.5	0.5	0.5	0.5	0.5	0.1	-0.9	0.3	0.7	0.5	0.5	0.5	393.35
0.5	0.5	0.5	0.5	0.5	0.1	-0.9	0.5	0.1	0.5	0.5	0.5	248.42
0.5	0.5	0.5	0.5	0.5	0.1	-0.9	0.5	0.3	0.5	0.5	0.5	304.52
0.5	0.5	0.5	0.5	0.5	0.1	-0.9	0.5	0.5	0.5	0.5	0.5	393.35
0.5	0.5	0.5	0.5	0.5	0.1	-0.9	0.7	0.1	0.5	0.5	0.5	304.52
0.5	0.5	0.5	0.5	0.5	0.1	-0.9	0.7	0.3	0.5	0.5	0.5	393.35
0.5	0.5	0.5	0.5	0.5	0.1	-0.7	0.7	0.7	0.5	0.5	0.5	209.77
0.5	0.5	0.5	0.5	0.5	0.3	-0.9	0.1	0.3	0.5	0.5	0.5	240.66
0.5	0.5	0.5	0.5	0.5	0.3	-0.9	0.3	0.1	0.5	0.5	0.5	240.66
0.5	0.5	0.5	0.5	0.5	0.3	-0.7	0.1	0.7	0.5	0.5	0.5	292.94
0.5	0.5	0.5	0.5	0.5	0.3	-0.7	0.3	0.5	0.5	0.5	0.5	292.94
0.5	0.5	0.5	0.5	0.5	0.3	-0.7	0.5	0.3	0.5	0.5	0.5	292.94
0.5	0.5	0.5	0.5	0.5	0.3	-0.7	0.7	0.1	0.5	0.5	0.5	292.94
0.5	0.5	0.5	0.5	0.5	0.3	-0.5	0.3	0.7	0.5	0.5	0.5	204.21
0.5	0.5	0.5	0.5	0.5	0.3	-0.5	0.5	0.5	0.5	0.5	0.5	204.21
0.5	0.5	0.5	0.5	0.5	0.3	-0.5	0.5	0.7	0.5	0.5	0.5	374.24
0.5	0.5	0.5	0.5	0.5	0.3	-0.5	0.7	0.3	0.5	0.5	0.5	204.21
0.5	0.5	0.5	0.5	0.5	0.3	-0.5	0.7	0.5	0.5	0.5	0.5	374.24
0.5	0.5	0.5	0.5	0.5	0.3	-0.3	0.7	0.7	0.5	0.5	0.5	240.66
0.5	0.5	0.5	0.5	0.5	0.5	-0.9	0.1	0.5	0.5	0.5	0.5	456.61
0.5	0.5	0.5	0.5	0.5	0.5	-0.9	0.3	0.3	0.5	0.5	0.5	456.61
0.5	0.5	0.5	0.5	0.5	0.5	-0.9	0.5	0.1	0.5	0.5	0.5	456.61

Appendix

Table 3.1 (continued): Relative efficiency of the proposed ratio estimator with respect to the BBB model for different choice of parameters with $P = P_1 = P_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	Cx_1	Cx_2	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	θ	θ_1	θ_2	RE
0.5	0.5	0.5	0.5	0.5	0.5	-0.7	0.1	0.7	0.5	0.5	0.5	456.61
0.5	0.5	0.5	0.5	0.5	0.5	-0.7	0.3	0.5	0.5	0.5	0.5	456.61
0.5	0.5	0.5	0.5	0.5	0.5	-0.7	0.5	0.3	0.5	0.5	0.5	456.61
0.5	0.5	0.5	0.5	0.5	0.5	-0.7	0.7	0.1	0.5	0.5	0.5	456.61
0.5	0.5	0.5	0.5	0.5	0.5	-0.5	0.3	0.7	0.5	0.5	0.5	456.61
0.5	0.5	0.5	0.5	0.5	0.5	-0.5	0.5	0.5	0.5	0.5	0.5	456.61
0.5	0.5	0.5	0.5	0.5	0.5	-0.5	0.7	0.3	0.5	0.5	0.5	456.61
0.5	0.5	0.5	0.5	0.5	0.5	-0.3	0.5	0.7	0.5	0.5	0.5	456.61
0.5	0.5	0.5	0.5	0.5	0.5	-0.3	0.7	0.5	0.5	0.5	0.5	456.61
0.5	0.5	0.5	0.5	0.5	0.5	-0.1	0.7	0.7	0.5	0.5	0.5	456.61
0.1	0.1	0.1	0.1	0.1	0.1	-0.9	0.1	0.7	1.0	1.0	1.0	255.99
0.1	0.1	0.1	0.1	0.1	0.1	-0.9	0.3	0.5	1.0	1.0	1.0	255.99
0.1	0.1	0.1	0.1	0.1	0.1	-0.9	0.5	0.3	1.0	1.0	1.0	255.99
0.1	0.1	0.1	0.1	0.1	0.1	-0.9	0.7	0.1	1.0	1.0	1.0	255.99
0.1	0.1	0.1	0.1	0.1	0.1	-0.7	0.3	0.7	1.0	1.0	1.0	255.99
0.1	0.1	0.1	0.1	0.1	0.1	-0.7	0.5	0.5	1.0	1.0	1.0	255.99
0.1	0.1	0.1	0.1	0.1	0.1	-0.7	0.7	0.3	1.0	1.0	1.0	255.99
0.1	0.1	0.1	0.1	0.1	0.1	-0.5	0.5	0.7	1.0	1.0	1.0	255.99
0.1	0.1	0.1	0.1	0.1	0.1	-0.5	0.7	0.5	1.0	1.0	1.0	255.99
0.1	0.1	0.1	0.1	0.1	0.1	-0.3	0.7	0.7	1.0	1.0	1.0	255.99
0.1	0.1	0.1	0.3	0.1	0.1	-0.9	0.1	0.7	1.0	1.0	1.0	205.76
0.1	0.1	0.1	0.3	0.1	0.1	-0.9	0.3	0.1	1.0	1.0	1.0	205.76
0.1	0.1	0.1	0.3	0.1	0.1	-0.9	0.3	0.3	1.0	1.0	1.0	225.67
0.1	0.1	0.1	0.3	0.1	0.1	-0.9	0.3	0.5	1.0	1.0	1.0	249.85
0.1	0.1	0.1	0.3	0.1	0.1	-0.9	0.3	0.7	1.0	1.0	1.0	279.84
0.1	0.1	0.1	0.3	0.1	0.1	-0.9	0.5	0.1	1.0	1.0	1.0	279.84
0.1	0.1	0.1	0.3	0.1	0.1	-0.9	0.5	0.3	1.0	1.0	1.0	318.00
0.1	0.1	0.1	0.3	0.1	0.1	-0.9	0.5	0.5	1.0	1.0	1.0	368.22
0.1	0.1	0.1	0.3	0.1	0.1	-0.9	0.5	0.7	1.0	1.0	1.0	437.27
0.1	0.1	0.1	0.3	0.1	0.1	-0.9	0.7	0.1	1.0	1.0	1.0	437.27
0.1	0.1	0.1	0.3	0.1	0.1	-0.7	0.3	0.7	1.0	1.0	1.0	205.76
0.1	0.1	0.1	0.3	0.1	0.1	-0.7	0.5	0.1	1.0	1.0	1.0	205.76
0.1	0.1	0.1	0.3	0.1	0.1	-0.7	0.5	0.3	1.0	1.0	1.0	225.67
0.1	0.1	0.1	0.3	0.1	0.1	-0.7	0.5	0.5	1.0	1.0	1.0	249.85
0.1	0.1	0.1	0.3	0.1	0.1	-0.7	0.5	0.7	1.0	1.0	1.0	279.84
0.1	0.1	0.1	0.3	0.1	0.1	-0.7	0.7	0.1	1.0	1.0	1.0	279.84
0.1	0.1	0.1	0.3	0.1	0.1	-0.7	0.7	0.3	1.0	1.0	1.0	318.00
0.1	0.1	0.1	0.3	0.1	0.1	-0.7	0.7	0.5	1.0	1.0	1.0	368.22
0.1	0.1	0.1	0.3	0.1	0.1	-0.7	0.7	0.7	1.0	1.0	1.0	437.27
0.1	0.1	0.1	0.3	0.1	0.1	-0.5	0.5	0.7	1.0	1.0	1.0	205.76

Appendix

Table 3.1 (continued): Relative efficiency of the proposed ratio estimator with respect to the BBB model for different choice of parameters with $P = P_1 = P_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	Cx_1	Cx_2	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	θ	θ_1	θ_2	RE
0.1	0.1	0.1	0.3	0.1	0.1	-0.5	0.7	0.1	1.0	1.0	1.0	205.76
0.1	0.1	0.1	0.3	0.1	0.1	-0.5	0.7	0.3	1.0	1.0	1.0	225.67
0.1	0.1	0.1	0.3	0.1	0.1	-0.5	0.7	0.5	1.0	1.0	1.0	249.85
0.1	0.1	0.1	0.3	0.1	0.1	-0.5	0.7	0.7	1.0	1.0	1.0	279.84
0.1	0.1	0.1	0.3	0.1	0.1	-0.3	0.7	0.7	1.0	1.0	1.0	205.76
0.1	0.1	0.1	0.3	0.1	0.3	-0.9	0.5	0.3	1.0	1.0	1.0	248.26
0.1	0.1	0.1	0.3	0.1	0.3	-0.9	0.5	0.5	1.0	1.0	1.0	364.76
0.1	0.1	0.1	0.3	0.1	0.3	-0.7	0.5	0.5	1.0	1.0	1.0	248.26
0.1	0.1	0.1	0.3	0.1	0.3	-0.7	0.5	0.7	1.0	1.0	1.0	364.76
0.1	0.1	0.1	0.3	0.1	0.3	-0.7	0.7	0.1	1.0	1.0	1.0	364.76
0.1	0.1	0.1	0.3	0.1	0.3	-0.5	0.5	0.7	1.0	1.0	1.0	248.26
0.1	0.1	0.1	0.3	0.1	0.3	-0.5	0.7	0.1	1.0	1.0	1.0	248.26
0.1	0.1	0.1	0.3	0.1	0.3	-0.5	0.7	0.3	1.0	1.0	1.0	364.76
0.1	0.1	0.1	0.3	0.1	0.3	-0.3	0.7	0.3	1.0	1.0	1.0	248.26
0.1	0.1	0.1	0.3	0.1	0.3	-0.3	0.7	0.5	1.0	1.0	1.0	364.76
0.1	0.1	0.1	0.3	0.1	0.3	-0.1	0.7	0.5	1.0	1.0	1.0	248.26
0.1	0.1	0.1	0.3	0.1	0.3	-0.1	0.7	0.7	1.0	1.0	1.0	364.76
0.1	0.1	0.1	0.3	0.1	0.3	0.1	0.7	0.7	1.0	1.0	1.0	248.26
0.1	0.1	0.1	0.3	0.1	0.5	-0.9	0.7	0.5	1.0	1.0	1.0	202.54
0.1	0.1	0.1	0.3	0.1	0.5	-0.9	0.7	0.7	1.0	1.0	1.0	358.04
0.1	0.1	0.1	0.3	0.1	0.5	-0.7	0.7	0.7	1.0	1.0	1.0	245.12
0.1	0.1	0.1	0.3	0.3	0.1	-0.9	0.1	0.1	1.0	1.0	1.0	364.76
0.1	0.1	0.1	0.3	0.3	0.1	-0.7	0.1	0.5	1.0	1.0	1.0	248.26
0.1	0.1	0.1	0.3	0.3	0.1	-0.7	0.1	0.7	1.0	1.0	1.0	364.76
0.1	0.1	0.1	0.3	0.3	0.1	-0.7	0.3	0.3	1.0	1.0	1.0	248.26
0.1	0.1	0.1	0.3	0.3	0.1	-0.7	0.3	0.5	1.0	1.0	1.0	364.76
0.1	0.1	0.1	0.3	0.3	0.1	-0.7	0.5	0.1	1.0	1.0	1.0	248.26
0.1	0.1	0.1	0.3	0.3	0.1	-0.7	0.5	0.3	1.0	1.0	1.0	364.76
0.1	0.1	0.1	0.3	0.3	0.1	-0.7	0.7	0.1	1.0	1.0	1.0	364.76
0.1	0.1	0.1	0.3	0.3	0.1	-0.5	0.5	0.7	1.0	1.0	1.0	248.26
0.1	0.1	0.1	0.3	0.3	0.1	-0.5	0.7	0.5	1.0	1.0	1.0	248.26
0.1	0.1	0.1	0.3	0.3	0.1	-0.5	0.7	0.7	1.0	1.0	1.0	364.76
0.1	0.1	0.1	0.3	0.3	0.3	-0.9	0.1	0.3	1.0	1.0	1.0	203.60
0.1	0.1	0.1	0.3	0.3	0.3	-0.9	0.3	0.1	1.0	1.0	1.0	203.60
0.1	0.1	0.1	0.3	0.3	0.3	-0.7	0.1	0.5	1.0	1.0	1.0	203.60
0.1	0.1	0.1	0.3	0.3	0.3	-0.7	0.3	0.3	1.0	1.0	1.0	203.60
0.1	0.1	0.1	0.3	0.3	0.3	-0.7	0.5	0.1	1.0	1.0	1.0	203.60
0.1	0.1	0.1	0.3	0.3	0.3	-0.5	0.1	0.7	1.0	1.0	1.0	203.60
0.1	0.1	0.1	0.3	0.3	0.3	-0.5	0.3	0.5	1.0	1.0	1.0	203.60
0.1	0.1	0.1	0.3	0.3	0.3	-0.5	0.5	0.3	1.0	1.0	1.0	203.60

Appendix

Table 3.1 (continued): Relative efficiency of the proposed ratio estimator with respect to the BBB model for different choice of parameters with $P = P_1 = P_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	Cx_1	Cx_2	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	θ	θ_1	θ_2	RE
0.1	0.1	0.1	0.3	0.3	0.3	-0.5	0.7	0.1	1.0	1.0	1.0	203.60
0.1	0.1	0.1	0.3	0.3	0.3	-0.3	0.3	0.7	1.0	1.0	1.0	203.60
0.1	0.1	0.1	0.3	0.3	0.3	-0.3	0.5	0.5	1.0	1.0	1.0	203.60
0.1	0.1	0.1	0.3	0.3	0.3	-0.3	0.7	0.3	1.0	1.0	1.0	203.60
0.1	0.1	0.1	0.3	0.3	0.3	-0.1	0.5	0.7	1.0	1.0	1.0	203.60
0.1	0.1	0.1	0.3	0.3	0.3	-0.1	0.7	0.5	1.0	1.0	1.0	203.60
0.1	0.1	0.1	0.3	0.3	0.3	0.1	0.7	0.7	1.0	1.0	1.0	203.60
0.1	0.1	0.1	0.3	0.3	0.5	-0.9	0.1	0.7	1.0	1.0	1.0	243.59
0.1	0.1	0.1	0.3	0.3	0.5	-0.9	0.3	0.5	1.0	1.0	1.0	243.59
0.1	0.1	0.1	0.3	0.3	0.5	-0.9	0.5	0.3	1.0	1.0	1.0	243.59
0.1	0.1	0.1	0.3	0.3	0.5	-0.9	0.7	0.1	1.0	1.0	1.0	243.59
0.1	0.1	0.1	0.3	0.3	0.5	-0.3	0.5	0.7	1.0	1.0	1.0	354.77
0.1	0.1	0.1	0.3	0.3	0.5	-0.3	0.7	0.5	1.0	1.0	1.0	354.77
0.1	0.1	0.1	0.3	0.3	0.5	0.1	0.7	0.7	1.0	1.0	1.0	243.59
0.1	0.1	0.1	0.3	0.5	0.1	-0.9	0.1	0.5	1.0	1.0	1.0	273.92
0.1	0.1	0.1	0.3	0.5	0.1	-0.9	0.3	0.3	1.0	1.0	1.0	221.81
0.1	0.1	0.1	0.3	0.5	0.1	-0.9	0.3	0.5	1.0	1.0	1.0	422.99
0.1	0.1	0.1	0.3	0.5	0.1	-0.9	0.5	0.3	1.0	1.0	1.0	310.38
0.1	0.1	0.1	0.3	0.5	0.1	-0.9	0.7	0.1	1.0	1.0	1.0	245.12
0.1	0.1	0.1	0.3	0.5	0.1	-0.7	0.7	0.7	1.0	1.0	1.0	245.12
0.1	0.1	0.1	0.3	0.5	0.3	-0.9	0.3	0.3	1.0	1.0	1.0	354.77
0.1	0.1	0.1	0.3	0.5	0.3	-0.7	0.3	0.5	1.0	1.0	1.0	354.77
0.1	0.1	0.1	0.3	0.5	0.3	-0.5	0.3	0.7	1.0	1.0	1.0	354.77
0.1	0.1	0.1	0.3	0.5	0.5	-0.9	0.5	0.3	1.0	1.0	1.0	303.12
0.1	0.1	0.1	0.3	0.5	0.5	-0.7	0.7	0.3	1.0	1.0	1.0	303.12
0.1	0.1	0.1	0.5	0.1	0.1	-0.9	0.5	0.7	1.0	1.0	1.0	201.69
0.1	0.1	0.1	0.5	0.1	0.1	-0.9	0.7	0.1	1.0	1.0	1.0	215.39
0.1	0.1	0.1	0.5	0.1	0.1	-0.9	0.7	0.3	1.0	1.0	1.0	222.96
0.1	0.1	0.1	0.5	0.1	0.1	-0.9	0.7	0.5	1.0	1.0	1.0	231.08
0.1	0.1	0.1	0.5	0.1	0.1	-0.9	0.7	0.7	1.0	1.0	1.0	239.82
0.1	0.1	0.1	0.5	0.1	0.1	-0.7	0.7	0.7	1.0	1.0	1.0	201.69
0.1	0.1	0.1	0.5	0.1	0.3	-0.9	0.5	0.1	1.0	1.0	1.0	222.49
0.1	0.1	0.1	0.5	0.1	0.3	-0.9	0.5	0.3	1.0	1.0	1.0	248.65
0.1	0.1	0.1	0.5	0.1	0.3	-0.9	0.5	0.5	1.0	1.0	1.0	281.79
0.1	0.1	0.1	0.5	0.1	0.3	-0.9	0.5	0.7	1.0	1.0	1.0	325.11
0.1	0.1	0.1	0.5	0.1	0.3	-0.9	0.7	0.1	1.0	1.0	1.0	469.47
0.1	0.1	0.1	0.5	0.1	0.3	-0.7	0.5	0.3	1.0	1.0	1.0	207.91
0.1	0.1	0.1	0.5	0.1	0.3	-0.7	0.5	0.5	1.0	1.0	1.0	230.58
0.1	0.1	0.1	0.5	0.1	0.3	-0.7	0.5	0.7	1.0	1.0	1.0	258.80
0.1	0.1	0.1	0.5	0.1	0.3	-0.7	0.7	0.1	1.0	1.0	1.0	342.67

Appendix

Table 3.1 (continued): Relative efficiency of the proposed ratio estimator with respect to the BBB model for different choice of parameters with $P = P_1 = P_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	Cx_1	Cx_2	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	θ	θ_1	θ_2	RE
0.1	0.1	0.1	0.5	0.1	0.3	-0.7	0.7	0.3	1.0	1.0	1.0	408.94
0.1	0.1	0.1	0.5	0.1	0.3	-0.5	0.5	0.7	1.0	1.0	1.0	214.95
0.1	0.1	0.1	0.5	0.1	0.3	-0.5	0.7	0.1	1.0	1.0	1.0	269.80
0.1	0.1	0.1	0.5	0.1	0.3	-0.5	0.7	0.3	1.0	1.0	1.0	309.26
0.1	0.1	0.1	0.5	0.1	0.3	-0.5	0.7	0.5	1.0	1.0	1.0	362.24
0.1	0.1	0.1	0.5	0.1	0.3	-0.5	0.7	0.7	1.0	1.0	1.0	437.12
0.1	0.1	0.1	0.5	0.1	0.3	-0.3	0.7	0.1	1.0	1.0	1.0	222.49
0.1	0.1	0.1	0.5	0.1	0.3	-0.3	0.7	0.3	1.0	1.0	1.0	248.65
0.1	0.1	0.1	0.5	0.1	0.3	-0.3	0.7	0.5	1.0	1.0	1.0	281.79
0.1	0.1	0.1	0.5	0.1	0.3	-0.3	0.7	0.7	1.0	1.0	1.0	325.11
0.1	0.1	0.1	0.5	0.1	0.3	-0.1	0.7	0.3	1.0	1.0	1.0	207.91
0.1	0.1	0.1	0.5	0.1	0.3	-0.1	0.7	0.5	1.0	1.0	1.0	230.58
0.1	0.1	0.1	0.5	0.1	0.3	-0.1	0.7	0.7	1.0	1.0	1.0	258.80
0.1	0.1	0.1	0.5	0.1	0.3	0.1	0.7	0.7	1.0	1.0	1.0	214.95
0.1	0.1	0.1	0.5	0.1	0.5	-0.9	0.5	0.7	1.0	1.0	1.0	229.58
0.1	0.1	0.1	0.5	0.1	0.5	-0.9	0.7	0.1	1.0	1.0	1.0	359.78
0.1	0.1	0.1	0.5	0.1	0.5	-0.7	0.7	0.1	1.0	1.0	1.0	280.29
0.1	0.1	0.1	0.5	0.1	0.5	-0.7	0.7	0.3	1.0	1.0	1.0	359.78
0.1	0.1	0.1	0.5	0.1	0.5	-0.5	0.7	0.1	1.0	1.0	1.0	229.58
0.1	0.1	0.1	0.5	0.1	0.5	-0.5	0.7	0.3	1.0	1.0	1.0	280.29
0.1	0.1	0.1	0.5	0.1	0.5	-0.5	0.7	0.5	1.0	1.0	1.0	359.78
0.1	0.1	0.1	0.5	0.1	0.5	-0.3	0.7	0.3	1.0	1.0	1.0	229.58
0.1	0.1	0.1	0.5	0.1	0.5	-0.3	0.7	0.5	1.0	1.0	1.0	280.29
0.1	0.1	0.1	0.5	0.1	0.5	-0.3	0.7	0.7	1.0	1.0	1.0	359.78
0.1	0.1	0.1	0.5	0.1	0.5	-0.1	0.7	0.5	1.0	1.0	1.0	229.58
0.1	0.1	0.1	0.5	0.1	0.5	-0.1	0.7	0.7	1.0	1.0	1.0	280.29
0.1	0.1	0.1	0.5	0.1	0.5	0.1	0.7	0.7	1.0	1.0	1.0	229.58
0.1	0.1	0.1	0.5	0.3	0.1	-0.9	0.1	0.1	1.0	1.0	1.0	342.67
0.1	0.1	0.1	0.5	0.3	0.1	-0.9	0.1	0.3	1.0	1.0	1.0	408.94
0.1	0.1	0.1	0.5	0.3	0.1	-0.9	0.3	0.1	1.0	1.0	1.0	469.47
0.1	0.1	0.1	0.5	0.3	0.1	-0.7	0.1	0.3	1.0	1.0	1.0	207.91
0.1	0.1	0.1	0.5	0.3	0.1	-0.7	0.1	0.5	1.0	1.0	1.0	230.58
0.1	0.1	0.1	0.5	0.3	0.1	-0.7	0.1	0.7	1.0	1.0	1.0	258.80
0.1	0.1	0.1	0.5	0.3	0.1	-0.7	0.3	0.1	1.0	1.0	1.0	222.49
0.1	0.1	0.1	0.5	0.3	0.1	-0.7	0.3	0.3	1.0	1.0	1.0	248.65
0.1	0.1	0.1	0.5	0.3	0.1	-0.7	0.3	0.5	1.0	1.0	1.0	281.79
0.1	0.1	0.1	0.5	0.3	0.1	-0.7	0.3	0.7	1.0	1.0	1.0	325.11
0.1	0.1	0.1	0.5	0.3	0.1	-0.7	0.5	0.1	1.0	1.0	1.0	269.80
0.1	0.1	0.1	0.5	0.3	0.1	-0.7	0.5	0.3	1.0	1.0	1.0	309.26
0.1	0.1	0.1	0.5	0.3	0.1	-0.7	0.5	0.5	1.0	1.0	1.0	362.24

Appendix

Table 3.1 (continued): Relative efficiency of the proposed ratio estimator with respect to the BBB model for different choice of parameters with $P = P_1 = P_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	Cx_1	Cx_2	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	θ	θ_1	θ_2	RE
0.1	0.1	0.1	0.5	0.3	0.1	-0.7	0.5	0.7	1.0	1.0	1.0	437.12
0.1	0.1	0.1	0.5	0.3	0.1	-0.7	0.7	0.1	1.0	1.0	1.0	342.67
0.1	0.1	0.1	0.5	0.3	0.1	-0.7	0.7	0.3	1.0	1.0	1.0	408.94
0.1	0.1	0.1	0.5	0.3	0.1	-0.5	0.5	0.7	1.0	1.0	1.0	214.95
0.1	0.1	0.1	0.5	0.3	0.1	-0.5	0.7	0.3	1.0	1.0	1.0	207.91
0.1	0.1	0.1	0.5	0.3	0.1	-0.5	0.7	0.5	1.0	1.0	1.0	230.58
0.1	0.1	0.1	0.5	0.3	0.1	-0.5	0.7	0.7	1.0	1.0	1.0	258.80
0.1	0.1	0.1	0.5	0.3	0.3	-0.9	0.1	0.1	1.0	1.0	1.0	207.50
0.1	0.1	0.1	0.5	0.3	0.3	-0.9	0.1	0.3	1.0	1.0	1.0	294.07
0.1	0.1	0.1	0.5	0.3	0.3	-0.9	0.3	0.1	1.0	1.0	1.0	407.37
0.1	0.1	0.1	0.5	0.3	0.3	-0.7	0.1	0.5	1.0	1.0	1.0	230.08
0.1	0.1	0.1	0.5	0.3	0.3	-0.7	0.1	0.7	1.0	1.0	1.0	341.57
0.1	0.1	0.1	0.5	0.3	0.3	-0.7	0.3	0.1	1.0	1.0	1.0	207.50
0.1	0.1	0.1	0.5	0.3	0.3	-0.7	0.3	0.3	1.0	1.0	1.0	294.07
0.1	0.1	0.1	0.5	0.3	0.3	-0.7	0.5	0.1	1.0	1.0	1.0	407.37
0.1	0.1	0.1	0.5	0.3	0.3	-0.5	0.3	0.5	1.0	1.0	1.0	230.08
0.1	0.1	0.1	0.5	0.3	0.3	-0.5	0.3	0.7	1.0	1.0	1.0	341.57
0.1	0.1	0.1	0.5	0.3	0.3	-0.5	0.5	0.1	1.0	1.0	1.0	207.50
0.1	0.1	0.1	0.5	0.3	0.3	-0.5	0.5	0.3	1.0	1.0	1.0	294.07
0.1	0.1	0.1	0.5	0.3	0.3	-0.5	0.7	0.1	1.0	1.0	1.0	407.37
0.1	0.1	0.1	0.5	0.3	0.3	-0.3	0.5	0.5	1.0	1.0	1.0	230.08
0.1	0.1	0.1	0.5	0.3	0.3	-0.3	0.5	0.7	1.0	1.0	1.0	341.57
0.1	0.1	0.1	0.5	0.3	0.3	-0.3	0.7	0.1	1.0	1.0	1.0	207.50
0.1	0.1	0.1	0.5	0.3	0.3	-0.3	0.7	0.3	1.0	1.0	1.0	294.07
0.1	0.1	0.1	0.5	0.3	0.3	-0.1	0.7	0.5	1.0	1.0	1.0	230.08
0.1	0.1	0.1	0.5	0.3	0.3	-0.1	0.7	0.7	1.0	1.0	1.0	341.57
0.1	0.1	0.1	0.5	0.3	0.5	-0.9	0.1	0.7	1.0	1.0	1.0	358.56
0.1	0.1	0.1	0.5	0.3	0.5	-0.9	0.3	0.3	1.0	1.0	1.0	279.55
0.1	0.1	0.1	0.5	0.3	0.5	-0.9	0.5	0.1	1.0	1.0	1.0	499.80
0.1	0.1	0.1	0.5	0.3	0.5	-0.7	0.3	0.5	1.0	1.0	1.0	279.55
0.1	0.1	0.1	0.5	0.3	0.5	-0.7	0.5	0.1	1.0	1.0	1.0	229.08
0.1	0.1	0.1	0.5	0.3	0.5	-0.7	0.5	0.3	1.0	1.0	1.0	499.80
0.1	0.1	0.1	0.5	0.3	0.5	-0.5	0.3	0.7	1.0	1.0	1.0	279.55
0.1	0.1	0.1	0.5	0.3	0.5	-0.5	0.5	0.3	1.0	1.0	1.0	229.08
0.1	0.1	0.1	0.5	0.3	0.5	-0.5	0.5	0.5	1.0	1.0	1.0	499.80
0.1	0.1	0.1	0.5	0.3	0.5	-0.5	0.7	0.1	1.0	1.0	1.0	358.56
0.1	0.1	0.1	0.5	0.3	0.5	-0.3	0.5	0.5	1.0	1.0	1.0	229.08
0.1	0.1	0.1	0.5	0.3	0.5	-0.3	0.5	0.7	1.0	1.0	1.0	499.80
0.1	0.1	0.1	0.5	0.3	0.5	-0.3	0.7	0.3	1.0	1.0	1.0	358.56
0.1	0.1	0.1	0.5	0.3	0.5	-0.1	0.5	0.7	1.0	1.0	1.0	229.08

Appendix

Table 3.1 (continued): Relative efficiency of the proposed ratio estimator with respect to the BBB model for different choice of parameters with $P = P_1 = P_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	Cx_1	Cx_2	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	θ	θ_1	θ_2	RE
0.1	0.1	0.1	0.5	0.3	0.5	-0.1	0.7	0.5	1.0	1.0	1.0	358.56
0.1	0.1	0.1	0.5	0.3	0.5	0.1	0.7	0.7	1.0	1.0	1.0	358.56
0.1	0.1	0.1	0.5	0.5	0.1	-0.7	0.1	0.5	1.0	1.0	1.0	229.58
0.1	0.1	0.1	0.5	0.5	0.1	-0.7	0.1	0.7	1.0	1.0	1.0	280.29
0.1	0.1	0.1	0.5	0.5	0.1	-0.7	0.3	0.3	1.0	1.0	1.0	229.58
0.1	0.1	0.1	0.5	0.5	0.1	-0.7	0.3	0.5	1.0	1.0	1.0	280.29
0.1	0.1	0.1	0.5	0.5	0.1	-0.7	0.3	0.7	1.0	1.0	1.0	359.78
0.1	0.1	0.1	0.5	0.5	0.1	-0.7	0.5	0.1	1.0	1.0	1.0	229.58
0.1	0.1	0.1	0.5	0.5	0.1	-0.7	0.5	0.3	1.0	1.0	1.0	280.29
0.1	0.1	0.1	0.5	0.5	0.1	-0.7	0.5	0.5	1.0	1.0	1.0	359.78
0.1	0.1	0.1	0.5	0.5	0.1	-0.7	0.7	0.1	1.0	1.0	1.0	280.29
0.1	0.1	0.1	0.5	0.5	0.1	-0.7	0.7	0.3	1.0	1.0	1.0	359.78
0.1	0.1	0.1	0.5	0.5	0.3	-0.9	0.1	0.1	1.0	1.0	1.0	279.55
0.1	0.1	0.1	0.5	0.5	0.3	-0.7	0.1	0.5	1.0	1.0	1.0	358.56
0.1	0.1	0.1	0.5	0.5	0.3	-0.7	0.3	0.3	1.0	1.0	1.0	358.56
0.1	0.1	0.1	0.5	0.5	0.3	-0.7	0.5	0.1	1.0	1.0	1.0	358.56
0.1	0.1	0.1	0.5	0.5	0.3	-0.5	0.1	0.7	1.0	1.0	1.0	229.08
0.1	0.1	0.1	0.5	0.5	0.3	-0.5	0.3	0.5	1.0	1.0	1.0	229.08
0.1	0.1	0.1	0.5	0.5	0.3	-0.5	0.3	0.7	1.0	1.0	1.0	499.80
0.1	0.1	0.1	0.5	0.5	0.3	-0.5	0.5	0.3	1.0	1.0	1.0	229.08
0.1	0.1	0.1	0.5	0.5	0.3	-0.5	0.5	0.5	1.0	1.0	1.0	499.80
0.1	0.1	0.1	0.5	0.5	0.3	-0.5	0.7	0.1	1.0	1.0	1.0	229.08
0.1	0.1	0.1	0.5	0.5	0.3	-0.5	0.7	0.3	1.0	1.0	1.0	499.80
0.1	0.1	0.1	0.5	0.5	0.3	-0.3	0.5	0.7	1.0	1.0	1.0	279.55
0.1	0.1	0.1	0.5	0.5	0.3	-0.3	0.7	0.5	1.0	1.0	1.0	279.55
0.1	0.1	0.1	0.5	0.5	0.5	-0.9	0.1	0.3	1.0	1.0	1.0	228.09
0.1	0.1	0.1	0.5	0.5	0.5	-0.9	0.3	0.1	1.0	1.0	1.0	228.09
0.1	0.1	0.1	0.5	0.5	0.5	-0.7	0.1	0.5	1.0	1.0	1.0	228.09
0.1	0.1	0.1	0.5	0.5	0.5	-0.7	0.3	0.3	1.0	1.0	1.0	228.09
0.1	0.1	0.1	0.5	0.5	0.5	-0.7	0.5	0.1	1.0	1.0	1.0	228.09
0.1	0.1	0.1	0.5	0.5	0.5	-0.5	0.1	0.7	1.0	1.0	1.0	228.09
0.1	0.1	0.1	0.5	0.5	0.5	-0.5	0.3	0.5	1.0	1.0	1.0	228.09
0.1	0.1	0.1	0.5	0.5	0.5	-0.5	0.5	0.3	1.0	1.0	1.0	228.09
0.1	0.1	0.1	0.5	0.5	0.5	-0.5	0.7	0.1	1.0	1.0	1.0	228.09
0.1	0.1	0.1	0.5	0.5	0.5	-0.3	0.3	0.7	1.0	1.0	1.0	228.09
0.1	0.1	0.1	0.5	0.5	0.5	-0.3	0.5	0.5	1.0	1.0	1.0	228.09
0.1	0.1	0.1	0.5	0.5	0.5	-0.3	0.7	0.3	1.0	1.0	1.0	228.09
0.1	0.1	0.1	0.5	0.5	0.5	-0.1	0.5	0.7	1.0	1.0	1.0	228.09
0.1	0.1	0.1	0.5	0.5	0.5	-0.1	0.7	0.5	1.0	1.0	1.0	228.09
0.1	0.1	0.1	0.5	0.5	0.5	0.1	0.7	0.7	1.0	1.0	1.0	228.09

Appendix

Table 3.1 (continued): Relative efficiency of the proposed ratio estimator with respect to the BBB model for different choice of parameters with $P = P_1 = P_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	Cx_1	Cx_2	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	θ	θ_1	θ_2	RE
0.3	0.3	0.3	0.3	0.1	0.3	-0.9	0.7	0.7	1.0	1.0	1.0	220.64
0.3	0.3	0.3	0.3	0.3	0.1	-0.9	0.3	0.7	1.0	1.0	1.0	220.64
0.3	0.3	0.3	0.3	0.3	0.1	-0.9	0.5	0.5	1.0	1.0	1.0	220.64
0.3	0.3	0.3	0.3	0.3	0.1	-0.9	0.5	0.7	1.0	1.0	1.0	283.48
0.3	0.3	0.3	0.3	0.3	0.1	-0.9	0.7	0.3	1.0	1.0	1.0	220.64
0.3	0.3	0.3	0.3	0.3	0.1	-0.9	0.7	0.5	1.0	1.0	1.0	283.48
0.3	0.3	0.3	0.3	0.3	0.1	-0.9	0.7	0.7	1.0	1.0	1.0	396.38
0.3	0.3	0.3	0.3	0.3	0.3	-0.9	0.1	0.7	1.0	1.0	1.0	228.40
0.3	0.3	0.3	0.3	0.3	0.3	-0.9	0.3	0.5	1.0	1.0	1.0	228.40
0.3	0.3	0.3	0.3	0.3	0.3	-0.9	0.5	0.3	1.0	1.0	1.0	228.40
0.3	0.3	0.3	0.3	0.3	0.3	-0.9	0.7	0.1	1.0	1.0	1.0	228.40
0.3	0.3	0.3	0.3	0.3	0.3	-0.7	0.3	0.7	1.0	1.0	1.0	228.40
0.3	0.3	0.3	0.3	0.3	0.3	-0.7	0.5	0.5	1.0	1.0	1.0	228.40
0.3	0.3	0.3	0.3	0.3	0.3	-0.7	0.7	0.3	1.0	1.0	1.0	228.40
0.3	0.3	0.3	0.3	0.3	0.3	-0.5	0.5	0.7	1.0	1.0	1.0	228.40
0.3	0.3	0.3	0.3	0.3	0.3	-0.5	0.7	0.5	1.0	1.0	1.0	228.40
0.3	0.3	0.3	0.3	0.3	0.3	-0.3	0.7	0.7	1.0	1.0	1.0	228.40
0.3	0.3	0.3	0.3	0.3	0.5	-0.7	0.5	0.7	1.0	1.0	1.0	326.22
0.3	0.3	0.3	0.3	0.3	0.5	-0.7	0.7	0.5	1.0	1.0	1.0	326.22
0.3	0.3	0.3	0.3	0.3	0.5	-0.3	0.7	0.7	1.0	1.0	1.0	245.69
0.3	0.3	0.3	0.3	0.5	0.1	-0.9	0.7	0.7	1.0	1.0	1.0	204.33
0.3	0.3	0.3	0.3	0.5	0.3	-0.9	0.1	0.7	1.0	1.0	1.0	485.29
0.3	0.3	0.3	0.3	0.5	0.3	-0.9	0.3	0.5	1.0	1.0	1.0	245.69
0.3	0.3	0.3	0.3	0.5	0.3	-0.9	0.7	0.3	1.0	1.0	1.0	326.22
0.3	0.3	0.3	0.3	0.5	0.3	-0.7	0.3	0.7	1.0	1.0	1.0	245.69
0.3	0.3	0.3	0.3	0.5	0.3	-0.7	0.7	0.5	1.0	1.0	1.0	326.22
0.3	0.3	0.3	0.3	0.5	0.3	-0.5	0.7	0.7	1.0	1.0	1.0	326.22
0.3	0.3	0.3	0.3	0.5	0.5	-0.9	0.1	0.7	1.0	1.0	1.0	323.40
0.3	0.3	0.3	0.3	0.5	0.5	-0.9	0.7	0.3	1.0	1.0	1.0	209.78
0.3	0.3	0.3	0.3	0.5	0.5	-0.7	0.3	0.7	1.0	1.0	1.0	323.40
0.3	0.3	0.3	0.3	0.5	0.5	-0.5	0.5	0.7	1.0	1.0	1.0	323.40
0.3	0.3	0.3	0.3	0.5	0.5	-0.3	0.7	0.7	1.0	1.0	1.0	323.40
0.3	0.3	0.3	0.5	0.1	0.3	-0.9	0.7	0.1	1.0	1.0	1.0	211.04
0.3	0.3	0.3	0.5	0.1	0.3	-0.9	0.7	0.3	1.0	1.0	1.0	231.73
0.3	0.3	0.3	0.5	0.1	0.3	-0.9	0.7	0.5	1.0	1.0	1.0	256.90
0.3	0.3	0.3	0.5	0.1	0.3	-0.9	0.7	0.7	1.0	1.0	1.0	288.22
0.3	0.3	0.3	0.5	0.1	0.3	-0.7	0.7	0.5	1.0	1.0	1.0	217.52
0.3	0.3	0.3	0.5	0.1	0.3	-0.7	0.7	0.7	1.0	1.0	1.0	239.55
0.3	0.3	0.3	0.5	0.1	0.3	-0.5	0.7	0.7	1.0	1.0	1.0	204.95
0.3	0.3	0.3	0.5	0.1	0.5	-0.9	0.7	0.3	1.0	1.0	1.0	210.54
0.3	0.3	0.3	0.5	0.1	0.5	-0.9	0.7	0.5	1.0	1.0	1.0	247.23

Appendix

Table 3.1 (continued): Relative efficiency of the proposed ratio estimator with respect to the BBB model for different choice of parameters with $P = P_1 = P_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	Cx_1	Cx_2	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	θ	θ_1	θ_2	RE
0.3	0.3	0.3	0.5	0.1	0.5	-0.9	0.7	0.7	1.0	1.0	1.0	299.41
0.3	0.3	0.3	0.5	0.1	0.5	-0.7	0.7	0.5	1.0	1.0	1.0	210.54
0.3	0.3	0.3	0.5	0.1	0.5	-0.7	0.7	0.7	1.0	1.0	1.0	247.23
0.3	0.3	0.3	0.5	0.1	0.5	-0.5	0.7	0.7	1.0	1.0	1.0	210.54
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.1	0.5	1.0	1.0	1.0	217.52
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.1	0.7	1.0	1.0	1.0	239.55
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.3	0.1	1.0	1.0	1.0	211.04
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.3	0.3	1.0	1.0	1.0	231.73
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.3	0.5	1.0	1.0	1.0	256.90
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.3	0.7	1.0	1.0	1.0	288.22
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.5	0.1	1.0	1.0	1.0	247.92
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.5	0.3	1.0	1.0	1.0	276.96
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.5	0.5	1.0	1.0	1.0	313.71
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.5	0.7	1.0	1.0	1.0	361.70
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.7	0.1	1.0	1.0	1.0	300.42
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.7	0.3	1.0	1.0	1.0	344.15
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.7	0.5	1.0	1.0	1.0	402.77
0.3	0.3	0.3	0.5	0.3	0.1	-0.9	0.7	0.7	1.0	1.0	1.0	485.46
0.3	0.3	0.3	0.5	0.3	0.1	-0.7	0.5	0.7	1.0	1.0	1.0	204.95
0.3	0.3	0.3	0.5	0.3	0.1	-0.7	0.7	0.5	1.0	1.0	1.0	217.52
0.3	0.3	0.3	0.5	0.3	0.1	-0.7	0.7	0.7	1.0	1.0	1.0	239.55
0.3	0.3	0.3	0.5	0.3	0.3	-0.9	0.1	0.5	1.0	1.0	1.0	213.97
0.3	0.3	0.3	0.5	0.3	0.3	-0.9	0.1	0.7	1.0	1.0	1.0	293.71
0.3	0.3	0.3	0.5	0.3	0.3	-0.9	0.3	0.3	1.0	1.0	1.0	261.26
0.3	0.3	0.3	0.5	0.3	0.3	-0.9	0.3	0.5	1.0	1.0	1.0	390.79
0.3	0.3	0.3	0.5	0.3	0.3	-0.9	0.5	0.1	1.0	1.0	1.0	335.36
0.3	0.3	0.3	0.5	0.3	0.3	-0.7	0.3	0.5	1.0	1.0	1.0	213.97
0.3	0.3	0.3	0.5	0.3	0.3	-0.7	0.3	0.7	1.0	1.0	1.0	293.71
0.3	0.3	0.3	0.5	0.3	0.3	-0.7	0.5	0.3	1.0	1.0	1.0	261.26
0.3	0.3	0.3	0.5	0.3	0.3	-0.7	0.5	0.5	1.0	1.0	1.0	390.79
0.3	0.3	0.3	0.5	0.3	0.3	-0.7	0.7	0.1	1.0	1.0	1.0	335.36
0.3	0.3	0.3	0.5	0.3	0.3	-0.5	0.5	0.5	1.0	1.0	1.0	213.97
0.3	0.3	0.3	0.5	0.3	0.3	-0.5	0.5	0.7	1.0	1.0	1.0	293.71
0.3	0.3	0.3	0.5	0.3	0.3	-0.5	0.7	0.3	1.0	1.0	1.0	261.26
0.3	0.3	0.3	0.5	0.3	0.3	-0.5	0.7	0.5	1.0	1.0	1.0	390.79
0.3	0.3	0.3	0.5	0.3	0.3	-0.3	0.7	0.5	1.0	1.0	1.0	213.97
0.3	0.3	0.3	0.5	0.3	0.3	-0.3	0.7	0.7	1.0	1.0	1.0	293.71
0.3	0.3	0.3	0.5	0.3	0.5	-0.9	0.3	0.5	1.0	1.0	1.0	242.67
0.3	0.3	0.3	0.5	0.3	0.5	-0.9	0.3	0.7	1.0	1.0	1.0	498.42
0.3	0.3	0.3	0.5	0.3	0.5	-0.9	0.5	0.1	1.0	1.0	1.0	207.22
0.3	0.3	0.3	0.5	0.3	0.5	-0.9	0.5	0.3	1.0	1.0	1.0	368.84

Appendix

Table 3.1 (continued): Relative efficiency of the proposed ratio estimator with respect to the BBB model for different choice of parameters with $P = P_1 = P_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	Cx_1	Cx_2	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	θ	θ_1	θ_2	RE
0.3	0.3	0.3	0.5	0.3	0.5	-0.7	0.3	0.7	1.0	1.0	1.0	242.67
0.3	0.3	0.3	0.5	0.3	0.5	-0.7	0.5	0.3	1.0	1.0	1.0	207.22
0.3	0.3	0.3	0.5	0.3	0.5	-0.7	0.5	0.5	1.0	1.0	1.0	368.84
0.3	0.3	0.3	0.5	0.3	0.5	-0.7	0.7	0.1	1.0	1.0	1.0	292.74
0.3	0.3	0.3	0.5	0.3	0.5	-0.5	0.5	0.5	1.0	1.0	1.0	207.22
0.3	0.3	0.3	0.5	0.3	0.5	-0.5	0.5	0.7	1.0	1.0	1.0	368.84
0.3	0.3	0.3	0.5	0.3	0.5	-0.5	0.7	0.3	1.0	1.0	1.0	292.74
0.3	0.3	0.3	0.5	0.3	0.5	-0.3	0.5	0.7	1.0	1.0	1.0	207.22
0.3	0.3	0.3	0.5	0.3	0.5	-0.3	0.7	0.5	1.0	1.0	1.0	292.74
0.3	0.3	0.3	0.5	0.3	0.5	-0.1	0.7	0.7	1.0	1.0	1.0	292.74
0.3	0.3	0.3	0.5	0.5	0.1	-0.9	0.1	0.1	1.0	1.0	1.0	210.54
0.3	0.3	0.3	0.5	0.5	0.1	-0.9	0.1	0.3	1.0	1.0	1.0	247.23
0.3	0.3	0.3	0.5	0.5	0.1	-0.9	0.1	0.5	1.0	1.0	1.0	299.41
0.3	0.3	0.3	0.5	0.5	0.1	-0.9	0.1	0.7	1.0	1.0	1.0	379.50
0.3	0.3	0.3	0.5	0.5	0.1	-0.9	0.3	0.1	1.0	1.0	1.0	247.23
0.3	0.3	0.3	0.5	0.5	0.1	-0.9	0.3	0.3	1.0	1.0	1.0	299.41
0.3	0.3	0.3	0.5	0.5	0.1	-0.9	0.3	0.5	1.0	1.0	1.0	379.50
0.3	0.3	0.3	0.5	0.5	0.1	-0.9	0.5	0.1	1.0	1.0	1.0	299.41
0.3	0.3	0.3	0.5	0.5	0.1	-0.9	0.5	0.3	1.0	1.0	1.0	379.50
0.3	0.3	0.3	0.5	0.5	0.1	-0.9	0.7	0.1	1.0	1.0	1.0	379.50
0.3	0.3	0.3	0.5	0.5	0.1	-0.7	0.5	0.7	1.0	1.0	1.0	210.54
0.3	0.3	0.3	0.5	0.5	0.1	-0.7	0.7	0.5	1.0	1.0	1.0	210.54
0.3	0.3	0.3	0.5	0.5	0.1	-0.7	0.7	0.7	1.0	1.0	1.0	247.23
0.3	0.3	0.3	0.5	0.5	0.3	-0.9	0.1	0.3	1.0	1.0	1.0	242.67
0.3	0.3	0.3	0.5	0.5	0.3	-0.9	0.1	0.5	1.0	1.0	1.0	498.42
0.3	0.3	0.3	0.5	0.5	0.3	-0.9	0.3	0.1	1.0	1.0	1.0	242.67
0.3	0.3	0.3	0.5	0.5	0.3	-0.9	0.3	0.3	1.0	1.0	1.0	498.42
0.3	0.3	0.3	0.5	0.5	0.3	-0.9	0.5	0.1	1.0	1.0	1.0	498.42
0.3	0.3	0.3	0.5	0.5	0.3	-0.7	0.1	0.7	1.0	1.0	1.0	292.74
0.3	0.3	0.3	0.5	0.5	0.3	-0.7	0.3	0.5	1.0	1.0	1.0	292.74
0.3	0.3	0.3	0.5	0.5	0.3	-0.7	0.5	0.3	1.0	1.0	1.0	292.74
0.3	0.3	0.3	0.5	0.5	0.3	-0.7	0.7	0.1	1.0	1.0	1.0	292.74
0.3	0.3	0.3	0.5	0.5	0.3	-0.5	0.3	0.7	1.0	1.0	1.0	207.22
0.3	0.3	0.3	0.5	0.5	0.3	-0.5	0.5	0.5	1.0	1.0	1.0	207.22
0.3	0.3	0.3	0.5	0.5	0.3	-0.5	0.5	0.7	1.0	1.0	1.0	368.84
0.3	0.3	0.3	0.5	0.5	0.3	-0.5	0.7	0.3	1.0	1.0	1.0	207.22
0.3	0.3	0.3	0.5	0.5	0.3	-0.5	0.7	0.5	1.0	1.0	1.0	368.84
0.3	0.3	0.3	0.5	0.5	0.3	-0.3	0.7	0.7	1.0	1.0	1.0	242.67
0.3	0.3	0.3	0.5	0.5	0.5	-0.9	0.1	0.5	1.0	1.0	1.0	280.25
0.3	0.3	0.3	0.5	0.5	0.5	-0.9	0.3	0.3	1.0	1.0	1.0	280.25
0.3	0.3	0.3	0.5	0.5	0.5	-0.9	0.5	0.1	1.0	1.0	1.0	280.25

Appendix

Table 3.1 (continued): Relative efficiency of the proposed ratio estimator with respect to the BBB model for different choice of parameters with $P = P_1 = P_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	Cx_1	Cx_2	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	θ	θ_1	θ_2	RE
0.3	0.3	0.3	0.5	0.5	0.5	-0.7	0.1	0.7	1.0	1.0	1.0	280.25
0.3	0.3	0.3	0.5	0.5	0.5	-0.7	0.3	0.5	1.0	1.0	1.0	280.25
0.3	0.3	0.3	0.5	0.5	0.5	-0.7	0.5	0.3	1.0	1.0	1.0	280.25
0.3	0.3	0.3	0.5	0.5	0.5	-0.7	0.7	0.1	1.0	1.0	1.0	280.25
0.3	0.3	0.3	0.5	0.5	0.5	-0.5	0.3	0.7	1.0	1.0	1.0	280.25
0.3	0.3	0.3	0.5	0.5	0.5	-0.5	0.5	0.5	1.0	1.0	1.0	280.25
0.3	0.3	0.3	0.5	0.5	0.5	-0.5	0.7	0.3	1.0	1.0	1.0	280.25
0.3	0.3	0.3	0.5	0.5	0.5	-0.3	0.5	0.7	1.0	1.0	1.0	280.25
0.3	0.3	0.3	0.5	0.5	0.5	-0.3	0.7	0.5	1.0	1.0	1.0	280.25
0.3	0.3	0.3	0.5	0.5	0.5	-0.1	0.7	0.7	1.0	1.0	1.0	280.25
0.5	0.5	0.5	0.3	0.5	0.3	-0.9	0.7	0.7	1.0	1.0	1.0	211.38
0.5	0.5	0.5	0.5	0.3	0.3	-0.9	0.5	0.7	1.0	1.0	1.0	243.36
0.5	0.5	0.5	0.5	0.3	0.3	-0.9	0.7	0.3	1.0	1.0	1.0	224.31
0.5	0.5	0.5	0.5	0.3	0.3	-0.9	0.7	0.5	1.0	1.0	1.0	293.18
0.5	0.5	0.5	0.5	0.3	0.3	-0.9	0.7	0.7	1.0	1.0	1.0	423.08
0.5	0.5	0.5	0.5	0.3	0.3	-0.7	0.7	0.7	1.0	1.0	1.0	243.36
0.5	0.5	0.5	0.5	0.3	0.5	-0.9	0.5	0.7	1.0	1.0	1.0	265.96
0.5	0.5	0.5	0.5	0.3	0.5	-0.9	0.7	0.3	1.0	1.0	1.0	230.32
0.5	0.5	0.5	0.5	0.3	0.5	-0.9	0.7	0.5	1.0	1.0	1.0	385.15
0.5	0.5	0.5	0.5	0.3	0.5	-0.7	0.7	0.5	1.0	1.0	1.0	230.32
0.5	0.5	0.5	0.5	0.3	0.5	-0.7	0.7	0.7	1.0	1.0	1.0	385.15
0.5	0.5	0.5	0.5	0.3	0.5	-0.5	0.7	0.7	1.0	1.0	1.0	230.32
0.5	0.5	0.5	0.5	0.5	0.3	-0.9	0.1	0.7	1.0	1.0	1.0	203.10
0.5	0.5	0.5	0.5	0.5	0.3	-0.9	0.3	0.5	1.0	1.0	1.0	203.10
0.5	0.5	0.5	0.5	0.5	0.3	-0.9	0.3	0.7	1.0	1.0	1.0	314.65
0.5	0.5	0.5	0.5	0.5	0.3	-0.9	0.5	0.3	1.0	1.0	1.0	203.10
0.5	0.5	0.5	0.5	0.5	0.3	-0.9	0.5	0.5	1.0	1.0	1.0	314.65
0.5	0.5	0.5	0.5	0.5	0.3	-0.9	0.7	0.1	1.0	1.0	1.0	203.10
0.5	0.5	0.5	0.5	0.5	0.3	-0.9	0.7	0.3	1.0	1.0	1.0	314.65
0.5	0.5	0.5	0.5	0.5	0.3	-0.7	0.5	0.7	1.0	1.0	1.0	230.32
0.5	0.5	0.5	0.5	0.5	0.3	-0.7	0.7	0.5	1.0	1.0	1.0	230.32
0.5	0.5	0.5	0.5	0.5	0.3	-0.7	0.7	0.7	1.0	1.0	1.0	385.15
0.5	0.5	0.5	0.5	0.5	0.5	-0.9	0.3	0.7	1.0	1.0	1.0	423.08
0.5	0.5	0.5	0.5	0.5	0.5	-0.9	0.5	0.5	1.0	1.0	1.0	423.08
0.5	0.5	0.5	0.5	0.5	0.5	-0.9	0.7	0.3	1.0	1.0	1.0	423.08
0.5	0.5	0.5	0.5	0.5	0.5	-0.7	0.5	0.7	1.0	1.0	1.0	423.08
0.5	0.5	0.5	0.5	0.5	0.5	-0.7	0.7	0.5	1.0	1.0	1.0	423.08
0.5	0.5	0.5	0.5	0.5	0.5	-0.5	0.7	0.7	1.0	1.0	1.0	423.08

Appendix

Table 4.1: Relative efficiency of the proposed power transformation estimator with respect to the BBB model for different choice of parameters with $P = P_1 = P_2 = 0.7$

C_γ	C_{γ_1}	C_{γ_2}	C_y	Cx_1	Cx_2	ρ_{yx_1}	ρ_{yx_2}	$\rho_{x_1x_2}$	θ	θ_1	θ_2	α_{opt}	RE
0.1	0.1	0.1	0.5	0.3	0.5	0.1	0.7	0.7	0.5	0.5	0.5	1.00	221.88
0.1	0.1	0.1	0.5	0.5	0.3	0.7	0.1	0.7	0.5	0.5	0.5	-1.00	221.88
0.1	0.1	0.1	0.5	0.5	0.5	0.1	0.7	0.7	0.5	0.5	0.5	1.04	215.47
0.1	0.1	0.1	0.5	0.5	0.5	0.7	0.1	0.7	0.5	0.5	0.5	-1.04	215.47
0.1	0.1	0.1	0.1	0.3	0.5	0.1	0.7	0.7	1.0	1.0	1.0	0.23	234.49
0.1	0.1	0.1	0.1	0.5	0.3	0.7	0.1	0.7	1.0	1.0	1.0	-0.23	234.49
0.1	0.1	0.1	0.3	0.1	0.3	0.1	0.7	0.5	1.0	1.0	1.0	0.79	202.37
0.1	0.1	0.1	0.3	0.1	0.3	0.1	0.7	0.7	1.0	1.0	1.0	0.93	250.17
0.1	0.1	0.1	0.3	0.1	0.5	0.1	0.7	0.5	1.0	1.0	1.0	0.47	206.00
0.1	0.1	0.1	0.3	0.1	0.5	0.1	0.7	0.7	1.0	1.0	1.0	0.52	230.87
0.1	0.1	0.1	0.3	0.1	0.5	0.3	0.7	0.7	1.0	1.0	1.0	0.49	200.86
0.1	0.1	0.1	0.3	0.3	0.1	0.7	0.1	0.5	1.0	1.0	1.0	-0.79	202.37
0.1	0.1	0.1	0.3	0.3	0.1	0.7	0.1	0.7	1.0	1.0	1.0	-0.93	250.17
0.1	0.1	0.1	0.3	0.3	0.3	0.1	0.7	0.7	1.0	1.0	1.0	0.89	206.79
0.1	0.1	0.1	0.3	0.3	0.3	0.7	0.1	0.7	1.0	1.0	1.0	-0.89	206.79
0.1	0.1	0.1	0.3	0.3	0.5	0.1	0.7	0.5	1.0	1.0	1.0	0.49	200.61
0.1	0.1	0.1	0.3	0.3	0.5	0.1	0.7	0.7	1.0	1.0	1.0	0.70	358.60
0.1	0.1	0.1	0.3	0.5	0.1	0.7	0.1	0.5	1.0	1.0	1.0	-0.47	206.00
0.1	0.1	0.1	0.3	0.5	0.1	0.7	0.1	0.7	1.0	1.0	1.0	-0.52	230.87
0.1	0.1	0.1	0.3	0.5	0.1	0.7	0.3	0.7	1.0	1.0	1.0	-0.49	200.86
0.1	0.1	0.1	0.3	0.5	0.3	0.7	0.1	0.5	1.0	1.0	1.0	-0.49	200.61
0.1	0.1	0.1	0.3	0.5	0.3	0.7	0.1	0.7	1.0	1.0	1.0	-0.70	358.60
0.1	0.1	0.1	0.3	0.5	0.5	0.1	0.7	0.7	1.0	1.0	1.0	0.57	222.91
0.1	0.1	0.1	0.3	0.5	0.5	0.7	0.1	0.7	1.0	1.0	1.0	-0.57	222.91
0.1	0.1	0.1	0.5	0.1	0.3	0.1	0.7	0.5	1.0	1.0	1.0	1.31	206.82
0.1	0.1	0.1	0.5	0.1	0.3	0.1	0.7	0.7	1.0	1.0	1.0	1.56	258.32
0.1	0.1	0.1	0.5	0.1	0.5	0.1	0.7	0.5	1.0	1.0	1.0	0.78	210.69
0.1	0.1	0.1	0.5	0.1	0.5	0.1	0.7	0.7	1.0	1.0	1.0	0.86	237.40
0.1	0.1	0.1	0.5	0.1	0.5	0.3	0.7	0.7	1.0	1.0	1.0	0.81	205.21
0.1	0.1	0.1	0.5	0.3	0.1	0.7	0.1	0.5	1.0	1.0	1.0	-1.31	206.82
0.1	0.1	0.1	0.5	0.3	0.1	0.7	0.1	0.7	1.0	1.0	1.0	-1.56	258.32
0.1	0.1	0.1	0.5	0.3	0.3	0.1	0.7	0.7	1.0	1.0	1.0	1.49	211.54
0.1	0.1	0.1	0.5	0.3	0.3	0.7	0.1	0.7	1.0	1.0	1.0	-1.49	211.54
0.1	0.1	0.1	0.5	0.3	0.5	0.1	0.7	0.5	1.0	1.0	1.0	0.81	204.94
0.1	0.1	0.1	0.5	0.3	0.5	0.1	0.7	0.7	1.0	1.0	1.0	1.17	379.21
0.1	0.1	0.1	0.5	0.5	0.1	0.7	0.1	0.5	1.0	1.0	1.0	-0.78	210.69
0.1	0.1	0.1	0.5	0.5	0.1	0.7	0.1	0.7	1.0	1.0	1.0	-0.86	237.40
0.1	0.1	0.1	0.5	0.5	0.1	0.7	0.3	0.7	1.0	1.0	1.0	-0.81	205.21
0.1	0.1	0.1	0.5	0.5	0.3	0.7	0.1	0.5	1.0	1.0	1.0	-0.81	204.94
0.1	0.1	0.1	0.5	0.5	0.3	0.7	0.1	0.7	1.0	1.0	1.0	-1.17	379.21
0.1	0.1	0.1	0.5	0.5	0.5	0.1	0.7	0.7	1.0	1.0	1.0	0.95	228.82
0.1	0.1	0.1	0.5	0.5	0.5	0.7	0.1	0.7	1.0	1.0	1.0	-0.95	228.82

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 443-451

1538 – 9472/10/$95.00

A Flexible Method for Testing Independence in Two-Way Contingency Tables

Peyman Jafari
Shiraz University of Medical Sciences, Shiraz, Iran

Noori Akhtar-Danesh
McMaster University, Hamilton, Ontario Canada

Zahra Bagheri
Shiraz University of Medical Sciences, Shiraz, Iran

A flexible approach for testing association in two-way contingency tables is presented. It is simple, does not assume a specific form for the association and is applicable to tables with nominal-by-nominal, nominal-by-ordinal, and ordinal-by-ordinal classifications.

Key words: Monte-Carlo simulation, log-linear models, row-effect models.

Introduction

In many social and medical studies a crucial question is whether the categorical variables forming a contingency table are independent. Suppose that a sample of N observations is classified with respect to two categorical variables, one with r levels and the other with c levels. Using the notation in Table 1 for this two-dimensional table, n_{ij} denotes the observed frequency for cell (i, j), and $n_{i.}$ and $n_{.j}$ denote the row and column totals, respectively. Also, $\mathrm{P_{ij}}$ is estimated by $\hat{P}_{ij} = \frac{n_{ij}}{N}$.

Table 1: Notation for a Two-Way Contingency Table

Row Variable	Column Variable					Total
	1		j		c	
1	n_{11} p_{11}	...	n_{1j} p_{1j}	...	n_{1c} p_{1c}	$n_{1.}$ $p_{1.}$
i	n_{i1} p_{i1}	...	n_{ij} p_{ij}	...	n_{ic} p_{ic}	$n_{i.}$ $p_{i.}$
r	n_{r1} p_{r1}	...	n_{rj} p_{rj}	...	n_{rc} p_{rc}	$n_{r.}$ $p_{r.}$
Total	$n_{.1}$ $p_{.1}$	...	$n_{.j}$ $p_{.j}$	...	$n_{.c}$ $p_{.c}$	N 1

Peyman Jafari is an Assistant Professor of Biostatistics in the Department of Biostatistics, Faculty of Medicine. His research interests include: sequential clinical trials, design and analysis of quality of life studies. Email: jafarip@sums.ac.ir. Noori Akhtar-Danesh, is an Associate Professor of Biostatistics in the Department of Epidemiology and Biostatistics at the School of Nursing. His research interests include: survival analysis (including analysis of recurrent data and competing risks), multilevel modeling and longitudinal data analysis, structural equation modeling, modeling risk factors of obesity and depression, and meta-analysis. Email: daneshn@mcmaster.ca. Zahra Bagheri is a Ph.D. student in the Department of Biostatistics. Her research interests include: Longitudinal studies and categorical mixed models. Email: zbagheri@sums.ac.ir.

Log-linear models are a general approach for the analysis of contingency tables. The major advantages of log-linear models are that they provide a systematic approach to the analysis of complex multidimensional tables and estimate the magnitude of effects of interest; consequently, they identify the relative importance of different effects (Agresti, 2002). Let m_{ij} denote the expected frequencies in a two-way contingency table with nominal row and column classifications. In addition, let x and y represent the row and column variables, respectively. In the standard system of hierarchical log-linear models, there are two possible models. The saturated model

$$\log(m_{ij}) = \lambda + \lambda_i^x + \lambda_j^y + \lambda_{ij}^{xy} \qquad (1)$$

has *rc* parameters and zero degree of freedom (d.f.). Hence, this model describes the data perfectly, however, it is not useful because it does not provide data reduction. The model only serves as a baseline for comparison with the independence model.

The independence model

$$\log(m_{ij}) = \lambda + \lambda_i^x + \lambda_j^y \qquad (2)$$

has $r + c - 1$ parameters and $(r-1)(c-1)$ d.f. for testing lack of fit. Thus, the hypothesis of independence can be tested by comparing the saturated and independence models. The deviation from independence can be measured by the likelihood ratio statistic (LR)

$$D_I = 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \log\left(\frac{n_{ij}}{\hat{m}_{ij}}\right)$$

where $\hat{m}_{ij} = n_{i.}n_{.j} / N$ is the estimation of the expected frequency in the i^{th} category of the row and the j^{th} category of the column variable under the hypothesis of independence (H_0). If H_0 is true, D_I has an asymptotic Chi-square distribution with $(r-1)(c-1)$ degrees of freedom.

The log-linear method presented has a number of limitations. First, it often has low power to detect departures from independence, especially when the dimension of the table increases (Davis, 1991). Second, it treats all classifications as nominal; therefore if the order of categories changes for a variable in any way, the fit remains the same (Agresti, 2002). Instead, if the row and column variables are both ordinal with known scores, the Linear-by-Linear association model can be used. On the other hand, when scoring is used only for one of the row or column variables, the row-effect or column-effect association model can be used (Agresti, 1984).

In practice it may not be possible to choose obvious scores for both the row and column categories. One alternative is Goodman's RC model, in which the row and column scores are treated as parameters to be estimated (Goodman, 1969). Although the RC model can be used if the two variables are nominal, which does not impose any restriction on the type of the variables, calculation of the conditional test of independence is complicated and the distribution of the test statistic is not Chi-square (Agresti, 2002). In all of these models the researcher needs to specify the functional form for the association and, if the association form is chosen incorrectly, then the power of the model will decrease.

It should be noted that, some methods used for testing interaction in two-way ANOVA can also be applied to two-way contingency tables for testing association (Alin & Kurt, 2006). For example, Davis (1991) tested association in two-way contingency tables based on Tukey's model (Tukey, 1949). Also Christensen (1990) tested interaction in log-linear and logit models for categorical data with the logit version of Mandel's models (Mandel, 1961). Milliken and Graybill (1970) established a two-stage fitting procedure using Tukey's model (Tukey, 1949). Recently, Kharati and Sadooghi (2007) have proposed a new method for testing interaction in two-way ANOVA.

In this study, the same method used by Kharati and Sadooghi (2007) will be applied for testing association in two-way contingency tables. It is a flexible approach for testing independence that does not assume a special form for the association model. The method was applied to detect association in tables with nominal-by-nominal and nominal-by-ordinal data.

Methodology

Row Effect Model

If either the row or the column variable (but not both of them) is ordinal, then a row-effect or column-effect model can be fitted (Agresti, 1984; Agresti, 2002). The row effects model has the form

$$\log(m_{ij}) = \lambda + \lambda_i^x + \lambda_j^y + \mu_i v_j. \qquad (4)$$

This model is appropriate for two-way tables with ordered columns, using scores $v_1 < v_2 < \ldots < v_c$. Because the rows are unordered, the model treats them as parameters and denotes them by μ_i. The μ_i's are called the row effects. This model has *r-1* more parameters

than the independence model, which is a special case where $\mu_1 = \mu_2 = ... = \mu_r$.

The LR test of independence requires maximum likelihood (ML) estimates $\hat{m}_{ij}$ of expected cell frequencies under model (4). Let D_R denote the LR goodness of fit statistic for model (4) and let D_I denote the classical test of independence given by (2). A (r-1) (c-2) degrees of freedom test of $H_0 : \mu_1 = \mu_2 = ... = \mu_r$ can then be based on the LR statistic $D_{I|R} = D_I - D_R$.

We used the same method proposed by Kharati and Sadooghi (2007) for testing association in two-way contingency tables. Assume a $r \times c$ contingency table and simultaneously $r \geq 4$ (so the method excludes only 2×2, 3×2 and 3×3 tables). Divide the table according to the rows, into two sub-tables. The sub-tables are two contingency tables with $r_1 \times c$ and $r_2 \times c$ dimensions in which $r_1 + r_2 = r$. In the absence of any association in each sub-table, then the independence model

$$\log m_{ij} = \lambda + \lambda_i^x + \lambda_j^y \qquad (5)$$

can fit both datasets well. Let D_{I1} and D_{I2} denote the deviances for the two sub-tables, respectively. In generalized linear models if the response variables are normally distributed then D has a Chi-square distribution exactly. However, for responses with a Poisson distribution, the sampling distribution of D may have an approximate Chi-square distribution (Dobson, 2002). Therefore, under the independence log-linear model, D_{I1} and D_{I2} are independent and have approximate Chi-square distributions with $df_1 = (r_1 - 1)(c - 1)$ and $df_2 = (r_2 - 1)(c - 1)$ degrees of freedom, respectively. A new statistic for testing independence in two-way contingency tables is now defined.

If $t_1 = \frac{D_{I1}}{df_1}$ and $t_2 = \frac{D_{I2}}{df_2}$, then the new variable $F^* = \frac{Max(t_1, t_2)}{Min(t_1, t_2)}$ has the F distribution with d.f. = (df_1, df_2) where $t_1 > t_2$ or d.f. = (df_2, df_1) where $t_2 > t_1$. In the presence of any association, the F^* statistic tends to be large, thus, the hypothesis of no association if $F^* > F_\alpha(df_1, df_2)$ is rejected when $t_1 > t_2$ or $F^* > F_\alpha(df_2, df_1)$ where $t_2 > t_1$.

However, in this approach the most important question is how a table can be split into two separate tables. In some cases, based on a priori information, there may be a natural division of the table. In the absence of a-priori information, drawing a profile plot is suggested. Based on such a profile plot those lines which are parallel or have the same pattern will be put in the same group and the remaining in the other group. Additional details are provided in the examples and readers are also referred to Kharati and Sadooghi (2007) for more information.

Simulation Study

The programming for the Monte Carlo simulation was written in SAS version 9.1. The RANTBL function was used for generating and simulating contingency tables in SAS (Fan, Felsovalyi, Sivo & Keenan, 2002). Contingency table data may result from one of several possible sampling models. The test of independence discussed in this study is based on sampling in which a single random sample of size $\underline{N}$ is classified with respect to two characteristics simultaneously (Dobson, 2002). In the resulting contingency table, both sets of marginal total frequencies are random variables. The empirical power of each test was determined by simulating contingency tables under the dependence structure, and computing the proportion of times the independence hypothesis was rejected at a given significance level α. Under the dependence structure, P_{ij} is estimated by $\hat{P}_{ij} = \frac{n_{ij}}{N}$ (Table 1).

For each studied situation, 5,000 contingency tables were generated in which cell frequencies were drawn under the dependent structure. The influence of the total sample size (*N*) on the statistical properties of all tests was also evaluated. The choice of the proper total sample size for simulation depends on dimensions of the table. The power of the $D_{I|R}$ and F statistics for testing independence in two-way contingency tables (nominal-by-ordinal) were investigated and compared. In order to find the maximum F in each simulated table, all combinations of rows and columns to classify each table into two subtables were considered. The power of the D_I and *F* statistics for testing independence in two-way contingency tables (nominal-by- nominal) were also computed and compared.

Example 1: The Location of Prehistoric Artifact

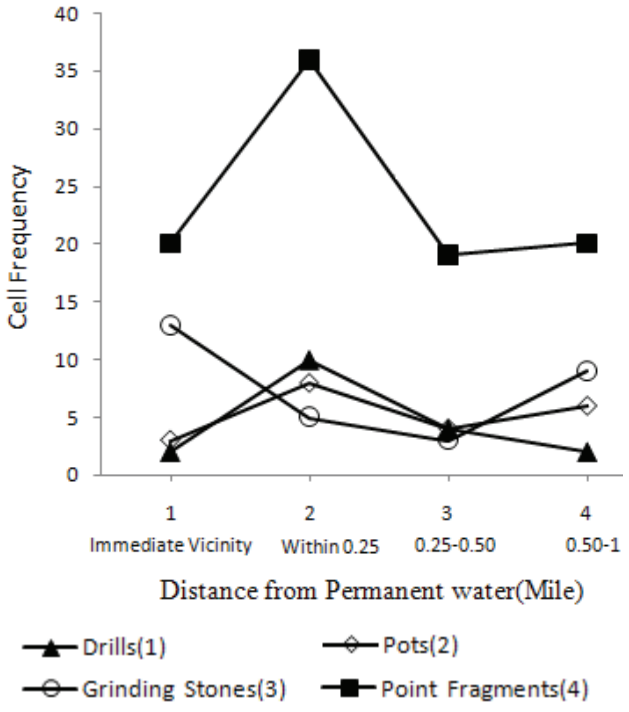
This example is based on the data provided in Simonoff (2003). As a result of archaeological excavations in Ruby Valley, Nevada, various prehistoric artifacts were discovered. Archaeologists were interested in the relationship between the type of artifacts found and the distance to permanent water, because the type of artifact discovered describes the type of site used by prehistoric hunters (Table 2). It was presumed that some tools were more difficult to move place to place and would thus be more likely to be discovered near permanent water. The following table is based on a subset of the artifacts discovered in Nevada (Simonoff, 2003).

In this example the row variable is nominal and the column variable is ordinal. Using the row-effect model (4), $D_R = 14.85$, $D_I = 16.26$ and $D_{I|R} = D_I - D_R = 1.40$. With respect to the asymptotic Chi-square distribution, $\chi^2_3 = 7.815$, there is no evidence of departure from independence. A similar result was obtained based on the F statistic. In the profile plot for these data (shown in Figure 1), the lines corresponding to rows 2, 4 are parallel. Thus, these rows were placed in the first sub-table and the remaining rows in the second sub-table. In this situation, F (3, 3) = 14.94 and P = .002 which is significant at the nominal level of 0.05. The result of our simulation showed that the F statistic is considerably more powerful than the row-effect model. The power of the F and $D_{I|R}$ are 0.43 and 0.15 respectively.

Table 2: Frequencies for Artifact Type and Distance from Permanent Water

Artifact Type	Distance from Permanent Water			
	Immediate Vicinity	Within 0.25 Miles	0.25-0.50 Miles	0.50-1 Miles
Drills	2	10	4	2
Pots	3	8	4	6
Grinding Stones	13	5	3	9
Point Fragments	20	36	19	20

Figure 1: Profile Plot of Data in Example 1



Example 2.1: Malignant Melanoma

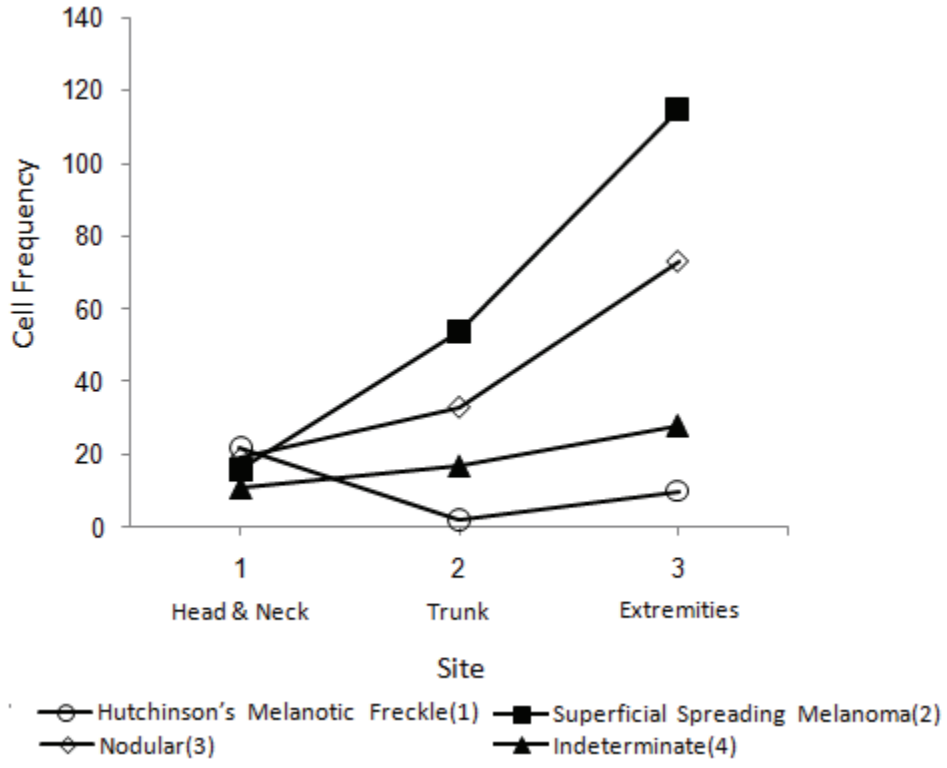
For the data in Table 3 the question of interest is whether there is any association between tumor type and site. These data are from a cross-sectional study of patients with a form of skin cancer called malignant melanoma (Dobson, 2002). For a sample of *N*=400 patients the site of the tumor and its histological type were recorded.

In testing the null hypothesis that tumor type and tumor site are independent, $D_I = 51.79$ and $P < .001$, which indicate that the association between type and site of tumor is highly significant. A similar result was obtained based on the proposed F statistic. In the profile plot for these data in Figure 2, the lines corresponding to rows 3 and 4 are nearly parallel which suggests that these rows can be placed in the one sub-table and the remaining rows in the other sub-table. The F statistic value for this division is statistically significant, $F(2, 2) = 43.41$, $p = .02$.

Table 3: Frequencies for Tumor Type and Site

Tumor Type	Site		
	Head and Neck	Trunk	Extremities
Hutchinson's Melanotic Freckle	22	2	10
Superficial Spreading Melanoma	16	54	115
Nodular	19	33	73
Indeterminate	11	17	28

Figure 2: Profile Plot of Data in Malignant Melanoma Example 2.1



Example 2.2: Malignant Melanoma

Next, substitute the frequencies 2, 16, 115, 73 and 28 in the cells (1, 2), (2, 1), (2, 3), (3, 3) and (4, 3) by 18, 45, 60, 38 and 20, respectively. In this situation, the null hypothesis is tested again. The new results, based on the likelihood ratio statistic, show that there is no significant association between tumor site and tumor type, $D_I = 11.80$, $P = .067$. However, a different result was obtained based on the F statistic at the α=0.05 level. In the profile plot for these data (Figure 3), the lines corresponding to rows 3, 4 are nearly parallel and close to each other. Therefore, these rows were placed in one table and the remaining rows in another table. The value of the F statistic for this division is highly significant, $F(2, 2) = 108.42$, $p < .01$.

Simulation Results

The results of the simulations showed that the power of the F and LR statistics in Malignant Melanoma Example 2.1 are 0.653 and 1, respectively, and in Malignant Melanoma Example 2.2 are 0.425 and 0.736, respectively.

This study also evaluated the influence of the total sample size (N) on the statistical properties for the above two examples. Table 4 shows the results of the estimation of power of the proposed F statistic and row-effect model ($D_{I|R}$) based on 5,000 simulated tables for the nominal-by-ordinal association model in Example 1. Table 5 shows these results for the proposed F statistic and the likelihood ratio statistic (D_I) based on 5,000 simulated tables for the nominal-by-nominal association model in Examples 2.1 and 2.2.

Table 4 shows that, for $N \leq 800$, especially when $N \leq 500$, the estimated power for the F statistic is considerably higher compared to the row-effect model ($D_{I|R}$). However for $N > 900$ the power of the row-effect model is dramatically higher than the F statistic.

Figure 3: Profile Plot of Data in Malignant Melanoma 2.2 Example

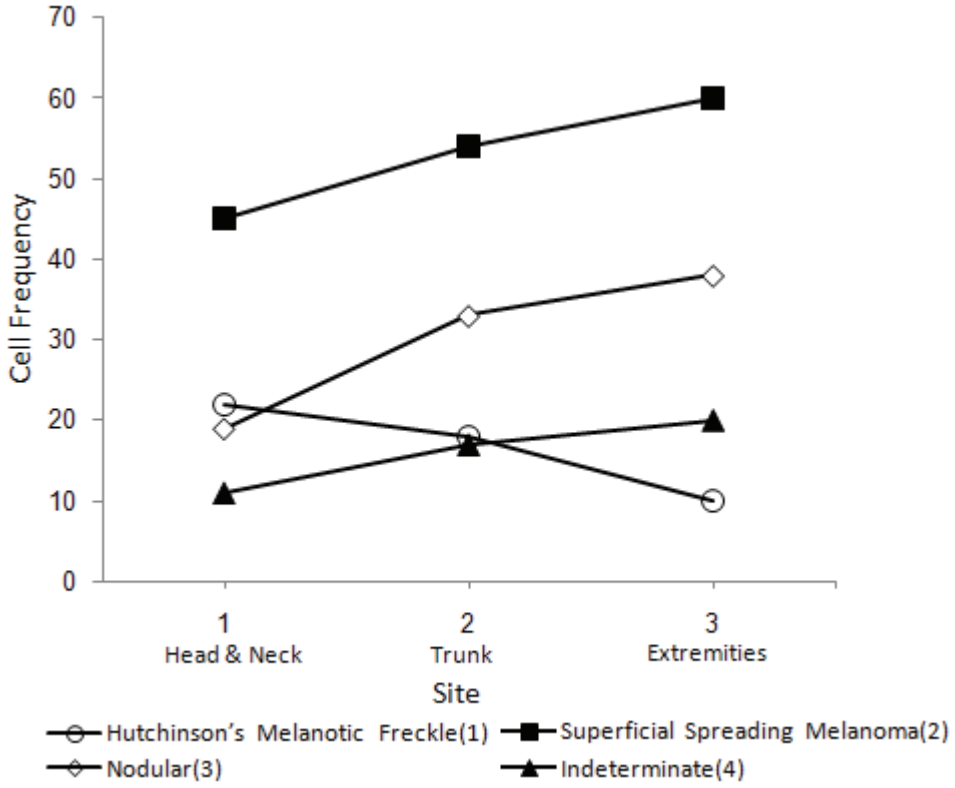


Table 4: Nominal-by-Ordinal Association: Estimation of Power for the F Statistic and the Row-Effect Model ($D_{I|R}$) Based on 5,000 Simulations in Example 1

N	F	$D_{I\|R}$
200	0.440	0.172
300	0.488	0.242
400	0.514	0.311
500	0.553	0.372
600	0.575	0.442
700	0.576	0.499
800	0.606	0.564
900	0.617	0.616
1,000	0.620	0.671
2,000	0.731	0.937

Table 5: Nominal-by-Nominal Association: Estimation of Power and for the F Statistic and the Likelihood Ratio Statistics (D_I) Based on 5,000 Simulations

Example 2.1		
N	F	D_I
200	0.532	0.976
300	0.602	0.999
400	0.641	1.000
500	0.669	1.000
600	0.697	1.000
700	0.716	1.000
800	0.736	1.000
900	0.759	1.000
1,000	0.769	1.000
2,000	0.872	1.000

Example 2.2		
N	F	D_I
200	0.343	0.474
300	0.387	0.649
400	0.455	0.806
500	0.490	0.903
600	0.525	0.947
700	0.564	0.979
800	0.583	0.988
900	0.610	0.996
1,000	0.629	0.998
2,000	0.772	1.000

Regarding Example 2.1, Table 5 shows that, for all *N*, the likelihood ratio statistic is considerably more powerful compared with the F statistic. When *N* increases, the power of the F statistic steadily increases, but the power of the likelihood ratio statistic converges to 1 for $N > 400$. Conversely, in Example 2.2, although the power of the likelihood ratio statistic is higher than the power of the F statistic, the rate of power increase is lower compared to Example 2.1.

Conclusion

A new statistic is proposed for testing independence in two-way contingency tables by dividing a table into two sub-tables. This method has been constructed based on the independence model so there is no need to specify any functional form for the association terms. Therefore, it could be applicable to any type of contingency tables, including nominal-by-nominal, nominal-by-ordinal and ordinal-by-ordinal.

The idea of partitioning contingency tables was first introduced by Kullback, et al. (1962) and Lancaster (1951). They showed that the overall Chi-square statistic for a contingency table can always be partitioned into as many components as the table's degrees of freedom. The Chi-square value of each component corresponds to a particular 2×2 table arising from the original table, and each component is independent of the others. Consequently a detailed examination of departures from independence can be made, thus enabling identification of those categories responsible for a significant overall Chi-square value. However, in this article the same technique was used for partitioning contingency tables that was applied to two-way ANOVA by Kharati and Sadooghi (2007). In the present work, this method was used for analyzing nominal-by-nominal and nominal-by-ordinal data.

It is notable that in a two-way ANOVA data are assumed to be normally distributed and the proposed F for testing interaction has an exact F distribution which leads to a two-sided test for equality of variances. In this study the response variable had Poisson distribution, so the proposed one-sided test has an asymptotic F distribution. Profile plots were also used as a preliminary tool to divide one table into two separate tables, which was the first step before applying the proposed method. However, there are other graphical methods such as corresponding analysis (Blasius & Greenarce, 2006), mosaic (Friendly, 1998) and z-plot (Choulakian & Allard, 1998), all of which can be helpful to visualizing and screening contingency tables before conducting any formal statistical analysis.

The power of the F statistic was compared with D_I and $D_{I|R}$. In Example 1, in which the row and column were nominal and ordinal respectively, it was believed that the row-effect model would be the best method for testing the association between row and column. Surprisingly, the proposed F statistic worked much better than expected. The results showed that while $D_{I|R}$ could not find any association between rows and columns; the proposed F was strongly significant. In this case the power simulation showed that the F statistic is more powerful than $D_{I|R}$ (0.43 vs. 0.15). Also the simulation results in Table 4 showed that for $N \leq 500$ the power of the F statistic was considerably higher than $D_{I|R}$. In this example, the results of the proposed F demonstrated that, despite the simplicity of its computations, it is more powerful than the row-effect model. These findings may encourage researchers to use the proposed F statistic for testing association in contingency tables.

In the Malignant Melanoma Example 2.1 when the row and column were nominal and there was a significant association between them, the simulation results showed that the D_I statistic was more powerful than F. In contrast, in the Malignant Melanoma Example 2.2, although D_I could not find any association between row and column, the proposed F was strongly significant. However, simulation showed that D_I was more powerful than F (0.76 vs. 0.44). In this case it should be noted that although the F statistic was often less powerful than the D_I, it was able to detect some special types of departures from the null hypothesis which could not be detected by D_I.

In conclusion, it is suggested that the F statistic serves as an alternative method for testing association in two-way contingency tables, in particular, if one variable is in ordinal scale. It is easy to use because it does not need any functional form for the association term. It is simple to compute and has good power. In addition to simplicity and flexibility, this test could be helpful in detecting the part of a table which contributes the association between row and column. It seems that, in some cases, this method enables us to detect an association in contingency tables that cannot be found by a row-effect model or likelihood ratio statistics.

References

Agresti, A. (1984). *Analysis of ordinal categorical data*. New York, NY: Wiley.

Agresti, A. (2002). *Categorical data analysis*. New York, NY: Wiley.

Alin, A., & Kurt, S. (2006). Testing non-additivity (interaction) in two-way ANOVA tables with no replication. *Statistical Methods in Medical Research*, *15*, 63-85.

Blasius, J., & Greenarce, M. (2006). Correspondence analysis and related methods in practice. In J. Blasius, & M. Greenarce, Eds., *Multiple correspondence analysis and related methods* (*Statistics in the Social and Behavioral Sciences*), 3-41. London: Chapman & Hall.

Choulakian, V., & Allard, J. (1998). The z-plot: a graphical procedure for contingency tables with an ordered response variable. In J.Blasius, & M. Greenarce, Eds., *Visualization of categorical data*, 99-106. San Diego, CA: Academic Press.

Christensen, R. (1990). *Log-linear models*. New York, NY: Springer-Verlag.

Davis, C. S. (1991). A one degree of freedom nominal association model for testing independence in two-way contingency tables. *Statistics in Medicine*, *10*, 1555-1563.

Dobson, A. (2002). *An Introduction to generalized linear models*. London: Chapman & Hall.

Fan, X., Felsovalyi, A., Sivo, S. A., & Keenan, S. C. (2002). *SAS for Monte Carlo studies a guide for quantitative researchers*. North Carolina: SAS Institute.

Friendly, M. (1998). Conceptual models for visualizing contingency table data. In J.Blasius, & M. Greenarce, Eds., *Visualization of categorical data*, 17-36. San Diego, CA: Academic Press.

Goodman, L. A. (1969). On partitioning chi-square and detecting partial association in three-way contingency tables. *Journal of the Royal Statistical Society*, *Series B*, *31*, 486-498.

Kharati, M., & Sadooghi, S. M. (2007). A new method for testing interaction in unreplicated two-way analysis of variance. *Communications in Statistics-Theory and Methods*, *36*, 2787-2803.

Kullback, S., Kupperman, M., & Ku, H. H. (1962). Tests for contingency tables and Markov chains. *Technometrics*, *4*, 573-608.

Lancaster, H. O. (1951). Complex contingency tables treated by partition of Chi-square. *Journal of the Royal Statistical Society*, *Series B*, *13*, 242-249.

Mandel, J. (1961). Non-additivity in two-way analysis of variance. *Journal of the American Statistical Association*, *56*, 878-888.

Milliken, G. A., & Graybill, F. A. (1970). Extensions of the general linear hypothesis model. *Journal of the American Statistical Association*, *65*, 797-807.

Simonoff, J. F. (2003). *Analyzing categorical data*. New York, NY: Springer.

Tukey, J. W. (1949). One degree of freedom for non-additivity. *Biometrics*, *5*, 232-242.

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 452-460

1538 – 9472/10/$95.00

Neighbor Balanced Block Designs for Two Factors

Seema Jaggi Cini Varghese
Indian Agricultural Statistics Research Institute,
New Delhi, India

N. R. Abeynayake
Wayamba University,
Gonawila (NWP), Sri Lanka

The concept of Neighbor Balanced Block (NBB) designs is defined for the experimental situation where the treatments are combinations of levels of two factors and only one of the factors exhibits a neighbor effect. Methods of constructing complete NBB designs for two factors in a plot that is strongly neighbor balanced for one factor are obtained. These designs are variance balanced for estimating the direct effects of contrasts pertaining to combinations of levels of both the factors. An incomplete NBB design for two factors is also presented and is found to be partially variance balanced with three associate classes.

Key words: Circular design, neighbor balanced, strongly neighbor balanced, variance balanced, partially variance balanced.

Introduction

In many agricultural experiments, the response from a given plot is affected by treatments applied to neighboring plots provided the plots are adjacent with no gaps. For example, when treatments are varieties, neighbor effects may be caused due to differences in height or date of germination, especially on small plots. Treatments such as fertilizer, irrigation, or pesticide may spread to adjacent plots causing neighbor effects. In order to avoid the bias in comparing the effects of treatments in such a situation, it is important to ensure that no treatment is unduly disadvantaged by its neighbor. Neighbor Balanced Designs, wherein the allocation of treatments is such that every treatment occurs equally often with every other treatment as neighbors, are used for these situations. These designs permit the estimation of direct and neighbor effects of treatments.

Azais, et al. (1993) developed a series of circular neighbor balanced block (NBB) designs for single factor experiments. A NBB design for a single factor with border plots is circular if the treatment in the left border is the same as the treatment in the right-end inner plot and the treatment in the right border is the same as the treatment in the left-end inner plot. Tomar, et al. (2005) also obtained some incomplete NBB designs for single factor experiments.

In certain experimental situations, the treatments are the combination of levels of two factors and only one of the factors exhibits neighbor effects. For example, agroforestry experiments consist of tree and crop combination in a plot and, because trees are much taller than the crop, it is suspected that the tree species of one plot may affect the response from the neighboring plots. The effect of the crop species in neighboring plots is assumed to be negligible. Under this situation, it is therefore desirable that designs allowing the estimation of direct effects of treatment combinations free of

Seema Jaggi is a Senior Scientist in Agricultural Statistics. Her research interests include: design of experiments, statistical computing, statistical techniques in agricultural research, web solutions to experimental designs and analysis. Email: seema@iasri.res.in. Cini Varghese is a Senior Scientist in Agricultural Statistics. Her research interests include: construction and analysis of experimental designs, characterization properties of experimental designs, web designs and software. Email: cini_v@iasri.res.in. N. R. Abeynayake is a Senior Lecturer in the Department of Agribusiness Management. Her research interests include: design of experiments, sample survey, applied statistics and teaching statistics. Email: rabeynayake@hotmail.com.

neighbor effects are developed. Langton (1990) advocated the use of both NBBs and guard areas in agroforestry experiments. Monod and Bailey (1993) presented two factor designs balanced for the neighbor effect of one factor.

NBB designs are defined for the experimental situation where the treatments are the combinations of levels of two factors and only one of the factors exhibits a neighbor effect in a block design with no gaps or guard areas between the plots. Some methods of constructing these designs balanced for the effects of one factor in the adjacent neighboring plots are presented.

Model

Let F_1 and F_2 be two factors in an experiment with f_1 and f_2 levels, respectively. The f_1 levels are represented as (*1, 2,* ...) and the f_2 levels as (*a, b,* ...). Consider an inner plot i (i = 1, 2, …, k) in the block $\theta(i)$ [= 1, 2, …, b] of a block design with a left neighbor plot i−1 and a right neighbor plot i+1. Let $\phi(i)$ and $\varphi(i)$ denote the levels of F_1 and F_2, respectively, on i. The general fixed effects model (Monod and Bailey, 1993) for Y_i, the response from plot i considered is

$$Y_i = \mu + \beta_{\theta(i)} + \tau_{\phi(i),\varphi(i)} + \delta_{\phi(i-1)} + \rho_{\phi(i+1)} + e_i \quad (1)$$

where μ is the general mean, $\beta_{\theta(i)}$ is the effect of block $\theta(i)$ to which plot i belongs, $\tau_{\phi(i),\varphi(i)}$ is the direct effect of the treatment combination $\phi(i)\varphi(i)$, $\delta_{\phi(i-1)}$ is the left neighbor effect of $\phi(i-1)$, $\rho_{\phi(i+1)}$ is the right neighbor effect of $\phi(i+1)$ and e_i is a random error term assumed to be identically and independently distributed with mean zero and constant variance.

Definitions

The following definitions for a block design with two factors in a plot and with neighbor effects for one factor (for example, F_1) from adjacent neighboring plots are provided.

Definition 1: circular block design. A block containing plots with treatment combinations and border plots is said to be left circular if the level of F_1 on the left border is the same as the level of F_1 on the right end inner plot. It is right circular if the level of F_1 on the right border is the same as the level of F_1 on the left end inner plot. A circular block is a left as well as right circular. A design with all circular blocks is called a circular block design. Note that the observations are not recorded from the border plots; these plots are taken only to have the neighbor effects of factor F_1.

Definition 2: strongly neighbor balanced. A circular block design with two factors F_1 and F_2 is called strongly neighbor balanced for factor F_1 if every combination of the two factors has each of the levels of factor F_1 as a right as well as a left neighbor a constant number of times, for example, μ_1'.

Definition 3: neighbor balanced. A circular block design with two factors F_1 and F_2 is neighbor balanced for factor F_1 if every combination of the two factors has levels of factor F_1 (except the level appearing in the combination) appearing μ_1'' times as a right and as a left neighbor.

Definition 4: variance balanced. A block design for two factors with left and right neighbor effects of factor F_1 is said to be variance balanced if the contrasts in the direct effects of $f_1 \times f_2$ combinations are estimated with the same variance, for example, V.

Definition 5: partially variance balanced. A block design for two factors with neighbor effects is partially variance balanced following some association scheme if the contrasts pertaining to the $f_1 \times f_2$ combinations from F_1 and F_2 factors are estimated with different variances, depending upon the order of association scheme.

Methodology

Complete NBB Designs for Two Factors: Method 1

Let f_1 be a prime number with its primitive root as x and $f_2 = f_1 - s$, $s = 1, 2, \ldots, f_1 - 2$. Obtain a basic array of f_2 columns each of size f_1 from the following initial sequence for values of $i = 1, 2, \ldots, f_2$:

f_1
x^i
x^{i+1}
.
.
.
x^{i+f_1-2}

Develop the columns of this array cyclically, mod f_1 to obtain f_1 sets of f_2 columns each. Allocate f_2 symbols denoted by a, b, … to each of the sets in such a way that symbol a occurs with all entries of column 1 in each set, b with all entries of column 2 of each set and so on. Considering the rows as blocks and making the blocks circular by adding appropriate border plots results in a complete block design for f_1f_2 treatment combinations with f_1 blocks each of size f_1f_2 which is strongly neighbor balanced for factor F_1. It is observed that each of the f_1f_2 combinations of factor F_1 and F_2 has each level of factor F_1 as left and right neighbor once, that is, $\mu_1' = 1$, and the design is complete in the sense that all the f_1f_2 combinations appear in a block. The designs obtained are variance balanced for estimating the direct effects of contrasts in f_1f_2 treatment combinations as the corresponding information matrix ($\mathbf{C}_\tau$) is:

$$\mathbf{C}_\tau = f_1\mathbf{I} - \frac{1}{f_2}\mathbf{J}, \qquad (2)$$

where **I** is an identity matrix of order f_1f_2 and **J** is the matrix of all unities.

Example 1

Let $f_1 = 5$ be the number of level of first factor F_1 represented by 1, 2, 3, 4, 5. Further let $s = 3$ resulting in $f_2 = f_1 - s = 2$ level of second factor denoted by a, b. If the rows represent the blocks and 5×2 (= 10) treatment combinations in rows the block contents, then the following arrangement forms a circular complete block design for 10 treatment combinations in five blocks each of size 10 strongly neighbor balanced for five levels of F_1:

4	*5a 5b*	*1a 1b*	*2a 2b*	*3a 3b*	*4a 4b*		*5*
3	*2a 4b*	*3a 5b*	*4a 1b*	*5a 2b*	*1a 3b*		*2*
2	*4a 3b*	*5a 4b*	*1a 5b*	*2a 1b*	*3a 2b*		*4*
5	*3a 1b*	*4a 2b*	*5a 3b*	*1a 4b*	*2a 5b*		*3*
1	*1a 2b*	*2a 3b*	*3a 4b*	*4a 5b*	*5a 1b*		*1*

It may be observed from the above that all the 10 combinations of factor F_1 and F_1 are balanced for factor F_1 as each combination has each of the levels of factor F_1 as left and right neighbor exactly once.

Example 2

If $f_1 = 5$ and $s = 2$, then $f_2 = 3$ and the design for 15 treatment combinations is as follows:

4	*5a 5b 5c*	*1a 1b 1c*	*2a 2b 2c*	*3a 3b 3c*	*4a 4b 4c*		*5*
2	*2a 4b 3c*	*3a 5b 4c*	*4a 1b 5c*	*5a 2b 1c*	*1a 3b 2c*		*2*
5	*4a 3b 1c*	*5a 4b 2c*	*1a 5b 3c*	*2a 1b 4c*	*3a 2b 5c*		*4*
1	*3a 1b 2c*	*4a 2b 3c*	*5a 3b 4c*	*1a 4b 5c*	*2a 5b 1c*		*3*
3	*1a 2b 4c*	*2a 3b 5c*	*3a 4b 1c*	*4a 5b 2c*	*5a 1b 3c*		*1*

Table 1 presents a list of designs consisting of the variance of contrast between different treatments combinations (V) along with other parameters for number of level of first factor (F_1) ≤ 13.

Complete NBB Designs for Two Factors: Method 2

Let f_1 be an even number and $f_2 = 2$. Obtain a square array **L** of order f_1 by developing the following initial sequence mod f_1 (replacing *0* by f_1):

$$1 \quad f_1 \quad 2 \quad f_1\text{-}1 \ldots \frac{f_1}{2} \quad \frac{f_1}{2}+1$$

Table 1: Parameters and Variance of Strongly Complete NBB Designs for Two Factors

f_1	s	f_2	$k = f_1 f_2$	$b = f_1$	μ_1'	V
5	1	4	20	5	1	0.40
	2	3	15		1	0.40
	3	2	10		1	0.40
7	1	6	42	7	1	0.29
	2	5	35		1	0.29
	3	4	28		1	0.29
	4	3	21		1	0.29
	5	2	14		1	0.29
11	1	10	110	11	1	0.18
	2	9	99		1	0.18
	3	8	88		1	0.18
	4	7	77		1	0.18
	5	6	66		1	0.18
	6	5	55		1	0.18
	7	4	44		1	0.18
	8	3	33		1	0.18
	9	2	22		1	0.18
13	1	12	156	13	1	0.15
	2	11	143		1	0.15
	3	10	130		1	0.15
	4	9	117		1	0.15
	5	8	104		1	0.15
	6	7	91		1	0.15
	7	6	78		1	0.15
	8	5	65		1	0.15
	9	4	52		1	0.15
	10	3	39		1	0.15
	11	2	26		1	0.15

Juxtapose the mirror image $\mathbf{L}'$ of $\mathbf{L}$ to the right hand side of $\mathbf{L}$ to obtain an arrangement of f_1 rows and $2f_1$ columns, and allocate the first level of F_2 to all the units of $\mathbf{L}$ and second level to all the units of $\mathbf{L}'$. Considering the rows as blocks and making the blocks circular results in a complete NBB design with block size $2f_1$ which is strongly neighbor balanced for factor F_1. Each of the $2f_1$ combinations of factor F_1 and F_2 have each level of factor F_1 as left and right neighbor exactly once, that is, $\mu_1'=1$ and the design is also variance balanced.

In general, for any even number of levels of F_2 ($f_2 = 2n$), the squares may be juxtaposed in the following manner:

$$\mathbf{L} \quad \mathbf{L}' \quad \mathbf{L} \quad \mathbf{L}' \quad \ldots$$

Allocating the first level of F_2 to each unit in $\mathbf{L}$, second level to units in $\mathbf{L}'$, third level to units in $\mathbf{L}$ again and so on, a complete NBB design in f_1 blocks of size $2f_1n$ balanced for factor F_1 is obtained. The designs obtained are also variance balanced for estimating the direct effects of

contrasts in $2f_1n$ treatment combinations as the corresponding information matrix ($\mathbf{C_\tau}$) is of the following form:

$$\mathbf{C_\tau} = f_1 \mathbf{I} - \frac{1}{2n}\mathbf{J}. \qquad (3)$$

Example 3

Let $f_1 = 6$ and $f_2 = 2$. Figure 1 shows a circular complete NBB block design for 6×2 (= 12) combinations in six blocks of size 12 balanced for six levels of F_1. For $f_2 = 4$, the design obtained for 6×4 (= 24) combinations in six blocks of size 24 strongly balanced for six levels of F_1 is shown in Figure 2.

Figure 1: Circular Complete NBB Block Design for 6×2 (= 12) Combinations

	L						**L′**						
1	*1a*	*6a*	*2a*	*5a*	*3a*	*4a*	*4b*	*3b*	*5b*	*2b*	*6b*	*1b*	*1*
2	*2a*	*1a*	*3a*	*6a*	*4a*	*5a*	*5b*	*4b*	*6b*	*3b*	*1b*	*2b*	*2*
3	*3a*	*2a*	*4a*	*1a*	*5a*	*6a*	*6b*	*5b*	*1b*	*4b*	*2b*	*3b*	*3*
4	*4a*	*3a*	*5a*	*2a*	*6a*	*1a*	*1b*	*6b*	*2b*	*5b*	*3b*	*4b*	*4*
5	*5a*	*4a*	*6a*	*3a*	*1a*	*2a*	*2*	*1b*	*3b*	*6b*	*4b*	*5b*	*5*
6	*6a*	*5a*	*1a*	*4a*	*2a*	*3a*	*3*	*2b*	*4b*	*1b*	*5b*	*6b*	*6*

Figure 2: Block design for 6×4 (= 24) combinations in six blocks of size 24

1	*1a*	*6a*	*2a*	*5a*	*3a*	*4a*	*4b*	*3b*	*5b*	*2b*	*6b*	*1b*	
2	*2a*	*1a*	*3a*	*6a*	*4a*	*5a*	*5b*	*4b*	*6b*	*3b*	*1b*	*2b*	
3	*3a*	*2a*	*4a*	*1a*	*5a*	*6a*	*6b*	*5b*	*1b*	*4b*	*2b*	*3b*	
4	*4a*	*3a*	*5a*	*2a*	*6a*	*1a*	*1b*	*6b*	*2b*	*5b*	*3b*	*4b*	
5	*5a*	*4a*	*6a*	*3a*	*1a*	*2a*	*2b*	*1b*	*3b*	*6b*	*4b*	*5b*	
6	*6a*	*5a*	*1a*	*4a*	*2a*	*3a*	*3b*	*2b*	*4b*	*1b*	*5b*	*6b*	
	L						**L′**						

1c	*6c*	*2c*	*5c*	*3c*	*4c*	*4d*	*3d*	*5d*	*2d*	*6d*	*1d*	*1*
2c	*1c*	*3c*	*6c*	*4c*	*5c*	*5d*	*4d*	*6d*	*3d*	*1d*	*2d*	*2*
3c	*2c*	*4c*	*1c*	*5c*	*6c*	*6d*	*5d*	*1d*	*4d*	*2d*	*3d*	*3*
4c	*3c*	*5c*	*2c*	*6c*	*1c*	*1d*	*6d*	*2d*	*5d*	*3d*	*4d*	*4*
5c	*4c*	*6c*	*3c*	*1c*	*2c*	*2d*	*1d*	*3d*	*6d*	*4d*	*5d*	*5*
6c	*5c*	*1c*	*4c*	*2c*	*3c*	*3d*	*2d*	*4d*	*1d*	*5d*	*6d*	*6*
L						**L′**						

Incomplete NBB Designs for Two Factors

Let f_1 be a prime or prime power and be denoted by *1,2,....* Develop f_1 - 1 mutually orthogonal Latin squares (MOLS) of order f_1. Juxtapose these MOLS so that we obtain an arrangement of f_1 symbols in $f_1(f_1 - 1)$ rows and f_1 columns. Delete the last q columns (q =0, 1, 2, …, f_1 - 4) and consider rows as blocks along with border plots, to make the blocks circular. To all the units in l^{th} column (l = 1, …, f_1 - q) of this arrangement attach the f_2 (f_2 = *a, b,* …) levels of F_2, i.e. *a* to column 1, *b* to column 2 and so on. Considering the rows as blocks and making the blocks circular results in an incomplete NBB design in $f_1(f_1 - 1)$ blocks of size f_1 - q each and $\mu_1'' = 1$ balanced for factor F_1. The design is incomplete because all the combinations are not appearing in a block. For q = 0, the design has all the levels of F_1 and F_2 appearing in all the blocks. The design obtained is combinatorially neighbor balanced but in the terms of variance, the design is partially balanced with three associate class association scheme.

Example 4

For $f_1 = f_2$ = 5 i.e. q = 0, NBB design in 25 combinations is as follows:

5	*1a*	*2b*	*3c*	*4d*	*5e*	*1*
1	*2a*	*3b*	*4c*	*5d*	*1e*	*2*
2	*3a*	*4b*	*5c*	*1d*	*2e*	*3*
3	*4a*	*5b*	*1c*	*2d*	*3e*	*4*
4	*5a*	*1b*	*2c*	*3d*	*4e*	*5*
4	*1a*	*3b*	*5c*	*2d*	*4e*	*1*
5	*2a*	*4b*	*1c*	*3d*	*5e*	*2*
1	*3a*	*5b*	*2c*	*4d*	*1e*	*3*
2	*4a*	*1b*	*3c*	*5d*	*2e*	*4*
3	*5a*	*2b*	*4c*	*1d*	*3e*	*5*
3	*1a*	*4b*	*2c*	*5d*	*3e*	*1*
4	*2a*	*5b*	*3c*	*1d*	*4e*	*2*
5	*3a*	*1b*	*4c*	*2d*	*5e*	*3*
1	*4a*	*2b*	*5c*	*3d*	*1e*	*4*
2	*5a*	*3b*	*1c*	*4d*	*2e*	*5*
2	*1a*	*5b*	*4c*	*3d*	*2e*	*1*
3	*2a*	*1b*	*5c*	*4d*	*3e*	*2*
4	*3a*	*2b*	*1c*	*5d*	*4e*	*3*
5	*4a*	*3b*	*2c*	*1d*	*5e*	*4*
1	*5a*	*4b*	*3c*	*2d*	*1e*	*5*

After randomization the design may have the following layout:

2	*3c*	*4d*	*5e*	*1a*	*2b*	*3*
2	*3b*	*4c*	*5d*	*1e*	*2a*	*3*
4	*5c*	*1d*	*2e*	*3a*	*4b*	*5*
1	*2d*	*3e*	*4a*	*5b*	*1c*	*2*
4	*5a*	*1b*	*2c*	*3d*	*4e*	*5*
3	*5c*	*2d*	*4e*	*1a*	*3b*	*5*
1	*3d*	*5e*	*2a*	*4b*	*1c*	*3*
3	*5b*	*2c*	*4d*	*1e*	*3a*	*5*
2	*4a*	*1b*	*3c*	*5d*	*2e*	*4*
2	*4c*	*1d*	*3e*	*5a*	*2b*	*4*
5	*3e*	*1a*	*4b*	*2c*	*5d*	*3*
2	*5b*	*3c*	*1d*	*4e*	*2a*	*5*
1	*4c*	*2d*	*5e*	*3a*	*1b*	*4*
5	*3d*	*1e*	*4a*	*2b*	*5c*	*3*
5	*3b*	*1c*	*4d*	*2e*	*5a*	*3*
2	*1a*	*5b*	*4c*	*3d*	*2e*	*1*
1	*5c*	*4d*	*3e*	*2a*	*1b*	*5*
1	*5d*	*4e*	*3a*	*2b*	*1c*	*5*
1	*5e*	*4a*	*3b*	*2c*	*1d*	*5*
1	*5a*	*4b*	*3c*	*2d*	*1e*	*5*

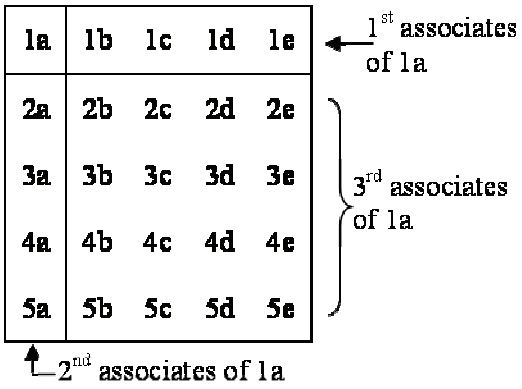
Association Scheme

Two treatment combinations $\phi\varphi$ and $\phi'\varphi'$ are said to be first associates if $\phi = \phi'$ i.e. the combinations with same F_1 level and different F_2 level are first associates. Two treatment combinations $\phi\varphi$ and $\phi'\varphi'$ are said to be second associate if $\varphi = \varphi'$ i.e. the combinations with same F_2 level and different F_1 level are second associates, and remaining are third associates.

For the Example 4, the arrangement of 25 treatment combinations arising from 5 levels of the first factor and 5 levels of second factor are shown in Figure 3. For the given association scheme $v = f_1 f_2$, number of first associates = f_2 – 1, number of second associates = f_1 – 1 and number of third associates = $f_1 f_2 - f_1 - f_2 + 1$. The two treatment combinations that are first and second associates do not appear together in the design whereas the third associates appear once in the design. The above association scheme

may be also called a rectangular association scheme.

Figure 3: 25 Treatment Combinations Arising From 5 Levels of the First Factor and 5 Levels of the Second Factor



The information matrix for estimating twenty five combinations of the above design obtained using SAS (PROC IML) is shown in (4). The matrix has three distinct off -diagonal elements due to the three class association scheme. The design obtained by Monod (1992) becomes a special case of this for $q = 0$.

Example 5

For $f_1 = 5$, $q = 1$ and $f_2 = 4$, that is, a NBB design in $f_1 f_2 = 20$ combinations is as follows:

4	*1a*	*2b*	*3c*	*4d*	*1*
5	*2a*	*3b*	*4c*	*5d*	*2*
1	*3a*	*4b*	*5c*	*1d*	*3*
2	*4a*	*5b*	*1c*	*2d*	*4*
3	*5a*	*1b*	*2c*	*3d*	*5*
2	*1a*	*3b*	*5c*	*2d*	*1*
3	*2a*	*4b*	*1c*	*3d*	*2*
4	*3a*	*5b*	*2c*	*4d*	*3*
5	*4a*	*1b*	*3c*	*5d*	*4*
1	*5a*	*2b*	*4c*	*1d*	*5*
5	*1a*	*4b*	*2c*	*5d*	*1*
1	*2a*	*5b*	*3c*	*1d*	*2*
2	*3a*	*1b*	*4c*	*2d*	*3*
3	*4a*	*2b*	*5c*	*3d*	*4*
4	*5a*	*3b*	*1c*	*4d*	*5*
3	*1a*	*5b*	*4c*	*3d*	*1*
4	*2a*	*1b*	*5c*	*4d*	*2*
5	*3a*	*2b*	*1c*	*5d*	*3*
1	*4a*	*3b*	*2c*	*1d*	*4*
2	*5a*	*4b*	*3c*	*2d*	*5*

Variances of all estimated elementary contrasts pertaining to direct effects of various treatment combinations that are mutually first associate (V_1), second associates (V_2) and third associate (V_3) were computed using a SAS program developed in IML. A list of designs consisting of these variances along with other parameters is shown in Table 2 for a practical range of parameter values, that is, for the number of level of first factor (F_1) and second factor (F_2) ≤ 13.

$$\mathbf{C} = \begin{bmatrix} 3.20\mathbf{I}_5 - 0.11\mathbf{J}_5 & 0.20\mathbf{I}_5 - 0.17\mathbf{J}_5 & 0.20\mathbf{I}_5 - 0.17\mathbf{J}_5 & 0.20\mathbf{I}_5 - 0.17\mathbf{J}_5 & 0.20\mathbf{I}_5 - 0.17\mathbf{J}_5 \\ 0.20\mathbf{I}_5 - 0.17\mathbf{J}_5 & 3.20\mathbf{I}_5 - 0.11\mathbf{J}_5 & 0.20\mathbf{I}_5 - 0.17\mathbf{J}_5 & 0.20\mathbf{I}_5 - 0.17\mathbf{J}_5 & 0.20\mathbf{I}_5 - 0.17\mathbf{J}_5 \\ 0.20\mathbf{I}_5 - 0.17\mathbf{J}_5 & 0.20\mathbf{I}_5 - 0.17\mathbf{J}_5 & 3.20\mathbf{I}_5 - 0.11\mathbf{J}_5 & 0.20\mathbf{I}_5 - 0.17\mathbf{J}_5 & 0.20\mathbf{I}_5 - 0.17\mathbf{J}_5 \\ 0.20\mathbf{I}_5 - 0.17\mathbf{J}_5 & 0.20\mathbf{I}_5 - 0.17\mathbf{J}_5 & 0.20\mathbf{I}_5 - 0.17\mathbf{J}_5 & 3.20\mathbf{I}_5 - 0.11\mathbf{J}_5 & 0.20\mathbf{I}_5 - 0.17\mathbf{J}_5 \\ 0.20\mathbf{I}_5 - 0.17\mathbf{J}_5 & 0.20\mathbf{I}_5 - 0.17\mathbf{J}_5 & 0.20\mathbf{I}_5 - 0.17\mathbf{J}_5 & 0.20\mathbf{I}_5 - 0.17\mathbf{J}_5 & 3.20\mathbf{I}_5 - 0.11\mathbf{J}_5 \end{bmatrix} \quad (4)$$

Table 2: Parameters and Variances of Incomplete NBB Designs for Two Factors

f_1	f_2	$k = f_2$	b	μ_1''	V_1	V_2	V_3	$\overline{V}$
5	5	5	20	1	6.40	6.13	6.53	6.44
5	4	4	20	1	6.00	5.37	5.87	5.79
7	7	7	42	1	10.29	10.17	10.46	10.40
7	6	6	42	1	10.00	9.81	10.14	10.10
7	5	5	42	1	9.60	9.23	9.63	9.56
7	4	4	42	1	9.00	8.13	8.62	8.55
8	8	8	56	1	12.25	12.17	12.42	12.40
8	7	7	56	1	12.00	11.87	12.16	12.10
8	6	6	56	1	11.67	11.44	11.78	11.70
8	5	5	56	1	11.20	10.77	11.17	11.10
8	4	4	56	1	10.50	9.50	10.00	9.94
9	9	9	72	1	15.97	15.95	15.94	15.90
9	8	8	72	1	14.00	13.91	14.16	14.10
9	7	7	72	1	13.71	13.57	13.85	13.80
9	6	6	72	1	13.33	13.08	13.42	13.40
9	5	5	72	1	12.80	12.32	12.72	12.70
9	4	4	72	1	12.00	10.87	11.37	11.30
11	11	11	110	1	18.18	17.98	18.17	18.20
11	10	10	110	1	17.89	17.47	17.99	17.90
11	9	9	110	1	17.72	17.58	17.75	17.70
11	8	8	110	1	17.44	17.19	17.45	17.40
11	7	7	110	1	17.07	16.72	17.04	17.00
11	6	6	110	1	16.51	16.07	16.49	16.40
11	5	5	110	1	15.84	15.06	15.57	15.50
11	4	4	110	1	14.23	13.33	13.73	13.70
13	13	13	156	1	22.15	22.13	22.17	22.20
13	12	12	156	1	22.00	22.00	22.17	22.10
13	11	11	156	1	21.82	21.77	21.95	21.90
13	10	10	156	1	21.60	21.53	21.73	21.70
13	9	9	156	1	21.33	21.24	21.46	21.40
13	8	8	156	1	21.00	20.86	21.12	21.10
13	7	7	156	1	20.57	20.35	20.64	20.60
13	6	6	156	1	20.00	19.64	19.97	19.90
13	5	5	156	1	19.20	18.51	18.91	18.90
13	4	4	156	1	18.00	16.37	16.87	16.80

References

Azais, J. M., Bailey, R. A., & Monod, H. (1993). A catalogue of efficient neighbor designs with border plots. *Biometrics*, *49*, 1252-1261.

Langton, S. (1990). Avoiding edge effects in agroforestry experiments: the use of neighbor-balanced designs and guard areas. *Agroforestry Systems*, *12*, 173-185.

Monod, H. (1992). Two factor neighbor designs in incomplete blocks for intercropping experiments. *The Statistician*, *41*(*5*), 487-497.

Monod, H., & Bailey, R. A. (1993). Two factor designs balanced for the neighbor effects of one factor. *Biometrika*, *80*, 643-659.

Tomar, J. S., Jaggi, S., & Varghese, C. (2005). On totally balanced block designs for competition effects. *Journal of Applied Statistics*, *32*(*1*), 87-97.

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 461-469

1538 – 9472/10/$95.00

Adjusted Confidence Interval for the Population Median of the Exponential Distribution

Moustafa Omar Ahmed Abu-Shawiesh
Hashemite University,
Zarqa Jordan

The median confidence interval is useful for one parameter families, such as the exponential distribution, and it may not need to be adjusted if censored observations are present. In this article, two estimators for the median of the exponential distribution, *MD*, are considered and compared based on the sample median and the maximum likelihood method. The first estimator is the sample median, MD_1, and the second estimator is the maximum likelihood estimator of the median, MD_{MLE}. Both estimators are used to propose a modified confidence interval for the population median of the exponential distribution, *MD*. Monte Carlo simulations were conducted to evaluate the performance of the proposed confidence intervals with respect to coverage probability, average width and standard error. A numerical example using a real data set is employed to illustrate the use of the modified confidence intervals; results are shown.

Key words: Exponential distribution, maximum likelihood estimator, sample median, confidence interval, coverage probability, average width.

Introduction

In most situations, researchers are interested in the estimate of the median of the population from which the sample data was drawn. Point estimates, such as the sample median, are of limited value because it is not possible to attach statements regarding the amount of confidence in their estimation of an unknown parameter. Of great value is an interval estimate, an estimate about which a researcher can make statements of confidence called the confidence interval (Daniel, 1990). A confidence interval provides much more information about the population value of the quantity of interest than does a point estimate (Smithson, 2001). Furthermore, the confidence intervals provide a way to report an estimate of a population parameter along with some information about the estimates precision. Although different settings lead to different formulas for computing confidence intervals, the basic interpretation is always the same. A two-sided confidence interval is the probability that a given parameter lies between a certain lower bound and upper bound (Kececioglu, 2002). According to Lewis (1996, page 216), confidence intervals are important because they are "the primary means by which the precision of a point estimator can be determined" and provide "lower and upper confidence limits to indicate how tightly the sampling distribution is compressed around the true value of the estimated quantity". The median confidence interval is useful for one parameter families, such as the exponential distribution, and it may not need to be adjusted if censored observations are present (Patel, et al., 1976).

The objective of this study is to modify the confidence interval for the population median of the exponential distribution, *MD*, based on two methods; the first method is based on the sample median, MD_1, while the second method is based on the maximum likelihood estimator of the median, MD_{MLE}. It is assumed that the underlying random sample $X_1, X_2, ..., X_n$ comes from an exponential distribution. The performance of the proposed

Moustafa Omar Ahmed Abu-Shawiesh is a member of the Faculty of Science in the Department of Mathematics. Email him at: mabushawiesh@hu.edu.jo.

modified confidence intervals is evaluated and compared using a Monte Carlo simulation to calculate the estimated coverage probability, the average width and the standard error; the use of these newly proposed methods is illustrated by a numerical example.

The Exponential Distribution

The exponential distribution is one of the most important and widely used continuous probability distributions in statistical practice. It possesses several important statistical properties, and yet exhibits great mathematical tractability (Balakrishnan & Basu, 1996). It is the most frequently used distribution in such fields as queuing theory, reliability theory and reliability engineering where in this case it is provide models which are used to study many industrial phenomena such as time between machine breakdowns, length of queues or waiting time problems, at repair or processing facilities and the reliability of electronic systems, for example how long it takes for a bank teller to serve a customer (Maguire, et al., 1952; Betteley, et al., 1994 ; Montgomery, 2005). The exponential distribution also plays an important part in life testing problems; it would be an adequate choice for a situation where the failure rate appears to be more or less constant (Sinha & Bhattacharjee, 2004). The exponential distribution may be viewed as a continuous counterpart of the geometric distribution, which describes the number of Bernoulli trials necessary for a discrete process to change state. In contrast, the exponential distribution describes the time for a continuous process to change state (Trivedi, 2001). Furthermore, the exponential distribution is related to Poisson in much the same way as the geometric is to binomial, where in a Poisson process the time between events has an exponential distribution (Betteley, et al., 1994). The exponential distribution is also the only continuous distribution having what is called the memoryless property, that is, the future lifetime of an individual has the same distribution no matter how it is at present.

The random variable X has an exponential distribution with the rate parameter λ, that is, $X \sim Exp(1/\lambda)$, if and only if the density of it can be written as follows:

$$f(x;\lambda) = \lambda e^{-\lambda x}, \quad x \geq 0, \lambda > 0 \qquad (1)$$

The parameter λ , represents the mean number of events per unit time (e.g., the rate of arrivals or the rate of failure). The exponential distribution is supported on the interval [0, ∞). The mean (expected value) of an exponentially distributed random variable X with rate parameter λ is given by:

$$\mu = E(X) = \frac{1}{\lambda} \qquad (2)$$

In light of the examples given above, this makes sense: if a person receives phone calls at an average rate of 2 per hour, then they can expect to wait one-half hour for every call. Also, note that approximately 63% of the possible values lie below the mean for any exponential distribution (Betteley, et al., 1994).

The median of an exponentially distributed random variable X with rate parameter λ is given by:

$$MD = \frac{\ln(2)}{\lambda} = \frac{0.69315}{\lambda} \qquad (3)$$

The maximum likelihood estimator (MLE) for the rate parameter λ, given an independent and identically distributed random sample of size n, $X_1, X_2, ..., X_n$, drawn from the exponential distribution, $Exp(1/\lambda)$, is given by:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\overline{X}} . \qquad (4)$$

While this estimate is the most likely reconstruction of the true parameter λ, it is only an estimate, and as such, the more data points available the better the estimate will be. Also, the MLEs are consistent estimators of their parameters and are asymptotically efficient (Casella & Berger, 2002).

The Used Estimators

The sample mean, $\overline{X}$, and the

sample median, MD_1, which are used in this study for constructing the proposed modified confidence intervals for the exponential distribution median, *MD*, are now introduced.

The Sample Mean, $\overline{X}$

The sample mean is the most well known example of a measure of location, or average. It is defined for a set of values as the sum of values divided by the number of values and is denoted by $\overline{X}$. The sample mean for a random sample of size *n* observations $X_1, X_2, ..., X_n$ can be defined as follows:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad (5)$$

The main advantages of the sample mean, $\overline{X}$, are: it is easy to compute, easy to understood and takes all values into account. Its main disadvantages are: it is influenced by outliers, can be considered unrepresentative of data where outliers occur because many values may be well away from it and it requires all values in order to calculate its value (Betteley, et al., 1994; Francis, 1995).

The Sample Median, MD_1

The sample median is perhaps the best known of the resistant location estimators. It is insensitive to behavior in the tails of the distribution. The sample median is defined for a set of values as the middle value when the values are arranged in order of magnitude and it is denoted herein by MD_1. The sample median for a random sample of size *n* observations $X_1, X_2, ..., X_n$ can be defined as follows:

$$MD_1 = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & if \quad n \quad is \quad odd \\ \dfrac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2} & if \quad n \quad is \quad even \end{cases} \qquad (6)$$

The main advantages of the sample median, MD_1, is that, it is easy to determine, requires only the middle values to calculate, can be used when a distribution is skewed - as in the case of the exponential distribution, is not affected by outliers and has a maximal 50% breakdown point. The main disadvantages of the sample median, MD_1, are that it is difficult to handle in mathematical equations, it does not use all available values and it can be misleading in a distribution with a long tail because it discards so much information. The sample median, though, is considered as an alternative average to the sample mean (Betteley, et al., 1994; Francis, 1995). However, the sample median, MD_1, has become as a good general purpose estimator and is generally considered as an alternative average to the sample mean, $\overline{X}$.

Estimating the Exponential Distribution Median:

Two techniques are now introduced for finding estimates, the method of sample median and the method of maximum likelihood which is the most widely used.

The Method of Sample Median

Given a random sample of size *n* observations, $X_1, X_2, ..., X_n$, the estimator of the exponential population median, *MD*, based on the method of sample median, MD_1, is denoted by MD_{MD1}. Now, from equation (3):

$$\mathrm{MD} = \frac{1}{\lambda}\ln(2) \;\Rightarrow\; \lambda = \frac{\ln(2)}{\mathrm{MD}} \qquad (8)$$

Thus, if the exponential population median *MD* in (8) is estimated by the sample median MD_1, results in the following approximation:

$$\hat{\lambda} = \frac{\ln(2)}{MD_1}. \qquad (9)$$

Therefore, equating the results in (4) and (9) and solving, the following approximation is obtained:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i} \cong \frac{\ln(2)}{MD_1} \Rightarrow \sum_{i=1}^{n} X_i \cong \frac{n\,MD_1}{\ln(2)} \tag{10}$$

The Maximum Likelihood Estimator of the Median

Given a random sample of size n observations, $X_1, X_2, ..., X_n$, the estimator of the exponential population median based on the maximum likelihood method is denoted herein by MD_{MLE} and can be defined as follows:

$$MD_{MLE} = \frac{1}{\hat{\lambda}} \ln(2) = \frac{\sum_{i=1}^{n} X_i}{n} \ln(2) \tag{11a}$$

where

$$\hat{\lambda} = \frac{\mathrm{n}}{\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{X_i}} = \frac{1}{\overline{\mathrm{X}}} \,. \tag{11b}$$

Comparing the Two Estimators of the Exponential Distribution Median

It is known that the maximum likelihood estimators are asymptotically unbiased and efficient. Concretely, the estimator MD_{MLE} is unbiased and $Var(MD_{MLE}) = \frac{\ln^2(2)}{n\lambda^2}$. Moreover, the sample median estimator, MD_1, is asymptotically normal distributed with asymptotic variance $AVar(MD_1) = \frac{1}{4nf^2(MD)}$, where $f(.)$ is the corresponding density and MD is the theoretical median (Vann deer Vaart, 1998). Asymptotically unbiased means that the average value over many random samples for the two estimators MD_1 or MD_{MLE} is the exponential distribution median, MD. To compare the two estimators MD_1 and MD_{MLE} in terms of how far they are from the exponential distribution median (MD) on the average for many random samples, it is necessary to compare their root mean square error, $RMSE$, given as follows:

$$RMSE = \sqrt{\frac{1}{r} \sum_{i=1}^{r} (MD_i - MD)^2} \tag{12}$$

where $MD_1, MD_2, ..., MD_r$ are the values of the estimators MD_1 and MD_{MLE} for r replications and MD is the value of the exponential distribution true median.

The Confidence Interval for the Exponential Distribution Median

Next, the confidence interval for the median of the exponential distribution, MD, is derived by modifying the confidence interval for the mean of the exponential distribution, μ. Let $X_1, X_2, ..., X_n$ be a random sample of size n from the exponential distribution with parameter λ, that is, $X \sim Exp(1/\lambda)$, then the exact two sided $100(1-\alpha)\%$ confidence interval for the exponential distribution mean, μ, is given by (Trivedi, 2001):

$$P\left(\frac{2\sum_{i=1}^{n} X_i}{\chi^2_{(2n,\alpha/2)}} < \mu = \frac{1}{\lambda} < \frac{2\sum_{i=1}^{n} X_i}{\chi^2_{(2n,1-\alpha/2)}}\right) = 1-\alpha \tag{13}$$

where the $\chi^2_{(v,p)}$ is the p^{th} percentile of a Chi-square distribution with v degrees of freedom. The χ^2-value can be read off from the χ^2-table for the distribution (for example, see Kinney, 1997, page 506). Now, from equation (3) the exponential distribution median, MD, is given as follows:

$$\mathrm{MD} = \frac{1}{\lambda} \ln(2) = \mu \ln(2) \Rightarrow \frac{\mathrm{MD}}{\ln(2)} = \frac{1}{\lambda} = \mu \tag{14}$$

The $\mathrm{MD_{MLE}}$ Confidence Interval

This confidence interval is obtained by substituting the result from (14) into equation (13); this gives the exact $100(1-\alpha)\%$ confidence interval for the exponential distribution median, MD, as follows:

$$P\left(\frac{2\sum_{i=1}^{n} X_i}{\chi^2_{(2n,\alpha/2)}} < \frac{MD}{\ln(2)} = \frac{1}{\lambda} < \frac{2\sum_{i=1}^{n} X_i}{\chi^2_{(2n,1-\alpha/2)}}\right)$$

$$= P\left(\frac{2\ln(2)\sum_{i=1}^{n} X_i}{\chi^2_{(2n,\alpha/2)}} < MD < \frac{2\ln(2)\sum_{i=1}^{n} X_i}{\chi^2_{(2n,1-\alpha/2)}}\right)$$

$$= 1-\alpha \qquad (15)$$

The MD_{MLE} confidence interval is exact. It is based on the fact that $2\lambda \sum_{i=1}^{n} X_i$ follows a $\chi^2_{(2n)}$ distribution. The coverage probability must be exactly 95%.

The MD_{MD1} Confidence Interval

This confidence interval is obtained by substituting the result from (10) into equation (13) to give the exact $100(1-\alpha)\%$ confidence interval for the exponential distribution median, *MD*, as follows:

$$P\left(\frac{2(n\,MD_1/\ln(2))}{\chi^2_{(2n,\alpha/2)}} < \frac{MD}{\ln(2)} < \frac{2(n\,MD_1/\ln(2))}{\chi^2_{(2n,1-\alpha/2)}}\right)$$

$$= P\left(\frac{2n\,MD_1}{\chi^2_{(2n,\alpha/2)}} < MD < \frac{2n\,MD_1}{\chi^2_{(2n,1-\alpha/2)}}\right)$$

$$= 1-\alpha \qquad (16)$$

The MD_1 confidence interval is not exact. Its expression in (16) is based on equation (10) which is only an approximation of the statistics In order to see that, the performance of the MD_1 confidence interval is studied by calculating the coverage probability, the average width and the standard error using Monte-Carlo simulations. Actual, approximate and exact confidence intervals based on the sample median MD_1 can be also constructed using standard methods.

The Monte Carlo Simulation Study

A Monte Carlo simulation was designed to compare and study the behavior of the two estimators MD_1 and MD_{MLE} and investigate the behavior of the proposed approximate confidence intervals for the exponential distribution median, *MD*. *FORTRAN* programs were used to generate the data from the exponential distribution and run the simulations and to make the necessary tables. Results are from the exponential distribution with parameter λ which was set to 1 and 0.5, to increase skewness. The more the repetition, the more accurate are simulated results, therefore 10,000 random samples of sizes n = 10, 15, 20, 30, 40, 50 and 100 were generated.

Table (1) shows the simulated results for the root mean square error, *RMSE*, and the average of MD_1's and MD_{MLE}'s (AVG) to illustrate that both estimators are approximately unbiased for the true median of the exponential distribution, *MD*. The simulated results for the coverage probability ($\hat{\mathbf{P}}$), average width (AW) and standard error (SE) of the exact confidence interval for the exponential mean, μ, and the two proposed approximate confidence intervals for the exponential distribution median, *MD*, are shown in tables (2-4). The criteria used to evaluate the exact and proposed approximate confidence intervals is the value of the coverage probability ($\hat{\mathbf{P}}$) and average width (AW); a good method should have an observed coverage probability ($\hat{\mathbf{P}}$) near to the nominal coverage probability and a small scaled average width (AW).

The simulation results in Table 1 show that the maximum likelihood estimator of the median, MD_{MLE}, is a much better estimator for the population median of the exponential distribution, *MD*, than the sample median, MD_1. While both estimators are approximately unbiased, the root mean square error, *RMSE*, for the sample median, MD_1, is larger than that of the maximum likelihood estimator of the median, MD_{MLE}. It should be noted that the accuracy of the maximum likelihood estimator of the median, MD_{MLE}, increases as the sample size, n, increases which clearly provides a very good estimator, even considering that the discrepancy of these two estimators is very small

-that is, these two estimators asymptotically coincide.

As shown in Tables 2-4, the simulation results show that the coverage probability ($\hat{\mathbf{P}}$) for the confidence interval of the mean and the approximate confidence interval of the median based on the *MLE* method for the exponential distribution are the same and very close to the nominal confidence coefficient.

The approximate confidence interval of the median based on the sample median method for the exponential distribution provides the lower coverage probability ($\hat{\mathbf{P}}$) and gives the lowest width among the three methods. The average widths (AW) for the two proposed confidence interval methods are approximately the same for moderate and large sample sizes. However, the estimated average width (AW) for the sample median method is the shortest among all considered methods, but it has poor coverage probability. Furthermore, as sample sizes increases, the performance of the proposed confidence interval based on the *MLE* method improves, but is still much lower than the nominal confidence.

Table 1: The Root Mean Square Error and Average of the Two Estimators for the Exponential Distribution Median

Sample Size (n)	$\lambda = 1$ (True Median = 0.69315)			
	RMSE(MD_1)	AVG(MD_1)	RMSE(MD_{MLE})	AVG(MD_{MLE})
10	0.31575	0.74851	0.22369	0.69487
15	0.26754	0.72676	0.18040	0.69332
20	0.22459	0.72003	0.15715	0.69394
30	0.18276	0.71180	0.12808	0.69335
40	0.15907	0.70649	0.11106	0.69342
50	0.14108	0.70459	0.09887	0.69439
100	0.10005	0.69894	0.06961	0.69265
Sample Size (n)	$\lambda = 0.5$ (True Median = 1.38629)			
	RMSE(MD_1)	AVG(MD_1)	RMSE(MD_{MLE})	AVG(MD_{MLE})
10	0.63150	1.49703	0.44739	1.38973
15	0.53509	1.45352	0.36080	1.38663
20	0.44918	1.44005	0.31430	1.38788
30	0.36553	1.42359	0.25615	1.38671
40	0.31814	1.41299	0.22211	1.38684
50	0.28215	1.40917	0.19773	1.38877
100	0.20010	1.39788	0.13922	1.38530

Table 2: Coverage Probabilities, Average Width and Standard Error for the Confidence Interval of the Mean of the Exponential Distribution

n	$1-\alpha=0.95$					
	$\lambda=1$			$\lambda=0.5$		
	$\hat{\mathbf{P}}$	AW	SE	$\hat{\mathbf{P}}$	AW	SE
10	94.55	1.504	0.484	94.55	3.007	0.968
15	94.84	1.148	0.299	94.84	2.297	0.598
20	94.51	0.964	0.218	94.51	1.928	0.437
30	94.58	0.762	0.141	94.58	1.524	0.282
40	94.94	0.650	0.104	94.94	1.299	0.208
50	94.79	0.576	0.082	94.79	1.153	0.164
100	95.02	0.399	0.040	95.02	0.798	0.080

Table 3: Coverage Probabilities, Average Width and Standard Error for the Confidence Interval of the Median of the Exponential Distribution with $\lambda=1$

n	$1-\alpha=0.95$					
	Confidence Interval Method					
	MD_{MLE} Method			MD_{MD1} Method		
	$\hat{\mathbf{P}}$	AW	SE	$\hat{\mathbf{P}}$	AW	SE
10	94.55	1.123	0.336	85.11	1.042	0.466
15	94.84	0.834	0.207	82.76	0.796	0.305
20	94.51	0.693	0.151	83.92	0.668	0.215
30	94.58	0.542	0.098	83.41	0.528	0.139
40	94.94	0.459	0.072	83.00	0.450	0.103
50	94.79	0.405	0.057	83.28	0.400	0.081
100	95.02	0.279	0.028	82.95	0.277	0.040

Table 4: Coverage Probabilities, Average Width and Standard Error for the Confidence Interval of the Median of the Exponential Distribution with $\lambda=0.5$

n	$1-\alpha=0.95$					
	Confidence Interval Method					
	MD_{MLE} Method			MD_{MD1} Method		
	$\hat{\mathbf{P}}$	AW	SE	$\hat{\mathbf{P}}$	AW	SE
10	94.55	2.246	0.671	85.11	2.085	0.933
15	94.84	1.669	0.414	82.76	1.592	0.609
20	94.51	1.387	0.303	83.92	1.337	0.430
30	94.58	1.085	0.195	83.41	1.056	0.277
40	94.94	0.918	0.144	83.00	0.901	0.206
50	94.79	0.811	0.114	83.28	0.799	0.162
100	95.02	0.558	0.056	82.95	0.553	0.080

Numerical Example

This example is taken from Wilk, et al. (1962); the data set represents the lifetimes (in weeks) of 34 transistors in an accelerated life test. The transistors were tested and the test continued until all of them failed. The lifetimes for the 34 transistors (in weeks) were recorded as follows:

3,	4,	5,	6,	6,	7,	8,	8,	9,
9,	9,	10,	10,	11,	11,	11,	13,	13,
13,	13,	13,	17,	17,	19,	19,	25,	29,
33,	42,	42,	52,	52,	52,	52		

The sample mean $\overline{X} = 18.912$ weeks, the exponential median MD_1 = 13 weeks, the exponential median MD_2 = 13.108 weeks and the skewness is 1.265695, which is highly skewed distribution. Furthermore, $\hat{\lambda} = \frac{1}{\overline{X}} = \frac{1}{18.912} = 0.05287 \cong 0.053$ and - using the approximation derived earlier - results in $\hat{\lambda} = \frac{\ln(2)}{MD_1} = \frac{0.69315}{13} = 0.05332 \cong 0.053$ which indicates that the two values are very close and therefore the approximation is good. Based on Kibria (2006), the above data set are assumed to come from an exponential distribution with mean $\mu = 21$ weeks; using the Kolmogorov-Smirnov test, the K-S statistic = 0.1603 and the p-value = 0.3125, it indicates that the sample data are from an exponential distribution with mean $\mu = 21$ weeks, and therefore (from equation (3)) has a median *MD* = 14.556 weeks. The resulting 95% confidence interval and the corresponding confidence width for the exact confidence interval for the exponential mean and the two proposed methods for the exponential median are calculated and given in table (5).

Table 5: The 95% Confidence Intervals for the Lifetimes Data

Confidence Interval Method	95% Confidence Interval	Width
Exact for Mean	(13.874 , 27.308)	13.434
MD_{MLE}	(9.617 , 18.929)	9.312
MD_{MD1}	(9.537 , 18.772)	9.235

From table (5), it is observed that the exact confidence interval for the exponential mean, as expected, covered the hypothesized true population mean of $\mu = 21$ weeks and also the proposed confidence intervals for the exponential median, *MD*, covered the hypothesized true population median *MD* = 14.556 weeks. However, the proposed confidence interval for the exponential median, *MD*, based on the sample median, MD_1, provided the shortest confidence interval width.

Conclusion

The median - one of the most important and popular measures for location - has many good features. The median confidence interval is useful for one parameter families, such as the exponential distribution, and it may not need to be adjusted if censored observations are present. The maximum likelihood estimation is a popular statistical method used to make inferences about parameters of the underlying probability distribution from a given data set. This study proposed an approximate confidence interval for the median of the exponential distribution, *MD*, based on two estimators, the sample median, MD_1, and the maximum likelihood estimator of the median, MD_{MLE}.

The results of this study show that using a maximum likelihood estimator, *MLE*, for the population median of the exponential distribution, *MD*, is better alternative to the classical estimator based on the sample median, MD_1. As shown by the study results, the maximum likelihood estimator of the median, MD_{MLE}, provides a good estimation for the population median of the exponential distribution, *MD*, and the proposed confidence interval based on this estimator had a good coverage probabilities compared to the sample median method. However, it produced slightly wider estimated width. It appears that the sample size, *n*, has significant effect on the two proposed confidence interval methods. Moreover, both of the proposed methods are computationally simpler. If scientists and

researchers are conservative about the smaller width, they might consider confidence interval based on sample median method as a possible interval estimator for the population median of the exponential distribution, *MD*. Finally, the results obtained from the simulation study coincided with that of the numerical example.

Acknowledgement

The author would like to thank the Hashemite University for their cooperation during the preparation of this article.

References

Balakrishnan, N., & Basu, A. P. (1996). *The exponential distribution: theory, methods and applications*, CRC press.

Bettely, G., Mettric, N., Sweeney, E., & Wilson, D. (1994). *Using statistics in industry: Quality improvement through total process control*, (*1st Ed.*). London: Prentice Hall International Ltd.

Casella, G., & Berger, R. L. (2002). *Statistical inference*. Pacific Grove, CA: Duxbury.

Daniel, W. W. (1990). *Applied nonparametric statistics* (*2nd Ed.*). Duxbury, Canada: Thomson Learning.

Francis, A. (1995). *Business mathematics and statistics*. London: DP Publications, Ltd.

Kececioglu, D. (2002). *Reliability engineering handbook*. Lancaster, UK: DEStech Publications.

Kibria, B. M. G. (2006). Modified confidence intervals for the mean of the asymmetric distribution. *Pakestanian Journal of Statistics*, *22*(*2*), 109-120.

Kinney, J. (1997). *Probability: An introduction with statistical applications*. New York: John Wiley and Sons, Inc.

Lewis, E. E. (1996). *Introduction to reliability engineering*, (*2nd Ed.*). New York: John Wiley and Sons, Inc.

Maguire, B. A., Pearson, E. S., & Wynn, A. H. A. (1952). The time intervals between industrial accidents. *Biometrika*, *39*, 168-180.

Montgomery, D. C. (2005). *Introduction to statistical quality control* (*5th Ed.*). New York: John Wiley and Sons, Inc.

Patel, J. K., Kapadia, C. H., & Owen, D. B. (1976). *Handbook of statistical distributions*. New York: Marcel Dekker.

Sinha, A., & Bhattacharjee, S. (2004). Test of parameter of an exponential distribution in predictive approach. *Pakestanian Journal of Statistics*, *20*(*3*), 409-414.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement*, *61*, 605-532.

Trivedi, K. S. (2001). *Probability and statistics with reliability, queuing, and computer science applications* (*2nd Ed.*). New York: John Wiley and Sons, Inc.

Van der Vaart, A.W., (1998). *Asymptotic Statistics*, 1-496. Cambridge University Press, Cambridge.

Wilk, M. B., Gnanadesikan, R., & Huyett, M. J. (1962). Estimation of parameters of the gamma distribution using order statistics. *Biometrika*, *49*, 525-545

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 470-479

1538 – 9472/10/$95.00

Nonlinear Trigonometric Transformation Time Series Modeling

K. A. Bashiru
Osun State University,
Osogbo, Osun State

O. E. Olowofeso S. A. Owabumoye
Federal University Of Technology,
Akure Ondo State, Nigeria

The nonlinear trigonometric transformation and augmented nonlinear trigonometric transformation with a polynomial of order two was examined. The two models were tested and compared using daily mean temperatures for 6 major towns in Nigeria with different rates of missing values. The results were used to determine the consistency and efficiency of the models formulated.

Key words: Nonlinear time series, polynomial, consistency, efficiency, missing value, model and forecasting.

Introduction

Time series analysis is an important technique used in many disciplines, including physics, engineering, finance, economics, meteorology, biology, medicine, hydrology, oceanography and geomorphology (Terasvirta & Anderson, 1992). This technique is primarily used to infer properties of a system by the analysis of a measured time record (data) (Priestley, 1988); this is accomplished by fitting a representative model to the data with an aim of discovering the underlying structure as closely as possible.

Traditional time series analysis is based on assumptions of linearity and stationarity. However, time series analysis (Box & Jenkins, 1970; Brock & Potter, 1993) has nonlinear features such as cycles, asymmetries, bursts, jumps, chaos, thresholds and heteroscedasticity, and mixtures of these must also be taken into account. Thus, a problem arises regarding a suitable definition of a nonlinear model because not every time series analysis is purely linear: the nonlinear class clearly encompasses a large number of possible choices. For these reasons, non-linear time series analysis is a rapidly developing area and there have been major developments in model building and forecasting (De Gooijer & Kumar, 1992).

The growing interest in studying nonlinear and non-stationary time series models in many practical problems stems from the inherently non-linear nature of many phenomena in physics, engineering, meteorology, medicine, hydrology, oceanography, economics and finance, that is, many real world problems do not satisfy the assumptions of linearity and/or stationarity (Bates & Watts, 1988; DeGooijer & Kumar, 1992; Sugihara & May, 1990). Therefore, for many real time series data, nonlinear models are more appropriate than linear models for accurately describing the dynamic of the series and making multi-step-ahead forecast (Tsay, 1986; Barnett, Powell & Tauchen, 1991; Olowofeso, 2006). For example, financial markets and trends are influenced by climatic factors like daily temperature, amount of rainfall and intensity of sun, these are areas where a need exists to explain behaviors that are far from being even approximately linear. Nonlinear models would be more appropriate for forecasting and accurately describing returns and

K. A. Bashiru is a lecturer in the Department of Mathematical and Physical Sciences. His areas of interest are Geostatistics and Time series Econometrics. Email him at: kehindeadekunle2@yahoo.com. O. E. Olowofeso is an Associate Professor of Statistics in the Mathematical Sciences Department. He is also currently working in the Central Bank of Nigeria, Abuja. Email: olowofeso@yahoo.com. S. A. Owabumoye is a Master's student under the supervision of O. E Olowofeso in the Mathematical Sciences Department. Email: saowabumoye@yahoo.com.

volatility. Thus, the need for the further development of the theory and applications for nonlinear models is essential, and, because there are an enormous number of nonlinear models available for modeling and forecasting economic time series, research should help provide guidance for choosing the best model for a particular application (Robinson, 1983).

Methodology

The model proposed by Gallant (1981) called the Augmented Nonlinear Parametric Time Series Model (ANPTSM) was used in this study and a second model was formulated based on the Least Square Method Modified Nonlinear Trigonometric Transformation Time Series Model (MNTTTSM).

Data

Data used in this study were daily mean of temperatures from 1987 to 1996 for Ikeja, Ibadan, Ilorin, Minna and Zaria. The data were collected from the Meteorological Centre-Oshodi Lagos.

Model Formulation

Consider the format shown in Table 1. In this model, up to 9 years were considered and the model is formulated based on the data as shown in Table 2.

Assumption and Notation for the Models

Let:

$X_{t,i,k}$ = value of occurrence for day t of Month i in the year k;

$X_{t,k}$ = mean occurrence for day t of year k;

$X^*_{i,k}$ = mean occurrence for month i of year k;

X^{*y}_K = overall yearly mean for the sampled month;

X^{*m}_i = overall monthly mean for the sampled year;

t = the position of the day from the first day of the Month. $1 \leq t \leq 31$;

t_i = the sum of days in month i for $1 \leq i \leq 12$;

t_{ik} = the sum of days from the initial sampled month of initial sampled year to month i of year k;

t_i* = the sum of days from the initial sampled month to month I;

k = the position of a particular year from an initial sample year for $-\infty \leq k \leq \infty$;

n = the number of sampled years;

m = the number of sampled months; and

X_* = Grand Mean occurrence for k year(s) examined.

The first model was reviewed based on the assumption that the sum of the occurrences were presented monthly, where i^{th} month represents the month i for $1 \leq i \leq 12$ which is to be modeled using the number of days in each month (see Table 3).

Table 1: Model Formulation for a Particular Year

t/i	1	2	3	4	5	6	7	8	9	10	11	12	$\Sigma x_i/12$
1	$x_{1,1,k}$	$x_{1,2,k}$	$x_{1,3,k}$	$x_{1,4,k}$	$x_{1,5,k}$	$x_{1,6,k}$	$x_{1,7,k}$	$x_{1,8,k}$	$x_{1,9,k}$	$x_{1,10,k}$	$x_{1,11,k}$	$x_{1,12,k}$	$\bar{X}_1$
2	$x_{2,1,k}$	$x_{2,2,k}$	$x_{2,3,k}$	$x_{2,4,k}$	$x_{2,5,k}$	$x_{2,6,k}$	$x_{2,7,k}$	$x_{2,8,k}$	$x_{2,9,k}$	$x_{2,10,k}$	$x_{2,11,k}$	$x_{2,12,k}$	$\bar{X}_2$
t	$x_{t,1,k}$	$x_{t,2,k}$	$x_{t,3,k}$	$x_{t,4,k}$	$x_{t,5,k}$	$x_{t,6,k}$	$x_{t,7,k}$	$x_{t,8,k}$	$x_{t,9,k}$	$x_{t,10,k}$	$x_{t,11,k}$	$x_{t,12,k}$	$\bar{X}_t$
	$x^*_{1,k}$	$x^*_{2,k}$	$x^*_{3,k}$	$x^*_{4,k}$	$x^*_{5,k}$	$x^*_{6,k}$	$x^*_{7,k}$	$x^*_{8,k}$	$x^*_{9,k}$	$x^*_{10,k}$	$x^*_{11,k}$	$x^*_{12,k}$	$\Sigma x^*_{i,k}/12$

Table 2: Model Data Formulation

t/i_k	1	2	3	…	i_k	$\Sigma x_i/ik$
1	$x_{1,1,1}$	$x_{1,2,1}$	$x_{1,3,1}$	…	$x_{1,i,k}$	$\bar{X}_{1,k}$
2	$x_{2,1,1}$	$x_{2,2,1}$	$x_{2,3,1}$	…	$x_{2,i,k}$	$\bar{X}_{2,k}$
t	$x_{t,1,1}$	$x_{t,2,1}$	$x_{t,3,1}$	…	$x_{t,i,k}$	$\bar{X}_{t,k}$
	$x^*_{1,k}$	$x^*_{2,k}$	$x^*_{3,k}$	…	$x^*_{i,k}$	$\Sigma x^*_{i,k}/ik = X$

Table 3: Months and Sums of Occurrences Modeled

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan		
i	1	2	3	4	5	6	7	8	9	10	11	12	13	…	i_k
t	31	$28^1/_4$	31	30	31	30	31	31	30	31	30	31	31	…	t_i
t_i	31	$59^1/_4$	$90^1/_4$	$120^1/_4$	$151^1/_4$	$181^1/_4$	$212^1/_4$	$243^1/_4$	$273^1/_4$	$304^1/_4$	$334^1/_4$	$365^1/_4$	$396^1/_4$	…	t

Augmented Nonlinear Parametric Time Series Model (ANPTSM)

Trigonometric (sine and cosine) transformation augmented with polynomial of order two was applied to formulate the model across the year, that is, the monthly mean sample and the least square methods were used for estimating the model's parameters as follows. Let the equation be of the form

$$X_{t,i,k} = a_1 + a_2 t \mathrm{Sin}(t_{ik}) + a_3 t^2 \mathrm{Cos}(t_{ik}) + \varepsilon_i \quad 1 \le i \le 12 \tag{3.0}$$

The expected value of $X_{t,i,k}$ is $X^*_{i,k}$ then the equation can be reformed as below to estimate the parameters; a_1, a_2 and a_3 using Least Square Method.

$$X^*_{i,k} = a_1 + a_2 t_i \mathrm{Sin}(t_{ik}) + a_3 t_i^2 \mathrm{Cos}(t_{ik}) + \varepsilon_i \quad 1 \le i \le 12 \tag{3.1}$$

$$\therefore \varepsilon_i = X^*_{i,k} - \left(a_1 + a_2 t_i \mathrm{Sin}(t_{ik}) + a_3 t_i^2 \mathrm{Cos}(t_{ik})\right) \tag{3.2}$$

Let $\Sigma\varepsilon_i^2 = S$

$$S = \Sigma(X^*_{i,k} - \left(a_1 + a_2 t_i \mathrm{Sin}(t_{ik}) + a_3 t_i^2 \mathrm{Cos}(t_{ik})\right))^2 \tag{3.3}$$

Differentiating 3.3 with respect to a_1, a_2, a_3, a_3, as

$$\frac{\partial S}{\partial a_1} \to 0$$

results in

$$\Sigma X^*_{i,k} = ma_1 + a_2 \Sigma t_i \mathrm{Sin}(t_{ik}) + a_3 \Sigma t_i^2 \mathrm{Cos}(t_{ik}) \tag{3.4}$$

where m is the number of the monthly sample mean examined. Similarly, as

$$\frac{\partial S}{\partial a_2} \to 0$$

then

$$\Sigma t_i \mathrm{Sin}(t_{ik}) X^*_{i,k} = a_1 \Sigma t_i \mathrm{Sin}(t_{ik}) + a_2 \Sigma t_i^2 \mathrm{Sin}^2(t_{ik}) + a_3 \Sigma t_i^3 \mathrm{Sin}(t_{ik}) \mathrm{Cos}(t_{ik}) \tag{3.5}$$

and as

$$\frac{\partial S}{\partial a_3} \to 0$$

then

$$\Sigma t_i^2 \mathrm{Cos}(t_{ik}) X^*_{i,k} = a_1 \Sigma t_i^2 \mathrm{Cos}(t_{ik}) + a_2 \Sigma t_i^3 \mathrm{Sin}(t_{ik}) \mathrm{Cos}(t_{ik}) + a_3 \Sigma t_i^4 \mathrm{Cos}^2(t_{ik}) \tag{3.6}$$

Simultaneously solving equations 3.4, 3.5 and 3.6 using Cramer's Rule results in equations 3.7-3.10.

$$\Delta_0 = m\{\Sigma t_i^2 \mathrm{Sin}^2(t_{ik})\Sigma t_i^4 \mathrm{Cos}^2(t_{ik}) - (\Sigma t_i^3 \mathrm{Sin}(t_{ik})\mathrm{Cos}(t_{ik}))^2\}$$
$$-\Sigma t_i \mathrm{Sin}(t_{ik})\{\Sigma t_i \mathrm{Sin}(t_{ik})\Sigma t_i^4 \mathrm{Cos}^2(t_{ik}) - \Sigma t_i^2 \mathrm{Cos}(t_{ik})\Sigma t_i^3 \mathrm{Sin}(t_{ik})\mathrm{Cos}(t_{ik})\} \quad (3.7)$$
$$+\Sigma t_i^2 \mathrm{Cos}(t_{ik})\{\Sigma t_i \mathrm{Sin}(t_{ik})\Sigma t^3 \mathrm{Sin}(t_i)\mathrm{Cos}(t_{ik}) - \Sigma t_i^2 \mathrm{Cos}(t_i)\Sigma t_i^2 \mathrm{Sin}^2(t_{ik})\}$$

$$\Delta_1 = \Sigma X^*_{i,k}\{\Sigma t_i^2 \mathrm{Sin}^2(t_{ik})\Sigma t_i^4 \mathrm{Cos}^2(t_{ik}) - (\Sigma t_i^3 \mathrm{Sin}(t_{ik})\mathrm{Cos}(t_{ik}))^2\}$$
$$-\Sigma t_i \mathrm{Sin}(t_{ik})\{\Sigma t_i \mathrm{Sin}(t_{ik}) X^*_{i,k}\Sigma t^4 \mathrm{Cos}^2(t_{ik}) - \Sigma t_i^2 \mathrm{Cos}(t_{ik}) X^*_{i,k}\Sigma t_i^3 \mathrm{Sin}(t_{ik})\mathrm{Cos}(t_{ik})\} \quad (3.8)$$
$$+\Sigma t_i^2 \mathrm{Cos}(t_{ik})\{\Sigma t_i \mathrm{Sin}(t_{ik})\ X^*_{i,k}\Sigma t_i^3 \mathrm{Sin}(t_{ik})\mathrm{Cos}(t_{ik}) - \Sigma t_i^2 \mathrm{Cos}(t_{ik})\ X^*_{i,k}\Sigma t_i^2 \mathrm{Sin}^2(t_{ik})\}$$

$$\Delta_2 = m\{\Sigma t_i \mathrm{Sin}(t_{ik}) X^*_{i,k}\Sigma t_i^4 \mathrm{Cos}^2(t_{ik}) - (\Sigma t_i^2 \mathrm{Cos}(t_{ik}) X^*_{i,k}\Sigma t_i^3 \mathrm{Sin}(t_{ik})\mathrm{Cos}(t_{ik})\}$$
$$-\Sigma X^*_{i,k}\{\Sigma t_i \mathrm{Sin}(t_{ik})\Sigma t_i^4 \mathrm{Cos}^2(t_{ik}) - \Sigma t_i^2 \mathrm{Cos}(t_{ik})\Sigma t_i^3 \mathrm{Sin}(t_{ik})\mathrm{Cos}(t_{ik})\} \quad (3.9)$$
$$+\Sigma t_i^2 \mathrm{Cos}(t_{ik})\{\Sigma t_i \mathrm{Sin}(t_{ik})\Sigma t_i^2 \mathrm{Cos}(t_{ik}) X^*_{i,k} - \Sigma t_i^2 \mathrm{Cos}(t_{ik})\Sigma t_i \mathrm{Sin}(t_{ik}) X^*_{i,k}\}$$

$$\Delta_3 = m\{\Sigma t_i^2 \mathrm{Sin}^2(t_{ik})\Sigma t_i^2 \mathrm{Cos}(t_{ik}) X^*_{i,k} - (\Sigma t_i^3 \mathrm{Sin}(t_{ik})\mathrm{Cos}(t_i)\Sigma t_i \mathrm{Sin}(t_{ik}) X^*_{i,k}\}$$
$$-\Sigma t_i \mathrm{Sin}(t_{ik})\{\Sigma t_i \mathrm{Sin}(t_{ik})\Sigma t_i^2 \mathrm{Cos}(t_{ik}) X^*_{i,k} - \Sigma t_i^2 \mathrm{Cos}(t_{ik})\Sigma t_i \mathrm{Sin}(t_{ik}) X^*_{i,k}\} \quad (3.10)$$
$$+\Sigma X^*_{i,k}\{\Sigma t_i \mathrm{Sin}(t_{ik})\Sigma t_i^3 \mathrm{Sin}(t_{ik})\mathrm{Cos}(t_{ik}) - \Sigma t_i^2 \mathrm{Cos}(t_{ik})\Sigma t_i^2 \mathrm{Sin}^2(t_{ik})$$

Therefore, from equations 3.7, 3.8, 3.9 and 3.10, the following result:

$$a_1 = \frac{\Delta_1}{\Delta_0} \quad (3.11)$$

$$a_2 = \frac{\Delta_2}{\Delta_0} \quad (3.12)$$

$$a_3 = \frac{\Delta_3}{\Delta_0} \quad (3.13)$$

Next, substituting 3.11, 3.12 and 3.13 into 3.1 gives:

$$\hat{X}^*_{i,k} = \frac{\Delta_1}{\Delta_0} + \frac{\Delta_2}{\Delta_0} t\mathrm{Sin}(t_{ik}) + \frac{\Delta_3}{\Delta_0} t^2 \mathrm{Cos}(t_{ik}). \quad (3.14)$$

Because $X^*_{i,k}$ is the expected value of $X_{t,i,k}$, equation 3.14 can be rewritten as

$$\hat{X}_{t,i,k} = \frac{\Delta_1}{\Delta_0} + \frac{\Delta_2}{\Delta_0} t\mathrm{Sin}(t_{ik}) + \frac{\Delta_3}{\Delta_0} t^2 \mathrm{Cos}(t_{ik}). \quad (3.15)$$

Models 3.14 and 3.15 would only be visible provided there is an occurrence within a month of any sampled year.

Modified Nonlinear Trigonometric Transformation Time Series Model (MNTTTSM)

In a situation where a whole month of data is missing, the above model may be difficult to apply and a different model would be needed. The model for such occurrence is formulated as follows. If the data in 3.2 are reformed such that the monthly means are those shown in Table 4. Consider:

$$X^*_{i,k} = a + b\sin(t_i^*) + \varepsilon_i \quad (3.16)$$

where

$$1 \le i \le 12\ ,\ 1 \le t_i \le 365\frac{1}{4}$$

If the expected value of $X^*_{i,k}$ is X^{*m}_i, then equation 3.16 can take the form

$$X^{*m}_i = a + b\ \sin(t_i^*) + \varepsilon_i \qquad (3.17)$$

where

$$1 \le i \le 12\ ,\ 1 \le t_i^* \le 365\frac{1}{4}$$

An ordinary least square method was used in estimating the parameters a and b. If $S_m = \varepsilon_i^2 = \Sigma (X^{*m}_i - (a+b\sin(t_i^*))^2$, then differentiating with respect to a and b

$$\frac{\partial S_m}{\partial a} = -2\Sigma\left(X^{*m}_i - a + b\sin\left(t_i^*\right)\right)$$

as

$$\frac{\partial S_m}{\partial a} \to 0$$

$$\Rightarrow\ \delta\Sigma X^{*m}_i = 12a + b\Sigma\sin\left(t_i^*\right) \qquad (3.18)$$

Also,

$$\frac{\partial S_m}{\partial b} = -2\Sigma(\sin\left(t_i^*\right)\left(X^{*m}_i - \left(a + b\sin\left(t_i^*\right)\right)\right)$$

as

$$\frac{\partial S_m}{\partial b} \to 0$$

$$\Rightarrow\ \Sigma\sin\left(t_i*\right)\ X^{*m}_i = a\Sigma\sin\left(t_i^*\right) + b\Sigma\sin^2\left(t_i^*\right) \qquad (3.19)$$

Using Cramer's Rule to solve equations 3.18 and 3.19 simultaneously, results in:

$$\Delta_4 = 12\Sigma\mathrm{Sin}^2(t_i^*) - \left(\Sigma\mathrm{Sin}(t_i^*)\right)^2$$

$$\Delta_5 = \Sigma X^{*m}_i \Sigma\mathrm{Sin}^2(t_i^*) - \Sigma\mathrm{Sin}(t_i^*) X^{*m}_i \Sigma\mathrm{Sin}(t_i^*)$$

$$\Delta_6 = 12\Sigma\mathrm{Sin}(t_i^*) X^{*m}_i - \Sigma X^{*m}_i \Sigma\mathrm{Sin}(t_i^*)$$

where parameters

$$a = \frac{\Delta_5}{\Delta_4} = \frac{\Sigma X^{*m}_i \Sigma\mathrm{Sin}^2(t_i^*) - \Sigma\mathrm{Sin}(t_i^*) X^{*m}_i \Sigma\mathrm{Sin}(t_i^*)}{12\Sigma\mathrm{Sin}^2(t_i^*) - \left(\Sigma\mathrm{Sin}(t_i^*)\right)^2} \qquad (3.20)$$

and

$$b = \frac{\Delta_6}{\Delta_4} = \frac{12\Sigma\mathrm{Sin}(t_i^*) X^{*m}_i - \Sigma X^{*m}_i \Sigma\mathrm{Sin}(t_i^*)}{12\Sigma\mathrm{Sin}^2(t_i^*) - \left(\Sigma\mathrm{Sin}(t_i^*)\right)^2} \qquad (3.21)$$

Therefore, the model for monthly occurrence is

Table 4: Modified Nonlinear Trigonometric Transformation Time Series Model Data

t/i	1	2	3	4	5	6	7	8	9	10	11	12	$\Sigma x_i/12$
1	$X^*_{1,1}$	$X^*_{2,1}$	$X^*_{3,1}$	$X^*_{4,1}$	$X^*_{5,1}$	$X^*_{6,1}$	$X^*_{7,1}$	$X^*_{8,1}$	$X^*_{9,1}$	$X^*_{10,1}$	$X^*_{11,1}$	$X^*_{12,1}$	X^{*y}_1
2	$X^*_{1,2}$	$X^*_{2,2}$	$X^*_{3,2}$	$X^*_{4,2}$	$X^*_{5,2}$	$X^*_{6,2}$	$X^*_{7,2}$	$X^*_{8,2}$	$X^*_{9,2}$	$X^*_{10,2}$	$X^*_{11,2}$	$X^*_{12,2}$	X^{*y}_2
k	$X^*_{1,k}$	$X^*_{2,k}$	$X^*_{3,k}$	$X^*_{4,k}$	$X^*_{5,k}$	$X^*_{6,k}$	$X^*_{7,k}$	$X^*_{8,k}$	$X^*_{9,k}$	$X^*_{10,k}$	$X^*_{11,k}$	$X^*_{12,k}$	X^{*y}_k
	x^{*m}_1	x^{*m}_2	x^{*m}_3	x^{*m}_4	x^{*m}_5	x^{*m}_6	x^{*m}_7	x^{*m}_8	x^{*m}_9	x^{*m}_{10}	x^{*m}_{11}	x^{*m}	$\Sigma x^{*m}_{,i=} X_{*12}$

$$X^{*m}{}_i = \frac{\Delta_5}{\Delta_4} + \frac{\Delta_6}{\Delta_4} Sin(t_i^*) \qquad (3.22)$$

Because $X^{*m}{}_i$ is an expected value for $X^*{}_{i,k}$ then equation 3.22 can be rewritten as

$$X^*{}_{i,k} = \frac{\Delta_5}{\Delta_4} + \frac{\Delta_6}{\Delta_4} Sin(t_i^*) \qquad (3.23)$$

Similarly, along the sampled year $X^{*y}{}_K = c + d\ Sin(\lambda k)$ for $-\infty \leq k \leq +\infty$, $15 \leq \lambda \leq 75$. The λ must be chosen such that $\Sigma\ \varepsilon_i = 0$, $\Sigma\varepsilon_i^2$ is as small as possible.

If $S_y = \varepsilon_i^2 = \Sigma\ (X^{*y}{}_K - (c+d\sin(\lambda k))^2$ then

$$\frac{\partial S_y}{\partial c} = -2\Sigma(X^{*y}{}_K - (c + d\sin(\lambda k))$$

as

$$\frac{\partial S_y}{\partial c} \to 0$$

$$\Rightarrow\ \delta\Sigma X^{*y}{}_K = nc + d\Sigma\sin(\lambda k) \qquad (3.24)$$

Also,

$$\frac{\partial S_y}{\partial d} = -2\Sigma(\sin(\lambda k)\ X^{*y}{}_K - (c + d\sin(\lambda k)))$$

as

$$\frac{\partial S_y}{\partial d} \to 0$$

$$\Rightarrow\ \delta\Sigma\sin(\lambda k)\ X^{*y}{}_K = c\Sigma\sin(\lambda k) + d\Sigma\sin^2(\lambda k) \qquad (3.25)$$

Solving equations 3.24 and 3.25 simultaneously using Cramer's Rule results in

$$\Delta_7 = n\Sigma\mathrm{Sin}^2(\lambda k) - (\Sigma\mathrm{Sin}(\lambda k))^2$$
$$\Delta_8 = \Sigma X^{*y}{}_k \Sigma\mathrm{Sin}^2(\lambda k) - \Sigma\mathrm{Sin}(\lambda k) X^{*y}{}_k \Sigma\mathrm{Sin}(\lambda k)$$
$$\Delta_9 = n\Sigma\mathrm{Sin}(\lambda k) X^{*y}{}_k - \Sigma X^{*y}{}_k \Sigma\mathrm{Sin}(\lambda k)$$

Where the parameters

$$c = \frac{\Delta_8}{\Delta_7} = \frac{\Sigma X^{*y}{}_k \Sigma\mathrm{Sin}^2(\lambda k) - \Sigma\mathrm{Sin}(\lambda k) X^{*y}{}_k \Sigma\mathrm{Sin}(\lambda k)}{n\Sigma\mathrm{Sin}^2(\lambda k) - (\Sigma\mathrm{Sin}(\lambda k))^2} \qquad (3.26)$$

and

$$d = \frac{\Delta_9}{\Delta_7} = \frac{n\Sigma\mathrm{Sin}(\lambda k) X^{*y}{}_k - \Sigma X^{*y}{}_k \Sigma\mathrm{Sin}(\lambda k)}{n\Sigma\mathrm{Sin}^2(\lambda k) - (\Sigma\mathrm{Sin}(\lambda k))^2} \qquad (3.27)$$

$$\therefore X^{*y}{}_k = \frac{\Delta_8}{\Delta_7} + \frac{\Delta_9}{\Delta_7}\mathrm{Sin}(\lambda k) \qquad (3.28)$$

The method of placing expected occurrences in a contingency table of a Chi-square was applied using equations 3.23 and 3.28 to obtain the model to find the daily occurrences for a particular month of a particular year. Therefore, the model for expected daily occurrences is

$$X_{t,i,k} = \frac{n(X^{*m}{}_i)(X^{*y}{}_k)}{\Sigma X^{*y}{}_k} \qquad (3.29)$$

Substituting 3.23 and 3.28 into 3.29, results in

$$X_{t,i,k} = \frac{n\left(\frac{\Delta_5}{\Delta_4} + \frac{\Delta_6}{\Delta_4} Sin(t_i^*)\right)\left(\frac{\Delta_8}{\Delta_7} + \frac{\Delta_9}{\Delta_7} Sin(\lambda k)\right)}{\sum\left(\frac{\Delta_8}{\Delta_7} + \frac{\Delta_9}{\Delta_7} Sin(\lambda k)\right)} \qquad (3.30)$$

Results

Model Analysis and Discussion

The data on the daily mean temperature for Ikeja, Ibadan, Ilorin, Minna and Zaria collected from the Meteorological Centre-Oshodi Lagos were used. The parameters of the models were estimated and the fitted models for each zone are shown in Table 5 for Ikeja, Ibadan, Ilorin and Minna for ANPTSM. Data for the daily mean temperature was used to estimate the parameters. The fitted model for Zaria could

Table 5: The Fitted Models for ANPTSM

Zones	Augmented Nonlinear Parametric Time Series Model (ANPTSM)
IKEJA	$26.88642582 + 0.047971536\mathrm{tSin}(\mathrm{t_{ik}}) - 0.000143793\mathrm{t^2Cos}(\mathrm{t_{ik}})$
IBADAN	$26.36612286 + 0.054847742\mathrm{tSin}(\mathrm{t_{ik}}) - 0.0000344912\mathrm{t^2Cos}\mathrm{t_{ik}}$
ILORIN	$26.2476883 + 0.048115874\mathrm{tSin}(\mathrm{t_{ik}}) - 0.000833551t^2 Cos(t_{ik})$
MINNA	$25.72428 + 0.062853tSin(t_{ik}) - 0.00073t^2 Cos(t_{ik})$
ZARIA	–

Table 6: The Fitted Models for MNTTTSM

Zones	Modified Nonlinear Trigonometric Transformation Time Series Model (MNTTTSM)
IKEJA	$\frac{10(26.87226 + 1.420072Sin\,t_i^*)(26.88996 + 0.13116Sin60k)}{\sum_{K=1}^{10}(26.88996 + 0.1311Sin60k)}$
IBADAN	$\frac{10(26.36749 + 1.591834Sin\,t_i^*)(26.36761 + 0.13535Sin90k)}{\sum_{K=1}^{10}(26.36761 + 0.1311Sin90k)}$
ILORIN	$\frac{10(26.45708 + 1.816182Sin\,t_i^*)(26.40106 + 0.409024Sin45k)}{\sum_{K=1}^{10}(26.40106 + 0.409024Sin45k)}$
MINNA	$\frac{10(27.56143 + 2.508736Sin\,t_i^*)(27.67047 + 0.112148Sin90k)}{\sum_{K=1}^{10}(27.67047 + 0.112148Sin90k)}$
ZARIA	$\frac{10(24.98532 + 1.210108Sin\,t_i^*)(25.00445 + 0.222282Sin90k)}{\sum_{K=1}^{10}(25.00445 + 0.222282Sin90k)}$

not be formulated due to the fact that many months of data were missing.

Table 6 shows the fitted models for Ikeja, Ibadan,Ilorin, Minna and Zaria for MNTTTSM using the daily mean temperature data to estimate their parameters. The fitted model for Zaria was formulated because MNTTTSM has the strength of addressing the problem of missing values. Thus, although many months' data were missing from Zaria's daily mean temperature, MNTTTSM parameters could still be estimated. This is one of the advantages of MNTTTSM over ANPTSM.

Table 7 shows that the results of the Pearson Product Moment Correlation coefficients and Spearman Brown's rank Order Correlation coefficients for Ikeja, Ibadan, Ilorin and Minna are highly and positively correlated, indicating a strong relationship between the actual data and estimated data for the daily mean temperature. In Zaria the correlation coefficient for MNTTTSM is positive but low which may indicate a weak relationship between the actual and estimated daily mean temperatures.

Apart from Ibadan, in which the correlation coefficient in ANPTSM is greater than MNTTTSM and Ikeja which has equal correlation coefficients, all other Zones, the correlation coefficient in MNTTTSM is greater than ANPTSM. This indicates that MNTTTSM shows a stronger relationship between the actual and estimated values than does ANPTSM. Although the relationship between actual and estimated values of MNTTTSM in Zaria is weak but positive, that of ANPTSM could not be estimated due to the large number of missing values in the data. Also, all of the correlations are significant at the 0.01 level (2-tailed).

As shown in Table 8, the mean of the actual and estimated values for each zones of all models are almost equal; differences are due to approximation (truncation error) during calculation. Also, the mean of the actual and estimated values of MNTTTSM are closer than those of ANPTSM, which implies that MNTTTSM estimates better than ANPTSM. It was also discovered from results in Table 8 that the more missing values in the data, the weaker the ANPTSM is in estimating, while in MNTTTSM, the model maintains its precision.

Table 7: Correlation Coefficients

Zones	Types	ANPTSM		MNTTTSM	
		Coefficients	Sig.	Coefficients	Sig.
IKEJA	Pearson's r	0.607	.000	0.607	.000
	Spearman's Rho	0.620	.000	0.620	.000
IBADAN	Pearson's r	0.594	.000	0.575	.000
	Spearman's Rho	0.622	.000	0.584	.000
ILORIN	Pearson's r	0.503	.000	0.589	.000
	Spearman's Rho	0.560	.000	0.612	.000
MINNA	Pearson's r	0.596	.000	0.676	.000
	Spearman's Rho	0.656	.000	0.686	.000
ZARIA	Pearson's r	-	-	0.419	.000
	Spearman's Rho	-	-	0.445	.000

Table 8: Comparison of ANPTSM and MNTTSM Means

Zones	N	ANPTSM		MNTTTSM	
		Actual	Estimated	Actual	Estimated
IKEJA	3,660	26.9077	26.8759	26.9077	26.8759
IBADAN	3,601	26.3749	26.3756	26.3749	26.3791
ILORIN	3,580	26.4558	26.2443	26.4558	26.4593
MINNA	3,362	27.5489	26.3611	27.5489	27.5559
ZARIA	3,588	-	-	25.0514	25.0172

Table 9 shows that the standard deviations for MNTTTSM are less than those of ANPTSM which indicates that MNTTTSM is better in estimating and forecasting than ANPTSM. Similarly, apart from the standard error of ANPTSM and MNTTTSM of Ikeja, which are equal, it may be observed that the standard errors for MNTTTSM were also smaller than those of ANPTSM, which indicates that MNTTTSM is better in estimating and forecasting than ANPTSM for time series data with missing values.

Table 10 shows that at Ikeja, there is a 95% chance that the differences between the actual and estimated daily mean temperature would lie between -0.00749 and 0.07119 in ANPTSM and -0.00748 and 0.07118 in MNTTTSM. Similarly, at Ibadan; -0.0462 and 0.04473 in ANPTSM and -0.0496 and 0.04114 in MNTTTSM, at Ilorin; 0.1505 and 0.2723 in ANPTSM and -0.0592 and 0.05218 in MNTTTSM, at Minna; 1.1155 and 1.2601 in model I and -0.0689 and 0.05482 in MNTTTSM while in Zaria is between -0.0546 and 0.12310.

It was also discovered that the range of the confidence interval for MNTTTSM is less than that of ANPTSM for Ikeja and Ibadan. In Ilorin and Minna, the lower confidence intervals of differences for ANPTSM are positive which indicates a 95% chance that the differences between their actual and estimated daily temperature (actual – estimate) are positive while those of MNTTTSM are not. This implies that the estimated daily temperatures for ANPTSM at Ilorin and Minna were under-estimated. Hence MNTTTSM is better in estimating and forecasting than ANPTSM when there are missing values in the time series.

Table 9: Comparison of ANPTSM and MNTTSM'S Standard Deviation and Standard Error of Differences

Zones	ANPTSM		MNTTTSM	
	Std. Dev.	Std. Error of the Mean	Std. Dev.	Std. Error of the Mean
IKEJA	1.2138	0.02006	1.2137	0.02006
IBADAN	1.3913	0.02319	1.3882	0.02313
ILORIN	1.8585	0.03106	1.6996	0.02841
MINNA	2.1381	0.03688	1.8293	0.03155
ZARIA	-	-	2.7152	0.04533

Table 10: Comparison of ANPTSM and MNTTTSM's 95 % Confidence Interval of the Difference

Zones	ANPTSM		MNTTTSM	
	Lower	Upper	Lower	Upper
IKEJA	-0.00749	0.07119	-0.00748	0.07118
IBADAN	-0.0462	0.04473	-0.0496	0.04114
ILORIN	0.1505	0.2723	-0.0592	0.05218
MINNA	1.1155	1.2601	-0.0689	0.05482
ZARIA	-	-	-0.0546	0.12310

Conclusion

The two models tested in this study were the Augmented Nonlinear Parametric Time Series Model (ANPTSM) and the Modified Nonlinear Trigonometric Transformation Time Series Model (MNTTTSM). Both models were tested using daily mean temperatures at Ikeja, Ibadan, Ilorin, Minna and Zaria, and the results were analyzed. It was discovered that ANPTSM could be used in forecasting provided the data is having few missing values. However MNTTTSM estimates forecasts better than ANPTSM in estimating missing values and forecasting. Based on results of this study, MNTTTSM is more efficient in estimating missing values and forecasts better than ANPTSM.

The beauty of a good model developed for nonlinear time series modeling is the ability to forecast better, the new method MNTTTTSM is therefore recommended for numerical solutions for a nonlinear model with missing values due to its higher capacity to address missing values. It was also noted that the mathematical derivative of MNTTTSM is simpler than ANPTSM which did not forecast better. Further research could be conducted by placing a condition in which data having a year or more of missing values is taken into consideration.

References

Barnett, W. A., Powell, J., & Tauchen, G. E. (1991). *Nonparametric and semi parametric methods in econometrics and statistics*. Proceedings of the 5th International Symposium in Economic Theory and Econometrics. Cambridge: Cambridge University Press.

Bates, D. M., & Watts, D. G. (1988). *Nonlinear regression analysis and its applications*. New York: Wiley.

Box, G. E. P., Jenkins, G. M. (1970). *Time series analysis, forecasting and control*. San Francisco: Holden-Day.

Brock, W. A., Potter, S. M. (1993). Nonlinear time series and macroeconometrics. In: G.S. Maddala, C. R. Rao & H. R. Vinod, Eds., *Handbook of Statistics*, *Vol. 11*, 195-229. Amsterdam: North-Holland.

De Gooijer, J. G., & Kumar, K. (1992). Some recent developments in non-linear time series modeling, testing and forecasting. *International Journal of Forecasting*, *8*, 135-156.

Friedman, J. H., & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, *76*, 817-823.

Gallant, A. R. (1987). *Nonlinear statistical models*. New York: Wiley.

Gallant, A. R., & Tauchen, G. (1989). Semi nonparametric estimation of conditionally constrained heterogeneous processes: asset pricing application. *Econometrica*, *57*, 1091-1120.

Granger, C. W. J., & Hallman, J. J. (1991a). Nonlinear transformations integrated time series. *Journal of Time Series Analysis*, *12*, 207-224.

Olowofeso, O. E. (2006). *Varying co-efficient regression models for non-linear time series: Estimation and testing*.13th Meeting of the Forum for Interdisciplinary Mathematical and Statistical Techniques.

Priestley, M. (1988). *Non-linear and non-stationary time series analysis*. London: Academic Press.

Robinson, P. M. (1983). Non-parametric estimation for time series models. *Journal of Time Series Analysis*, *4*, 185-208.

Sugihara, G., & May, R. M. (1990). Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, *344*, 734-741.

Terasvirta, T., & Anderson, H. M. (1992). Modelling nonlinearities in business cycles using smooth transition autoregressive models. *Journal of Applied Econometrics*, *7*, S119-S136.

Tsay, R. S. (1986). Nonlinearity tests for time series. *Biometrika*, *73*, 461-466.

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 480-487

1538 – 9472/10/$95.00

Incidence and Prevalence for A Triply Censored Data

Hilmi F. Kittani
The Hashemite University,
Jordan

The model introduced for the natural history of a progressive disease has four disease states which are expressed as a joint distribution of three survival random variables. Covariates are included in the model using Cox's proportional hazards model with necessary assumptions needed. Effects of the covariates are estimated and tested. Formulas for incidence in the preclinical, clinical and death states are obtained, and prevalence formulas are obtained for the preclinical and clinical states. Estimates of the sojourn times in the preclinical and clinical states are obtained.

Key words: Progressive disease model, prevalence, incidence, trivariate hazard function, censored data, proportional hazards model, sojourn times, chronic habitué.

Introduction

Louis, et al. (1978) introduced a natural history model for a progressive disease in a set of three articles: Albert, Gertman and Louis (1978), Albert, Gertman, Louis and Liu(1978) and Louis, Albert and Heghinian (1978). This model was extended by Kittani (1995a). Clayton (1978) also developed a model for association for the bivariate case and Oakes (1982) made inferences about the association parameter in Clayton's model. Clayton and Cuzick (1985) introduced the bivariate survival function for two failure times and made inferences about the association parameter, γ. Kittani (1995a, 1996, 1997, 1997-1998) considered the model for the bivariate case – that is, a case with two failure times (X, T) – by including covariates and by using Cox's proportional hazards model.

The motivation for this research lies in the fact that it is necessary to identify a three dimensional survival function for three failure times (X, Y, D) with four disease states (disease free state, preclinical state, clinical state and death state). In the model, X is the age upon entering the preclinical state (tumor onset or first heart attack), T is the age when entering the clinical state (symptoms first appear or second heart attack) and D is the age upon entering the death state (dying of cancer or acute myocardial infarction). Kittani (2010) considered estimating the parameters using nonparametric approach for a triply censored data.

Hilmi Kittani is a Professor of Statistics in the College of Science, Department of Mathematics. Email: kittanih@hu.edu.jo.

Background and Assumptions

As in the Louis, et al. (1978) model, it is assumed that $f_{XYZ}(x,y,z,a)$ is continuous – that is, $X = Y = Z = \infty$ is not allowed – and Y and Z are termed the sojourn times in the preclinical and clinical states respectively. The model proposed by Louis, et al. (1978) makes the assumption of no cohort effect, meaning that the distribution of the random variables (X, Y, Z) is independent of the age distribution A, or

$$f_{XYZA}(x, y, z, a) = f_{XYZ}(x, y, z) \times f_A(a)$$

and

$$f_{XYA}(x, y, a) = f_{XY}(x, y) \times f_A(a)$$

where $f_{XYZ}(x, y, z)$ is the joint pdf of (X, Y, Z) , $f_{XY}(x,y)$ is the joint pdf of X, Y and $f_A(a)$ is the pdf of A(the age distribution of the subject population). In addition, a subject is a chronic habitué of the PCS if, for that subject, $X < \infty$, $Y = \infty$, for example, subject never leaves PCS. According to the model, there will be no chronic habitués of the PCS or CS because, if a subject

lives long enough, then he/she will progress to the next state eventually (Louis et al., 1978).

The X, Y and Z axes are partitioned into I, J and K intervals according to Chiang, et al. (1989) and Hollford (1976); they assumed constant baseline hazards in each subinterval, $\lambda_{1i}(x) = \mu_{1i}$, $x \in I_i$, $\lambda_{2j}(y) = \mu_{2j}$, $y \in I_j$ and $\lambda_{3k}(z) = \mu_{3k}$, $z \in I_k$ in the i^{th}, j^{th} and k^{th} intervals respectively. The hazard functions for the n^{th} individual whose (X, Y, Z) values fall in the cube $I_i x I_j x I_k$ are modeled by assuming Cox's (1972) proportional hazards model and holds for each X, Y and Z in each respective I_i, I_j and I_k interval.

Assuming $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ (regression parameters) for the covariate $\boldsymbol{\omega}$ (p-dimensional) are constant (the same) for all intervals to be estimated**.** The hazard functions λ_1, λ_2 and λ_3 in the I^{th}, J^{th} and K^{th} intervals for n^{th} individual whose observed (X, Y, Z) value is (X_n, Y_n, Z_n) will be defined as

$$\lambda_{1i}(x_n) = \mu_{1i} e^{\alpha' \omega_n}, x_n \in I_i = (a_i, a_{i+1}], \lambda_{2j}(y_n) = \mu_{2j} e^{\beta' \omega_n},$$

and

$$y_n \in I_j = (b_j, b_{j+1}], \lambda_{3k}(z_n) = \mu_{3k} e^{\eta' \omega_n}, z_n \in I_k = (c_k, c_{k+1}].$$

Where μ_{1i}, μ_{2j} and μ_{3k} are baseline hazard functions associated with X, Y and Z respectively. Assuming α, β and η are constant (the same) regression parameters for the covariate $\boldsymbol{\omega}$ for all intervals and to be estimated along with the association parameter γ.

The joint survival function for the three non-negative random variables (X, Y, Z) given by Kittani (1995b) is:

$$F(x, y, z) = [e^{\gamma \Lambda_1(x)} + e^{\gamma \Lambda_2(y)} + e^{\gamma \Lambda_3(z)} - 2]^{-\frac{1}{\gamma}}. \quad (2.1)$$

Where $\gamma > 0$, $x > 0$, $y > 0$, $z > 0$, and Λ_1, Λ_2, Λ_3 are the cumulative hazard functions associated with X, Y and Z respectively. For example, to compute $\Lambda_{1i}(x)$, which is the cumulative hazard function for the n^{th} individual whose x value falls in the i^{th} interval (assuming a constant hazard over each interval) is as follows:

$$\Lambda_{1i}(x_n) = \int_0^{x_k} \lambda_1(u)\,du = \left[\sum_{r=1}^{i-1} \mu_{1r}(a_{r+1} - a_r) + \mu_{1i}(x_n - a_i) \right] e^{\alpha' \omega} \quad (2.2)$$

where $\Lambda_{2j}(y_n)$ and $\Lambda_{3k}(z_n)$ are defined in a similar way. Thus, the joint density function (X, Y, Z) is

$$f(x,y,z) = \left[(\gamma+1)(\gamma+2)\lambda_1(x)\lambda_2(y) \; \lambda_3(z)\, e^{\gamma[\Lambda_1(x)+\Lambda_2(y)+\Lambda_3(z)]}\, U^{(-\frac{1}{\gamma}-3)} \right] \quad (2.3)$$

where $\gamma > 0$, $x > 0$, $y > 0$, $z > 0$, and λ_1, λ_2 and λ_3 are base line hazard functions associated with X, Y and Z respectively as

$$U = e^{\gamma \Lambda_1(x)} + e^{\gamma \Lambda_2(y)} + e^{\gamma \Lambda_3(z)} - 2.$$

Kittani (1996) derived the likelihood function for the uncensored and censored cases in order to estimate the regression parameters by maximizing the likelihood function, that is, the n^{th} individual that generates data vector $\underline{\mathbf{w}}_n$, and $L(\underline{\mathbf{w}}_n)$ is the likelihood function contribution for the n^{th} individual as:

$$L(w_1, w_2, \ldots, w_N) = \prod_{n=1}^{N} L(w_n).$$

This likelihood function is maximized with respect to the unknown parameter vector; $\underline{\theta} = (\gamma, \underline{\alpha}, \underline{\beta}, \underline{\eta}, \underline{\mu}_1, \underline{\mu}_2, \underline{\mu}_3)$ with dimension (3p + I + J+ K+1) where p is the number of covariates.

To apply the Kittani (1995b) formula in the likelihood function it is first modeled for (X, Y, Z), then the transformations X = X, T = X + Y and D = X + Y + Z are performed to obtain the joint density function g(x, t, d) of (X, T, D) as:

$$g(x,t,d) = f(x,t-x,d-t) = (\gamma+1)(\gamma+2)\lambda_1(x)\lambda_2(t-x)\lambda_3(d-t).$$

$$e^{\gamma\left[\begin{matrix}\Lambda_1(x)+\Lambda_2(t-x)\\ +\Lambda_3(d-t)\end{matrix}\right]}\left[\begin{matrix}[e^{\gamma\Lambda_1(x)}+e^{\gamma\Lambda_2(t-x)}\\ +e^{\gamma\Lambda_3(d-t)}-2]^{(-\frac{1}{\gamma}-3)}\end{matrix}\right] \tag{2.4}$$

Preclinical, Clinical and Death Incidence

According to Louis, et al. (1978) under the assumption of no cohort effect, that is, (X, Y, Z) is independent of A, assuming $I_{PC}(a) = f_X(a)$, then preclinical incidence among those aged A is defined in terms of this model as:

$$I_{PC}(a) = f_X(a) = \sum_{i=1}^{I} \mu_{1i}\, e^{\alpha'\omega} e^{-\Lambda_{1i}(a)}, \quad a \in I_i \tag{3.1}$$

where f_X is the marginal density of X. In order to define the overall preclinical incidence, I_{PC} in terms of this model, the distribution of A must be defined. It is assumed throughout this article that if A is uniformly distributed over an interval I as

$$f_A(a) = 1/Id_i, \quad a \in I_i \tag{3.2}$$

where I is the number of intervals on the x-axis and d_i is the length of interval i, then the overall preclinical incidence in terms of this model is

$$I_{PC} = \int_0^\infty f_X(x).f_A(x)\,dx = \sum_{i=1}^{I} \frac{1}{I d_i} \int_{I_i} \mu_{1i}.e^{\alpha'\omega}.e^{-\Lambda_{1i}(x)}\,dx = \sum_{i=1}^{I} \frac{1}{I d_i}\left(e^{-\Lambda_{1i}(a_i)} - e^{-\Lambda_{1i}(a_{i+1})}\right). \tag{3.3}$$

Similarly, if there is no cohort effect, then the clinical incidence among those aged A is defined in terms of this model as

$$I_{CL}(a) = f_T(a) = \int_0^a f(x, a-x)dx = \int_0^a (\gamma+1)\mu_{1i}\mu_{2j}\, e^{(\alpha'+\beta')\omega} e^{\gamma[\Lambda_{1i}(x)+\Lambda_{2j}(a-x)]} U^{(-\frac{1}{\gamma}-2)}\,dx$$

Where $U = e^{\gamma\Lambda_{1i}(x)} + e^{\gamma\Lambda_{2j}(a-x)} - 1$ and f_T is the marginal density of T = X + Y. This integral cannot be obtained in a closed form and should be evaluated numerically. The overall clinical incidence in terms of this model is

$$I_{CL} = \int_0^\infty f_T(t)\, f_A(t)dt = \int_0^\infty \left(\int_0^t f(x,t-x)dx\right) f_A(t)\,dt = \sum_{j=1}^{J}\left[\frac{1}{Jd_j}\int_{J_j}\left(\int_0^t f(x,t-x)dx\right)dt\right] \tag{3.5}$$

where J is the number of intervals on the y-axis, d_j is the length of interval j and

$$f(x,t-x) = \left[\begin{array}{l} (\gamma+1)\,\mu_{1i}\,\mu_{2j}\,e^{(\alpha'+\beta')}\omega\, e^{\gamma(\Lambda_{1i}(x)+\Lambda_{2j}(a-x))} \\ \left(e^{\gamma(\Lambda_{1i}(x)} + e^{\gamma\Lambda_{2j}(a-x)} - 1 \right)^{(-\frac{1}{\gamma}-2)} \end{array} \right]$$

The above integral cannot be obtained in a closed form and should be evaluated numerically. Equations (3.1) – (3.5) are similar to those given by Kittani (1997).

Similarly, if there is no cohort effect, then death incidence among those aged A is defined in terms of this model as

$$I_{DI}(a) = f_D(a) = \int_0^a \left(\int_0^t f(x,\ t-x,\ a-t)\,dx \right) dt \tag{3.6}$$

This integral cannot be obtained in closed form and should be evaluated numerically. The overall death incidence in terms of this model is

$$\begin{aligned} I_{DI} &= \int_0^\infty f_D(t)\, f_A(t)dt \\ &= \int_0^\infty \left(\int_0^a \left(\int_0^t f(x,t-x,a-t)\,dx \right) f_A(t)\,dt \right) da \\ &= \sum_{k=1}^{K} \left[\frac{1}{Kd_k} \int_{I_k} \int_0^a \left(\int_0^t f(x,t-x,a-t)\,dx \right) dt \right] \end{aligned} \tag{3.7}$$

where K is the number of intervals on the z-axis and d_k is the length of interval.

Preclinical and Clinical Prevalence

According to Louis, et al. (1978) under the assumption of no cohort effect, (X, Y, Z) is independent of A, and the assumption of no chronic habitués of the PCS, then preclinical prevalence among those aged a is

$$\Phi_{PC}(a) = \int_0^a \left(\int_{a-x}^\infty f_{XY}(x,y)dy \right) dx, \tag{4.1}$$

then preclinical prevalence among those aged a is defined in terms of this model as

$$\begin{aligned} &\Phi_{PC}(a) = \\ &\int_0^a \mu_{1i}\, e^{\alpha'}\omega e^{\gamma\Lambda_{1i}(x)} \\ &\left(e^{\gamma\Lambda_{1i}(x)} + e^{\gamma\Lambda_{2j}(a-x)} - 1 \right)^{(-\frac{1}{\gamma}-1)} dx. \end{aligned} \tag{4.2}$$

The integral cannot be obtained in a closed form and should be evaluated numerically.

The overall preclinical prevalence according to Louis, et al. (1978) is

$$\Phi_{PC} = \int_0^\infty \Phi_{PC}(a) f_A(a)\,da \tag{4.3}$$

and, under the assumption of no cohort effect and no chronic habitué s of the PCS, the overall preclinical prevalence is defined in terms of this model as

$$\begin{aligned} \Phi_{PC} &= \int_0^\infty \Phi_{PC}(a) f_A(a)\,da \\ &= \sum_{i=1}^{M} \left(\frac{1}{Md_i} \int_{J_i} \Phi_{PC}(a)da \right) \\ &= \sum_{i=1}^{M} \left[\frac{1}{Md_i} \right. \\ &\left. \int_{J_i} \left(\int_0^a \mu_{1i}\, e^{\alpha'}\omega e^{\gamma\Lambda_{1i}(x)} \left(e^{\gamma\Lambda_{1i}(x)} + e^{\gamma\Lambda_{2j}(a-x)} - 1 \right)^{(-\frac{1}{\gamma}-1)} dx \right) da \right] \end{aligned} \tag{4.4}$$

The integral cannot be obtained in a closed form and should be evaluated numerically. Thus, the

clinical prevalence among those aged a in terms of this model is

$$\Phi_{CS}(a)=\int_0^a\int_{a-x}^{\infty}\int_{d-a}^{\infty} f_X(x,y,z)\,dz\,dy\,dx \tag{4.5}$$

where f(x, y, z) is given by

$$f(x,y,z)$$

$$=(\gamma+1)(\gamma+2)\lambda_1(x)\lambda_2(y)\lambda_3(z)e^{\gamma[\Lambda_1(x)+\Lambda_2(y)+\Lambda_3(z)]}U^{(-\frac{1}{\gamma}-3)} \tag{4.6}$$

where

$$U=[e^{\gamma\Lambda_{1i}(x)}+e^{\gamma\Lambda_{2j}(y)}+e^{\gamma\Lambda_{3k}(z)}-2].$$

Therefore the overall clinical prevalence in terms of this model is

$$\begin{aligned}\Phi_{CS}&=\int_0^{\infty}\Phi_{CS}(a)f_A(a)\,da\\&=\int_0^{\infty}\int_0^{a}\int_{a-x}^{\infty}\int_{d-a}^{\infty} f_{XYZ}(x,y,z)\,dz\,dy\,dx\\&=\sum_{k=1}^{K}\left[\frac{1}{Kd_k}\int_{I_k}\left(\int_0^{a}\int_{a-x}^{\infty}\int_{d-a}^{\infty} f_{XYZ}(x,y,z)\,dz\,dy\,dx\,da\right)\right]\end{aligned} \tag{4.7}$$

where f(x,y,z) is given by equation (4.6); the integral cannot be obtained in a closed form and should be evaluated numerically.

Estimation of the Sojourn Times in the Preclinical and Clinical States

Louis, et al. (1978) defined the mean duration of a disease in the preclinical state as

$$E(Y|X<\infty)=\frac{\int_0^{\infty}\Phi_{PC}(a)da}{\int_0^{\infty}I_{PC}(a)da}. \tag{5.1}$$

However, according to this model, no cohort effect and no chronic habitué s of the PCS are assumed, thus, the quantity E[Y|X < ∞] will be E(Y) because P[X < ∞] is 1. Therefore, substituting for $I_{PC}(a)$ and $\Phi_{PC}(a)$ in the above formula results in

$$E(Y)=\frac{\sum_{j=1}^{N}\int_{I_j}\left(\int_0^{a}\mu_{1i}e^{\alpha'\omega}e^{\gamma\Lambda_{1i}(x)}U^{(-\frac{1}{\gamma}-1)}\right)da}{\sum_{i=1}^{M}\left(e^{\Lambda_{1i}(a_i)}-e^{\Lambda_{1i}(a_{i+1})}\right)} \tag{5.2}$$

This integral cannot be obtained in a closed form and should be evaluated numerically.

Defining the mean duration of the disease in the clinical state as

$$E(Z|Y<\infty)=\frac{\int_0^{\infty}\Phi_{CS}(a)da}{\int_0^{\infty}I_{CS}(a)da} \tag{5.3}$$

and, assuming no cohort effect and no chronic habitué s of the CS, the quantity E[Z| Y < ∞] will be E(Z), because P[Y < ∞] is 1. Thus, substituting for $I_{CS}(a)$ and $\Phi_{CS}(a)$ in the above formula results in

$$E(Z)=\frac{\int_0^{\infty}\left(\int_0^{a}\left[\int_{a-x}^{\infty}\left\{\int_{d-a}^{\infty}f_{XYZ}(x,y,z)\,dz\right\}dy\right]dx\right)da}{\int_0^{\infty}\left(\int_0^{a}(\gamma+1)\mu_{1i}\mu_{2j}e^{(\alpha'+\beta')\omega}e^{\gamma(\Lambda_{1i}(x)+\Lambda_{2j}(a-x))}U^{(-\frac{1}{\gamma}-2)}dx\right)da} \tag{5.4}$$

where f(x, y, z) is given by equation (4.6). The integrals cannot be obtained in a closed form and should be evaluated numerically.

Asymptotic Distributions of the Epidemiological Measures

In order to make inferences about the epidemiological measures obtained, it is necessary to find their distributions; the Delta Method (Bishop, et al., 1975) is applied to determine means and variances. The parameter vector to be estimated is $\underline{\boldsymbol{\theta}} = (\gamma, \underline{\boldsymbol{\alpha}}, \underline{\boldsymbol{\beta}}, \underline{\boldsymbol{\eta}}, \underline{\boldsymbol{\mu}}_1, \underline{\boldsymbol{\mu}}_2, \underline{\boldsymbol{\mu}}_3)$ with dimension (3p + M+ N+ K+1) where p is the number of covariates, dim($\underline{\boldsymbol{\mu}}_1$) = **M,** dim($\underline{\boldsymbol{\mu}}_2$) = **N,** dim($\underline{\boldsymbol{\mu}}_3$) = K and dim($\underline{\boldsymbol{\alpha}}$) = dim($\underline{\boldsymbol{\beta}}$) = dim($\underline{\boldsymbol{\eta}}$) = p. Because $\underline{\hat{\theta}}$ is the MLE for $\underline{\theta}$ and, from the properties of the MLE's, $\underline{\hat{\theta}}$ is approximately normal with mean $\underline{\theta}$ and the covariance matrix $I^{-1}[\ \underline{\theta}\ ]$ is the inverted covariance matrix of $\underline{\theta}$ obtained from maximizing the log likelihood function for the censored case.

If $g(\underline{\hat{\theta}})$ is any function of $\underline{\hat{\theta}}$, the approximate distribution of $g(\underline{\hat{\theta}})$ may be found by applying the Delta Method as

$$g(\hat{\theta}) \approx N\left(g(\theta), [\frac{\partial g(\theta)}{\partial \theta}]\,[I(\theta)]^{-1}\,[\frac{\partial g(\theta)}{\partial \theta}]' \right). \tag{6.1}$$

where

$$\frac{\partial g(\theta)}{\partial \theta} = \left(\frac{\partial g(\theta)}{\partial \theta_1}, \frac{\partial g(\theta)}{\partial \theta_2}, \ldots, \frac{\partial g(\theta)}{\partial \theta_\upsilon} \right),$$

$$dim(\theta) = \upsilon = 3p + M + N + K + 1$$

and the estimated variance of $g(\underline{\hat{\theta}})$ is

$$\left([\frac{\partial g(\theta)}{\partial \theta}][I(\theta)]^{-1}\,[\frac{\partial g(\theta)}{\partial \theta}]' \right)_{\theta=\hat{\theta}}.$$

As an example, the formulas for the derivatives of the preclinical incidence are derived as follows. The estimate of preclinical incidence among those aged a

$$g_\theta(a) = I_{PC}(a) = \sum_{i=1}^{M} \mu_{li} e^{\alpha'\omega} e^{\Lambda_{li}(a)}, \quad a \in I_i \tag{6.2}$$

$$g_{\hat{\theta}}(a) = \hat{I}_{PC}(a) = \sum_{i=1}^{M} \hat{\mu}_{li} e^{\hat{\alpha}'\omega} e^{\hat{\Lambda}_{li}(a)}, \quad a \in I_i \tag{6.3}$$

where

$$\hat{\Lambda}_{li}(a) = \left[\sum_{r=1}^{i-1} \hat{\mu}_{lr}(a_{r+1} - a_r) + \hat{\mu}_{li}(a - a_i) \right] e^{\hat{\alpha}'\omega}$$

Differentiating g with respect to $\underline{\boldsymbol{\theta}}$, results in

$$\frac{\partial g(\theta)}{\partial \gamma} = \frac{\partial g(\theta)}{\partial \beta_m} = \frac{\partial g(\theta)}{\partial \mu_{2r}} = 0, \tag{6.4}$$

and

$$\frac{\partial g}{\partial \alpha_m} = \sum_{i=1}^{M} \mu_{li}\, z\, e^{\alpha'\omega} e^{-\Lambda_{li}(a)} (1 - \Lambda_{li}(a)), \quad a \in I_i \tag{6.5}$$

$$\frac{\partial g}{\partial \mu_{ls}} = \begin{cases} \sum_{s=1}^{M} e^{\alpha'\omega} e^{-\Lambda_{ls}(a)} [1 - \mu_{ls}(a - a_s) e^{\alpha'\omega}], & s = i, a \in I_i \\ -\sum_{i=1}^{M} \mu_{li}\, e^{2\alpha'\omega} e^{-\Lambda_{li}(a)} (a_{s+1} - a_s), & s < i,\ a \in I_i \\ 0, & s > i,\ a \in I_i \end{cases} \tag{6.6}$$

To test the effect of the covariates on morbidity and mortality (getting into the PCS, CS and DS):

$$H_0 : \underline{\boldsymbol{\alpha}} = 0 \text{ vs. } H_1 : \underline{\boldsymbol{\alpha}} \neq 0,$$

$$H_0 : \underline{\boldsymbol{\beta}} = 0 \text{ vs. } H_1 : \underline{\boldsymbol{\beta}} \neq 0,$$

and

$$H_0 : \underline{\boldsymbol{\eta}} = 0 \text{ vs. } H_1 : \underline{\boldsymbol{\eta}} \neq 0. \tag{6.7}$$

then the standard errors of the estimates are obtained from $I^{-1}[\ \underline{\boldsymbol{\theta}}\ ]$ which is the inverted

Hessian matrix obtained by numerical integration from special software, such as IMSL routines. From the properties of the MLE estimates, under H_0: $\underline{\boldsymbol{\theta}}_i = 0$, $\hat{\theta}_i$ is approximately normal with mean zero and standard error SE($\underline{\boldsymbol{\theta}}_i$). The test statistic

$$Z_{\theta_i} = \frac{\theta_i}{\mathrm{SE}(\theta_i)}. \tag{6.8}$$

is used to test the previous hypotheses and confidence intervals for $\underline{\boldsymbol{\theta}}_i$ can be obtained.

The estimate of the overall preclinical incidence

$$g(\theta) = I_{PC} = \sum_{i=1}^{M} \frac{1}{M d_i} \left(e^{-\Lambda_{1i}(a_i)} - e^{-\Lambda_{1i}(a_{i+1})} \right) \tag{6.9}$$

is

$$g(\hat{\theta}) = \hat{I}_{PC} = \sum_{i=1}^{M} \frac{1}{M d_i} \left(e^{-\hat{\Lambda}_{1i}(a_i)} - e^{-\hat{\Lambda}_{1i}(a_{i+1})} \right) \tag{6.10}$$

Differentiating g with respect to $\underline{\boldsymbol{\theta}}$, results in

$$\frac{\partial g(\theta)}{\partial \gamma} = \frac{\partial g(\theta)}{\partial \beta_m} = \frac{\partial g(\theta)}{\partial \mu_{2r}} = 0, \tag{6.11}$$

and

$$\frac{\partial g}{\partial \alpha_m} = \sum_{i=1}^{M} \frac{1}{M d_i} z e^{\alpha' \omega} \begin{pmatrix} \Lambda_{1i}(a_{i+1}) e^{-\Lambda_{1i}(a_{i+1})} \\ -\Lambda_{1i}(a_i) e^{-\Lambda_{1i}(a_i)} \end{pmatrix} \tag{6.12}$$

$$\frac{\partial g}{\partial \mu_{ls}} = \begin{cases} \sum_{i=1}^{M} \frac{1}{M} e^{\alpha' \omega} e^{-\Lambda_{1i}(a_{i+1})} & s = i,\ a \in I_i \\ \sum_{i=1}^{M} \frac{1}{M} e^{\alpha' \omega} \left(e^{-\Lambda_{1i}(a_{i+1})} - e^{-\Lambda_{1i}(a_i)} \right), & s < i,\ a \in I_i \\ 0, & s > i,\ a \in I_i \end{cases} \tag{6.13}$$

The covariance matrix for I_{PC} is

$$\left[\frac{\partial g(\theta)}{\partial \theta}\right] [I(\theta)]^{-1} \left[\frac{\partial g(\theta)}{\partial \theta}\right]'$$

where $I^{-1}[\underline{\theta}]$ the inverted covariance matrix of $\underline{\theta}$ obtained from maximizing the log likelihood function for the censored case.

References

Albert, A., Gertman, P. M., & Louis, T. A. (1978). Screening for the early detection of cancer I. The temporal natural history of a progressive of the disease state. *Mathematical Biosciences*, *40*, 61-59.

Albert, A., Gertman, P. M., Louis, T. A., & Liu, S. (1978). Screening for the early detection of cancer II. The temporal natural history of a progressive of the disease state. *Mathematical Biosciences*, *40*, 61-109.

Bishop, M., Feinberg, S., & Holland, P. (1975). *Discrete multivariate analysis*. Cambridge, MA: MIT Press.

Chaing, Y. K., Hardy, R. J., Hawkins, C. M., & Kapadia, A. S. (1989). An illness-death process with time dependent covariates. *Biometrics*, *45*, 669-681.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, *65*, 141-151.

Clayton, D. G., & Cuzick, J. (1985). Multivariate generalization of proportional hazards model. *Journal of the Royal Statistical Society*, *Series A*, *44*, 82-117.

Cox, D. R. (1972). Regression model and life tables (with discussion). *Journal of the Royal Statistical Society*, *Series B*, *34*, 187-220.

Hollford, T. R. (1976). Life tables with concomitant information (with discussion). *Biometrics*, *23*, 587-598.

Kittani, H. F. (1995a). Likelihood function for a progressive disease model for a bivariate survival function. *Qatar University Science Journal*, *15*(*2*), 287- 290.

Kittani, H. F. (1995b). Trivariate hazard function for censored survival data. *Journal of Mathematical Sciences*, *6*(*2*), 67- 73.

Kittani, H. F. (1996). Likelihood function for a progressive disease model for a trivariate survival function with covariates. *Journal of Mathematical Sciences*, *7*(*1*), 45-51.

Kittani, H. F. (1997). Epidemiological measures for a progressive disease model with a bivariate survival function with covariates. *Journal of Mathematical Sciences*, *8*(*1*), 41-50.

Kittani, H. F. (1997-1998). Likelihood function for the censored case for a progressive disease model for a bivariate survival function. *Aligarth Journal of Statistics*, *17* & *18*, 72-79.

Kittani, H.F. (2010). Estimation of Transition Functions for an Illness Death process. International Journal of Applied Mathematics and Statistics., Vol. 18, S10, 41-48.

Louis, T. A., Albert, A., & Heghinian, S. (1978). Screening for the early detection of cancer II. The temporal natural history of a progressive of the disease state. *Mathematical Biosciences*, *40*, 111-144.

Oakes, D. (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society*, *Series B*, *44*, 412-422.

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 488-494

1538 – 9472/10/$95.00

A Comparison between Unbiased Ridge and Least Squares Regression Methods Using Simulation Technique

Mowafaq M. Al-Kassab Omar Q. Qwaider
Al-al Bayt University,
Mafraq, Jordan

The parameters of the multiple linear regression are estimated using least squares ($\hat{B}_{LS}$) and unbiased ridge regression methods ($\hat{B}(KI,J)$). Data was created for fourteen independent variables with four different values of correlation between these variables using Monte Carlo techniques. The above methods were compared using the mean squares error criterion. Results show that the unbiased ridge method is preferable to the least squares method.

Key words: Least squares, prior information, unbiased ridge estimation, mean squares error.

Introduction

Consider the linear regression model:

$$Y^* = X^* B + U \qquad (1.1)$$

where X^* is a $(n\times(p+1))$ matrix of predictor variables of full rank, Y^* is a $(n\times 1)$ response vector, B is a $((p+1)\times 1)$ vector of parameters and U is a $(n\times 1)$ vector of errors with $E(U)=0$ and $Cov(U)=\sigma^2 I$. When multicollinearity exists, the least squares estimate $\hat{\mathrm{B}}_{\mathrm{LS}}=\left(\mathrm{X}^{*\mathrm{T}}\mathrm{X}^{*}\right)^{-1}\mathrm{X}^{*\mathrm{T}}\mathrm{Y}^{*}$ is unstable, and many different methods have been proposed to control multicollinearity (Hoerl & Kennard, 1970).

M. M. T. Al-Kassab is a Professor of Mathematical Statistics and Deputy Dean of the College of Science. His research interests include optimum stratum boundaries and biased and unbiased methods in regression. Email: mowafaq2002@yahoo.co.uk. Omar Q. Qwaider has his M.Sc. in statistics. His research interests are in regression estimation methods. Email: omar_qwaider_81@yahoo.com.

An alternative to the linear regression method is the unbiased ridge estimate

$$\hat{B}(KI,J)=\left(X^{*T}X^{*}+KI_P\right)^{-1}\left(X^{*T}Y^{*}+KJ\right)$$

where

$$J=\frac{\sum_{i=1}^{P}\hat{B}_{ils}}{P}$$

and

$$\hat{K}=\frac{P\sigma^2}{\left(\hat{B}-J\right)^T\left(\hat{B}-J\right)-\sigma^2 tr\left(X^TX\right)^{-1}}.$$

The unbiased ridge estimate regression, $\hat{B}(KI,J)$, has advantages and disadvantages. It is effective in practice but it is a complicated function of K, thus it is necessary to use rather complicated equations when employing some popular methods such as the Crouse, Jin and Hanumare (1995) criterion to select K (Swindel, 1976).

The General Multiple Linear Regression Model

The general multiple linear regression model is

$$Y_i^* = B_0 + B_1 X_{i1}^* + B_2 X_{i2}^* + ... + B_p X_{ip}^* + U_i \quad i = 1,2,...,n \qquad (2.1)$$

where $B_0, B_1, B_2, ..., B_P$ are the regression coefficients and $U_i \sim N(0,\sigma^2)$ is the random error associated with the observations. In matrix notation model (2-1) can be written as

$$\begin{bmatrix} y_1^* \\ y_2^* \\ \vdots \\ y_n^* \end{bmatrix} = \begin{bmatrix} 1 & x_{11}^* & x_{12}^* & \dots & x_{1p}^* \\ 1 & x_{21}^* & x_{22}^* & \dots & x_{2p}^* \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1}^* & x_{n2}^* & \dots & x_{np}^* \end{bmatrix} \begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_p \end{bmatrix} + \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_n \end{bmatrix}$$

$$Y^* = X^* B + U,$$

where Y^* is a $(n \times 1)$ column vector of observations on the dependent variable, X^* is a $((p+1)\times 1)$ matrix resulting from n observations on P explanatory variables $X_1^*, X_2^*, ..., X_p^*$ where the first column of 1's represent the intercept term, that is, $X_0^* = 1$, and $U \sim N(0,\sigma^2)$ is $(n \times 1)$ column vector of errors.

Assumptions of the standardized model are:

1. $E(U) = 0$
2. $Var(U) = E(UU^T) = \sigma^2 I$
3. Rank (X^*) = P where $p < n$

The ordinary least squares estimators are given by $\hat{B}_{LS} = \left(X^{*T}X^*\right)^{-1} X^{*T} Y^*$.

Properties of Ordinary Least Squares Estimators

1. Unbiasedness:
   An estimator, $\widehat{B}$, is said to be unbiased estimator of B if the expected value of $\widehat{B}$ equals B, that is, $E\left(\hat{B}_{LS}\right) = B$. (Casella & Berger, 2002)
2. Variance:
$$Var\left(\hat{B}_{LS}\right) = \sigma^2 \left(X^{*T} X^*\right)^{-1}$$
3. Mean squared error:

$$MSE\left(\hat{B}\right) = \sum_{i=1}^{P} Var\left(\hat{B}_i\right) + \sum_{i=1}^{P} \left(Bias(\hat{B}_i)\right)^2$$

$$\Rightarrow MSE\left(\hat{B}_{LS}\right) = \sum_{i=1}^{P} Var\left(\hat{B}_i\right) = \sigma^2 tr\left(X^T X\right)^{-1}$$

$$MSE\left(\hat{B}_{LS}\right) = \sigma^2 \sum_{i=1}^{P} \frac{1}{\ell_i} \qquad (2.2)$$

Unbiased Ridge Estimator

Ridge regression, which was proposed by Horel and Kennard (1970), suggests the use of $X^T X + K$, where K is a diagonal matrix rather than $X^T X$, so that the resulting estimators of B are known as the ridge regression estimators and are given by:

$$\hat{B} = (X^T X + K)^{-1} X^T Y \qquad (3.1)$$

Horel and Kennard (1970) suggested two forms for K. First, if $K = kI_p$, $0 < k < 1$. Substituting this in equation (3.1), results in

$$\hat{B}(k) = (X^T X + kI_p)^{-1} X^T Y \qquad (3.2)$$

and, using eigenvalues and eigenvectors, $\hat{B}(k)$ can be expressed as

$$\hat{B}(k) = \sum_{j=1}^{p} (\ell_j + k)^{-1} V_j V_j^T X^T Y. \qquad (3.3)$$

Second, if $K = diag(k_i),\ k_i > 0\ i = 1,2,....,p$, then

$$\hat{B}(k)=\sum_{j=1}^{p}(\ell_j+k_j)^{-1}V_jV_j^TX^TY \quad (3.4)$$

Swindle (1976) illustrated a technique for combining prior information with ridge regression that extended Hoerl and Kennard's model as follows:

$$B(kI,J)=(X^TX+kI)^{-1}(X^TY+kJ) \quad (3.5)$$

with J being a fixed vector of prior estimate of B. Swindle showed that there exists a value k which gives a smaller MSE than the least squares estimator for any fixed prior information, J.

Definition (1): A prior mean J is said to be good if the difference $\mathrm{MSE}(\hat{\mathrm{B}}(\mathrm{K}))-\mathrm{MSE}(\hat{\mathrm{B}}(\mathrm{kI},\mathrm{J}))$ is positive for all positive values k when both $\hat{B}(k)$ and $\hat{B}(kI,J)$ are computed by using the same value of k (Pliskin, 1987).

Remark: The restriction $k>0$ is made because, if $k=0$ then

$$\hat{B}_{LS}=\hat{B}(k)=\hat{B}(kI,J)=(X^TX)^{-1}X^TY$$

for all J, thereby implying that all three estimators have the same risk. In this study, it was found that the vector of prior information J depends on the arithmetic mean of the least squares estimators multiplying by a vector whose elements are ones, that is

$$J=\left[\frac{\sum_{i=1}^{p}\hat{B}_{iLS}}{p}\right]I_{p\times 1} \quad (3.6)$$

Unbiasedness of Ridge Estimators: Theorem (1)

Consider the standard linear regression model (2.1), where U is normally distributed $N(0,\sigma^2I)$, and the least square estimator, $\hat{B}$ is normally distributed $N(B,\sigma^2(X^TX)^{-1})$. The prior information J is independent of $\hat{B}_{LS}$, and J is normally distributed $N(B,V)$. Also assume that V has full rank covariance matrix and that the convex estimator is $B(C,J)=C\hat{B}_{LS}+(I-C)J$, where I is the $P\times P$ identity matrix and C is a $P\times P$ matrix. The optimal C in terms of minimum MSE is then

$$\mathrm{C}=\mathrm{V}(\sigma^2(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}+\mathrm{V})^{-1} \quad (3.7)$$

Corollary (1): Suppose $\hat{B}$ is an estimator of B with mean B and covariance matrix $\sum$, and J is prior information with mean B and covariance matrix V. Further assume that if J is uncorrelated with $\hat{B}$, and V and $\sum$ are of full rank, then the convex estimator $B(C,J)$ has a minimum MSE of optimal value

$$C=V(V+\sum)^{-1} \quad (3.8)$$

Theorem (2): Unbiased Ridge Estimate of B (Crouse, et al., 1995)

Let $\hat{B}_{LS}$ have a distribution with mean B and covariance $\sigma^2(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}$, denoted by $N(B,\sigma^2(X^TX)^{-1})$, as in the linear model. Similarly, let J be distributed $\mathrm{N}(\mathrm{B},(\frac{\sigma^2}{\mathrm{k}})\mathrm{I})$ for $\mathrm{k}>0$, and define $\mathrm{B}(\mathrm{C},\mathrm{J})=\mathrm{C}\hat{\mathrm{B}}_{\mathrm{LS}}+(\mathrm{I}-\mathrm{C})\mathrm{J}$; then, for the optimal value C in terms of minimum MSE $\mathrm{B}(\mathrm{C},\mathrm{J})=\hat{\mathrm{B}}(\mathrm{kI},\mathrm{J})=(\mathrm{X}^{\mathrm{T}}\mathrm{X}+\mathrm{kI})^{-1}(\mathrm{X}^{\mathrm{T}}\mathrm{Y}+\mathrm{KJ})$, and $B(C,J)$ is an unbiased estimate of B.

Proof: Assuming that $J \sim N(B,(\frac{\sigma^2}{k})I)$ and, from corollary (1), $\hat{B}$ has a distribution with mean B and covariance $\sum = \sigma^2 (X^T X)^{-1}$, that is, $\hat{B} \sim N(B, \sum)$, it is found that J is distributed with mean B and covariance $V = (\frac{\sigma^2}{k})I$ denoted by $J \sim N(B, V)$. Substituting this into equation (3.8) results in

$$\begin{aligned}\hat{C} &= \frac{\sigma^2}{k}\left(\sigma^2\left(X^T X\right)^{-1} + \frac{\sigma^2}{k} I\right)^{-1} \\ &= \frac{I}{k}\left((X^T X)^{-1} + \left(\frac{I}{k}\right) I\right)^{-1} \\ &= \left[k(X^T X)^{-1} + I\right]^{-1}\end{aligned}$$

Substituting $B(C,J) = C\hat{B} + (I-C)J$, results in

$$\begin{aligned}B(C,J) = &\left(k\left(X^T X\right)^{-1} + I\right)^{-1}\left(X^T X\right)^{-1} X^T Y \\ &+ \left(I - \left(k\left(X^T X\right)^{-1} + I\right)^{-1}\right) J\end{aligned}$$

and

$$\begin{aligned}B(C,J) = &\left(\left(X^T X\right) + kI\right)^{-1} X^T Y \\ &+ \left(I - \left(k\left(X^T X\right)^{-1} + I\right)^{-1}\right) J .\end{aligned}$$

Multiplying $\left(K(X^T X)^{-1} + I\right)^{-1}$ by $X^T X (X^T X)^{-1}$, results in

$$\begin{aligned}B(C,J) = &\left(\left(X^T X\right) + kI\right)^{-1} X^T Y + \\ &\left(I - X^T X\left(X^T X\right)^{-1}\left(k\left(X^T X\right)^{-1} + I\right)^{-1}\right) J \\ = &\left(\left(X^T X\right) + kI\right)^{-1} X^T Y \\ &+ \left(I - X^T X\left((X^T X) + kI\right)^{-1}\right) J\end{aligned}$$

Adding and subtracting kI to $X^T X$,

$$\begin{aligned}B(C,J) &= \begin{pmatrix}\left(X^T X + kI\right)^{-1} X^T Y + \\ \left(I - \left(X^T X + kI - kI\right)\left(X^T X + kI\right)^{-1}\right) J\end{pmatrix} \\ &= \begin{pmatrix}\left(X^T X + kI\right)^{-1} X^T Y + \\ \left(I + k\left(X^T X + kI\right)^{-1} - I\right) J\end{pmatrix}\end{aligned}$$

Simplifying the above results in:

$$\begin{aligned}B(C,J) &= \hat{B}(kI, J) \\ &= (X^T X + kI)^{-1}(X^T Y + kJ)\end{aligned} \tag{3.9}$$

Swindle (1976) did not propose a method for estimating the parameter k, however, Crouse, et al. (1995) proposed a procedure to estimate k, as follows:

$$\hat{k} = \begin{cases} \dfrac{P\sigma^2}{(\hat{B}-J)^T(\hat{B}-J) - \sigma^2 tr(X^T X)^{-1}}, \\ \quad \text{if} \quad (\hat{B}-J)^T(\hat{B}-J) - \sigma^2 tr(X^T X)^{-1} > 0 \\ \dfrac{P\sigma^2}{(\hat{B}-J)^T(\hat{B}-J)} \quad , \quad \text{o.w.} \end{cases} \tag{3.10}$$

If σ^2 is unknown, then σ^2 can be estimated by an unbiased estimator,

$$\hat{k}=\begin{cases}\dfrac{Ps^2}{(\hat{B}-J)^T(\hat{B}-J)-s^2tr(X^TX)^{-1}}, \\ \quad \text{if} \quad (\hat{B}-J)^T(\hat{B}-J)-\sigma^2tr(X^TX)^{-1}>0 \\ \dfrac{Ps^2}{(\hat{B}-J)^T(\hat{B}-J)} \quad , \quad \text{o.w.}\end{cases} \tag{3.11}$$

Properties of the Unbiased Ridge Estimators

1. Unbiasedness:

$$E\left(\hat{B}\left(kI,J\right)\right)=B$$

2. Variance:

$$Var\left(\hat{B}_i\left(kI,J\right)\right)=\sigma^2\frac{\ell_i}{\left(\ell_i+k_i\right)^2}$$

3. Mean Square's Error:

$$MSE\left(\hat{B}\left(kI,J\right)\right)= \sum_{i=1}^{P}\frac{I}{\left(\ell_i+k_i\right)^2}\left(\sum_{i=1}^{P}k_i\left(B_i-J\right)\right)^2+\sigma^2\sum_{i=1}^{P}\frac{\ell_i}{\left(\ell_i+k_i\right)^2} \tag{3.12}$$

Methodology

Model Description and Monte Carlo Simulation

This research used a Monte Carlo study to examine the properties of least squares and unbiased ridge methods. The properties were then compared in the sense of the MSE, which was evaluated using equations (2.2) and (3.12) respectively. Thirty observations (n=30) were generated for each of fourteen (p=14) explanatory variables; the explanatory variables were generated using the device:

$$\begin{cases}X_{ij}^*=(1-\alpha^2)^{1/2}Z_{ij}^*+\alpha Z_{i15}^* \ (j=1,2,...,m \ . \ i=1,2,...,30)\\ X_{ij}^*=Z_{ij}^* \ (j=m+1,m+2,...,14 \ . \ i=1,2,...,30)\end{cases}$$

Where Z_{ij} are independent standard normal pseudo-random numbers, Z_{i15} is the i^{th} element of the column vector of random error Z_{15}, α is specified so that the correlation between any two explanatory variables is given by α^2. The n observations for the dependent variable Y are determined by:

$$Y_i=\lambda_1X_{i1}+\lambda_2X_{i2}+.....+\lambda_{14}X_{i14}+U_i$$
$$i=1,2,....,30$$

where U_i are independent normal $(0,\sigma^2)$ pseudo-numbers evaluated by: $U_i=Z_{i15}-\overline{Z}_{15}$, and Y is standardized using unit length scale.

Results

The primary purpose of this research was to compare the MSE of the considered estimators, thus, the MSE for all estimators was evaluated. In addition, the efficiency of each estimator was evaluated. Thirteen experiments using Monte Carlo methods were conducted. The results of each experiment consist of five tables. The tables display the MSE of each estimator under one of five levels of correlation between explanatory variables. One set of experimental results is presented and consists of tables displaying the MSE of the least square and unbiased ridge methods for the desired correlation coefficients.

Conclusion

As shown in Tables 1-5, based on the thirteen experiments, it is concluded that the unbiased ridge method is preferable to the least square method because it results in smaller MSE values.

Table 1: Correlation Coefficient $\alpha^2 = 0.35$

Correlation Between	MSE Using Least Squares	MSE Using Unbiased Ridge
X1,X2	25.4000	0.122822
X1-X3	23.1838	0.0039352
X1-X4	30.7029	0.0368124
X1-X5	36.5401	0.144270
X1-X6	28.5695	0.0714341
X1-X7	25.3975	0.0241636
X1-X8	36.4954	0.128423
X1-X9	46.5005	0.0045159
X1-X10	1.57386	0.0355173
X1-X11	27.4589	0.0231471
X1-X12	38.3113	0.0382758
X1-X13	39.3052	0.0080928
X1-X14	46.2861	0.0331327

Table 2: Correlation Coefficient $\alpha^2 = 0.51$

Correlation Between	MSE Using Least Squares	MSE Using Unbiased Ridge
X1,X2	26.2279	0.0444524
X1-X3	24.3146	0.0111884
X1-X4	33.2860	0.0029957
X1-X5	45.0645	0.006178
X1-X6	33.4187	0.0064171
X1-X7	29.1076	0.0451311
X1-X8	44.4291	0.0554930
X1-X9	61.7260	0.0508783
X1-X10	29.6791	0.0113255
X1-X11	33.2142	0.0075011
X1-X12	47.9239	0.0162912
X1-X13	49.6498	0.0027323
X1-X14	68.5781	0.0129765

Table 3: Correlation Coefficient $\alpha^2 = 0.67$

Correlation Between	MSE Using Least Squares	MSE Using Unbiased Ridge
X1,X2	28.3239	0.0295791
X1-X3	26.6763	0.0084244
X1-X4	38.7922	0.0066016
X1-X5	61.6469	0.015094
X1-X6	42.9897	0.0242642
X1-X7	37.0921	0.0174133
X1-X8	59.5534	0.0087105
X1-X9	91.8063	0.0000733
X1-X10	38.3784	0.0059245
X1-X11	44.8721	0.0023228
X1-X12	66.4587	0.0110501
X1-X13	69.5876	0.0075445
X1-X14	113.142	0.0031203

Table 4: Correlation Coefficient $\alpha^2 = 0.84$

Correlation Between	MSE Using Least Squares	MSE Using Unbiased Ridge
X1,X2	35.6903	0.0079587
X1-X3	34.3758	0.0061240
X1-X4	57.3043	0.0004140
X1-X5	115.0238	0.042486
X1-X6	74.0384	0.0159949
X1-X7	64.4310	0.0030147
X1-X8	107.380	0.0007958
X1-X9	189.863	0.0025083
X1-X10	68.1341	0.0057765
X1-X11	83.3222	0.0018858
X1-X12	125.433	0.0037824
X1-X13	132.940	0.0039350
X1-X14	259.746	0.0037324

Table 5: Correlation Coefficient $\alpha^2 = 0.99$

Correlation Between	MSE Using Least Squares	MSE Using Unbiased Ridge
X1,X2	253.630	0.0064327
X1-X3	250.785	0.0011331
X1-X4	606.225	0.0098181
X1-X5	1645.7712	0.026003
X1-X6	974.1748	0.0078642
X1-X7	900.5158	0.0048502
X1-X8	1461.07	0.0158666
X1-X9	3049.69	0.0170372
X1-X10	976.803	0.0026540
X1-X11	1218.541	0.0011670
X1-X12	1787.02	0.0043619
X1-X13	1908.30	0.0040964
X1-X14	4586.19	0.0038209

References

Casella, G., & Berger, R. (2002). *Statistical Inference, 2nd Ed*. USA: Duxbury.

Crouse, R., Chun, J., & Hanumara, R. C. (1995). Unbiased ridge estimation with prior information and ridge trace. *Communications in Statistics*, *24*(*9*), 2341-2354.

Wichern, D. W., & Churchill, G. A. (1978). A comparison of ridge estimators. *Technometrics*, *20*(*3*), 301-310.

Gunst, R. F., Webster, J. T., & Mason, R. L. (1977). Biased estimation in regression: An evaluation using mean squared error. *Journal of the American Statistical Association*, *72*(*356*), 616-628.

Horel, A. E., & Kennard, R. W. (1970a). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55-83.

Mason, R. L., Gunst, R. F., & Webster, J. T. (1977). Regression analysis and problems of multicollinearty. *Communications in Statistics*, *4*(*3*), 279-292.

McDonald & Galarneau. (1975). A Monte Carlo evaluation of some ridge type estimators. *Journal of the American Statistical Association*, *70*, 407-416.

Murthy, K. P. N. (2003). *An introduction to Monte Carlo simulation in statistical physics*. India: Indira Gandhi Center for Atomic Research.

Pliskin, L. J. (1987). A ridge-type estimator and good prior means. *Communications in Statistics*, *16*(*12*), 3427-3429.

Swindel, B. F. (1976). Good ridge estimators based on prior information. *Communications in Statistics*, *A5*(*11*), 1065-1075.

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 495-501

1538 – 9472/10/$95.00

Ridge Regression Based on Some Robust Estimators

Hatice Samkar Ozlem Alpu
Eskisehir Osmangazi University,
Turkey

Robust ridge methods based on M, S, MM and GM estimators are examined in the presence of multicollinearity and outliers. $GM_{Walker,}$ using the LS estimator as the initial estimator is used. S and MM estimators are also used as initial estimators with the aim of evaluating the two alternatives as biased robust methods.

Key words: Multicollinearity, ridge regression, outlier, robust estimation, robust ridge methods.

Introduction

One of the main problems in regression estimation methods is multicollinearity. Multicollinearity is the term used to describe cases in which the regressors are correlated among themselves. The ridge regression model has been advocated in the literature as an alternative to LS estimation for the multicollinearity problem; in this method, which was proposed by Hoerl & Kennard (1970a, b), ridge estimators are used instead of LS estimators.

Another common problem in regression estimation methods is that of non-normal errors. The term simply means that the error distributions have fatter tails than the normal distribution. These fat-tailed distributions are more prone than the normal distribution to produce outliers, or extreme observations in the data. When outliers exist in the data, the use of robust estimators reduces their effects.

Ozlem Alpu is an Assistant Professor of Statistics. Her areas of research are computational statistics, statistical modelling and robust regression. E-mail her at: oalpu@ogu.edu.tr. Hatice Samkar is an Assistant Professor of Statistics. Her research interests include linear models, demography and regression analysis. E-mail her at: hfidan@ogu.edu.tr.

In the case of where both multicollinearity and outliers exist, the use of robust ridge regression is suggested. Robust ridge regression analysis has attracted the attention of some researchers in the literature. Holland (1973) gave the formulas for and derivation of ridge regression methods when weights are associated with each observation, and proposed the combination of ridge regression with robust regression methods. Askin & Montgomery (1980) presented a method based on augmented data sets for combining biased and robust regression techniques. Their estimates were constrained robust estimates, using an appropriately chosen ridge, Stein shrinkage or principal components constraint. Walker (1984) modified Askin and Montgomery's approach to allow the use of GM estimators instead of M estimators (Simpson & Montgomery, 1996). Silvapulle (1991) proposed a new class of ridge type M estimators obtained by using M estimators instead of LS estimators. In addition, he suggested a procedure for choosing the optimal value of the biasing parameter (k) adaptively.

Arslan & Billor (1996) proposed two alternative ridge type GM estimators to handle simultaneously multicollinearity and the existence of outliers. To reduce the effect of outliers, they computed robust estimates for k, and used these estimates to obtain robust ridge estimates for the regression coefficients. Another robust ridge regression estimator was suggested by Pfaffenberger & Dielman (1990). This estimator combines properties of the LAV

(Least Absolute Value) estimator and the ridge estimator, and is called RLAV (Ridge Least Absolute Value) estimator. Simpson & Montgomery (1996) proposed a biased-robust estimator that uses a multistage GM estimator with fully iterated ridge regression to control both influence and collinearity in the regression data set. Simpson & Montgomery (1998a) also evaluated existing and proposed robust methods relative to their performance on a comprehensive group of datasets with and without outliers. In addition, Simpson & Montgomery (1998b) developed and evaluated new robust regression procedures and compared their performance to the best alternatives currently available, in terms of efficiency, breakdown, and bounded influence. They offered the better performing alternatives as possible methods for use in a robust regression scenario.

Wisnowski, Simpson & Montgomery (2002) introduced a robust regression estimator that performs well regardless of the quantity and configuration of outliers. They show that the best available estimators are vulnerable when the outliers are extreme in the regressor space (high leverage). Their proposed compound estimator modified recently published methods with an improved initial estimate and measure of leverage.

In this study, robust ridge regression methods based on M, S, MM and GM estimators are examined in the presence of both outliers and multicollinearity. The computation of GM estimates requires two stages of parameter estimation, an initial estimate that provides a good starting point and a secondary estimate with iterations to a final estimate. LS is used as the initial estimator of GM in the study. In addition, S and MM estimators are used as initial estimators, with the aim of evaluating two alternatives as biased robust methods, as they are the top two robust estimation methods and are also highly efficient and effective against most types of outlier configurations. The performance of the robust ridge estimators is examined by using mean square error (MSE) on a hospital manpower dataset (Myers, 1990).

Methodology

Ridge Regression

Consider the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\mathbf{y}$ is a vector of n response values, $\mathbf{X}$ is an $n \times p$ matrix of rank p, $\boldsymbol{\beta}$ is a vector such that $\mathrm{E}(\boldsymbol{\varepsilon}) = 0$, and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_\mathbf{n}$. All variables in this model are corrected for their means and scaled to unit length, so that $\mathbf{X'X}$ is in correlation form.

If the columns of $\mathbf{X}$ are multicollinear, then the least-squares estimator of $\boldsymbol{\beta}$, namely $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, is an unreliable estimator due to the large variances associated with its elements. The most popular of the methods that can be used to cope with multicollinearity is ridge regression. This method, developed by Hoerl & Kennard (1970a, b), is based on adding a positive constant k to the diagonal element of $\mathbf{X'X}$. This leads to a biased estimator $\boldsymbol{\beta}_\mathbf{R}$ of $\boldsymbol{\beta}$, called the ridge estimator and given by:

$$\hat{\boldsymbol{\beta}}_\mathbf{R} = (\mathbf{X}'\mathbf{X} + k\mathbf{I}_\mathbf{n})^{-1}\mathbf{X}'\mathbf{y} \tag{2.2}$$

When both outliers and multicollinearity occur in a dataset, it would seem beneficial to combine methods designed to deal with these problems individually. Thus, robust ridge estimators will be resistant to multicollinearity problems and will be less affected by outliers.

Robust Ridge Regression

The following formula is used to compute robust ridge estimates:

$$\hat{\boldsymbol{\beta}}_{RobustRidge} = (\mathbf{X}'\mathbf{X} + k^*\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}_{Robust}, \tag{2.3}$$

where $\hat{\boldsymbol{\beta}}_{Robust}$ denotes the coefficient estimates from the robust estimators. Many methods of selecting appropriate k^* values have been proposed in the literature. In this study, the method proposed by Hoerl, Kennard & Baldwin (HKB) (1975), based on LS estimators, has been

used for the selection of the k^* value, building on robust estimators:

$$k^* = \frac{p.\hat{\sigma}^2_{Robust}}{\hat{\beta}'_{Robust}\hat{\beta}_{Robust}}, \qquad (2.4)$$

Where p is the number of regressors, and $\hat{\sigma}^2_{Robust}$ is the robust scale estimator.

Robust Estimations

The most popular of all robust estimation techniques is M estimation, proposed by Huber (1964). The M estimator minimizes the objective function

$$\min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x_i}'\hat{\beta}}{s}\right).$$

Differentiating the objective function with respect to the coefficients β, defining $\psi = \rho'$, and setting the partial derivates to 0, the system of equations can be written

$$\min_{\beta} \sum_{i=1}^{n} \psi\left(\frac{y_i - \mathbf{x_i}'\hat{\beta}}{s}\right).\mathbf{x_i} = 0,$$

where s is a robust estimate of scale.

GM estimators are a natural extension of M estimators (Walker, 1984). GM estimation is multistage estimation with two desirable properties, efficiency and bounded influence. These estimators bound the influence of the observations both in the x and y direction by using weight functions. The GM estimators are solutions to the normal equations

$$\sum_{i=1}^{n} \pi_i \psi\left(\frac{y_i - \mathbf{x_i}'\hat{\beta}}{s\pi_i}\right)\mathbf{x_i} = 0,$$

where the π_i denote the weights. This estimator was developed by Schweppe (Simpson & Montgomery, 1998a).

In the literature, several GM estimation approaches are suggested using various combinations of GM components (objective function, initial estimate, scale estimate, π-weight function, etc.). The GM estimation approach of Walker (1984) is one of the approaches. The GM approach of Walker uses the Schweppe objective function that downweights outliers with high leverage points only if the corresponding residual is large. It is recommended to use the LS as the initial estimator and a non-iterated MAD as the estimate of scale. Convergence to the final estimate is obtained by using iteratively reweighted LS (Wisnowski, Montgomery & Simpson, 2001).

In this study, Walker's (1984) GM method and two alternative GM estimation approaches have been used. In the first approach, the Schweppe function, Huber's Ψ, $\min(1,c|DFFITS|^{-1})$ and S estimation are used instead of LS for the objective function, leverage function, π-weight function, initial and scale estimation, respectively. Final parameter estimates are found by iteratively reweighted LS.

S estimators developed by Rousseeuw & Yohai (1984) are based on the minimization of the dispersion of the residuals. The S estimator is given by

$$\min_{\hat{\beta}} s(e_1(\beta),...,e_n(\beta)),$$

and the scale estimator is

$$\hat{\sigma} = s(e_1(\hat{\beta}),...,e_n(\hat{\beta})).$$

The dispersion function $s(e_1(\beta),...,e_n(\beta))$ is found as the solution to $\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{e_i}{s}\right)$=K, where K is a constant and $\rho(.)$ is the residual function. Rousseeuw & Yohai (1984) suggest using the following function:

$$\rho(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4} & |x| \le c \\ \frac{c^2}{6} & |x| > c \end{cases}.$$

The 50% breakdown point of the S estimators is achieved by taking c=1.548 and K=0.1995 (Rousseeuw & Leroy, 1987).

In the second GM approach considered, the objective, leverage and π-weight functions are calculated as in the first GM approach, and the MM estimator is used for the initial and scale estimation. The MM estimator is a high breakdown and high efficiency estimator with three stages. The initial estimate is a high breakdown estimate using an S estimate. The second stage computes an M estimate of the scale of the errors from the initial S estimate residuals.

The last step is an M estimate of the regression parameters using a redescending ψ function that assigns a weight of 0.0 to abnormally large residuals (Wisnowski, Montgomery & Simpson, 2001). Because MM estimation combines high breakdown value estimation and M estimation, it has both a high breakdown property and a higher statistical efficiency than S estimation (Chen, 2002). Although MM estimation does not theoretically bound the possible influence, it performs very well in some high leverage outlier situations (Simpson & Montgomery, 1998).

MSE Criterion for Robust Ridge Estimators

To illustrate the performance of robust ridge estimators, the MSE criterion proposed by Silvapulle (1991) is used for M estimation and that of Arslan & Billor (1996) for GM estimation. The MSE of the robust ridge estimators based on the M and GM estimators is as follows:

$$MSE(\hat{\boldsymbol{\alpha}}_{Robust}(k^*)) = \sum_{i=1}^{n}\lambda_i(\lambda_i+k^*)^{-2}\Omega_{ii} + \sum_{i=1}^{n}\left(\frac{k^*\alpha_i}{\lambda_i+k^*}\right)^2,$$

$$MSE(\hat{\boldsymbol{\alpha}}_{\text{Robust}}) = \sum_{i=1}^{p}\Omega_{\text{ii}},$$

where $\boldsymbol{\Omega}$ is a *(pxp)* $\text{cov}(\hat{\boldsymbol{\alpha}}_{Robust})$ matrix, and λ_i are the eigenvalues of $\mathbf{X'X}$. Any estimator $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ has a corresponding estimator $\hat{\boldsymbol{\beta}}(=\mathbf{P}\hat{\boldsymbol{\alpha}})$, such that $MSE(\hat{\boldsymbol{\alpha}}) = MSE(\hat{\boldsymbol{\beta}})$, where $MSE(\hat{\boldsymbol{\beta}})$ refers to the total *MSE,* $E\left\{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right\}$ (Silvapulle, 1991; Arslan & Billor, 1996).

Results

A hospital manpower dataset taken from Myers (1990) was examined as an example to compare the performance of the considered estimators. This example contains five regressors and one response variable. Because the data have been standardized, the model does not include the intercept term, thus, the $\mathbf{X'X}$ matrix is in the form of a correlation matrix:

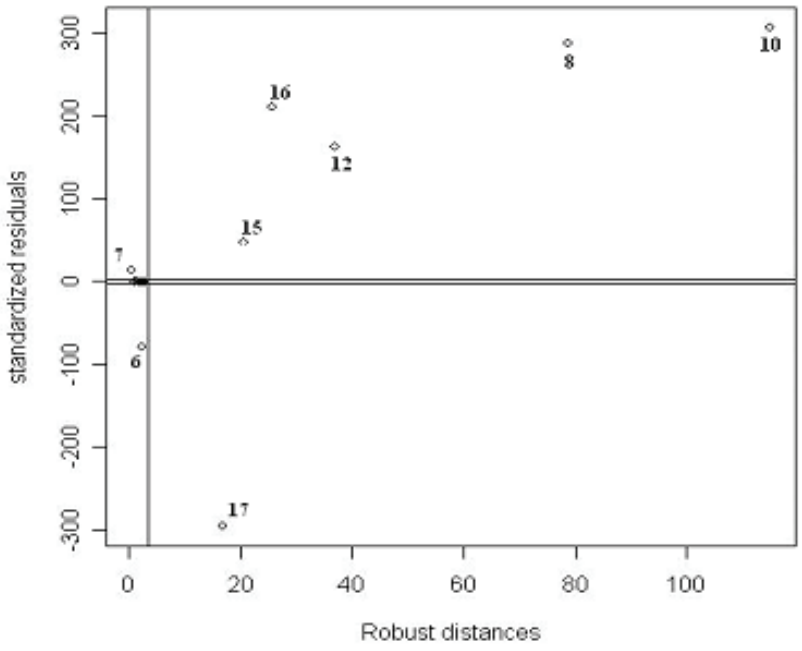
$$\mathbf{X'X} = \begin{bmatrix} 1.000 & 0.907 & 0.999 & 0.936 & 0.671 \\ 0.907 & 1.000 & 0.907 & 0.910 & 0.447 \\ 0.999 & 0.907 & 1.000 & 0.933 & 0.671 \\ 0.936 & 0.910 & 0.933 & 1.000 & 0.463 \\ 0.671 & 0.447 & 0.671 & 0.623 & 1.000 \end{bmatrix}.$$

The matrix $\mathbf{X'X}$ has eigenvalues $\lambda_1 = 4.197$, $\lambda_2 = 0.667$, $\lambda_3 = 0.095$, $\lambda_4 = 0.041$, $\lambda_5 = 0.0001$. It is observed that the regressors are moderately to highly correlated. Moreover, $(\lambda_1/\lambda_5) = (4.1971/0.0001) = 83942$, which implies the existence of multicollinearity in the dataset.

In addition, in Figure 1, $x_6, x_7, x_8, x_{10}, x_{12}, x_{15}, x_{16}$ and x_{17} are flagged as outliers or leverage points, and the points $(y_6, x_6), (y_7, x_7), (y_8, x_8), (y_{10}, x_{10}), (y_{12}, x_{12}), (y_{15}, x_{15}), (y_{16}, x_{16})$ and (y_{17}, x_{17}) are regression outliers. The points $(y_8, x_8), (y_{10}, x_{10}), (y_{12}, x_{12}), (y_{15}, x_{15}), (y_{16}, x_{16})$ and (y_{17}, x_{17}) are called bad leverage points. Regression outliers for which *x* values are not leverage points are called outliers in the *y* direction. In Figure 1, the points $(y_6, x_6), (y_7, x_7)$ are outliers in the *y* direction as well.

In the presence of outliers in the data, the use of robust methods provides more stable parameter estimates. With this aim, initial robust regression estimates have first been calculated to

Figure 1: Robust Residuals versus Distances for Hospital Manpower Data



obtain robust ridge estimates in the presence of both multicollinearity and outliers; these estimates are given in Table 1.

Table 1: Initial Robust Parameter Estimates

$\hat{\beta}_{Robust}$	M	S	MM
$\hat{\beta}_1$	-0.2159	0.4036	0.5948
$\hat{\beta}_2$	0.1782	0.6739	0.6543
$\hat{\beta}_3$	1.1891	0.3053	0.1099
$\hat{\beta}_4$	-0.0759	-0.3054	-0.2582
$\hat{\beta}_5$	-0.1281	-0.0909	-0.0853
$\hat{\sigma}$	0.0264	0.0199	0.0199

As shown in Table 1, the $\hat{\beta}_1$ value of the M estimator is found to have a negative sign. This value is inconsistent with the $\hat{\beta}_1$ values obtained from the S and MM estimators. In the presence of multicollinearity in a dataset, the signs of parameters can be found to be different from expectations. The sign of the $\hat{\beta}_1$ value can be said to occur inversely due to the potential effects of multicollinearity. In addition, the magnitudes of the parameter values for the M estimator are fairly different from those of the S and MM estimators. It is thought that the S and MM estimates are better than the M estimates because the scale estimates of S and MM are more efficient than the M estimates.

Second, biasing parameters (k^*) have been found by using the estimates in Table 1. Robust ridge estimates via the biasing parameters are calculated and shown in Table 2. In Table 2, the sign of $\hat{\beta}_1$ value of the M estimate is the same as that of the other robust ridge estimates. The effect of multicollinearity on the sign of the $\hat{\beta}_1$ value is removed by using ridge regression. The magnitudes of the parameter estimates are coherent with each other, except ridge regression estimates based on M estimates.

Table 2: Robust Ridge Parameter Estimates

$\hat{\beta}_{Robust\ Ridge}$	M	S	MM	$GM_{Standard}$	GM_S	GM_{MM}
$\hat{\beta}_1$	0.4621	0.3329	0.3343	0.3927	0.3600	0.3426
$\hat{\beta}_2$	0.1755	0.6663	0.6472	0.4351	0.4913	0.5117
$\hat{\beta}_3$	0.5239	0.3540	0.3444	0.4162	0.3874	0.3683
$\hat{\beta}_4$	-0.0857	-0.2801	-0.2299	-0.1491	-0.1408	-0.1263
$\hat{\beta}_5$	-0.1325	-0.0842	-0.0776	-0.0738	-0.0678	-0.0681
k^*	0.0032	0.0024	0.0023	0.0055	0.0030	0.0031
MSE	0.0398	0.0249	0.0242	0.0314	0.0209	0.0200

From Table 2, it is observed that the result of MSE based on M estimation is the worst among other robust methods. The worst value of the scale estimates in Table 1 belongs to the M estimates; thus, the results of Table 1 are consistent with those of Table 2. The result of MSE for GM_{Walker} is the second worst value. GM estimators were developed to overcome the deficiency of M estimators; Table 2 shows that the MSE value of GM is better than that of M. It has been noted that GM estimation is multistage, while the initial estimates of GM_{Walker} are based on LS. The method of LS is not robust in the presence of outliers in the data. For this reason, the MSE of GM_S and GM_{MM}, proposed in the study as alternatives to GM_{Walker}, are less than that of GM_{Walker}. The MSEs of the GM_S and GM_{MM} estimates are significantly less than the MSEs of the other robust ridge estimates. Furthermore, the results of MSEs for robust ridge estimates based on MM are less than those of the S estimates.

Conclusion

In this study, in the presence of both multicollinearity and outliers in a dataset, a biasing parameter k^* is calculated using the $\hat{\beta}_{Robust}$ and $\hat{\sigma}_{Robust}$ values obtained from several robust methods (M, S, MM, GM_{Walker}, GM_S and GM_{MM}), robust ridge estimates are then obtained. The performance of the robust estimators is affected by the percentage of data that are outliers, the location of the outliers in the *x* and *y* directions and their magnitudes. For this reason, the performance of the estimators considered must be interpreted in terms of these components.

The performance of ridge estimators based on M, S, MM, GM_{Walker}, GM_S and GM_{MM} estimation methods have been considered for the dataset in terms of the MSE criterion. For this dataset, the result of MSE from robust ridge regression based on M estimation is the worst among all robust techniques. Because the data includes outliers in both the *x* and *y* directions, the M estimators cannot bound the outliers in the *x* direction. In this situation, GM estimators, which bound the effects of outliers in both the *x* and *y* directions, are expected to have better performance than M estimators. Thus, under these circumstances, it has been shown that the ridge GM estimators would be preferred.

However, the result of MSE for GM_{Walker} is the second worst value. There are several outliers in the *x*-direction in the data and a few of them are extreme. On the other hand, the GM_{Walker} method uses LS estimates, which are not robust, as initial estimates. In this situation, it is expected that the MM and S estimators should have better performance in

terms of MSE, because MM and S estimators are high breakdown estimators.

GM_S and GM_{MM} estimates combine the properties of high breakdown, efficiency and robustness against outliers in the *x* and *y* directions. Consequently, the MSE of the GM_S and GM_{MM} estimates are somewhat smaller than that of the GM_{Walker} ridge estimates. According to this result, it can be said that using robust estimation methods as an initial estimator for GM give more efficient and high breakdown estimates when the dataset contains outliers in the *x* and *y* directions. As the result, the performance of robust ridge regression estimates based on GM_S and GM_{MM} estimators met expectations in terms of the MSE criterion in this dataset.

References

Arslan, O., & Billor, N. (1996). Robust ridge regression estimation based on the GM-estimators. *Journal of Mathematical and Computational Science*, *9*(*1*), 1-9.

Askin, G. R., & Montgomery, D. C. (1980). Augmented robust estimators. *Techonometrics*, *22*, 333-341.

Chen, C. (2002). Robust regression and outlier detection with the ROBUSTREG procedure. *Proceedings of the Twenty-seventh Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc.

Hoerl, A. E., & Kennard, R. W. (1970a). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55-67.

Hoerl, A. E., & Kennard, R. W. (1970b). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, *12*, 69-82.

Hoerl, A. E, Kennard, R. W., & Baldwin, K. F. (1975). Ridge regression: Some simulations. *Communications in Statistical Theory*, *4*(*2*), 105-123.

Holland, P. W. (1973). Weighted ridge regression: Combining ridge and robust regression methods. *NBER Working Paper Series*, *#11*, 1-19.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, *35*, 73-101.

Myers, R. (1990). *Classical and Modern Regression with Applications*. Boston, MA: Duxburry.

Pfaffenberger, R. C., & Dielman, T. E. (1990). A comparison of regression estimators when both multicollinearity and outliers are present. In *Robust regression: Analysis and applications*, K. Lawrence & J. Arthur (Eds.), 243-270. New York: Marcel Dekker.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.

Rousseeuw, P. J., & Yohai, V. J. (1984). Robust regression by means of S-estimators. In *Robust and nonlinear time series analysis*, J. Franke, W. Hardle & D. Martin (Eds.), 256-272. Heidelberg, Germany: Springer-Verlag.

Silvapulle, M. J. (1991). Robust ridge regression based on an M estimator. *Austrailian Journal of Statistics*, *33*, 319-333.

Simpson, J. R., & Montgomery, D. C. (1996). A biased robust regression technique for combined outlier-multicollinearity problem. *Journal of Statistical Computation Simulation*, *56*, 1-22.

Simpson, J. R., & Montgomery, D. C. (1998a). A performance based assessment of robust regression methods. *Communications in Statistical Simulations*, *27*(*4*), 1031-1049.

Simpson, J. R., & Montgomery, D. C. (1998b).The development and evaluation of alternative generalized M estimation techniques. *Communications in Statistical Simulations*, *27*(*4*), 999-1018.

Walker, E. (1984). *Influence, collinearity and robust estimation in regression*. Unpublished Ph.D. dissertation, Department of Statistics, Virginia Polytechnic Institute.

Wisnowski, J. W., Montgomery, D. C., & Simpson, J. R. (2001).A comparative analysis of multiple outlier detection procedures in the linear regression model. *Computational Statistics and Data Analysis*, *36*, 351-382.

Wisnowski, J. W., Simpson, J. R., & Montgomery, D. C. (2002). An improved compound estimator for robust regression. *Communications in Statistical Simulations*, *31*(*4*), 653-672.

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 502-511

1538 – 9472/10/$95.00

Robust Estimators in Logistic Regression: A Comparative Simulation Study

Sanizah Ahmad Norazan Mohamed Ramli
Universiti Teknologi MARA (UiTM),
Selangor Darul Ehsan, Malaysia

Habshah Midi
Universiti Putra Malaysia,
Selangor Darul Ehsan, Malaysia

The maximum likelihood estimator (MLE) is commonly used to estimate the parameters of logistic regression models due to its efficiency under a parametric model. However, evidence has shown the MLE has an unduly effect on the parameter estimates in the presence of outliers. Robust methods are put forward to rectify this problem. This article examines the performance of the MLE and four existing robust estimators under different outlier patterns, which are investigated by real data sets and Monte Carlo simulation.

Key words: Logistic regression, robust estimates, downweighting, leverage points.

Introduction

Logistic regression models are widely used in the field of medical and behavioral sciences. These models are used to describe the effect of explanatory variables on a binary response variable. The logistic regression model assumes independent Bernoulli distributed response variables with the probability of a positive response modeled as

$$P\left(Y_i = 1 \middle| X = x_i\right) = F\left(x_i^T \beta\right)$$

where F is the logistic distribution function, $x_i \in \Re^p$ are vectors of explanatory variables and $\beta \in \Re^p$ is unknown. Such models are usually estimated by the maximum likelihood estimator (MLE) due to its efficiency under a parametric model. Unfortunately, the MLE is very sensitive to outlying observations.

Pregibon (1981) stated that the estimated parameters in logistic regression may be severely affected by outliers; hence, several robust alternatives which are much less affected by outliers are proposed in the literature (for example, Pregibon, 1981; Copas, 1988; Kunsch, et al., 1989; Carroll & Pederson, 1993; Bianco & Yohai, 1996; Croux & Haesbroeck, 2003). The goal of this article is to demonstrate a formal comparison between the MLE and several robust methods for logistic regression through a simulation study and real data examples.

Background

The logistic regression model assumes an independent Bernoulli response variable Y which takes values 1(for success) or 0 (for failure). Let $X = \left(1, x_1, ..., x_p\right)$ be a vector of independent explanatory variables. Given a binary variable Y and a $p \times 1$ vector X of covariates, the logistic regression model is of the form:

$$P\left(Y = 1 \middle| X = x_i\right) = F\left(x_i^T \beta\right) = \frac{exp\left(x_i^T \beta\right)}{1 + exp\left(x_i^T \beta\right)},$$

$$i = 1, ..., n \qquad (1)$$

Sanizah Ahmad is a lecturer on the Faculty of Computer and Mathematical Sciences. Email him at: saniz924@salam.uitm.edu.my. Habshah Midi is an Associate Professor in the Department of Mathematics. Email him at: habshah@math.upm.edu.my. Norazan Mohamed Ramli is a lecturer on the Faculty of Computer and Mathematical Sciences. Email him at: norazan@tmsk.uitm.edu.my.

where $\beta^T = (\beta_0, \beta_1, ..., \beta_p)$ is a vector of parameters and F is assumed to be a continuous and increasing distribution function. For estimating the β parameters, the maximum likelihood estimator (MLE) is classically used and is defined by an objective function

$$\hat{\beta}_{MLE} = \arg\max_{\beta} \sum_{i=1}^{n} l(Y_i, x_i; \beta) \qquad (2)$$

where the log-likelihood contributions are

$$l(Y_i, x_i; \beta) = Y_i \ln F(x_i^T \beta) + (1 - Y_i) \ln[1 - F(x_i^T \beta)] \qquad (3)$$

which gives an asymptotically efficient procedure for estimating β. Alternatively, the MLE may be obtained by minimizing the deviance,

$$\hat{\beta}_{MLE} = arg\ min_{\beta} \sum_{i=1}^{n} D_i(\beta) \qquad (4)$$

where

$$D_i(\beta) = \{-Y_i \ln[P(x_i^T \beta)] - (1 - Y_i) \ln[1 - P(x_i^T \beta)]\}.$$

Differentiating (2) with respect to β results in the likelihood score equation

$$\sum_{i=1}^{n} [Y_i - F(x_i^T \beta)] x_i = 0. \qquad (5)$$

These equations are solved iteratively by using either the Newton-Raphson or Fisher Scoring method. It is important to point out that the MLE in logistic regression does not exist when the data has no overlap. The estimator can only be estimated if the data has overlap where the two parts of data given by the values of the dependent variable, $\{X = x_i | Y_i = 1\}$ and $\{X = x_i | Y_i = 0\}$ are not separated in the space of explanatory variables (Albert & Anderson, 1984). The MLE is asymptotically normal and is an efficient estimator, nonetheless, it is extremely sensitive to outliers and hence is said to not be robust. For this reason, several robust alternatives of the MLE have been created to remedy this problem.

Outliers in Logistic Regression

It is important to distinguish between the different cases of outlying observations in logistic regression. In a binary logistic model, outliers can occur in the *Y*-space, the *X*-space or in both spaces. For binary data, all the *y*'s are 0 or 1, hence an error in the *y* direction can only occur as a transposition 0→1 or 1→0 (Copas, 1988). This type of outlier is also known as residual outlier or misclassification-type error. An observation which is extreme in the design space *X* is called a leverage outlier or leverage point: a leverage point can be considered good or bad.

A good leverage point occurs when $Y = 1$ with a large value of $P(Y = 1 | x_i)$ or when $Y = 0$ with small value of $P(Y = 1 | x_i)$, and vice versa for a bad leverage point. Victoria-Feser (2002) showed that the MLE can be influenced by extreme values in the design space, and the case of misclassification errors has been studied by Pregibon (1982) and Copas (1988). Croux, et al. (2002) found that the most dangerous outliers, termed bad leverage points, are misclassified observations which are at the same time outlying in the design space of *x* variables.

Robust Estimators in Logistic Regression

In general, two alternative approaches to making MLE more robust in logistic regression exist. The first is based on weighting the likelihood score function in (5), the so-called Mallows-class (Mallows, 1975; Hampel, et al., 1986, §6.3). Two types of estimators fall in this category: the Mallows-type and Schweppe-type estimators. The former were introduced by Kunsch, et al. (1989) where the weights depend on the response as well as the covariates. Mallows-type estimators were also suggested by Kunsch, et al. (1989) but were analyzed more deeply by Carroll and Pederson (1993). This type of estimator downweights in terms of the relative position in the design space (leverage) and often uses Mahalanobis distance.

A general robust estimate for the logistic model (1) is given by the solution in β of (Carroll and Pederson, 1993)

$$\sum_{i=1}^{n} w_i x_i \left\{ Y_i - F\left(x_i^T \beta\right) - c\left(x_i \beta\right) \right\} = 0, \quad (6)$$

with w_i being the weights which may depend on x_i, y_i, or both and $c(x_i, \beta)$ is a correction function defined to ensure consistency. If $w_i \equiv 1$ and $c(x_i, \beta) = 0$, then (6) gives the usual logistic regression estimate. If $w_i = w\left(x_i, x_i^T \beta\right)$ and $c(x_i, \beta) = 0$, then the weights depend only on the design, and the estimator is called Mallows class. The estimator thus represents a weighted maximum likelihood estimator. Stefanski (1985) suggested downweighting via robust Mahalanobis distance for the covariate vector, $\boldsymbol{x}$. If $w_i = w\left(x_i, x_i^T \beta, Y_i\right)$, then the estimator is in the Schweppe class (Kunsch, et al., 1989) where the weights depend on the response as well as the covariates. This estimator is also known as the conditionally unbiased bounded influence function (CUBIF) estimator.

The second robust approach is proposed by Pregibon (1982) who worked directly with the objective function in (4). He replaced the deviance function in (4) with a robust estimator defined by

$$\beta = \underset{\beta}{arg\,min} \sum_{i=1}^{n} \lambda \left[D_i \left(x_i^T \beta, y_i \right) \right], \quad (7)$$

where λ is a strictly increasing Huber's type function. This estimator was designed to give less weight to observations poorly accounted for by the model, however, this estimator did not downweight influential observations in the design space and was not consistent. Bianco and Yohai (1996) improved this method which was consistent and more robust than Pregibon's estimator by defining

$$\beta = \underset{\beta}{arg\,min} \sum_{i=1}^{n} \left\{ \rho \left[\begin{array}{l} D\left(x_i^T \beta, y_i\right) + G\left(F\left(x_i^T \beta\right)\right) \\ + G\left(1 - F\left(x_i^T \beta\right)\right) \end{array} \right] \right\}. \quad (8)$$

The ρ chosen by Bianco and Yohai (1996) is a bounded, differentiable and a nondecreasing function defined by

$$\rho(x) = \begin{cases} x - \left(x^2 / 2k\right) & \text{if } x \le k \\ k/2 & \text{otherwise} \end{cases} \quad (9)$$

where k is a positive number, $G(x) = \int_0^x \psi(-\ln u)\, du$ and $\psi(x) = \rho'(x)$ but stressed that other choices of ρ are possible. Croux and Haesbroeck (2003) extended the Bianco and Yohai estimator (BY) by including weights to the BY estimator to reduce the influence of outlying observations in the covariate space. This weighted BY (WBY) estimator, also called the Croux and Haesbroeck (CH) estimator, can be defined as:

$$\beta = \underset{\beta}{arg\,min} \sum_{i=1}^{n} w\left(x_i\right) \left\{ \rho \left[\begin{array}{l} D\left(x_i^T \beta, y_i\right) + G\left(F\left(x_i^T \beta\right)\right) \\ + G\left(1 - F\left(x_i^T \beta\right)\right) \end{array} \right] \right\} \quad (10)$$

where the weights $w(x_i)$, in order to be a decreasing function of robust Mahalanobis distances, are distances computed using the Minimum Covariance Determinant (MCD) estimator (see Rousseeuw & Leroy, 1987) taken as:

$$w\left(x_i\right) = \begin{cases} 1 & \text{if} \quad RD_i^2 \le \chi^2_{p,0.975} \\ 0 & \text{else} \end{cases}. \quad (11)$$

This WBY estimator remains consistent because the weighting is only used on the *x*-variables. Unfortunately, the above weighting procedure also reduces the weight of the good leverage points, which is not necessary and may lead to a loss of efficiency.

Methodology

Simulation Study

A simulation study was carried out to compare the robustness of the estimators discussed. These estimators are: the MLE and the four robust estimators for logistic regression, the Mallows-type estimator (MALLOWS) with weights depending on a robust Mahalanobis distance (Carroll & Pederson, 1993) and the conditionally unbiased bounded influence (CUBIF) estimator (Kunsch, et al., 1989), both of which are computed by standard available routines in the Robust package of S-Plus, the Bianco & Yohai (BY) estimator (1996) with choice of objective function and implementation (Croux & Haesbroeck, 2003) and the weighted Bianco-Yohai (WBY) estimator, both S-plus programs available at www.econ.kuleuven.be/public/NDBAE06/programs/.

Following the simulation study carried out by Croux and Haesbroeck (2003), a logistic regression model is generated with two independent normally distributed covariates. The error terms ε_i are drawn from a logistic distribution defined as:

$$y = I\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \geq 0\right).$$

The true parameter values are $\beta = (0, 2, 2)$ with sample size $n = 200$; a large sample size is chosen to avoid separation problems.

The simulation study is reported under a variety of situations. Initially, data without contamination, having two explanatory variables independently and normally distributed with zero mean and unit variance is considered. Next, to examine the robust properties of all, the data is contaminated in three different ways, similar to the idea proposed by Victoria-Feser (2002). First, proportions (a certain percentage) are taken of the responses y chosen randomly and changed from either 0 to 1 or 1 to 0; this constitutes the misclassification-type error. For each contaminated case, 1%, 3%, 5%, 7% and 10% of the original data set are contaminated. Second, the same proportions are taken to contaminate both covariates and replace them by the value of 2 for moderate leverage points. The same process is then repeated and replaces the value by 6 for extreme leverage points. Finally, the same proportions are considered and the generated data are contaminated with both types of outliers simultaneously which constitutes bad leverage points.

To further investigate leverage points, following the idea suggested by Bondell (2005), the proportions of the explanatory variables x_1 and x_2 were taken simultaneously and their values were replaced with $x = 1,...,7$ gradually from moderate to extreme covariates in the design space with $Y = 1$. The proportions of the observations with bad leverage points were then contaminated by replacing the explanatory variables with values $x = 1,...,7$ gradually with response variable $Y = 0$.

The five methods were then applied to these data under different situations already mentioned. In each simulation run included 1,000 replications. The performances of the five methods are evaluated based on the bias and the mean squared error (MSE). The bias for each parameter and the mean squared error are respectively defined as:

$$Bias = \left\| \frac{1}{1000} \sum_{i=1}^{1000} \hat{\beta}_i - \beta \right\|$$

and

$$MSE = \left(\frac{1}{1000} \sum_{i=1}^{1000} \left\| \hat{\beta}_i - \beta \right\|^2 \right)$$

where $\|\cdot\|$ indicates the Euclidean norm (Croux & Haesbroeck, 2003).

Results

The bias and the MSE of the five estimates are shown in Tables 1-5. A good estimator is one that has bias and MSE which are relatively small or close to zero. It can be observed that, in clean data with zero percentage of outliers, the biases and MSEs of all five estimators are fairly close to each other.

Table 1 shows data with misclassified errors. The bias and MSE of the MLE estimates were immediately affected by 1% misclassified-type error. The results suggest that the MLE

becomes biased with 1% contamination, CUBIF with 3% contamination and BY with 7% contamination. The MALLOWS and WBY exhibit good robust estimators with the latter being the best method.

It can be observed from Table 2 that there is not much difference between the classical and the robust methods when contaminating data with extreme leverage points (replacing x by 6 and $Y = 1$). Similar results were obtained for moderate leverage points (replacing x by 2 and $Y = 1$); these results are not shown due to space limitations.

It is interesting to observe the results of Table 3 in the situation where 5% of the data was contaminated with leverage points by gradually increasing the distance of x. Similar conclusions to those from Table 2 can be made where the biases and MSEs for all methods are relatively small. Hence, it can be concluded that leverage points do not have much effect on the data because this type of contamination is considered as good leverage points.

Table 1: Bias and MSE of All Methods for Data with Various Percentages of Misclassified Errors

% of	MLE		MALLOWS		CUBIF		BY		WBY	
misc error	bias	MSE	bias	MSE	bias	MSE	bias	MSE	bias	MSE
0	0.0909	0.2781	0.0871	0.2774	0.0898	0.2782	0.1074	0.3089	0.1088	0.3204
1	1.4570	2.1918	0.1400	0.3073	0.5148	0.4482	0.1344	0.3765	0.1745	0.3842
3	2.4648	6.1013	0.3484	0.2098	1.1933	1.4598	0.2542	0.1716	0.0565	0.1120
5	2.7288	7.4773	0.4309	0.6467	1.6603	2.8031	0.7688	1.3257	0.0703	0.3217
7	2.8247	8.0053	0.4354	0.5614	2.0318	4.1658	2.8258	8.0112	0.3752	0.5560
10	2.8838	8.5320	0.7716	0.8849	2.4287	5.9337	2.8771	8.3148	0.0515	0.3961

Table 2: Bias and MSE of All Methods for Data with Various Percentages of Extreme Leverage Points

% of	MLE		MALLOWS		CUBIF		BY		WBY	
lev pt	bias	MSE	bias	MSE	bias	MSE	bias	MSE	bias	MSE
0	0.0840	0.2616	0.0817	0.2588	0.0833	0.2605	0.0845	0.2812	0.0875	0.2908
1	0.8072	0.8122	0.8115	0.8188	0.8122	0.8219	0.8138	0.8356	0.8198	0.8513
3	0.8035	0.8096	0.8083	0.8183	0.8060	0.8143	0.8121	0.8416	0.8118	0.8506
5	0.7910	0.7903	0.7954	0.7979	0.7911	0.7922	0.8019	0.7867	0.7867	0.8101
7	0.8089	0.8392	0.8124	0.8452	0.8111	0.8442	0.8150	0.8601	0.8109	0.8632
10	0.8162	0.8421	0.8216	0.8519	0.8186	0.8458	0.8454	0.9045	0.8463	0.9096

Table 3: Bias and MSE of All Methods for Data with 5% Leverage Points for Various Distances

	MLE		MALLOWS		CUBIF		BY		WBY	
distance	bias	MSE	bias	MSE	bias	MSE	bias	MSE	bias	MSE
clean	0.0909	0.2781	0.0871	0.2774	0.0898	0.2782	0.1074	0.3089	0.1088	0.3204
x=1	0.4882	0.4206	0.4909	0.423	0.4894	0.4233	0.5019	0.4601	0.4990	0.4758
x=2	0.5037	0.4219	0.5060	0.4255	0.5050	0.4243	0.5027	0.4366	0.4963	0.4434
x=3	0.5479	0.4861	0.5527	0.4923	0.5528	0.4929	0.5319	0.4824	0.5342	0.4887
x=4	0.5175	0.4495	0.5220	0.4549	0.5195	0.4518	0.5154	0.4639	0.5211	0.4732
x=5	0.4987	0.4156	0.5022	0.4183	0.5010	0.4178	0.4853	0.4191	0.4848	0.4275
x=6	0.4873	0.4059	0.4894	0.4087	0.4885	0.4093	0.4776	0.4119	0.4768	0.4160
x=7	0.4612	0.3972	0.4635	0.4004	0.4633	0.4009	0.4279	0.3925	0.4249	0.4103

The presence of moderate and extreme bad leverage points changes the picture dramatically. It can be observed from Tables 4 and 5 that for both cases, the CUBIF estimator can only withstand up to 3% contamination. The BY estimator can tolerate up to 3% contamination when $x = 2$, and 5% contamination when $x = 6$. The WBY estimator is better than the MALLOWS for the moderate bad leverage points. In this situation, the WBY and the MALLOWS can only withstand up to 3% and 1% contamination, respectively. Nevertheless, with data having extreme bad leverage points, the performances of the WBY and MALLOWS are equally good: both estimators are able to withstand up to 10% contamination.

Finally the results shown in Table 6 are discussed in the context of the situation where the data has 5% bad leverage points and is at various distances of the explanatory variables. By gradually increasing the distance of x and when $Y = 0$, the MLE is biased for all x; the bias worsens as x increases for MLE, but bias is consistent with the CUBIF estimators. By contrast, the bias of the MALLOWS estimator is small for $x = 6$ and $x = 7$. The BY estimator performs best when the bad leverage points are located at $x = 5$ and $x = 6$. Conversely, the biases and MSEs of the WBY estimates are consistently the smallest among other estimators for $x = 2,\ldots,7$. The results shown in Table 6 reveal that the WBY performs much better compared to the other estimators.

Numerical Examples

Two real data sets are considered to illustrate the behavior of the various robust estimates discussed. Results of the estimated coefficients, as well as their standard errors, are presented for the original and the modified data. The modified data refer to the original data with deleted outlier observation(s). A good estimator is one that has parameter estimates reasonably close to the MLE estimates of the modified data (clean data). Kordzakhia, et al. (2001) suggested another criterion for evaluating various estimators. They proposed comparing the various estimates using a goodness-of-fit discrepancy, the Chi-square statistic based on the arcsin transformation χ^2_{arc} defined as

Table 4: Bias and MSE of All Methods for Data with Moderate Bad Leverage Points (Replacing *x* by 2 and Y=0)

% of	MLE		MALLOWS		CUBIF		BY		WBY	
bad lev pt	bias	MSE	bias	MSE	bias	MSE	bias	MSE	bias	MSE
0	0.0909	0.2781	0.0871	0.2774	0.0898	0.2782	0.1074	0.3089	0.1088	0.3204
1	0.6339	0.5457	0.4976	0.3976	0.4249	0.3403	0.1222	0.2839	0.0072	0.3100
3	1.4107	2.0720	1.2084	1.5427	1.0922	1.2793	0.5695	0.5144	0.1954	0.3150
5	1.8501	2.0720	1.6461	2.7746	1.5235	2.3883	1.0211	1.1926	0.3895	0.4337
7	2.1888	4.8457	2.0127	4.1041	1.9247	3.7592	1.6166	2.7169	0.6992	0.7607
10	2.3917	5.7686	2.2550	5.1330	2.2226	4.9893	2.1789	4.8230	1.0665	1.3894

Table 5: Bias and MSE of All Methods for Data with Extreme Bad Leverage Points (Replacing *x* by 6 and Y=0)

% of	MLE		MALLOWS		CUBIF		BY		WBY	
bad lev pt	bias	MSE	bias	MSE	bias	MSE	bias	MSE	bias	MSE
0	0.0909	0.2781	0.0871	0.2774	0.0898	0.2782	0.1074	0.3089	0.1088	0.3204
1	1.4570	2.1918	0.1400	0.3073	0.5148	0.4482	0.1344	0.3765	0.1745	0.3842
3	2.4648	6.1013	0.3484	0.2098	1.1933	1.4598	0.2542	0.1716	0.0565	0.1120
5	2.7288	7.4773	0.4309	0.6467	1.6603	2.8031	0.7688	1.3257	0.0703	0.3217
7	2.8247	8.0053	0.4354	0.5614	2.0318	4.1658	2.8258	8.0112	0.3752	0.5560
10	2.8838	8.5320	0.7716	0.8849	2.4287	5.9337	2.8771	8.3148	0.0515	0.3961

Table 6: Bias and MSE of All Methods for Data with Bad Leverage Points at 5% Contamination for Various Distances

	MLE		MALLOWS		CUBIF		BY		WBY	
distance	bias	MSE	bias	MSE	bias	MSE	bias	MSE	bias	MSE
clean	0.0909	0.2781	0.0871	0.2774	0.0898	0.2782	0.1074	0.3089	0.1088	0.3204
x=1	1.2243	1.5730	1.2404	1.6134	1.2344	1.5975	1.0517	1.2093	1.0817	1.2803
x=2	1.8718	3.5497	1.7615	3.1523	1.6518	2.7846	1.1034	1.3547	0.4069	0.7711
x=3	2.2447	5.0795	1.8442	3.4507	1.6346	2.7267	0.9045	1.0110	0.1705	0.3836
x=4	2.4888	6.2345	1.6528	2.8113	1.6403	2.746	0.7273	0.7795	0.1515	0.3691
x=5	2.6367	6.9921	1.1466	1.4881	1.6387	2.7385	0.5689	0.6169	0.1243	0.3584
x=6	2.7193	7.4377	0.4851	0.5108	1.6465	2.7669	0.5183	0.8591	0.1290	0.3492
x=7	2.7635	7.6605	0.2009	0.1815	1.6542	2.7695	1.2914	3.2433	0.1693	0.2076

$$\chi^2_{arc} = \sum_{i=1}^{n} 4\left[arcsin\sqrt{y_i} - arcsin\sqrt{\pi_i} \right]^2,$$

where $\sqrt{\pi_i}$ represents the fitted probabilities for $i = 1, 2, ..., n$. The lower χ^2_{arc}, the better the goodness-of-fit.

Example: Leukemia Data

The Leukemia Data (Cook & Weisberg, 1982) was analyzed by Carroll and Pederson (1993), among others. The data set consists of measurements on 33 leukemia patients. The response variable is 1 if the patient survived more than 52 weeks and 0 otherwise. Two covariates are present in the model: white blood cell count (WBC) and AG status, which is the presence or absence of certain morphologic characteristic in the white cells. Cook and Weisberg (1982) considered these data to illustrate the identification of influential observation and they detected one observation (#15), corresponding to a patient with WBC = 100,000 who survived for a long period of time to be influential when the MLE was used. The plot in Figure 1 suggests that the observation looks like a bad leverage point.

Table 7 exemplifies the estimated parameters and estimated standard errors for the various procedures including MLE32. The MLE32 refers to the MLE estimates for the clean data after deleting observation (#15). A good estimator is one that has parameter estimates fairly close to the MLE32. It can be observed from Table 7 that the MALLOWS and WBY estimates are reasonably close to the MLE32 estimates. However, the Mallows Chi-square statistic is larger than the WBY, hence, the WBY is the best estimator for Leukemia Data because it gives the smallest χ^2_{arc} value and their estimates are closer to the MLE32. WBY is followed by the MALLOWS, BY and CUBIF estimators.

Example: Vaso-Constriction Data

The Vaso-constriction data is a well-known dataset referred to as skin data. It was introduced by Finney (1947) and was studied by Pregibon (1982) to illustrate the impact of potential influential observations in logistic regression. The binary outcomes (presence or absence of vaso-constriction of the skin of the digits after air inspiration) are explained by two explanatory variables: x_1 the volume of air inspired, and x_2 the inspiration rate (both in logarithms). The literature, which extensively uses this dataset, often reports observations (#4) and (#18) as outliers. As shown in Figure 2, a plot of the data based on the maximum likelihood fit shows that the two observations (#4 and #18) look more like misclassified errors rather than outlying observations.

Table 8 presents the estimated parameters, estimated standard errors and goodness-of-fit measures for the various procedures including MLE37 after removing the two influential observations. Several interesting points appear from Table 8. It is notable that the

Figure 1: Scatter Plot of Leukemia Data

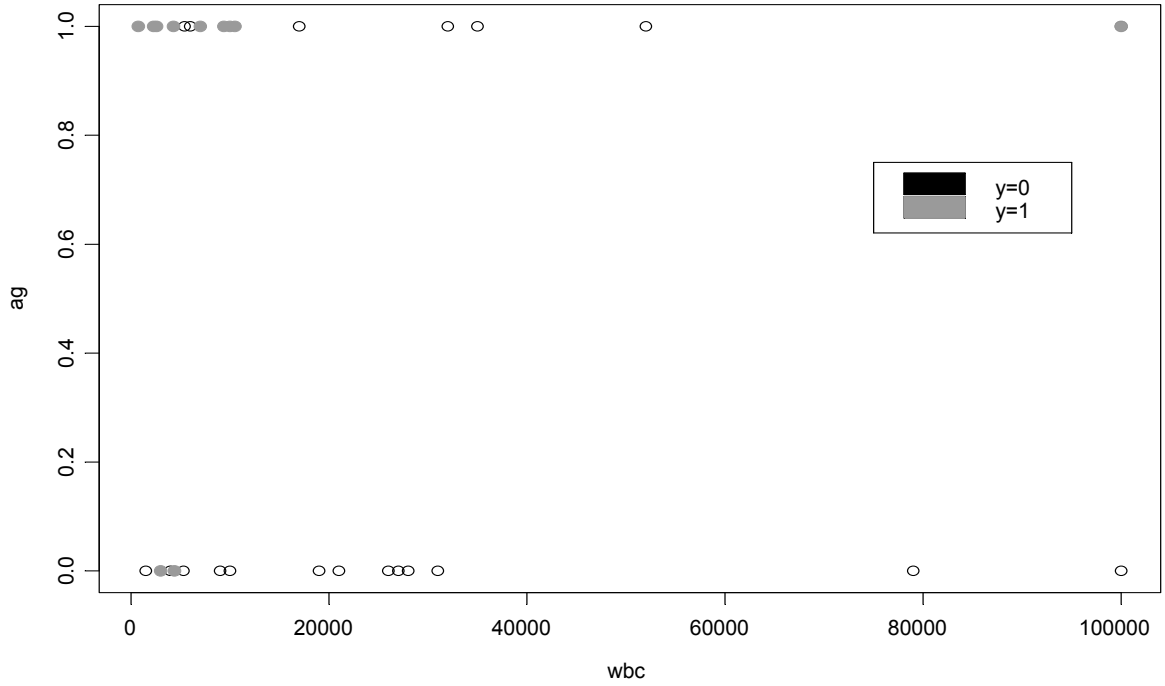


Table 7: Leukemia Data: Estimated Parameters, Standard Errors and Goodness-of-Fit Measures

Estimation	Intercept		WBC		AG		
Method	Est.	s.e.	Est.	s.e.	Est.	s.e.	χ^2_{arc}
MLE	-1.3073	0.2931	-0.3177	1.454	2.2611	2.2003	52.16
MLE32	0.2119	7.0996	-2.3545	6.9497	2.5581	4.9143	32.52
MALLOWS	0.1710	6.7568	-2.2535	6.7818	2.524	4.6589	42.46
CUBIF	-0.6763	1.7135	-0.9110	3.4500	2.2495	1.1712	46.73
BY	0.1595	5.0511	-1.7740	5.7623	1.9276	3.3011	44.05
WBY	0.1891	6.8884	-2.1927	6.7853	2.4003	4.6923	39.47

CUBIF and MALLOWS yield results reasonably close to the MLE. The results also show that the BY and WBY estimates have been strongly affected when the two influential observations are removed from the dataset. It may be observed that the parameter estimates and the standard errors of both estimates become large because, without the two observations, the remaining data set is in a situation of quasi-complete separation (Albert & Anderson, 1984), with little overlap between observations $y_i = 0$ and $y_i = 1$. Thus, the model is nearly undetermined. For this reason, the BY and WBY downweight these observations and have large increases of coefficients and standard errors. The parameter estimates and the standard errors of both estimators are considerably close to the MLE37 estimates. However, the BY has the smallest χ^2_{arc} value, therefore, the BY estimator gives the best result for this data set.

Figure 2: Scatter Plot of Vaso Data

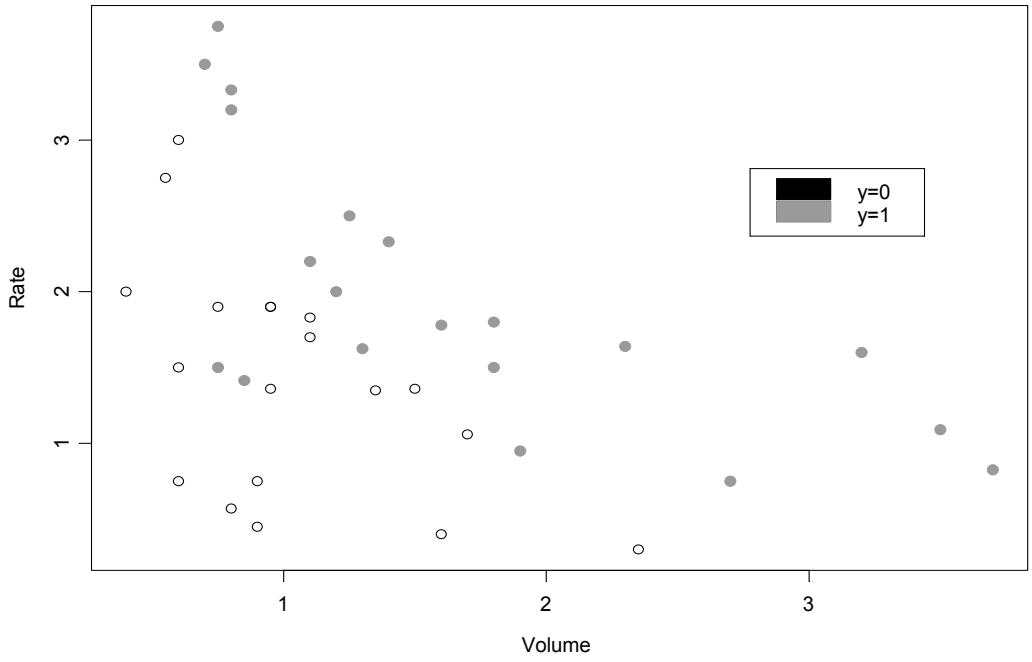


Table 8: Vaso Data: Estimated Parameters, Standard Errors and Goodness-of-Fit Measures

Estimation	Intercept		Log(Volume)		Log(Rate)		χ^2_{arc}
Method	Est.	s.e.	Est.	s.e.	Est.	s.e.	
MLE	-2.9239	1.2877	5.2205	1.8579	4.6312	1.7889	48.39
MLE37	-24.5812	14.0211	39.5498	23.2463	31.9352	17.7595	12.34
MALLOWS	-2.9207	1.2908	5.1673	1.8470	4.5967	1.7886	48.41
CUBIF	-2.8776	1.2707	5.1661	1.8364	4.5646	1.7644	48.47
BY	-6.8667	10.0507	10.7523	15.3086	9.381	12.7798	40.87
WBY	-6.8465	10.0672	10.7504	15.3346	9.3785	12.8014	40.91

Conclusion

The goal of this study was to compare the performance of the MLE and four robust estimators for the logistic model under both clean and contaminated data sets. The findings signify that the MLE can be biased in the presence of misclassified error and bad leverage points, whereas some robust estimators are better than others depending on the type of contamination. When the contamination data are leverage points, the simulation results indicate that all parameter estimates are not dramatically affected, because they have consistently small bias. Overall, the WBY estimator is preferred because it is more robust than other estimators tested in this study for any type of contamination in the data. This estimator is followed by the BY, MALLOWS and CUBIF. However, further investigation is needed to compare these robust estimators through an extensive simulation study involving different parameter values, sample sizes and parameter size. Further studies are also needed to investigate more suitable robust methods to cater outlying observations in logistic regression. Most robust methods unfortunately rely on simple downweighting of

distant observations in the design space regardless of whether or not they are misspecified, whether they are good or bad leverage points and what influence they have on the model.

References

Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, *71*, 1-10.

Bianco, A. M., & Yohai, V. J. (1996). Robust estimation in the logistic regression model. In *Robust statistics, Data analysis and computer intensive methods,* H. Reider, Ed., 17-34. New York: Springer Verlag.

Bondell, H. D. (2005). Minimum distance estimation for the logistic regression model. *Biometrika*, *92*, 724-731.

Carroll, R. J., & Pederson, S. (1993). On robust estimation in the logistic regression model. *Journal of the Royal Statistical Society*, *Series B*, *55*, 693-706.

Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression.* London: Chapman and Hall.

Copas, J. B. (1988). Binary regression model for contaminated data (with discussion). *Journal of the Royal Statistical Society*, *Series B*, *50*, 225-265.

Croux, C., Flandre, C., & Haesbroeck, G. (2002). The breakdown behavior of the maximum likelihood estimator in the logistic regression model. *Statistics & Probability Letters*, *60*, 377-386.

Croux, C., & Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computational Statistics & Data Analysis Journal*, *44*, 273-295.

Finney, D. J., (1947). The estimation from individual records of the relationship between dose and quantal response. *Biometrika*, *34*, 320-334.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust Statistics: The approach based on influence functions*. New York, NY: John Wiley.

Kordzakhia, N., Mishra, G. D., & Reiersolmoen, L. (2001). Robust estimation in the logistic regression model. *Journal of Statistical Planning and Inference*, *98*, 211-223.

Kunsch, H. R., Stefanski, L. A., & Carroll, R. J. (1989). Conditionally unbiased bounded influence estimation in general regression models, with applications to generalized linear models. *Journal of American Statistical Association*, *84*, 460-466.

Mallows, C. L. (1975). *On some topics in robustness*. Murray Hill, NJ: Bell Telephone Laboratories.

Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics*, *9*, 705-724.

Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrika*, *73*, 413-425.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection.* New York: Wiley.

Stefanski, L. A. (1985). The effects of measurement error on parameter estimation. *Biometrika*, *72*, 583-592.

Victoria-Feser, M.-P. (2002). Robust inference with binary data. *Psychometrica*, *67*, 21-32.

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 512-519

1538 – 9472/10/$95.00

A General Class of Chain-Type Estimators in the Presence of Non-Response Under Double Sampling Scheme

Sunil Kumar
University of Jammu,
(J & K), India

Housila P. Singh
Vikram University,
Ujjain, India

Sandeep Bhougal
Shri Mata Vaishno Devi University,
Kakryal, Jammu (J & K), India

General class chain ratio type estimators for estimating the population mean of a study variable are examined in the presence of non-response under a double sampling scheme using a factor-type estimator (FTE). Properties of the suggested estimators are studied and compared to those of existing estimators. An empirical study is carried out to demonstrate the performance of the suggested estimators; empirical results support the theoretical study.

Key words: Double sampling, factor-type estimator, chain ratio estimator, non-response.

Introduction

Over the last five decades one of the major developments in sample surveys is the use of an auxiliary variable x, correlated with the study variable y in order to obtain estimates of the population total or mean of the study variable. Various estimation procedures in sample surveys require advance knowledge of some auxiliary variable x_i, which is then used to increase the precision of estimates. When the population mean $\overline{X}$ is not known, it can be estimated from a preliminary large sample on which only the auxiliary characteristic x is observed. The value of $\overline{X}$ in the estimator is then replaced by its estimate, and a smaller second-phase sample of the variable of interest (study variable) y is taken. This technique, known as double sampling or two-phase sampling, is particularly appropriate if the x_i values are easily accessible and are much less expensive to collect than the y_i values (Sitter, 1997; Hidiroglou & Sarndal, 1998). Neyman (1938) was the first to describe the concept of double sampling in connection with collecting information on strata sizes is a stratified sampling (Singh & Espejo, 2007).

In some practical situations it is observed that, when conducting a sample survey, complete information for all the units selected in the sample is not obtained due to the occurrence of non-response. Hansen and Hurwitz (1946) considered the problem of non-response while estimating the population mean by taking a sub-sample from the non-response group with the help of an unbiased estimator; they suggested combining the information available from response and non-response groups. Further, rectification in the estimation procedure for the population mean in the presence of non-response using auxiliary variable was proposed by Cochran(1977), Rao (1986, 1987), Khare and Srivastava (1993, 1995, 1997), Okator and Lee (2000), Tabasum and Khan (2004, 2006), and Singh and Kumar (2008a, 2008b, 2008c, 2009a, 2009b) using the Hansen and Hurwitz (1946) technique. This article develops a one parameter family of chain ratio type estimators with two auxiliary variables in the presence of non-response. The proposed family is based on factor type estimators (FTE) developed by Singh and Shukla (1987) and Singh, et al. (1994) and empirical studies support the results.

Sunil Kumar is in the Department of Statistics in University of Jammu College. Housila P. Singh is a Professor in the School of Studies in Statistics. Email: hpsujn@rediffmail.com. Sandeep Bhougal is in the School of Mathematics.
Email: sandeep.bhougal@smvdu.ac.in.

The Proposed Strategy

Consider a finite population $U=(U_1,U_2,...,U_N)$ of size N. Let y be the study variable, x_1 be the main auxiliary variable with an unknown mean that is highly correlated with main character y, and x_2 be an additional auxiliary variable with known mean that is less correlated with y than is x_1. A large first phase sample of size n' from the finite population U is selected by simple random sampling without replacement (SRSWOR). A smaller second phase sample of size n is selected from n' by SRSWOR. Non-response occurs in the second phase sample of size n in which n_1 units respond and n_2 units do not. From the n_2 non-respondents, by SRSWOR a sample of size $r=n_2/k;\ k>1$ units is selected where k is the inverse sampling rate at the second phase sample of size n with all r units responding. Thus, (n_1+r) are the responding units on the study variable y, consequently the estimator for the population mean $\bar{Y}$ of the study variable y using a sub-sampling scheme envisaged by Hansen and Hurwitz (1946) is defined as

$$\bar{y}^{*}=w_1\bar{y}_1+w_2\bar{y}_{2r}, \tag{1}$$

where

$$w_1=n_1/n,\ w_2=n_2/n,\ \bar{y}_1=\sum_{i=1}^{n_1}y_i/n_1$$

and

$$\bar{y}_{2r}=\sum_{i=1}^{r}y_i/r.$$

It is known that the estimator $\bar{y}^{*}$ is an unbiased estimator of the population mean $\bar{Y}$ of the study variable y and has a variance as given by

$$Var(\bar{y}^{*})=\theta_1 S_y^2+\theta^{*}S_{y(2)}^2, \tag{2}$$

where

$$\theta_1=\left(\frac{1}{n}-\frac{1}{N}\right),\quad \theta^{*}=\frac{W_2(k-1)}{n},\quad W_2=N_2/N,$$

and S_y^2 and $S_{y(2)}^2$ are the population mean square of the variable y for the entire population and for the non-responding group of the population. Similarly, for estimating the population mean $\bar{X}_i$ of the auxiliary variable $x_i;\ (i=1,2)$, the unbiased estimator $\bar{x}_i^{*}$ is given by

$$\bar{x}_i^{*}=w_1\bar{x}_{i(1)}+w_2\bar{x}_{i(2r)}, \tag{3}$$

where $\bar{x}_{i(1)}$ and $\bar{x}_{i(2r)}$ are the sample means of the auxiliary variable $x_i;\ (i=1,2)$ based on n_1 and r units respectively.

The variance of $\bar{x}_i^{*}$ is given by

$$Var(\bar{x}_i^{*})=\theta_1 S_{x_i}^2+\theta^{*}S_{x_i(2)}^2, \tag{4}$$

where $S_{x_i}^2$ and $S_{x_i(2)}^2$ are the population mean square of $x_i;\ (i=1,2)$ for the entire population and the non-responding group of the population.

The Proposed Class of Strategy

Using an unknown constant $t>0$ and two auxiliary variables x_1 and x_2, a general class of chain ratio type of strategy $[D,\bar{y}_F^{*}(t)]$ is defined for estimating the population mean $\bar{Y}$ of the study variable y in the presence of non-response as follows:

$$\bar{y}_F^{*}(t)=\bar{y}^{*}g_3(1,0)\left[\frac{\varphi\{\lambda_1(t)\}}{\varphi\{\lambda_2(t)\}}\right], \tag{5}$$

where

$$\varphi\{\lambda_i(t)\}=\lambda_i(t)+\{1-\lambda_i(t)\}g_2(0,1);\ i=1,2,$$

$$\lambda_1(t)=\frac{\theta B}{A+\theta B+C},\ \lambda_2(t)=\frac{C}{A+\theta B+C},$$

$$A=(t-1)(t-2),\ B=(t-1)(t-4),$$

$$C=(t-2)(t-3)(t-4),\ \theta=n/N,$$

$$\theta_2=\left(\frac{1}{n'}-\frac{1}{N}\right),\ \theta_3=\left(\frac{1}{n}-\frac{1}{n'}\right),$$

$$g_1(\alpha,\beta)=\left(\frac{\bar{X}_1}{\bar{x}_1^*}\right)^{\alpha}\left(\frac{\bar{X}_2}{\bar{x}_2^*}\right)^{\beta},$$

$$g_2(\alpha,\beta)=\left(\frac{\bar{X}_1}{\bar{x}_1^{'}}\right)^{\alpha}\left(\frac{\bar{X}_2}{\bar{x}_2^{'}}\right)^{\beta},$$

$$g_3(\alpha,\beta)=\left(\frac{\bar{x}_1^{'}}{\bar{x}_1^*}\right)^{\alpha}\left(\frac{\bar{x}_2^{'}}{\bar{x}_2^*}\right)^{\beta},\ \bar{X}_1=\sum_{i=1}^{N_1}x_i/N_1,$$

$$\bar{X}_2=\sum_{i=1}^{N_2}x_i/N_2,\ \bar{x}_1^{'}=\sum_{i=1}^{n_1^{'}}x_i/n_1^{'}$$

and

$$\bar{x}_2^{'}=\sum_{i=1}^{n_2^{'}}x_i/n_2^{'}.$$

In order to identify some of the members of the proposed strategy and compare their efficiencies, certain classical strategies are put forth:

(i) $\left[D,\bar{y}_R^*\right];\ \bar{y}_R^*=\bar{y}^*g_3(1,0)$ (6)

by Khare and Srivastava (1993), Okafor and Lee (2000) and Tabasum and Khan (2004)

(ii) $\left[D,\bar{y}_P^*\right];\ \bar{y}_P^*=\bar{y}^*g_3(-1,0)$ (7)

by Khare and Srivastava (1993)

(iii) $\left[D,\bar{y}_C^*\right];\ \bar{y}_C^*=\bar{y}^*g_3(1,0)g_2(0,1)$. (8)

Some Strategies of the Class

For $t=1$ and 4 respectively,

(i) $\left[D,\bar{y}_F^*(1)\right]=\left[D,\bar{y}_C^*\right]$, (9)

(ii) $\left[D,\bar{y}_F^*(4)\right]=\left[D,\bar{y}_R^*\right]$. (10)

Further, for $t=2$ and 3,

$$\left[D,\bar{y}_F^*(2)\right];\ \bar{y}_F^*(2)=\bar{y}^*g_3(1,0)g_2(0,-1),$$ (11)

$$\left[D,\bar{y}_F^*(3)\right];\ \bar{y}_F^*(3)=\bar{y}^*g_3(1,0)\begin{Bmatrix}(1+h)\\-hg_2(0,-1)\end{Bmatrix},$$ (12)

where $h=n(N-n)^{-1}$, and $\bar{y}_F^*(2)$ is a chain type estimator in D in which $\bar{X}_1$ is estimated through the product estimator utilizing $\bar{X}_2$ where non-response on auxiliary variable x_1 and $\bar{y}_F^*(3)$ is a chain type estimator in D in which $\bar{X}_1$ is estimated utilizing a dual to ratio estimator with non-response on auxiliary variable x_1.

Properties of the Proposed Strategy

To obtain the bias and mean square error (*MSE*) of the proposed general class of strategy $\left[D,\bar{y}_F^*(t)\right]$, under the large sample approximation,

$$\bar{y}^*=\bar{Y}(1+\varepsilon_0),\ \bar{x}_1^*=\bar{X}_1(1+\varepsilon_1),$$
$$\bar{x}_2^*=\bar{X}_2(1+\varepsilon_2),\ \bar{x}_1^{'}=\bar{X}_1(1+\varepsilon_1^{'}),$$
$$\text{and }\bar{x}_2^{'}=\bar{X}_2(1+\varepsilon_2^{'}),$$

such that

$$E(\varepsilon_0)=E(\varepsilon_1)=E(\varepsilon_2)=E(\varepsilon_1^{'})=E(\varepsilon_2^{'})=0$$

and

$$E(\varepsilon_0^2)=\theta_1S_y^2+\theta^*S_{y(2)}^2,$$

$$E(\varepsilon_1^2)=\theta_1S_{x_1}^2+\theta^*S_{x_1(2)}^2,$$

$$E(\varepsilon_2^2)=\theta_1S_{x_2}^2+\theta^*S_{x_2(2)}^2,\ E(\varepsilon_1^{'2})=\theta_2S_{x_1}^2,$$

$$E(\varepsilon_2^{'2})=\theta_2S_{x_2}^2,\ E(\varepsilon_0\varepsilon_1)=\theta_1S_{yx_1}+\theta^*S_{yx_1(2)},$$

$$E(\varepsilon_0\varepsilon_1^{'})=\theta_2S_{yx_1},\ E(\varepsilon_1\varepsilon_1^{'})=\theta_2S_{x_1}^2,$$

$$E(\varepsilon_0\varepsilon_2^{'})=\theta_2S_{yx_2},\ E(\varepsilon_1\varepsilon_2^{'})=\theta_2S_{x_1x_2},$$
$$E(\varepsilon_1^{'}\varepsilon_2^{'})=\theta_2S_{x_1x_2},$$

where

$$S_{yx_1} = \frac{1}{(N-1)}\sum_{i=1}^{N}(y_i - \bar{Y})(x_{1i} - \bar{X}_1),$$

$$S_{yx_2} = \frac{1}{(N-1)}\sum_{i=1}^{N}(y_i - \bar{Y})(x_{2i} - \bar{X}_2),$$

$$S_{yx_{1(2)}} = \frac{1}{(N_2-1)}\sum_{i=1}^{N_2}(y_i - \bar{Y}_2)(x_{1i} - \bar{X}_{1(2)}),$$

$$S_{x_1x_2} = \frac{1}{(N-1)}\sum_{i=1}^{N}(x_{1i} - \bar{X}_1)(x_{2i} - \bar{X}_2),$$

$$\bar{X}_{1(2)} = \sum_{i=1}^{N_2} x_{1i}/N_2,$$

$$\bar{Y} = \sum_{i=1}^{N} y_i/N,$$

$$\bar{Y}_2 = \sum_{i=1}^{N_2} y_i/N_2,$$

N_1 and $N_2 (= N - N_1)$ are the sizes of the responding and non-responding units from the finite population N.

Expressing the proposed estimator $\bar{y}_F^*(t)$ in terms of $\varepsilon' s$,

$$\bar{y}_F^*(t) = \bar{Y}(1+\varepsilon_0)\frac{(1+\varepsilon_1')}{(1+\varepsilon_1)}\left[\frac{\lambda_1(t) + \{1-\lambda_1(t)\}(1-\varepsilon_2')^{-1}}{\lambda_2(t) + \{1-\lambda_2(t)\}(1-\varepsilon_2')^{-1}}\right]. \quad (13)$$

It is assumed that $|\lambda_2(t)\varepsilon_2'| < 1$, because $\lambda_2(t) = \frac{C}{A+\theta B+C}$, for any choice of t, $|\lambda_2(t)| < 1$. Thus, if $|\varepsilon_2'| < 1$, $|\lambda_2(t)\varepsilon_2'| < 1$ is a valid assumption, expanding the right hand side of (13) and neglecting the terms involving powers of $\varepsilon' s$ greater than two results in

$$\bar{y}_F^*(t) = \bar{Y}\begin{Bmatrix} 1+\varepsilon_0-\varepsilon_1+\varepsilon_1'+\varepsilon_1^2-\varepsilon_1\varepsilon_1'-\varepsilon_0\varepsilon_1+\varepsilon_0\varepsilon_1' \\ +\lambda(t)\left(\varepsilon_2'-\varepsilon_1\varepsilon_2'+\varepsilon_1'\varepsilon_2'+\varepsilon_0\varepsilon_2'-\lambda_2\varepsilon_2'^2\right)\end{Bmatrix},$$

$$\{\bar{y}_F^*(t) - \bar{Y}\} = \bar{Y}\begin{Bmatrix} \varepsilon_0-\varepsilon_1+\varepsilon_1'+\varepsilon_1^2-\varepsilon_1\varepsilon_1'-\varepsilon_0\varepsilon_1+\varepsilon_0\varepsilon_1' \\ +\lambda(t)\left(\varepsilon_2'-\varepsilon_1\varepsilon_2'+\varepsilon_1'\varepsilon_2'+\varepsilon_0\varepsilon_2'-\lambda_2\varepsilon_2'^2\right)\end{Bmatrix}, \quad (14)$$

where $\lambda(t) = \lambda_1(t) - \lambda_2(t)$.

Taking expectations of both sides of (14), results in the bias of $\bar{y}_F^*(t)$ to the first degree of approximation, as

$$B\left(\bar{y}_F^*(t)\right) = \begin{Bmatrix} \theta_3\left(1-K_{yx_1}\right)S_{x_1}^2 + \theta^*\left(1-K_{yx_{1(2)}}\right)S_{x_{1(2)}}^2 \\ -\lambda(t)\theta_2\left(\lambda_2(t)-K_{yx_2}\right)S_{x_2}^2 \end{Bmatrix}, \quad (15)$$

where

$$K_{yx_1} = \frac{S_{yx_1}}{S_{x_1}^2} = \rho_{yx_1}\frac{S_y}{S_{x_1}},\ K_{yx_2} = \frac{S_{yx_2}}{S_{x_2}^2} = \rho_{yx_2}\frac{S_y}{S_{x_2}},$$

$$\rho_{yx_1} = \frac{S_{yx_1}}{S_y S_{x_1}} \text{ and } \rho_{yx_2} = \frac{S_{yx_2}}{S_y S_{x_2}}.$$

Squaring both sides of (14) and neglecting terms of $\varepsilon' s$ involving power greater than two,

$$\left(\bar{y}_F^*(t) - \bar{Y}\right)^2 = \bar{Y}^2\{\varepsilon_0 - \varepsilon_1 + \varepsilon_1' + \lambda(t)\varepsilon_2'\}^2$$

$$\left(\bar{y}_F^*(t) - \bar{Y}\right)^2 = \bar{Y}^2\begin{Bmatrix} \varepsilon_0^2+\varepsilon_1^2+\varepsilon_1'^2+\lambda^2(t)\varepsilon_2'^2 \\ -2\varepsilon_0\varepsilon_1+2\varepsilon_0\varepsilon_1'+2\lambda(t)\varepsilon_0\varepsilon_2' \\ -2\varepsilon_1\varepsilon_1'-2\lambda(t)\varepsilon_1\varepsilon_2'+2\lambda(t)\varepsilon_1'\varepsilon_2' \end{Bmatrix}. \quad (16)$$

Taking expectations of both sides of (16), gives the mean square error of $\bar{y}_F^*(t)$ to the first degree of approximation as

$$MSE\left(\bar{y}_F^*(t)\right) = \begin{bmatrix} \theta_1 S_y^2 + \theta_3\left(1-2K_{yx_1}\right)S_{x_1}^2 \\ +\lambda(t)\theta_2\left(\lambda(t)+2K_{yx_2}\right)S_{x_2}^2 \\ +\theta^*\left\{S_{y(2)}^2+\left(1-2K_{yx_1(2)}\right)S_{x_1(2)}^2\right\} \end{bmatrix}, \quad (17)$$

where

$$K_{yx_1(2)} = \frac{S_{yx_1(2)}}{S_{x_1(2)}^2} = \rho_{yx_1(2)}\frac{S_{yx_1(2)}}{S_{x_1(2)}},$$

$$\rho_{yx_1(2)} = \frac{S_{yx_1(2)}}{S_{y(2)}S_{x_1(2)}}$$

Corollary

Letting $\lambda(t) = -1, \lambda_2(t) = 1$ for $t = 1$ in (15) and (17), the bias and MSE of $\bar{y}_C^*$, respectively given by

$$B\left(\bar{y}_F^*(1) = \bar{y}_C^*\right) = \begin{bmatrix} \theta_3\left(1-K_{yx_1}\right)S_{x_1}^2 \\ +\theta^*\left(1-K_{yx_1(2)}\right)S_{x_1(2)}^2 \\ +\theta_2\left(1-K_{yx_2}\right)S_{x_2}^2 \end{bmatrix} \quad (18)$$

and

$$MSE\left(\bar{y}_F^*(1) = \bar{y}_C^*\right) = \begin{bmatrix} \theta_1 S_y^2 + \theta_3\left(1-2K_{yx_1}\right)S_{x_1}^2 \\ +\theta_2\left(1-K_{yx_2}\right)S_{x_2}^2 \\ +\theta^*\left\{S_{y(2)}^2+\left(1-2K_{yx_1(2)}\right)S_{x_1(2)}^2\right\} \end{bmatrix}. \quad (19)$$

To obtain the bias and *MSE* of $\bar{y}_R^*$, assume that $\lambda(t) = \lambda_2(t) = 0$ for $t = 4$, in (15) and (17),

$$B\left(\bar{y}_F^*(4) = \bar{y}_R^*\right) = \begin{bmatrix} \theta_3\left(1-K_{yx_1}\right)S_{x_1}^2 \\ +\theta^*\left(1-K_{yx_1(2)}\right)S_{x_1(2)}^2 \end{bmatrix} \quad (20)$$

and

$$MSE\left(\bar{y}_F^*(4) = \bar{y}_R^*\right) = \begin{bmatrix} \theta_1 S_y^2 + \theta_3\left(1-2K_{yx_1}\right)S_{x_1}^2 \\ +\theta^*\left\{S_{y(2)}^2+\left(1-2K_{yx_1(2)}\right)S_{x_1(2)}^2\right\} \end{bmatrix}. \quad (21)$$

The $MSE\left(\bar{y}_F^*(t)\right)$ is minimized, when

$$\lambda(t) = -K_{yx_2}. \quad (22)$$

Thus, substituting (22) in (17), results in the optimum mean square error of $\bar{y}_F^*(t)$, as

$$MSE\left(\bar{y}_F^*(t)\right)_{opt} = \begin{bmatrix} \theta_1 S_y^2 - \theta_2 K_{yx_2}^2 S_{x_2}^2 + \theta_3\left(1-2K_{yx_1}\right)S_{x_1}^2 \\ +\theta^*\left\{S_{y(2)}^2+\left(1-2K_{yx_1(2)}\right)S_{x_1(2)}^2\right\} \end{bmatrix}. \quad (23)$$

Efficiency Comparisons

From (2), (19), (21) and (23),

$$Var\left(\bar{y}^*\right) - MSE\left(\bar{y}_F^*(t)\right)_{opt} = \begin{bmatrix} \theta_3\left(1-2K_{yx_1}\right)S_{x_1}^2 - \theta_2 K_{yx_2}^2 S_{x_2}^2 \\ +\theta^*\left(1-2K_{yx_1(2)}\right)S_{x_1(2)}^2 \end{bmatrix}, \quad (24)$$

$$MSE\left(\bar{y}_C^*\right) - MSE\left(\bar{y}_F^*(t)\right)_{opt} = \theta_2\left(1-K_{yx_2}^2\right)^2 S_{x_2}^2 > 0 \text{ when } K_{yx_2} < 1, \quad (25)$$

$$MSE\left(\bar{y}_k^*\right) - MSE\left(\bar{y}_F^*(t)\right)_{opt} = \theta_2 K_{yx_2}^2 S_{x_2}^2 \quad (26)$$

It is explicit from the equations (24)-(26) that the proposed class of estimator $\bar{y}_F^*(t)$ is more efficient than:

(i) The usual unbiased estimator $\bar{y}^*$;

(ii) The estimator $\bar{y}_C^*$ when $K_{yx_2} < 1$; and

(iii) The estimator $\bar{y}_R^*$, the ratio type estimator proposed by Khare and Srivastava (1993),

Tabasum and Khan (2004) and Okafor and Lee (2000).

Thus, it may be concluded that the general chain ratio type class of proposed strategy $\left[D, \bar{y}_F^*(t)\right]$ is more efficient than the usual unbiased estimator $\bar{y}^*$, the estimator $\bar{y}_C^*$ and the ratio type estimator $\bar{y}_R^*$.

Empirical Study

To examine the effectiveness of the suggested class of chain ratio types, data sets studied by Khare and Sinha (2007) are considered. The data, from the Department of Paediatrics, Banaras Hindu University during 1983-1984, is the physical growth of an upper socio economic group of 95 school age children of Varanasi under ICMR study. The first 25% (i.e., 24 children) have been considered as non-responding units. The descriptions of the variates are given below:

Population I:

y : Height (in cm.) of the children,

x_1 : Skull circumference (cm) of the children,

x_2 : Chest circumference (cm) of the children.

For this population:

$$\bar{Y} = 115.9526,$$
$$\bar{X}_1 = 51.1726,$$
$$\bar{X}_2 = 55.8611,$$
$$S_y^2 = 35.6041,$$
$$S_{x_1}^2 = 2.3662,$$
$$S_{x_2}^2 = 10.7155,$$
$$S_{y(2)}^2 = 26.0532,$$
$$S_{x_1(2)}^2 = 1.6079,$$
$$S_{x_2(2)}^2 = 9.1060,$$
$$\rho_{yx_1} = 0.3740,$$
$$\rho_{yx_2} = 0.620,$$
$$\rho_{yx_1(2)} = 0.571,$$
$$\rho_{yx_2(2)} = 0.401,$$
$$\rho_{x_1x_2} = 0.2970,$$
$$\rho_{x_1x_2(2)} = 0.570,$$
$$N = 95,$$
$$n = 35,$$
$$n' = 45$$

Population II:

y : Weight (kg) of the children,

x_1 : Chest circumference (cm) of the children,

x_2 : Mid-arm circumference (cm) of the children.

For this population,

$$\bar{Y} = 19.4968,$$
$$\bar{X}_1 = 55.8611,$$
$$\bar{X}_2 = 16.7968,$$
$$S_y^2 = 9.2662,$$
$$S_{x_1}^2 = 10.7155,$$
$$S_{x_2}^2 = 2.1115,$$
$$S_{y(2)}^2 = 5.5424,$$
$$S_{x_1(2)}^2 = 9.1060,$$
$$S_{x_2(2)}^2 = 1.4323,$$
$$\rho_{yx_1} = 0.846,$$
$$\rho_{yx_2} = 0.797,$$
$$\rho_{yx_1(2)} = 0.729,$$

$$\rho_{yx_2(2)} = 0.757,$$
$$\rho_{x_1x_2} = 0.725,$$
$$\rho_{x_1x_2(2)} = 0.641,$$
$$N = 95,$$
$$n = 35,$$
$$n' = 45$$

The percent relative efficiencies (PREs) of the estimators $\bar{y}_R^*$ and $\bar{y}_C^*$ have been computed along with the proposed estimator $\bar{y}_F^*(t)$ at its optimum with respect to the usual unbiased estimator $\bar{y}^*$ for two data sets for different values of k ; results are displayed in Table 1.

Results and Conclusion

Table 1 shows that the percent relative efficiency (*PRE*) of the proposed estimator $\bar{y}_F^*(t)$ at its optimum with respect to $\bar{y}^*$ is at its maximum over the ratio estimator $\bar{y}_R^*$ and the estimator $\bar{y}_C^*$ in both the populations. In population I, the *PREs* of all the estimators decreases with the increase in the value of k while in population II, the *PREs* of all the estimators increases with the increase in the value of k. Further, it is envisaged that the estimator $\left\{\bar{y}_F^*(t)\right\}_{opt}$ is the best estimator among $\bar{y}^*$, $\bar{y}_R^*$ and $\bar{y}_C^*$ in both the populations.

Table 1: Percent Relative Efficiencies (*PREs*) of the Different Estimators with Respect to $\bar{y}^*$ for Different Values of k

Estimator	Population - I				Population - II			
	$(1/k)$							
	$(1/5)$	$(1/4)$	$(1/3)$	$(1/2)$	$(1/5)$	$(1/4)$	$(1/3)$	$(1/2)$
$\bar{y}^*$	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
$\bar{y}_R^*$	124.27	121.21	117.27	112.01	129.65	129.71	129.79	129.92
$\bar{y}_C^*$	144.83	144.27	143.52	142.45	168.52	175.91	186.72	204.05
$\left\{\bar{y}_F^*(t)\right\}_{opt}$	145.16	144.66	143.97	142.99	178.81	188.82	203.89	229.16

References

Cochran, W. G. (1977). *Sampling techniques*, (*3rd Ed.*). New York: John Wiley & Sons.

Hansen, M. H., & Hurwitz, W. N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, *41*, 517-529.

Hidiroglou, M. A., & Särndal, C. E. (1998). Use of auxiliary information for two phase sampling. *Survey Methodology*, 873-878.

Khare, B. B., & Sinha, R. R. (2007). Estimation of the ratio of the two population means using multi auxiliary characters in the presence of non-response. *Statistical Technology* In *Life Testing, Reliability, Sampling Theory and Quality Control*, 163-171.

Khare, B. B., & Srivastava, S. (1993). Estimation of population mean using auxiliary character in presence of non-response. *National Academy of Science and Letters India*, *16*, 111-114.

Khare, B. B., & Srivastava, S. (1995). Study of conventional and alternative two-phase sampling ratio, product and regression estimators in presence of non-response. *Proceedings of the Indian National Science Academy*, *65*, 195-203.

Khare, B. B., & Srivastava, S. (1997). Transformed ratio type estimators for the population mean in the presence of non response. *Communications in Statistical Theory and Methods*, *26*, 1779-1791.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, *33*, 101-116.

Okafor, F. C., & Lee, H. (2000). Double sampling for ratio and regression estimation with sub-sampling the non-respondents. *Survey Methodology*, *26*(*2*), 183-188.

Rao, P. S. R. S. (1986). Ratio estimation with sub sampling the non- respondents. *Survey Methodology*, *12*, 217-230.

Rao, P. S. R. S. (1987). Ratio and regression estimates with sub sampling the non - respondents. *Paper presented at a special contributed session of the International Statistical Association Meeting, Sept., 2-16, 1987*, Tokyo, Japan.

Singh, H. P., & Kumar, S. (2008 a). Estimation of mean in presence of non-response using two phase sampling scheme. *Statistics Papers*, *DOI 10.1007/s00362-008-040-5.*

Singh, H. P., & Kumar, S. (2008 b). A regression approach to the estimation of finite population mean in presence of non-response. *Australian and New Zealand Journal of Statistics*, *50*(*4*), 395-408.

Singh, H. P., & Kumar, S. (2008 c). A general family of estimators of finite population ratio, product and mean using two phase sampling scheme in the presence of non-response. *Journal of Statistical Theory and Practice*, *2*(*4*), 677-692.

Singh, H. P., & Kumar, S. (2009 a). A general class of estimators of the population mean in survey sampling using auxiliary information with sub sampling the non-respondents. *Korean Journal of Applied Statistics*, *22*(*2*), 387-402.

Singh, H. P., & Kumar, S. (2009 b). A General Procedure of Estimating the Population Mean in the Presence of Non-Response under Double Sampling using Auxiliary Information. *SORT*, *33*(*1*), 71-84.

Singh, H. P., & Espejo, R. M. (2007). Double sampling ratio-product estimator of a finite population mean in sample survey. *Journal of Applied Statistics*, *34*(*1*), 71-85.

Singh, V. K., & Shukla, D. (1987). One parameter family of factor type ratio estimator. *Metron*, *45*(*1, 2*), 273-283.

Singh, V. K., Singh, G. N., & Shukla, D. (1994). A class of chain ratio type estimators with two auxiliary variables under double sampling scheme. *Sankhya*, *56*(*2B*), 209-221.

Sitter, R. R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, *37*, 95-102.

Tabasum, R., & Khan, I. A. (2004). Double sampling for ratio estimation with non-response. *Journal of the Indian Society of Agricultural Statistics*, *58*, 300-306.

Tabasum, R., & Khan, I. A. (2006). Double sampling ratio estimator for the population mean in presence of non-response. *Assam Statistics*, *20*, 73-83.

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 520-535

1538 – 9472/10/$95.00

A GA-Based Sales Forecasting Model Incorporating Promotion Factors

Li-Chih Wang
Tunghai University
Taichung, Taiwan ROC

Chin-Lien Wang
Ling Tung University
Taichung, Taiwan ROC

Because promotions are critical factors highly related to product sales of consumer packaged goods (CPG) companies, predictors concerning sales forecast of CPG products must take promotions into consideration. Decomposition regression incorporating contextual factors offers a method for exploiting both reliability of statistical forecasting and flexibility of judgmental forecasting employing domain knowledge. However, it suffers from collinearity causing poor performance in variable identification and parameter estimation with traditional ordinary least square (OLS). Empirical research evidence shows that - in the case of collinearity - in variable identification, parameter estimation, and out of sample forecasting, genetic algorithms (GA) as an estimator outperform OLS consistently and significantly based on a log-linear regression model concerning weekly sales forecasting of CPG products from a manufacturer in both busy and off seasons.

Key words: Sales forecasting, genetic algorithm, ordinary least square, collinearity, variance influence factor.

Introduction

Due to competition promotion has increasingly become a key factor of marketing in consumer packaged goods (CPG) industries because sales are highly related to promotion activities. To properly forecast unit sales of products for a particular company in the CPG industry, forecasters must take this contextual factor into account. Forecasts generated with most statistical models are consistent, but are usually devoid of the flexibility and comprehensiveness of contextual information. The lack of contextual information is exploited with judgmental forecasting, thus predictors and users of the forecasts are often tormented with the issue of inconsistency due to bias. These issues are clearly pointed out by Sanders and Ritzman (1992), Armstrong and Collopy (1998), Webby, et al. (2001) and De Gooijer and Hyndman (2006) among many others.

Regression is a natural choice to connect both methods (Edmunson, 1990; Bunn & Wright, 1991; Armstrong, et al., 2005; Nikolopoulos, et al., 2006), because regression is able to incorporate critical contextual factors into the model and produces consistent results. In regression modeling, the classical ordinary least square (OLS) still is one of the most widely used estimators to identify significant factors and estimate parameters in linear regression (Draper & Smith, 1998; Rawlings, et al., 1998). However, it suffers from limitations posed by issues of outliers (Cook, 1977; Rawlings, et al., 1998; Meloun & Militky, 2001), sample size (Belsley, et al., 1980; Belsley, 1982; Yu, 2000) and multi-collinearity.

Multi-collinearity is the condition of one predictor variable which can be expressed as the exact or near linear combination of other predictor variables (Gunst & Mason, 1977) in case of small size sample, regression models with highly correlated independent variables, and groups of dummy variables or sporadic

Li-Chih Wang is professor in the Department of Industrial Engineering and Enterprise information and the dean of the School of Engineering at Tonghai University, Taiwan, ROC. Email: wanglc@thu.edu.tw. Chin-Lien Wang is a lecturer in the department of Business Administration at Ling Tung University, Taiwan, ROC. Email: love123wang@gmail.com.

variables. As noted by Smith and Campbell (1980) it is ultimately caused by too little variation in predictor variables in the dataset which induce inflated variance of variable coefficients. Moreover, it usually causes many problems such as truly critical variables to become insignificant (Hendry, 2000) and incorrect parameter estimation in both sign and magnitude (Slinker, 1985), these problems will usually lead analysis and inference, as well as forecasting of the regression model to be out of track.

To address the issue of collinearity, an alternative parameter estimator called genetic algorithm (GA) is proposed; GA is an option to alleviate collinearity problems and obtain desired results with efficiency. This study begins with a log-linear regression model incorporating price and a group of non-price promotion related dummy variables (Kumar & Pereira, 1997; Heerde, et al., 2002a, 2002b). The model's effect parameters are assessed and decomposed with GA incorporating a fitness function of mean absolute percentage error (MAPE), which - without the square operation of errors. James and Stein (1961) exhibited an estimator which, under squared error loss, dominates the least squares estimator and, coupled with a realistic constraint on coefficient of variables, it is believed will - to some extent - avoid the issue of inflated influence of outliers and problems caused by collinearity in OLS.

This article proposes GA as an adequate alternative model estimator in regression modeling, particularly in situations of serious collinearity, through a comparative study of OLS and GA in in-sample parameters estimation and out-of-sample forecasting, respectively, with an empirical study on weekly unit sales forecasting of CPG products from a name brand manufacturer.

Methodology

Formulation of a Regression Model

Equation (1) of the multiplicative regression model is motivated by Wittink et al.'s analytical models in a series of articles (Foekens, et al., 1999; Heerde, et al., 2002a, 2002b). Regression modeling uses a stepwise method called backward elimination (Draper & Smith, 1998), starting from the model incorporating all critical factors considered, then removes insignificant variables one by one iteratively. The model can be formulated as

$$S_{it} = \lambda_{it} \left(P_{it} / \widehat{P}_i \right)^{\theta_{it}} \prod_{l=1}^{n} \mu_{l\,it}{}^{D_{l\,it}} \ell^{\varepsilon_{it}}, \ \forall t \in Q \quad (1)$$

where,

i denotes an item number, $i = 1, 2, 3, \ldots, I$;
t denotes specific number of period referenced;
$1 \le t \le T$, T is the total number of normal periods;
I is the total number of items involved;
Q denotes the set of referenced periods;
S_{it} is the total unit sales of the item i in period t under a retailer, for weekly sales, t actually represents a certain week in the referenced periods;
λ_{it} denotes the normal unit sales (base sale) of the item i in period t without any promotion under a retailer;
$\widehat{P}_i$ is the list price of item i;
P_{it} is the discount price of item i during period t under a retailer;
θ_{it} denotes the coefficient of price elasticity of item i during period t under a retailer;
D denotes an indicator parameter(or dummy variable) of non-price promotion mix;
$D_{l\,it}$ is the l-th component of a vector of n indicator parameters of non-price promotion mix ($D_{1\,it}, D_{2\,it}, \ldots, D_{n\,it}$) of item i in period t. $D_{l\,it} = 1$ denotes a promotion mix of type l arises, the default value of $D_{l\,it} = 0$;
$\mu_{l\,it}$ denotes the non-price promotion effect parameter (multiplier) of corresponding non-price promotion mix ($D_{l\,it}$) of item i during normal period t under a retailer; and
ε_{it} denotes the residual error.

Taking the natural logarithm in both sides of (1) results in the following:

$$ln S_{it} = ln \lambda_{it} + \theta_{it} ln\left(P_{it} / \hat{P}_i\right) + \sum_{l=1}^{n} D_{lit} ln \mu_{it} + \varepsilon_{it};$$
$$\forall t \in Q \qquad (2)$$

A nonlinear model such as (1) is transformed to a linear regression model (Carroll & Ruppert, 1988; Franses & McAleer, 1998), which is the underlying model to conduct model fitting and model checking in this study.

Model Fitting: Parameter Estimation with GA

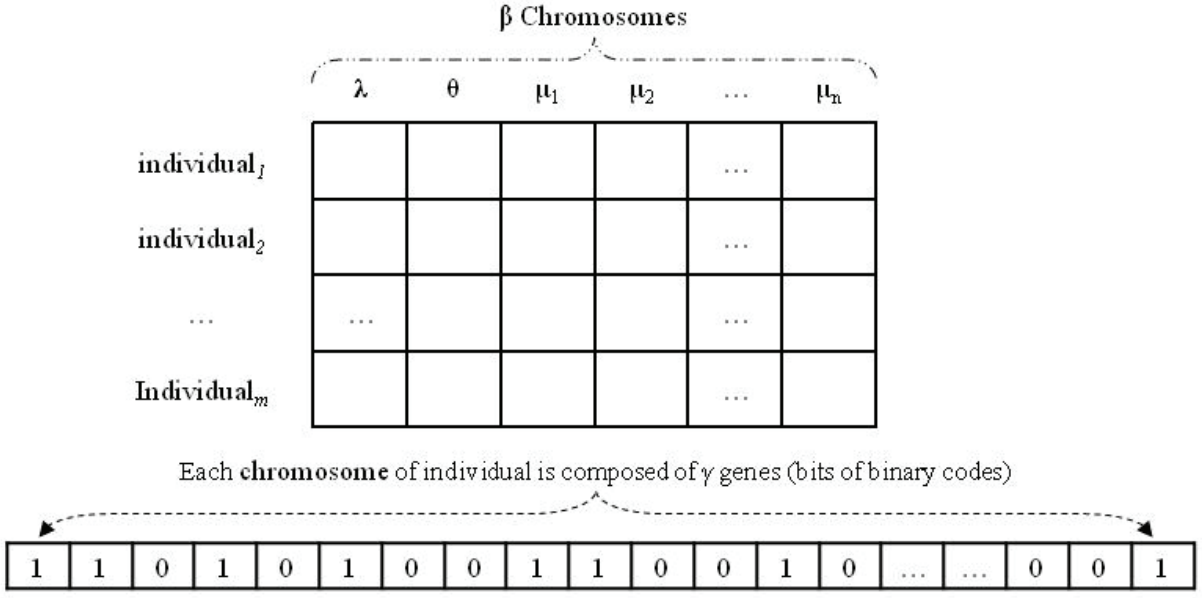
The genetic algorithm (GA) is proposed as a regression estimator to identify critical variables and estimate coefficients of variables as opposed to the widely employed least square type of estimators in situations of small sample size or a model mainly composed of dummy variables and sporadic variables.

Features of the Genetic Algorithm (GA)

GA simulates Darwin's biological evolution by selecting encoded individuals (solutions) in the population with higher fitness (via a fitness function) through stochastic crossover and mutation to generate a population of individuals (reproduction) more fitted to the environment (better solutions) from generation to generation (Holland, 1992; Goldberg, 1987, 1989). In estimating parameters of complicated multivariate nonlinear models, GA is generally considered to be better than other alternatives such as nonlinear least square and maximum likelihood estimation due to its parallel search capability (Schaffer, et al., 1989; Eiben & Michalewicz, 1999), even based on a small size dataset it is capable of deriving satisfying results.

The initial population is randomly created in the encoded form of a binary matrix, there exist *m* rows, each row of binary string in the matrix is an individual (solution) which encompasses β chromosomes, each chromosome represents a parameter and is composed of γ genes, each gene is represented by a binary code (See Figure 1). Each individual is evaluated by the fitness function as shown in Equation (3), in each generation, the best α% ($1 \leqq \alpha \leqq 6$) of the population are kept as elites to the next generation, the remaining population is created by randomly selected pairs of individuals conducting a one-point crossover within each chromosome of such pairs to reproduce offspring, forming a random recombination of individuals' ingredients of genes, to search for a new solution space and possibly a better solution. After this, a one-bit mutation is performed with a view to creating new pieces of

Figure 1: The Composition of Population Generated Randomly in GA



gene originally not possessed by members of the population through randomly selected genes within each individual; this occasional random change in genes could open the door to new possibilities of better solutions. Afterwards, each encoded individual in the population is decoded back to a string of real numbers of parameters and each individual is evaluated by the fitness function, this iterative process repeats until a termination condition is met.

Parameters such as crossover probability (P_c) and mutation probability (P_m) of GA are designed to vary with the number of generations processed or others, such as moving average percent of improvement (MAPI) in fitness function value within certain number of generations, to keep proper diversity of the population while retaining the convergence capability, to circumvent getting stuck too early at local solutions in its search process (Liu, et al, 2003; Pham & Karaboga, 1997).

Based on (2), the fitness function of GA may be formulated as

$$FV_i = MAPE_i = \frac{(\sum_{t=}^{T} \left| ln\, S_{it} - ln\, \tilde{S}_{it} \right| / ln\, S_{it})}{T}, \ \forall t \in Q \quad (3)$$

where the term $\left| \ln S_{it} - \ln \widetilde{S}_{it} \right|$ is the absolute value of difference between the natural logarithm of the actual sales volume ($\ln S_{it}$) of the i^{th} item and natural logarithm of the estimated sales volume ($\ln \widetilde{S}_{it}$) of the same item in period *t*. *T* denotes the number of normal periods. The objective of GA is to find a solution with the minimal $MAPE_i$. The smallest $MAPE_i$ found is updated once a smaller one is found in the solution search process. After model fitting, every effect parameter in (2) is derived in real value.

Model Checking

A regression diagnostics focused on normality and independence is performed to determine if critical assumptions of linear regression are violated, based on Equation (2). If these assumptions are severely violated, particularly if collinearity arises among predictor variables, bias may be a serious issue in model fitting or in model specification.

The normality test is conducted through a one-sample Kolmogorov-Smirnov test (Lilliefors, 1967) and a Q-Q plot (Berilant, et al., 2005). An independence test in this study consists of two parts, namely, a multi-collinearity test and an autocorrelation test. The former is performed via variance inflation factor (VIF), whereas the latter is performed via Durbin-Watson (D-W) test (Savin & White, 1977; Draper & Smith, 1998). VIF is one of the most popular measures used to detect collinearity in the literature (Belsley, et al., 1980; Belsley, 1982; Stine, 1995), which can be derived via regression of one predictor variable to all other predictors and can be formulated as

$$VIF_j = 1 / (1 - R^2_j). \quad j = n + 2. \quad (4)$$

where n denotes the number of types of non-price promotion mixes and R^2_j is the coefficient of determination from regression of the j^{th} predictor variable on the other predictor variables. As described in Theil (1971) and Berk (1977), estimated effect parameters can be directly proportional to VIF_j as the following equation:

$$s^2(\hat{\beta}_j) = VIF_j(\sigma^2 / (T-1)V^2{}_j) \quad (5)$$

where $\hat{\beta}_j$ denotes the j^{th} effect parameters in equation (2), $s^2(\hat{\beta}_j)$, σ^2, and $V^2{}_j$ is the variance of $\hat{\beta}_j$ .and variance of regression errors, as well as the variance of the j^{th} predictor variable, respectively. T denotes the number of periods in the training period and can be perceived as sample size.

The D-W test focuses on testing whether any autocorrelation exists among the following series of regression error terms in equation (2) : $\varepsilon_{it}, \varepsilon_{it-1}, ..., \varepsilon_{i1}$. The statistic can be formulated as

$$d = \sum_{t=2}^{T} (\varepsilon_{it} - \varepsilon_{it-1})^2 / \sum_{t=1}^{T} \varepsilon^2{}_{it} \quad (6)$$

In general, as the serial correlation increases, d decreases.

The Re-composition of Variable Coefficients Estimated

The cycle length of CPG industry is about 52 weeks long, thus, let $t' = t + 52$, denoting the corresponding week to be forecasted in a new year. A naïve sales forecasting method considering cycle length to forecast unit sales of item i of period t' in a new year (see Williams, 1987), based on sales data of week t in the referenced year, would be

$$\ln \hat{S}_{it'} = \ln \eta_i + \ln \pi_{it} + \theta_{it'} \ln\left(\frac{P_{it'}}{\hat{P}_i}\right) + \sum_{l=1}^{n} D_{lit'} \ln \mu_{it'},$$
$$t' = t + 52, \forall t' \in Z \tag{7}$$

where η_i denotes the average normal sale of item i across referenced periods, π_{it} denotes the seasonal index of item i in period t, and Z denotes the set of periods to be forecasted. All parameters in equation (7) are derived either with GA or OLS. Let $e_{1it'}$ denote the price effect multiplier of item i in forecasting period t' and $e_{2\ it'}$ denote the effect multiplier of a non-price promotion mix. In each group of indicator parameters one condition at most will arise in each period, resulting in

$$\ln \hat{S}_{it'} = \ln \eta_i + \ln \pi_{it} + \ln e_{1it} + \ln e_{2it},$$
$$t' = t + 52, \forall t' \in Z \tag{8}$$

In its re-composed form, equation (8) can be used to forecast weekly unit sales. Parameters estimated through GA or OLS - based on observations in the training periods - can be recombined as in equation (8) to respond to expected promotional campaigns in the forecasting horizon (as specified in the promotion proposals) to perform out of sample forecasting with $\ln \hat{S}_{it'}$ being transformed back to $\hat{S}_{it'}$ in the following empirical study.

Empirical Study: Background

This study focuses on the forecast of weekly sales volume for several series of CPG products, manufactured by Company F, under retailer B. Company F is a leading manufacturer specialized in dehumidifier and deodorizer products in Taiwan, and retailer B is an international outlet of DIY products. A sales data set of 10 items from 2007 and the first 4 months in 2008, aggregated from retailer B's outlets, coupled with price promotion, non-price promotion, and promotion proposals, are used to conduct the empirical study. The details of price rate and type of non-price promotion mix of these items are displayed in Tables 1a and 1 b. Each effect parameter is set to be constrained within a specific range in GA which was implemented in Matlab 6.5, for example, the price elasticity coefficient is set to be in the range of [-8, 0], while effect parameters of non-price promotion mixes are set to be between 1 and 5. However, the coefficients of predictor variables in OLS regression are estimated without any constraint in the statistical package SPSS 13.

Empirical Study: Experimental Design

In order to take both the busy season and off season into account and to have a proper assessment of the performance of both estimators, the forecasting horizon is designed to consist of two periods of equal duration, the first period includes the first 6 weeks of 2008 (one of the major busy seasons in that year) and the second period starts from the 11th week and ends at the16th week of 2008 (one of the off seasons in that same year). The 10 product items manufactured by a name brand company of CPG products in Taiwan, in retailer B's outlets are the forecasting target in the empirical research.

To properly evaluate the performance of parameter estimation via GA and OLS as well as that of out-of-sample forecasting based on parameters derived from GA and OLS, respectively, particularly the consistency of performance, model fitting and checking is conducted with GA first and then with OLS consecutively, all based on the dataset of the entire year of 2007 as the first training period; this is a small period, thus, the training dataset in this period can be denoted as a small sample. The dataset for 2007 combined with the first 10 weeks of 2008 is the second training period is longer than the first, thus, the training dataset in forecast weekly unit sales of items of concern in the forecasting horizon.

Table 1a: Summary of Promotion Proposals for Year 2007 of Company F's Products under Retailer B

Item	Product Type	2007 Promotion Sessions, Content Denoted as ($P_{it} / \widehat{P}_i, D_l$)*							
		12/29-2/28	3/29-4/24	4/26-6/12	6/14-8/07	8/09-9/11	9/13-11/13	11/15-12/15	10/01-12/31
1	Deodorizer	1, D_4	1, D_1	1, D_1	89/99, D_3	1, D2	85/99, D6	1, D_1	79.5/99, D_7
2	Deodorizer	59/65, D_5	1, D_1	1, D_1	59/65, D_3	1, D_1	1, D_1	59/65, D_6	49.5/65, D_7
3	Deodorizer	1, D_4	1, D_1	1, D_1	119/138, D_2	119/138, D_2	1, D_1	1, D_1	99.5/138, D_7
4	Dehumidifier	75/89, D_5	75/89, D_3	75/89, D_2	75/89, D_2	75/89, D_2	75/89, D_3	1, D_1	75/89, D_7
5	Dehumidifier	89/95, D_5	1, D_1	1, D_1	1, D_1	89/95, D_2	89/95, D_3	89/95, D_6	89/95, D_7
6	Cleaner	90/109, D_5	90/109, D_1	90/109, D_1	90/109, D_2	89/109, D_2	1, D_6	90/109, D_6	1, D_6
7	Cleaner	85/89, D_5	1, D_1	85/89, D_3	85/89, D_3	85/89, D_3	1, D_6	1, D_2	1, D_6
8	Cleaner	195/219, D_5	1, D_1	195/219, D_5	1, D_1	195/219, D_3	195/219, D_6	195/218. D_6	189/219, D_7
9	Insect Pest	79/99, D_5	79/99, D_1	70/99, D_1	70/99, D_1	1, D_1	1, D_1	79/99, D_6	1, D_1
10	Insect Pest	52/65, D_4	1, D_1	52/65, D_1	52/65, D_3	52/65, D_1	49/65, D_6	49/65, D_6	44.5/65, D_7

*Details of D_l, l = 1, 2, 3, 4, 5, 6, 7, can be checked in the Formulation of a Regression Model description

Table 1b: Summary of Promotion Proposals for Year 2008 of Company F's Products under Retailer B

Item	Product Type	2008 Promotion Sessions, Content Denoted as ($P_{it} / \widehat{P}_i$, D_l)	
		12/27-2/12	2/14-4/1
1	Deodorizer	85/99, D_4	1, D_1
2	Deodorizer	55/65, D_5	1, D_1
3	Deodorizer	1, D_4	1, D_1
4	Dehumidifier	1, D_5	1, D_1
5	Dehumidifier	75/95, D_5	1, D_1
6	Cleaner	89/109, D_5	1, D_1
7	Cleaner	85/89, D_5	1, D_1
8	Cleaner	169/219, D_5	1, D_1
9	Insect Pest	1, D_5	1, D_1
10	Insect Pest	52/65, D_4	1, D_1

this period can be denoted as a large sample. Parameters derived from either estimator based both small sample and large sample are used to

Results

Results of Model Fitting

The details of model fitting results are shown in Figure 4 and Tables A1-A4 in the Appendix. Tables A1-A2 are concerned with parameters estimated with GA on small sample and large sample, respectively, while Tables A3-A4 are concerned with parameters estimated with OLS on these two samples respectively. Most parameters derived from GA are consistent with expectations, such as, the effect parameters of μ_1 to μ_3 increase from 2.182 to 2.287 in busy-season periods and increase from 2.277 to 2.796 in off-season periods. This may be explained by more effort being made and more expenditure for promotions therein; μ_5 is larger than μ_4 because non-price promotion type 5 employs direct mail in addition to all aspects included in type 4, μ_7 is larger than μ_6 for the same reason. Effect parameters estimated by OLS also are roughly consistent with expectations; their magnitudes are much smaller than expected, however. For example, many are smaller than 1 which indicates a negative effect in promotion and seems unreasonable based on experience (see Tables A3-A4).

Nearly every intercept (normal sales) is inflated to the extent that it exceeds the unit sales of an item in a certain period and becomes difficult to explain based on daily life experience. However, the issue of difficult explanation for parameters derived (Mandel, 2007) is very common in least square type of estimators, including weighted least square and partial least square, in addition to OLS. Often critical variables are deleted from the model by OLS, for example, 3 variables are removed for item 8 based on small sample, price elasticity of item 5 is discarded in both samples, and in items 3 and 6 price elasticity coefficients are deleted in the model by OLS. These phenomena can lead to a dilemma of incapability to take advantage of certain domain knowledge or contextual information. Moreover, the price elasticity coefficients of item 2 from OLS in the large sample are positive (see Table A4) - a phenomenon which goes against common sense, but the underlying reasons are now investigated.

A Comparative Analysis of Results in Model Checking via VIF and T-W Tests

The normality test, consisting of the one-sample Kolmogorov-Smirnov test and Q-Q plot, in which both GA and OLS passed the test with data from both small and large samples without difficulty. The independence test measures of VIF and the results of D-W tests, however, showed complex but interesting consequences in two training periods of different length via GA and OLS and warranted further investigation. As shown in Tables 2 and 3 the number displayed in each cell of these tables is the average VIF of a specific effect parameter of a certain item. The number in the cell in the right hand side column in the table is the mean of the average VIF for each item concerned.

Note that the mean of the average VIF in the first training period is much larger than that of the second training period, even though not every mean of the average VIFs in the first training period is necessarily bigger than its counterpart in the second training period. Some outliers arising in the first training period considerably increase the relevant measure. However, as Smith and Campbell (1980) note, although VIF can identify the source of inadequate parameter estimation, it cannot measure the amount of imprecision.

Because the main difference of the two training datasets is the sample size, one is 47 (5 cases are discarded as outliers in mixed periods which include two different kinds of promotions in a single week), whereas the other one is 56 (6 cases are discarded). The large sample seems to enable the predictor variables to have more changes in values within the dataset to alleviate the collinearity issue arising in the small sample based model. For example, as shown in Tables 2-3, the mean of average VIF reduces from 6.196 to 3.286 and the standard deviation reduces from 7.662 to 1.242 as the sample size increases from 47 to 56.

Table 2: The Results of Average VIF of Each Predictor Variable for Each Item Based on Small Sample

Item	Average VIF								
	Price Elasticity θ	Pro-Mix μ_1	Pro-Mix μ_2	Pro-Mix μ_3	Pro-Mix μ_4	Pro-Mix μ_5	Pro-Mix μ_6	Pro-Mix μ_7	Mean
1	1.523	20.139	15.038	19.374	--	23.929	28.459	19.104	18.224
2	1.293	2.357	2.230	2.275	2.265	--	3.785	3.708	2.559
3	1.072	1.160	1.143	2.134	2.922	--	--	1.046	1.580
4	1.345	2.801	--	2.382	3.111	--	2.868	4.636	2.857
5	7.902	2.547	1.290	7.737	--	7.664	1.257	--	4.733
6	1.376	1.483	1.900	1.605	--	1.605	1.312	1.436	1.531
7	1.265	3.993	4.992	--	--	6.996	3.724	--	4.194
8	1.410	1.521	1.655	--	--	1.958	1.673	--	1.643
9	1.191	2.099	1.972	--	2.278	--	2.278	--	1.964
10	1.229	40.567	21.733	27.930	--	21.486	23.077	--	22.670

As a result, the problem of deletion of critical predictor variables in model fitting with OLS seems to have been improved. For example, from item 2 to item 10, the number of discarded predictors decreases row by row with the exception of the item 9 row in both tables. In addition, the quality of parameter estimation conducted by OLS also seems to improve particularly for items with mean VIF greater than 10 (Craney & Surles, 2002), such as items 1 and 10 in the small sample, reduced to around 3 or less based on large sample (see Tables A3-A4 in the Appendix).

The much more serious issue of deleting predictor variables and the downgrade of parameter estimation quality owing to sample size change did not occur in model fitting with GA (see Tables A1-A2 in the Appendix), however. Compared with OLS, GA shows better and more consistent behavior in model fitting. The problems caused by the occurrence of collinearity among predictor variables of models based on a smaller dataset did not affect GA to a great degree in its parameter estimation. The reason may be attributed to the flexibility GA has in dealing with the dataset to comply with its purposes through the formulation of a fitness function and the constraint of variable coefficients.

A comparative analysis was conducted as shown in Table 4 in which a D-W test was performed to check if any serious problem of serial correlation arising in the error terms of model fitting occurred. Roughly speaking, no big change concerning the condition of autocorrelation among sequential series of errors created in model fitting with GA based on different size of samples was observed. The number of cases rejected in the D-W test is 3 in the small sample, while in the large sample the number increases to 4. There exists an obvious change in the results of model fitting with OLS in this regard, the number of null hypotheses, H_0, rejected in the small sample is 6 out of 10, for large sample the number of rejected test cases reduces to 3. Apparently, for the small sample the condition of autocorrelation of regression errors created by OLS is more serious than that based on the large sample. However, regression errors created by GA did not show same kind of change between the small sample and large sample.

Table 3: The Results of Average VIF of Each Item

Item	Average VIF								
	Price Elasticity θ	Pro-Mix μ_1	Pro-Mix μ_2	Pro-Mix μ_3	Pro-Mix μ_4	Pro-Mix μ_5	Pro-Mix μ_6	Pro-Mix μ_7	Mean
1	1.337	4.413	2.896	2.544	--	3.775	4.863	3.091	3.274
2	1.269	1.907	1.923	2.038	1.685	--	3.692	3/211	2.086
3	1.326	7.514	8.089	6.148	6.295	--	--	1.882	5.209
4	1.322	2.906	--	2.694	2.969	--	3.362	5.597	3.142
5	8.327	2.005	1.349	8.273	1.956	8.208	1.313	--	4.490
6	1.305	3.007	2.913	2.646	2.913	2.646	5.488	2.434	2.919
7	1.271	5.149	6.260	--	--	7.917	4.522	--	5.024
8	1.396	1.538	1.796	--	--	1.690	1.882	--	1.660
9	1.203	2.369	2.253	--	2.151	--	2.635	--	2.122
10	1.217	3.456	3.076	3.836	--	2.927	2.820	3.197	2.933

Table 4: Results of the Durbin Watson Test for GA and OLS Respectively in Two Data Samples

Item	Durbin-Watson Test							
	GA				OLS			
	Small Sample		Large Sample		Small Sample		Large Sample	
	d	Test Result	d	Test Result	d	Test Result	d	Test Result
1	1.048	Inconclusive	1.030	Reject H_0	1.888	Reject H_0	1.085	Reject H_0
2	1.579	Inconclusive	1.282	Inconclusive	1.888	Reject H_0	1.576	Inconclusive
3	0.922	Reject H_0	0.797	Reject H_0	1.026	Reject H_0	1.097	Reject H_0
4	1.544	Inconclusive	1.106	Reject H_0	1.625	Reject H_0	1.225	Inconclusive
5	1.799	Reject H_0	1.616	Inconclusive	1.591	Reject H_0	1.350	Inconclusive
6	1.247	Inconclusive	1.117	Inconclusive	1.333	Inconclusive	1.223	Inconclusive
7	1.375	Inconclusive	1.744	Reject H_0	1.349	Inconclusive	1.689	Reject H_0
8	1.385	Inconclusive	1.367	Inconclusive	1.517	Inconclusive	1.429	Inconclusive
9	1.664	Reject H_0	1.460	Inconclusive	1.910	Reject H_0	1.910	Inconclusive
10	1.237	Inconclusive	1.286	Inconclusive	1.489	Inconclusive	1.508	Inconclusive

Figure 2: Comparative Forecasting Performance Based on Parameters Generated with GA and OLS on Small Sample

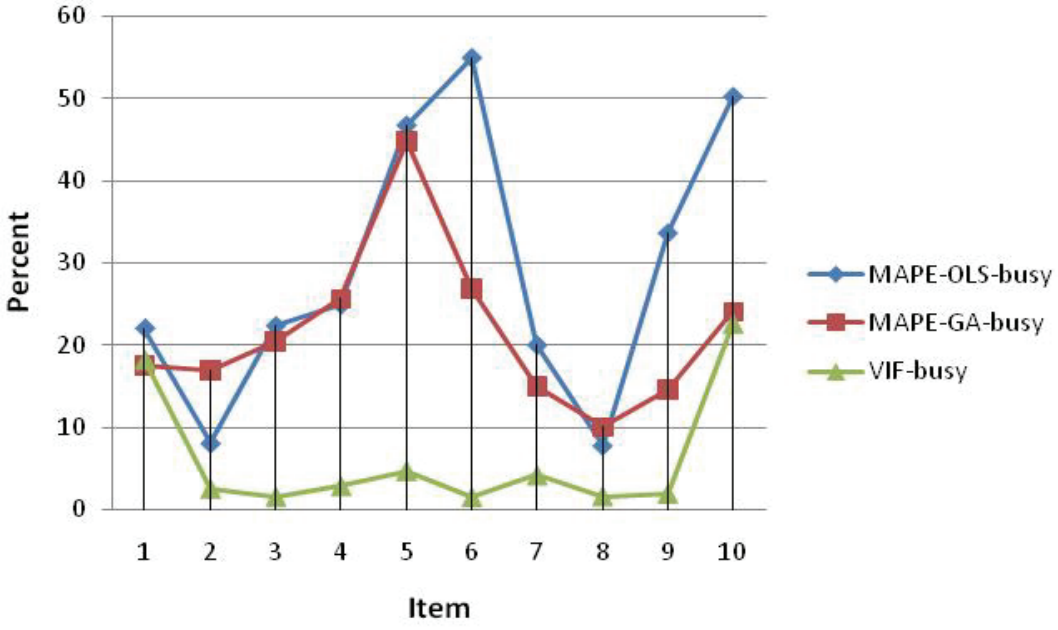


Table 5: Comparison of the Accuracy of Forecasting Based on Parameters Derived from GA and OLS

Item	MAPEs			
	Busy Season		Off Season	
	GA	OLS	GA	OLS
1	14.17%	22.09%	17.58%	26.53%
2	8.10%	8.10%	16.91%	20.94%
3	22.21%	22.38%	20.48%	20.68%
4	24.93%	24.93%	25.63%	26.89%
5	11.40%	46.67%	44.72%	47.58%
6	23.00%	54.86%	26.85%	37.33%
7	19.67%	20.05%	15.04%	43.75%
8	22.42%	7.81%	10.03%	17.21%
9	33.60%	33.60%	14.60%	14.50%
10	28.33%	50.17%	24.07%	27.14%
AVG	20.78%	29.07%	21.59%	28.26%

Predictor Variable for Each Item Based on Large Sample

The model, estimated with two different methods based on two different sized samples is shown in Table 5. The superiority of the forecasting performance of GA over that of its counterpart is obvious: for forecasts in the busy season with parameters estimated from a small sample, except 4 cases, in which 3 cases are ties, only in one case did forecasting based on effect parameters derived from GA lose its ground to forecasting based on parameters generated with OLS. For forecasts in the off season the margin widens, 9 out of 10 items with parameters estimated via GA have an edge over those assessed by OLS, in terms of MAPE, in both seasons (see Table 5).

In addition, a paired-samples t test was conducted between MAPEs of forecasting based on parameters derived from GA and OLS on small samples $t = -1.629$ at the $\alpha = 0.10$ significance level (1-tailed) and the critical value is 1.383. Thus, the null hypothesis that, on average, the MAPE of OLS is smaller than or equal to that of GA is rejected and the alternative hypothesis that, on average, the MAPE of OLS is greater than that of GA is supported. The same paired sample t test results in a $t = -2.459$, rejects H_0 and supports that, on average, the MAPE of OLS is greater than that of GA for a large sample at the 0.90 confidence level.

To further evaluate the effect of collinearity among predictor variables on the forecasting performance in either the busy or off season, two figures illustrate how and to what extent MAPE forecasting based on parameters derived from GA and OLS responds to the measure of VIF. For forecasting based on parameters derived from a small sample, on average, both GA and OLS show insignificant results between the MAPE of forecasts and the average VIF, 0.271 and 0.316, respectively, in the Pearson correlation test (2-tailed) with $\alpha = 0.05$ level (see Figure 2 and Table 6). Conversely, using the same test, forecasting GA A paired t test (1-tailed) was performed, with $\alpha = 0.05$, between parameters estimated with GA on the small sample and large sample,

the t value $= -0.547$ is greater than the critical value of -1.895, so it does not reject the H_0 that, on average, parameters estimated with GA based on small sample are less than or equal to parameters estimated with GA based on large sample. Conversely, a paired t test, with the same $\alpha=0.05$ between parameters estimated with OLS on the small and large sample results in $t = 7.551$, which is much greater than the critical value 1.895, thus it may be concluded at 95% confidence that, on average, parameters estimated with OLS on a small sample are greater than that for a large sample. Based on the above information, model parameters estimated with GA appear more stable than parameters and OLS based on parameters generated from large sample shows a significant result; the correlation coefficients are 0.618 and 0.649, respectively, (see Figure 3 and Table 6).

No significant difference exists between the performance of forecasting based on parameters derived on large or small sample of data for either GA or OLS. In sum, collinearity makes regression modeling with OLS more sensitive to a change in sample size so that the correlation between VIF and MAPE becomes less obvious in a small sample. Becuase low VIF is a necessary condition for good forecasting performance (Williams, 1987), a change from small to large sample does not create a significant difference in forecasting performance in terms of MAPE regardless of whether GA or OLS is used as estimator of model parameters.

Conclusion

If a regression model is based on a limited size sample or if the variation of values in the dataset pertain to a specific critical variable that is too small, then the issue of collinearity will arise and make the model very sensitive to the sample size change and may negatively and seriously affect proper variable identification and variable coefficient assessment. Under such a situation, any analysis, inference or forecast based on the parameters of the model can be questionable. An alternative estimator, the genetic algorithm (GA), can - with proper formulation in fitness function and realistic constraints regarding coefficients of critical variables - have better and more consistent performance in both critical variable identification and variable coefficient estimation,

Figure 3: Comparative Forecasting Performance Based on Parameters Generated with GA and OLS on Large Sample

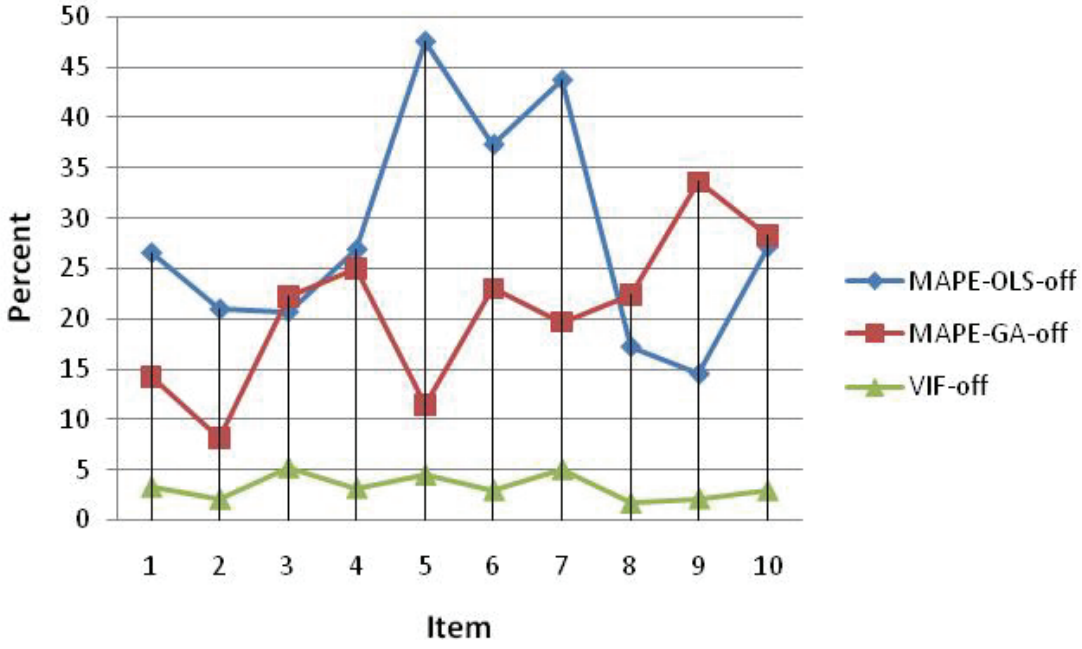
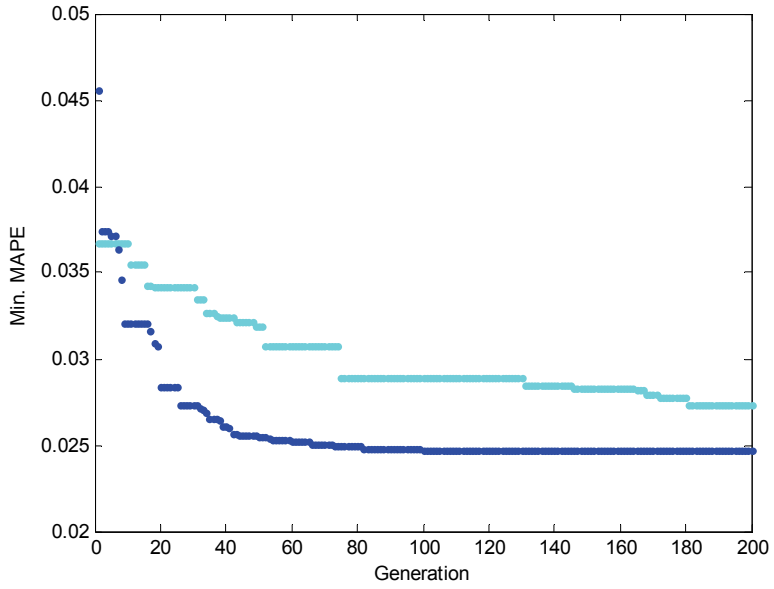


Table 6: Pearson Correlation Test Between VIF and MAPE

Estimator	Small Sample		Large Sample	
	Pearson Correlation	Significance (2 tailed) 0.05	Pearson Correlation	Significance (2 tailed) 0.05
GA	0.271	0.448	0.618	0.057
OLS	0.316	0.373	0.649	0.042

Figure 4: Typical Convergence Process of GA in this Study Compared to Generic GA



which can be verified via a series of measures, charts and model checking tests.

Empirical results support the points presented in this article via weekly unit sales forecasting based on a log-linear regression model of 10 CPG products from a name brand manufacturer in Taiwan in both a busy and an off season. More in-depth and wider investigations will be of great help to generalize points made in this article and to increase the amount of supporting data for use of the GA.

References

Armstrong, J. S., & Collopy, F. (1998). Integration of statistical methods and judgment for time series forecasting: principles from empirical research. In G. Wright & P. Goodwin (Eds.), *Forecasting with Judgment*. Chichester, England:John Willey.

Armstrong, J. S., Collopy, F., & Thomas, Y. J. (2005). Decomposition by causal forces: a procedure for forecasting complex time series. *International Journal of Forecasting*, *21*, 25-36.

Belsley, D.A. (1982) 'Assessing the presence of harmful collinearity and other forms of weak data through a test for signal-to-noise', *Journal of Econometrics*, 20, 211-253.

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980) *Regression Diagnostics: Identifying influential data and sources of collinearity*. New York: John Wiley & Sons.

Berilant, J., de Wet, T., & Goegebeur, Y. (2005). A goodness-of-fit statistic for Pareto-type behavior. *Journal of Computational and Applied Mathematics*, *186*(*1*), 99-116.

Bunn, D., & Wright, G. (1991). Interaction of judgmental and statistical forecasting methods: Issues and analysis. *Management Science*, *37*, 501-518.

Burk, K. N. (1977). Tolerance and condition in regression computations. *Journal of the American Statistical Association*, *72*, 863-866.

Carroll. R. J., & Ruppert. D. (1988). *Transformation and Weighting in Regression*. New York, NY: Chapman and Hall.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, *19*, 15-18.

Craney, T, A., & Surles, J, G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering*, *14(3)*, 391-403.

De Gooijer, J. G., & Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, *22*, 443-473.

Draper, N. R., & Smith, H. (1998). *Applied Regression Analysis*. New York, NY: John Wiley & Sons, Inc.

Edmundson, R. H. (1990). Decomposition: A strategy for judgmental forecasting, *Journal of Forecasting*, *9*, 305-314.

Eiben, A. E., Hinterding, R., & Michalewicz, Z. (1999). Parameter control in evolutionary algrithms. *IEEE Transactions on Evolutionary Computation*, *3(2)*, 124-141.

Foekens, E. W., Leeßang, P. H., & Wittink, D. R. (1999). Varying parameter models to accommodate dynamic promotion effects. *Journal of Econometric*s, *89*, 249-268.

Franses, P. H., & McAleer, M. (1998). Testing for unit roots and non-linear transformations. *Journal of Times Series Analysi*s, *19*, 147-164.

Goldberg, D. E. (1987). Simple genetic algorithms and the minimal deceptive problem, in: L. Davis (*Ed.*), *Genetic Algorithms and Simulated Annealing*. New York, NY: Hyperion Books.

Goldberg, D. E. (1989). *Genetic algorithms, in search, optimization & machine* learning. Boston, MA: Addison-Wesley.

Gunst, R. F., & Mason, R. L., (1977). Advantages of examining multicollinearitics in regression analysis . *Biometrics*, *33*, 249-260.

Hendry, D. F. (2000). Econometric modelling. Lecture notes for the PhD course in econometric modelling and economic forecasting, Department of Economics, University of Oslo.

Van Heerde, H. J., Leeflang, P. H., & Wittink, D. R. (2002a). How promotions work: Scan*Pro-based evolutionary model building. *Schmalenbach Business Review*, *54*, 198-220.

Van Heerde, H. J., Leeflang, P. H., & Wittink, D. R. (2002b). Flexible decomposition of price promotion effects using store-level scanner data. Schmalenbach Business Review, *54*, 198-220.

Holland, J. H. (1975). *Adaptation in natural and artificial systems*. Ann Arbor, MI: University of Michigan Press.

James, W., & Stein, C. (1961). Estimation with Quadratic Loss. In *Proceedings of the fourth Berkeley symposium mathematical statistics and probability, Vol. 1*. Berkeley, CA: University of California Press.

Kumar, V., & Pereira, A. (1997). Assessing the competitive impact of type, timing, frequency, and magnitude of retail promotions. *Journal of Business Research*, *40*, 1-13.

Lim, J. S., & O'Connor, M. (1996). Judgmental forecasting with time series and causal information. *International Journal of Forecasting*, *12*, 139-153.

Liu, Z., Zhou, J., & Lai, S. (2003). *New adaptive genetic algorithm based on ranking*. Proceedings of the second international conference on machine learning and cybernetics.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, *64*, 399-402.

Meloun, M., & Militky, J. (2001). Detection of single influential points in OLS regression model building. *Analytica Chimica Acta*, *439*, 169-191.

Mandel, Igor. (2007). *Statistical modeling and business expertise, or where is the truth?* Working paper, Advanced marketing models, New York, NY.

Nikolopoulos, K., Goodwin, P., Patelis, A., & Assimakopoulos, V. (2006). Forecasting with cue information: A comparison of multiple regression with alternative forecasting approaches. *European Journal of Operations Research*, *180*, 354-368.

Pham, D. T., & Karaboga, D. (1997). Genetic algorithms with variable mutation rates: Application to fuzzy logic controller design. Proceedings of the I MECH E Part I. *Journal of Systems & Control Engineering*, *211*(*2*), 157-167.

Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (1998). *Applied Regression Analysis-A Research Tool*. New York, NY: Springer-Verlag.

Sanders, N., & Ritzman, L. P. (1992). The need for contextual and technical knowledge in judgmental forecasting. *Journal of Behavioral Decision Making*, *5*, 39-52.

Savin, N. E., & White, K. J. (1977). The Durbin-Watson test for serial correlation with extreme sample sizes or many regressors. *Ecometrica*, *45*, 1989-1996.

Schaffer, J. D., Caruana, R. A., Eshelman, L. J., & Das, R. (1989). A study of control parameters affecting online performance of genetic algorithms for function optimization. In *Proceedings of the Third International Conference on Genetic Algorithms*, 51–60.

Slinker, B. K., & Glantz, S. A. (1985). Multiple regression for physiological data analysis: the problem of multi-collinearity. *American journal of Physiology*, *249*, R1-R12.

Smith, G., & Campbell, F. (1980). A critique of some ridge regression methods. *Journal of the American Statistical Association*, *75*(*369*), 74-81.

Stine, R. A. (1995). Graphical interpretation of variance inflation factors. *The American Statistician*, *45*, 53-56.

Theil, H. (1971). *Principles of econometrics*. New York, NY: John Wiley and Sons, Inc.

Webby, R., O'Connor, M., & Lawrence, M. (2001). Judgmental time series forecasting with domain knowledge. In Armstrong, J. S. (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners.* Norwall, MA: Kluwer Academic Publishers.

Williams, T. M. (1987). Adaptive Holt-Winters forecasting. *Journal of Operational Research Society*, *38 (6)*, 553-560.

Yu, C. H. (2000). An overview of remedial tools for collinearity in SAS. In *Proceedings of 2000 Western Users of SAS Software Conference*, 196-201.

Appendix

Table A1: Parameters Estimated via GA on Small Sample

Item	MAPE	Price Elas. θ	Normal Sales λ	Pro-Mix μ_1	Pro-Mix μ_2	Pro-Mix μ_3	Pro-Mix μ_4	Pro-Mix μ_5	Pro-Mix μ_6	Pro-Mix μ_7
1	0.124	-0.002	50.074	1.857	2.878	3.235	2.456	--	2.393	2.815
2	0.121	-0.0006	55.81	1.935	2.876	3.672	3.000	--	2.795	3.619
3	0.101	-0.273	110.62	2.190	--	2.193	--	4.163	--	2.340
4	0.107	-0.498	55.73	2.369	3.227	3.800	2.889	--	3.266	4.070
5	0.136	-0.488	200.00	2.750	2.300	1.614	--	3.436	1.969	3.951
6	0.121	-2.113	180.50	2.050	--	1.632	--	3.259	--	1.781
7	0.126	-2.749	79.09	1.808	--	2.698	3.132	--	2.720	2.834
8	0.092	-0.398	73.982	2.514	2.442	--	--	3.595	2.242	--
9	0.131	-8.768	91.265	1.709	1.911	--	2.614	2.951	1.854	--
10	0.236	-4.185	22.223	2.640	2.125	2.653	--	2.879	2.255	2.915
Mean	0.129	-1.947	91.929	2.182	2.537	2.687	2.818	3.381	2.437	3.041

Table A2: Parameters Estimated via GA on Large Sample

Item	MAPE	Price Elas. θ	Normal Sales λ	Pro-Mix μ_1	Pro-Mix μ_2	Pro-Mix μ_3	Pro-Mix μ_4	Pro-Mix μ_5	Pro-Mix μ_6	Pro-Mix μ_7
1	0.140	-0.048	42.83	2.171	3.362	3.764	2.872	--	--	3.257
2	0.121	-0.094	58.54	2.033	2.648	3.450	2.897	--	2.610	3.380
3	0.128	-1.078	78.035	2.845	--	2.875	--	3.170	--	2.665
4	0.130	-0.681	51.906	2.503	3.974	2.650	2.711	--	3.505	3.858
5	0.146	-0.100	199.68	2.414	2.304	1.728	--	3.680	2.938	1.559
6	0.126	-0.020	174.62	2.120	--	2.264	--	4.50	--	2.459
7	0.131	-2.765	87.99	1.625	--	3.809	2.967	--	2.436	1.083
8	0.097	-0.357	78.152	2.354	2.330	--	--	3.549	2.138	--
9	0.134	-7.110	74.700	2.236	2.520	--	3.408	--	2.443	--
10	0.237	-5.632	23.904	2.468	1.654	1.825	--	2.263	1.582	2.238
Mean	0.139	-1.789	87.036	2.277	2.685	2.796	2.971	3.432	2.522	2.562

Appendix (continued)

Table A3: Parameters Estimated via OLS on Small Sample

Item	MAPE	Price Elas. θ	Normal Sales λ	Pro-Mix μ_1	Pro-Mix μ_2	Pro-Mix μ_3	Pro-Mix μ_4	Pro-Mix μ_5	Pro-Mix μ_6	Pro-Mix μ_7
1	0.109	-0.098	430.30	0.214	0.3336	0.744	0.298	--	--	1.341
2	0.109	0.049	108	1	1.47	1.848	1.538	--	1.877	1.986
3	0.103	--	237	1	0.718	1.109	1.73	--	--	1.47
4	0.110	-0.61	132.44	1	--	1.545	1.241	--	1.367	1.543
5	0.129	--	309.00	1.836	1.433	1.230	--	2.487	1.959	--
6	0.099	--	375.00	--	1.370	1.182	--	2.457	1.321	1.333
7	0.129	-2.350	141.00	--	1.865	--	--	0.852	1.617	--
8	0.127	-0.210	188.00	--	1.000	--	--	--	--	--
9	0.133	-7.695	160.00	--	1.140	--	1.844	--	1.166	--
10	0.236	-6.54	61.00	--	0.600	0.687	--	1.041	0.617	0.804
Mean	0.128	-1.945	214.174	1.010	1.103	1.192	1.330	1.709	1.418	1.413

Table A4: Parameters Estimated via OLS on Large Sample

Item	MAPE	Price Elas. θ	Normal Sales λ	Pro-Mix μ_1	Pro-Mix μ_2	Pro-Mix μ_3	Pro-Mix μ_4	Pro-Mix μ_5	Pro-Mix μ_6	Pro-Mix μ_7
1	0.144	-0.098	121.80	0.768	1.178	1.329	1.103	--	--	1.169
2	0.116	-2.187	194.46	0.575	0.788	0.997	0.875	--	--	1.040
3	0.137	-2.908	265	0.852	0.642	0.725	--	--	--	0.480
4	0.142	-0.61	181.00	0.709	--	1.131	1.024	--	--	1.129
5	0.145	--	578.77	0.84	0.765	0.656	0.983	1.328	--	--
6	0.133	-0.69	495.00	0.757	0.912	0.787	1.514	1.637	--	0.888
7	0.159	-2.48	223.87	0.63	1.174	--	--	0.530	--	--
8	0.102	-0.233	175.75	1.077	1.068	--	--	1.706	--	--
9	0.138	-6.900	192.82	0.858	0.978	--	1.406	--	--	--
10	0.241	-4.065	51.310	1.197	0.943	1.170	--	1.630	--	1.350
Mean	0.146	-2.241	247.978	0.826	0.939	0.971	1.151	1.366	--	1.009

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 536-546

1538 – 9472/10/$95.00

Maximum Downside Semi Deviation Stochastic Programming for Portfolio Optimization Problem

Anton Abdulbasah Kamil Adli Mustafa
Universiti Sains Malaysia,
Penang, Malaysia

Khlipah Ibrahim
Universiti Teknologi Mara, Dungun,
Terengganu, Malaysia

Portfolio optimization is an important research field in financial decision making. The chief character within optimization problems is the uncertainty of future returns. Probabilistic methods are used alongside optimization techniques. Markowitz (1952, 1959) introduced the concept of risk into the problem and used a mean-variance model to identify risk with the volatility (variance) of the random objective. The mean-risk optimization paradigm has since been expanded extensively both theoretically and computationally. A single stage and two stage stochastic programming model with recourse are presented for risk averse investors with the objective of minimizing the maximum downside semi-deviation. The models employ the here-and-now approach, where a decision-maker makes a decision before observing the actual outcome for a stochastic parameter. The optimal portfolios from the two models are compared with the incorporation of the deviation measure. The models are applied to the optimal selection of stocks listed in Bursa Malaysia and the return of the optimal portfolio is compared between the two stochastic models. Results show that the two stage model outperforms the single stage model for the optimal and in-sample analysis.

Key words: Portfolio optimization, maximum semi-deviation measure, downside risk, stochastic linear programming.

Introduction

Portfolio optimization is an important research field in financial decision making. The most important character within optimization problems is the uncertainty of future returns. To handle such problems, probabilistic methods are utilized alongside optimization techniques. Stochastic programming is the approach employed in this study to deal with uncertainty. Stochastic programming is a branch of mathematical programming where the parameters are random, the objective of which is to find the optimum solution to problems with uncertain data. This approach can simultaneously deal with both the management of portfolio risk and the identification of the optimal portfolio. Stochastic programming models explicitly consider uncertainty in the model parameters and they provide optimal decisions which are hedged against such uncertainty.

In the deterministic framework, a typical mathematical programming problem could be stated as

$$\begin{aligned} &\min_{x} \quad f(x) \\ &s.t \quad g_i(x) \le 0, \quad i = 1,\dots m, \end{aligned} \tag{1.1}$$

where x is from R^n or Z^n. Uncertainty, which is usually described by a random element, $\xi(\omega)$, where ω is a random outcome from a space Ω, leads to situation where one has to deal with $f(x,\xi(\omega))$ and $g_i(x,\xi(\omega))$, as

Anton Abdulbasah Kamil is an Associate Professor in the School of Distance Education, Universiti Sains Malaysia, Malaysia. Email: anton@usm.my. Adli Mustafa is a Senior lecturer in the School of Mathematical Sciences, Universiti Sains Malaysia, Malaysia. Email: adli@usm.my. Khlipah Ibrahim is an Associate Professor in the Universiti Teknologi Mara, Malaysia.

opposed to just $f(x)$ and $g_i(x)$. Traditionally, the probability distribution of ξ is assumed to be known (or can be estimated) and is unaffected by the decision vector x. The problem becomes decision making under uncertainty where decision vector x must be chosen before the outcome from the distribution of $\xi(\omega)$ can be observed.

Markowitz (1952, 1959) incorporated the concept of risk into the problem and introduced the mean-risk approach, which identifies risk with the volatility (variance) of the random objective. Since 1952, the mean-risk optimization paradigm has been extensively developed both theoretically and computationally. Konno and Yamazaki (1991) proposed mean absolute deviation (MAD) from the mean as the risk measure to estimate the nonlinear variance-covariance of the stocks in the mean-variance (MV) model. It transforms the portfolio selection problem from a quadratic programming problem into a linear problem. The popularity of downside risk among investors is growing and mean-return-downside risk portfolio selection models seem to oppress the familiar mean-variance approach.

The reason mean-variance models are successful is because they separate return fluctuations into downside risk and upside potential. This is relevant for asymmetrical return distributions, for which the mean-variance model punishes the upside potential in the same fashion as the downside risk. Thus, Markowitz (1959) proposed downside risk measures, such as semi variance, to replace variance as the risk measure. Subsequently, downside risk models for portfolio selection have grown in popularity (Sortino & Forsey, 1996).

Young (1998) introduced another linear programming model to maximize the minimum return or minimize the maximum loss (minimax) over time periods and he applied it to stock indices of eight countries from January 1991 until December 1995. The analysis showed that the model performs similarly with the classical mean-variance model. In addition, Young argued that - when data is log-normally distributed or skewed - the minimax formulation might be a more appropriate method compared to the classical mean-variance formulation, which is optimal for normally distributed data. Ogryczak (2000) also considered the minimax model but analyzed it with the maximum semi deviation.

Dantzig (1955) and Beale (1955) independently suggested an approach to stochastic programming termed stochastic programming with recourse; recourse is the ability to take corrective action after a random event has taken place. Their innovation was to amend the problem to allow a decision maker the opportunity to make corrective actions after a random event has taken place. In the first stage, a decision maker makes a here and now decision. In the second stage the decision maker sees a realization of the stochastic elements of the problem but is allowed to make further decisions to avoid the constraints of the problem becoming infeasible.

Stochastic programming is becoming more popular in finance as computing power increases and there have been numerous applications of stochastic programming methodology to real life problems over the last two decades. The applicability of stochastic programs to financial planning problems was first recognized by Crane (1971). More recently Worzel, et al. (1994) and Zenios, et al. (1998) have developed multistage stochastic programs with recourse to address portfolio management problems with fixed-income securities under uncertainty in interest rates. Their models integrate stochastic programming for the selection of portfolios using Monte Carlo simulation models of the term structure of interest rates.

Hiller and Eckstein (1994), Zenios (1995) and Consiglo and Zenios (2001) also applied stochastic programs to fixed-income portfolio management problems. Chang, et al. (2002) modeled a portfolio selection problem with transaction costs as a two-stage stochastic programming problem and evaluated the model using historical data obtained from the Taiwan Stock Exchange; their results show that the model outperforms the market and the MV and MAD models.

In this article, a single stage and two stage stochastic programming model are developed with recourse for portfolio selection. The objective is to minimize the maximum

downside deviation measure of portfolio returns from the expected return. The so-called here-and-now approach is utilized: a decision-maker makes a decision (now) before observing the actual outcome for the stochastic parameter. The portfolio optimization problem considered follows the original Markowitz (1959) formulation and is based on a single period model of investment. At the beginning of a period, an investor allocates capital among various securities assuming that each security is represented by a variable; this is equivalent to assigning a nonnegative weight to each variable. During the investment period, a security generates a random rate of return. The change of invested capital observed at the end of the period is measured by the weighted average of the individual rates of return.

The objective of this study is to compare the optimal portfolio selected using two different stochastic programming models. The optimal portfolios are compared between the single stage and two stage models with the incorporation of deviation measure. This method is applied to the optimal selection of stocks listed in Bursa Malaysia and the return of the optimal portfolio from the two models is compared.

Methodology

Consider a set of securities $I=\{i : i = 1,2,...,n\}$ for an investment; at the end of a certain holding period the assets generate returns, $\tilde{r} = (\tilde{r}_1, \tilde{r}_2, ..., \tilde{r}_n)^T$. The returns are unknown at the beginning of the holding period, that is at the time of the portfolio selection, and are treated as random variables; their mean value is denoted by, $\bar{r} = E(\tilde{r}) = (\bar{r}_1, \bar{r}_2, ..., \bar{r}_n)^T$. At the beginning of a holding period an investor wishes to apportion his budget to these assets by deciding on a specific allocation $\boldsymbol{x} = (x_1, x_2, ..., x_n)^T$ such that $x_i \geq 0$ (i.e., short sales are not allowed) and $\sum_{i \in I} x_i = 1$ (budget constraint). In this article, boldface characters are used to denote vectors, and the symbol ~ denotes random variables.

The uncertain return of a portfolio at the end of a holding period is $\tilde{R} = R(\boldsymbol{x}, \tilde{\boldsymbol{r}}) = \boldsymbol{x}^T \tilde{\boldsymbol{r}}$. This is a random variable with a distribution function F, that is, $F(x, \mu) = P\{R(x, \tilde{r}) \leq \mu\}$. It is assumed that F does not depend on the portfolio composition $\boldsymbol{x}$. The expected return of the portfolio is

$$\bar{R} = E[\tilde{R}] = E[R(\boldsymbol{x}, \tilde{\boldsymbol{r}})] = \bar{R}(\boldsymbol{x}, \bar{\boldsymbol{r}}).$$

Suppose the uncertain returns of the assets, $\tilde{\boldsymbol{r}}$, are represented by a finite set of discrete scenarios $\Omega = \{\omega : \omega = 1,2,...,S\}$, whereby the returns under a particular scenario $\omega \in \Omega$ take the values $\boldsymbol{r}_\omega = (r_{1\omega}, r_{2\omega}, ..., r_{n\omega})^T$ with associated probability $p_\omega > 0$, $\sum_{\omega \in \Omega} p_\omega = 1$. The mean return of the assets is $\bar{\boldsymbol{r}} = \sum_{\omega \in \Omega} p_\omega \boldsymbol{r}_\omega$. The portfolio return under a particular realization of asset return $\boldsymbol{r}_\omega$ is denoted by $R_\omega = R(\boldsymbol{x}, \boldsymbol{r}_\omega)$. The expected portfolio return is expressed as:

$$\begin{aligned} \bar{\mathrm{R}} &= \bar{R}(\mathrm{x}, \mathrm{r}_\omega) \\ &= E[R(\mathrm{x}, \mathrm{r}_\omega)] \\ &= \sum_{\omega \in \Omega} p_\omega R(\mathrm{x}, \mathrm{r}_\omega). \end{aligned}$$

Let $M[R(\boldsymbol{x}, \boldsymbol{r}_\omega)]$ be the minimum of the portfolio return. The maximum (downside) semideviation measure is defined as

$$\begin{aligned} \kappa(\mathrm{x}) &= MM[R(\mathrm{x}, \mathrm{r}_\omega)] \\ &= [\mathrm{E}[R(\mathrm{x}, \mathrm{r}_\omega)] - Min\,[R(\mathrm{x}, \mathrm{r}_\omega)] \end{aligned} \quad (2.1)$$

Maximum downside deviation risk $MM[R(\boldsymbol{x}, \boldsymbol{r}_\omega)]$ is a very pessimistic risk measure related to the worst case analysis. It does not take into account any distribution of outcomes other than the worst one.

Properties of the $MM[R(\boldsymbol{x}, \tilde{\boldsymbol{r}})]$ Measures

Artzner, et al. (1999) introduced the axiomatic approach to construction of risk measures. This approach has since been repeatedly employed by many authors for the development of other types of risk measures

tailored to specific preferences and applications (see Rockafellar, et al., 2002, 2004; Acerbi, 2002; Ruszcynski & Shapiro, 2004).

Proposition 1: $MM[R(\boldsymbol{x},\tilde{\boldsymbol{r}})]$ measure is a deviation measure.

Proof:

1. Subadditivity:

$$\kappa(X_1 + X_2) \le \kappa(X_1) + \kappa(X_2).$$

$$\begin{aligned} MM[R_1(\mathrm{x},\tilde{\mathrm{r}}) + R_2(\mathrm{x},\tilde{\mathrm{r}})] &= \max\{E[R_1(\mathrm{x},\tilde{\mathrm{r}}) + R_2(\mathrm{x},\tilde{\mathrm{r}})] \\ &\quad -[R_1(\mathrm{x},\tilde{\mathrm{r}}) + R_2(\mathrm{x},\tilde{\mathrm{r}})]\} \\ &= \max\{(E[R_1(\mathrm{x},\tilde{\mathrm{r}})] - R_1(\mathrm{x},\tilde{\mathrm{r}})) \\ &\quad +(E[R_2(\mathrm{x},\tilde{\mathrm{r}})] - R_2(\mathrm{x},\tilde{\mathrm{r}})\} \\ &\le \max\{E[R_1(\mathrm{x},\tilde{\mathrm{r}})] - R_1(\mathrm{x},\tilde{\mathrm{r}})\} \\ &\quad +\max\{E[R_2(\mathrm{x},\tilde{\mathrm{r}})] - R_2(\mathrm{x},\tilde{\mathrm{r}})\} \\ &\le MM[R_1(\mathrm{x},\tilde{\mathrm{r}})] + MM[R_2(\mathrm{x},\tilde{\mathrm{r}})] \end{aligned}$$

2. Positive Homogeneity:

$$MM[0] = max(E[0] - 0) = 0.$$

$$\begin{aligned} MM[\lambda(R(x,\tilde{r})] &= max\{E[\lambda R(x,\tilde{r})] - \lambda R(x,\tilde{r})\} \\ &= \lambda\, max\{E[R(x,\tilde{r})] - R(x,\tilde{r})\} \\ &= \lambda MM[R(x,\tilde{r})], \quad for\ all\ \lambda > 0 \end{aligned}$$

3. Translation invariance:

$\kappa(X + \alpha) = \kappa(X) - \alpha$, for all real constants α.

$$\begin{aligned} MM[(R(x,\tilde{r}) + \alpha] &= max\{E([R(x,\tilde{r}) + \alpha] - [R(x,\tilde{r}) + \alpha])\} \\ &= max\{E[R(x,\tilde{r})] + \alpha - R(x,\tilde{r}) - \alpha\} \\ &= max\{E[R(x,\tilde{r})] - R(x,\tilde{r})\} \\ &= MM[(R(x,\tilde{r})] \end{aligned}$$

4. Convexity:

$\kappa[\lambda X_1 + (1-\lambda)X_2] \le \lambda\kappa(X_1) + (1-\lambda)\kappa(X_2)$ for all $\lambda \in [0,1]$.

$$\begin{aligned} MM[\lambda R_1(\mathrm{x},\tilde{\mathrm{r}}) + (1-\lambda)R_2(\mathrm{x},\tilde{\mathrm{r}})] &= \max\{E[\lambda R_1(\mathrm{x},\tilde{\mathrm{r}}) + (1-\lambda)R_2(\mathrm{x},\tilde{\mathrm{r}})] \\ &\quad -[\lambda R_1(\mathrm{x},\tilde{\mathrm{r}}) + (1-\lambda)R_2(\mathrm{x},\tilde{\mathrm{r}})]\} \\ &= \max\{(E[\lambda R_1(\mathrm{x},\tilde{\mathrm{r}})] + E[(1-\lambda)R_2(\mathrm{x},\tilde{\mathrm{r}}))] \\ &\quad -\lambda R_1(\mathrm{x},\tilde{\mathrm{r}}) + (1-\lambda)R_2(\mathrm{x},\tilde{\mathrm{r}})\} \\ &= \max\{\lambda(E[R_1(\mathrm{x},\tilde{\mathrm{r}})] - R_1(\mathrm{x},\tilde{\mathrm{r}})) \\ &\quad +(1-\lambda)(E[R_2(\mathrm{x},\tilde{\mathrm{r}})] - R_2(\mathrm{x},\tilde{\mathrm{r}}))\} \\ &\le \lambda \max\{(E[R_1(\mathrm{x},\tilde{\mathrm{r}})] - R_1(\mathrm{x},\tilde{\mathrm{r}}))\} + \\ &\quad (1-\lambda)\max\{E[R_2(\mathrm{x},\tilde{\mathrm{r}})] - R_2(\mathrm{x},\tilde{\mathrm{r}}))\} \\ &\le \lambda MM[R_1(\mathrm{x},\tilde{\mathrm{r}})] + (1-\lambda)MM[R_2(\mathrm{x},\tilde{\mathrm{r}})] \end{aligned}$$

Single Stage Stochastic Programming Portfolio Optimization Model with MM Deviation Measure

The portfolio selection optimization model is formulated as a single stage stochastic programming model as follows.

Definition 1: S_MM

The stochastic portfolio optimization problem where the difference between the expected portfolio return and the maximum of minimum portfolio returns is minimized and constraining the expected portfolio return is:

$$\underset{\mathrm{x}\in X}{Minimize}\ \max_{\omega\in\Omega}\ [\bar{R}(\mathrm{x},\mathrm{r}_\omega) - R(\mathrm{x},\mathrm{r}_\omega)] \tag{2.2a}$$

Subject to:

$$R(x,r_\omega) = \sum_{i\in I} x_i r_{\omega i} \quad \forall\,\omega\in\Omega \tag{2.2b}$$

$$\bar{R}(x,r_\omega) = \sum_{\omega\in\Omega} p_\omega R(x,r_\omega) \tag{2.2c}$$

$$\bar{R}(x,r_\omega) \ge \alpha \tag{2.2d}$$

$$\sum_{i\in I} x_i = 1 \tag{2.2e}$$

$$L_i \le x_i \le U_i \quad \forall\, i\in I \tag{2.2f}$$

Model S_MM minimizes the maximum semi deviation of portfolio returns from the expected portfolio return at the end of the investment horizon. Equation (2.2b) defines the total portfolio return under each scenario ω. Equation (2.2c) defines the expected return of the portfolio at the end of the horizon, while equation (2.2d) constrains the expected return by the target return α. Equation (2.2e) insures that the total weights of all investments sum to one, that is, budget constraints ensuring full investment of available budget. Finally equation (2.2f) insures that the weights on assets purchased are nonnegative, disallowing short sales and placing upper bounds on the weights. Solving the parametric programs (2.2) for different values of the expected portfolio return α yields the MM-efficient frontier.

Linear Programming Formulation for S_MM

Models S_MM have a non linear objective function and a set of linear constraints, thus the models are non linear stochastic programming. However, the models can be transformed to linear models as follows.

For every scenario $\omega \in \Omega$, let an auxiliary variable,

$$\eta = \max_{\omega \in \Omega} [\bar{R}(\boldsymbol{x}, \boldsymbol{r}_\omega) - R(\boldsymbol{x}, \boldsymbol{r}_\omega)] \quad (2.3)$$

subject to

$$\eta \geq \max_{\omega \in \Omega} [\bar{R}(x, r_\omega) - R(x, r_\omega)] \text{ for } \forall \omega \in \Omega,$$

then,

$$MM[R(\boldsymbol{x}, \boldsymbol{r}_\omega)] = \eta \quad (2.4)$$

subject to

$$\eta \geq \max_{\omega \in \Omega} [\bar{R}(x, r_\omega) - R(x, r_\omega)] \text{ for } \forall \omega \in \Omega.$$

Substituting (2.4) in the portfolio optimization models (2.2) results in the following stochastic linear programming model:

$$\text{Minimize } \eta, \quad (2.5a)$$

subject to:

$$R(x, r_\omega) = \sum_{i \in I} x_i r_{\omega i} \quad (2.5b)$$

$$\bar{R}(x, r_\omega) = \sum_{\omega \in \Omega} p_\omega R(x, r_\omega) \quad (2.5c)$$

$$\bar{R}(x, r_\omega) \geq \alpha \quad (2.5d)$$

$$\bar{R}(x, r_\omega) - R(x, r_\omega) \leq \eta \quad (2.5e)$$

$$\sum_{i \in I} x_i = 1 \quad (2.5f)$$

$$L_i \leq x_i \leq U_i \quad \forall i \in I \quad (2.5g)$$

Theorem 1

If $\boldsymbol{x}^*$ is an optimal solution to (2.2), then $(\boldsymbol{x}^*, \eta^*)$ is an optimal solution to (2.5), where $\eta = \max_{\omega \in \Omega} [\bar{R}(\boldsymbol{x}, \boldsymbol{r}_\omega) - R(\boldsymbol{x}, \boldsymbol{r}_\omega)]$. Conversely, if $(\boldsymbol{x}^*, \eta^*)$ where $\eta = \max_{\omega \in \Omega} [\bar{R}(\boldsymbol{x}, \boldsymbol{r}_\omega) - R(\boldsymbol{x}, \boldsymbol{r}_\omega)]$ is an optimal solution to (2.5), then $\boldsymbol{x}^*$ is an optimal solution to (2.2).

Proof:

If $\boldsymbol{x}^*$ is an optimal solution to (2.2), then $(\boldsymbol{x}^*, \eta^*)$ is a feasible solution to (2.5), where $\eta = \max_{\omega \in \Omega} [\bar{R}(\boldsymbol{x}^*, \boldsymbol{r}_\omega) - R(\boldsymbol{x}^*, \boldsymbol{r}_\omega)]$. If $(\boldsymbol{x}^*, \eta^*)$ is not an optimal solution to (2.5), then a feasible solution $(\boldsymbol{x}, \eta)$ exists to (2.5) where $\eta = \max_{\omega \in \Omega} [\bar{R}(\boldsymbol{x}, \boldsymbol{r}_\omega) - R(\boldsymbol{x}, \boldsymbol{r}_\omega)]$ such that $\eta \leq \eta^*$.

If $\max_{\omega \in \Omega} [\bar{R}(\boldsymbol{x}, \boldsymbol{r}_\omega) - R(\boldsymbol{x}, \boldsymbol{r}_\omega)] \leq \eta$,

then

$$\max_{\omega \in \Omega} [\bar{R}(x, r_\omega) - R(x, r_\omega)] \leq \eta < \eta^* < \max_{\omega \in \Omega} [\bar{R}(x^*, r_\omega) - R(x^*, r_\omega)]$$

which contradicts that $\boldsymbol{x}^*$ is an optimal solution to (2.2).

However, if $(\boldsymbol{x}^*, \eta^*)$ is an optimal solution to (2.5), where $\eta = \max_{\omega \in \Omega} [\bar{R}(\boldsymbol{x}, \boldsymbol{r}_\omega) - R(\boldsymbol{x}, \boldsymbol{r}_\omega)]$ then $\boldsymbol{x}^*$ is an optimal solution to (2.2). Otherwise, a feasible solution $\boldsymbol{x}$ to (2.2) exists such that

$$\max_{\omega \in \Omega} [\bar{R}(\mathrm{x}, \mathrm{r}_\omega) - R(\mathrm{x}, \mathrm{r}_\omega)] < \max_{\omega \in \Omega} [\bar{R}(\mathrm{x}^*, \mathrm{r}_\omega) - R(\mathrm{x}^*, \mathrm{r}_\omega)]$$

Denoting $\eta = \max_{\omega \in \Omega} [\bar{R}(\boldsymbol{x}, \boldsymbol{r}_\omega) - R(\boldsymbol{x}, \boldsymbol{r}_\omega)]$, leads to

$$\eta = \max_{\omega \in \Omega} [\bar{R}(\boldsymbol{x}, \boldsymbol{r}_\omega) - R(\boldsymbol{x}, \boldsymbol{r}_\omega)]$$

$$< \max_{\omega \in \Omega} [\bar{R}(\mathrm{x}^*, \mathrm{r}_\omega) - R(\mathrm{x}^*, \mathrm{r}_\omega)]$$

$$< \eta^*$$

which contradicts that $(\boldsymbol{x}^*, \eta^*)$ is an optimal solution to (2.5).

Two Stage Stochastic Programming Model with Recourse

A dynamic model where not only the uncertainty of the returns is included in the model but future changes, recourse, to the initial compositions are allowed is now introduced. The portfolio optimization is formulated by assuming an investor can make corrective action after the realization of random values by changing the composition of the optimal portfolio. This can be accomplished by formulating the single period stochastic linear programming models with the mean absolute negative deviation measure as a two-stage stochastic programming problem with recourse. The two-stage stochastic programming problem allows a recourse decision to be made after uncertainty of the returns is realized.

Consider the case when the investor is interested in a first stage decision $\boldsymbol{x}$ which hedges against the risk of the second-stage action. At the beginning of the investment period, the investor selects the initial composition of the portfolio, $\boldsymbol{x}$. The first stage decision, $\boldsymbol{x}$, is made when there is a known distribution of future returns. At the end of the planning horizon, after a particular scenario of return is realized, the investor rebalances the composition by either purchasing or selling selected stocks. In addition to the initial - or first stage - decision variables $\boldsymbol{x}$, let a set of second stage variables, $\boldsymbol{y}_{i,\omega}$ represent the composition of stock i after rebalancing is done, that is, $\boldsymbol{y}_{i,\omega} = \boldsymbol{x}_i + \boldsymbol{P}_{i,\omega}$ or $\boldsymbol{y}_{i,\omega} = \boldsymbol{x}_i - \boldsymbol{Q}_{i,\omega}$, where $\boldsymbol{P}_{i,\omega}$ and $\boldsymbol{Q}_{i,\omega}$ are the quantity purchased and sold respectively and $\boldsymbol{y}_{i,\omega}$ is selected after the uncertainty of returns is realized.

Linear Representation of MM

Before formulating the two stage stochastic programming models to minimize the second stage risk measure to address the portfolio optimization problem, the mean absolute negative deviation and maximum downside deviation of portfolio returns are formulated from the expected return in terms of the second stage variables $\boldsymbol{y}$.

$$\text{Let } \kappa(R(\boldsymbol{y}_\omega, \boldsymbol{r}_\omega)) = MM[R(\boldsymbol{y}_\omega, \boldsymbol{r}_\omega)] = \max_{\omega \in \Xi} [\bar{R}(\boldsymbol{y}_\omega, \boldsymbol{r}_\omega) - R(\boldsymbol{y}_\omega, \boldsymbol{r}_\omega)] \quad (2.6)$$

For every scenario $\omega \in \Omega$, if the auxiliary variable is

$$\eta = \max_{\omega \in \Omega} [\bar{R}(\boldsymbol{y}_\omega, \boldsymbol{r}_{,\omega}) - R(\boldsymbol{y}_\omega, \boldsymbol{r}_\omega)] \quad (2.7)$$

subject to

$$\eta \geq \max_{\omega \in \Omega} [\bar{R}(y_\omega, r_\omega) - R(y_\omega, r_\omega)] \text{ for } \forall \omega \in \Omega \quad (2.8)$$

then

$$MM[R(\boldsymbol{x}, \boldsymbol{r}_\omega)] = \eta \quad (2.9)$$

subject to

$$\eta \geq \max_{\omega \in \Omega} [\bar{R}(y_\omega, r_\omega) - R(y_\omega, r_\omega)] \text{ for } \forall \omega \in \Omega.$$

Two Stage Stochastic Linear Programming Formulation of 2S_MM

The two stage stochastic linear programming model is formulated for the portfolio optimization problem that hedges against second stage MM as follows.

Definition 2: 2S_MM

The stochastic portfolio optimization problem where the downside maximum semi-deviation of portfolio returns from the expected return is minimized and the expected portfolio return is constrained is:

$$\text{Minimize } \eta \qquad (2.10a)$$

$$\sum_{i \in I} x_i = 1 \qquad (2.10b)$$

$$\sum_{i \in I} y_{\omega i} = 1 \quad \forall \omega \in \Omega \qquad (2.10c)$$

$$\bar{R}(x, r_\omega) + R(y_\omega, r_\omega) \geq \alpha \quad \forall \omega \in \Omega \qquad (2.10d)$$

$$L_i \leq x_i \leq U_i \quad \forall i \in I \qquad (2.10e)$$

$$L_{\omega i} \leq y_{\omega i} \leq U_{\omega i} \quad \forall i \in I, \forall \omega \in \Omega \qquad (2.10f)$$

$$R(y_\omega, r_\omega) \geq \eta \quad \forall \omega \in \Omega \qquad (2.10g)$$

Model (2.10) minimizes the maximum downside semi deviation of the portfolio return from the expected portfolio return of the second stage variable, $\boldsymbol{y}$, at the end of the investment period. Equation (2.10b) insures that the total weights of all investments in the first stage sum to one, and equation (2.10c) insures that the total weights of all investments in the second stage under each scenario, ω, sum to one - that is, budget constraints ensuring full investment of available budget. Equation (2.10d) constrains the expected return by the target return, α, while equations (2.10e) and (2.10f) insure that the weights on assets purchased are nonnegative, disallowing short sales and placing an upper bound on the weights in the first stage and second stage respectively. Finally, equations (2.10g) and (2.10h) define the mean absolute negative deviation of portfolio returns from the expected portfolio return in the second stage and the auxiliary variables for the linear representation of the deviation measure.

Numerical Analysis

Models developed herein were tested on ten common stocks listed on the main board of Bursa Malaysia. These stocks were randomly selected from a set of stocks that were listed on December 1989 and were still in the list in May 2004; closing prices were obtained from Investors Digest. At first, sixty companies were selected at random, ten stocks were then selected and the criterion used to select the ten stocks in the analysis is as follows:

i. Those companies which do not have a complete closing monthly price during the analysis period were excluded.

ii. Because the portfolios were examined on the basis of historical data, those with negative average returns over the analysis period were excluded.

Empirical distributions computed from past returns were used as equiprobable scenarios. Observations of returns over N_S overlapping periods of length Δt are considered as the N_S possible outcomes (or scenarios) of future returns and a probability of $\frac{1}{N_s}$ is assigned to each of them. Assume T historical prices, P_t, $t = 1,2,...,T$ of the stocks under consideration. For each point of time, the realized return vector over the previous period of 1 month is computed, which will be further considered as one of the N_S scenarios for future returns on the assets. Thus, for example, a scenario r_{is} for the return on asset i is obtained as:

$$r_{is} = \frac{P_i(t+1) - P_i(t)}{P_i(t)}. \qquad (3.1)$$

For each stock, 100 scenarios of the overlapping periods of length 1 month were obtained, that is, N_S.

To evaluate the performance of the two models, the portfolio returns resulting from applying the two stochastic optimization models were examined. A comparison is made between the S_MM and 2S_MM models by analyzing the optimal portfolio returns in-sample portfolio returns and out-of-sample portfolio returns over a 60-month period from June 1998 to May 2004. At each month, the historical data from the previous 100 monthly observations is used to solve the resulting optimization models and record the return of the optimal portfolio. The in-sample realized portfolio return is then calculated. The clock is advanced one month and the out-of-sample realized return of the portfolio is determined from the actual return of the assets. The same procedure is repeated for the next period and the average returns are computed for in-sample and out-of-sample realized portfolio return. The minimum monthly required return α is equal to one in the analysis for both the S_MM and 2S_MM models.

Results

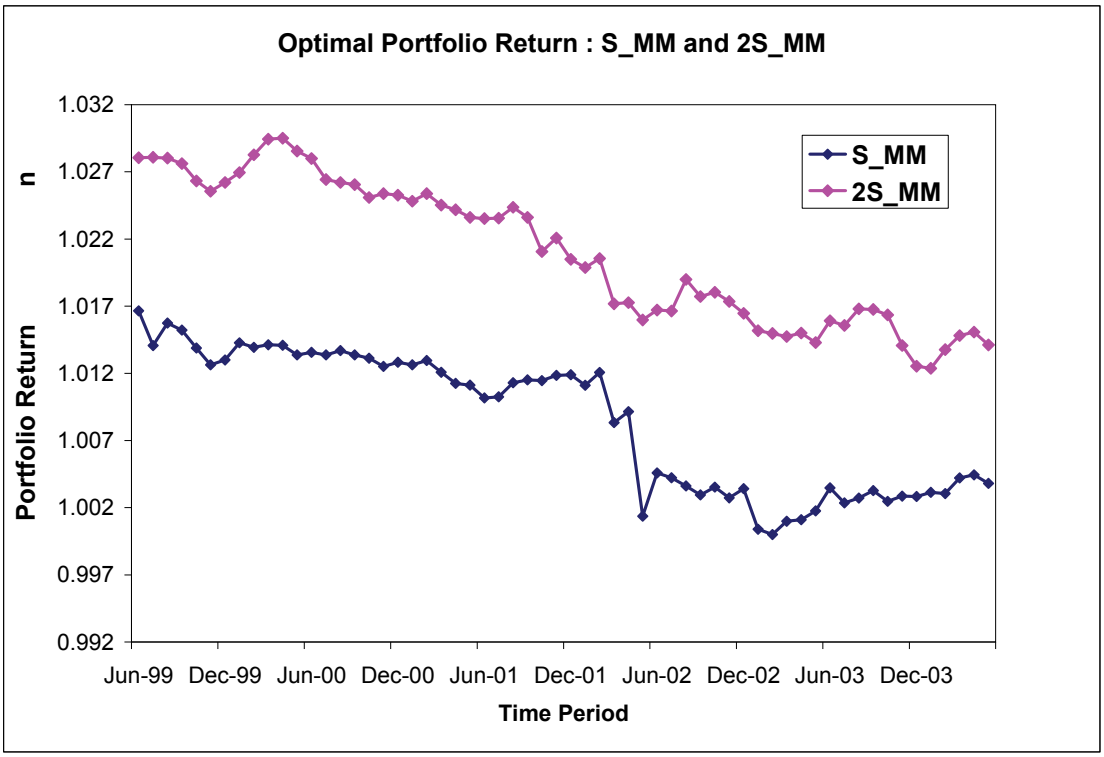
Comparison of Optimal Portfolio Returns between S_MM and 2S_MM

Figure 1 presents the graphs of optimal portfolio returns resulting from solving the two models; S_MM and 2S_MM. The optimal portfolio returns of the two models exhibit a similar pattern: a decreasing trend is observed in the optimal returns for both models. However, as illustrated in Figure 1, the optimal portfolio returns from the two stage stochastic programming with recourse model (2S_MM) are higher than the optimal portfolio returns from the single stage stochastic programming model (S_MM) in all testing periods. This shows that an investor can make a better decision regarding the selection of stocks in a portfolio when taking into consideration both making decision facing the uncertainty and the ability of making corrective actions when the uncertain returns are realized compared to considering only making decisions facing the uncertainty alone.

Comparison of Average In-Sample Portfolio returns between S_MM and 2S_MM

The average realized returns were used

Figure 1: Comparison of Optimal Portfolio Returns S_MM and 2S_MM Models



to compare in-sample portfolio returns between the S_MM model and 2S_MM model; results are presented in Figure 2. An increasing trend is observed in the months from December 1999 until April 2000, and then a decreasing trend is noted until June 2001. From June 2001 until May 2004 both averages show an increasing trend. The average in-sample portfolio returns of 2S_MM are higher than the average in-sample portfolio returns in all testing periods.

Comparison of Out-Of-Sample Portfolio Returns between S_MM and 2S_MM Models

In a real-life environment, model comparison is usually accomplished by means of ex-post analysis. Several approaches can be used to compare models. One of the most commonly applied methods is based on the representation of the ex-post returns of selected portfolios over a given period and on comparing them against a required level of return. The comparison of out-of-sample portfolio returns between the single stage stochastic programming model S_MM and the two stage stochastic programming with recourse model 2S_MM is also accomplished using the average return. The results of the out-of-sample analysis are presented in Figure 3.

Throughout the testing periods, the average returns from the two models show similar patterns. An increasing trend is observed in the months from December 1999 until December 2000, and then a decreasing trend is observed until June 2001. Starting from June 2001, both averages show an increasing trend. The average out-of-sample of the two-stage model 2S_MM is higher than those of single stage model S_MM. The models have been applied directly to the original historical data treated as future returns scenarios, thus loosening the trend information. Possible application of forecasting procedures prior to the portfolio optimization models considered may be an interesting direction for future research. For references on scenario generation see Carino, et al., (1998).

Figure 2: Comparison of Average In-Sample Portfolio Return between S_MM and 2S_MM Models

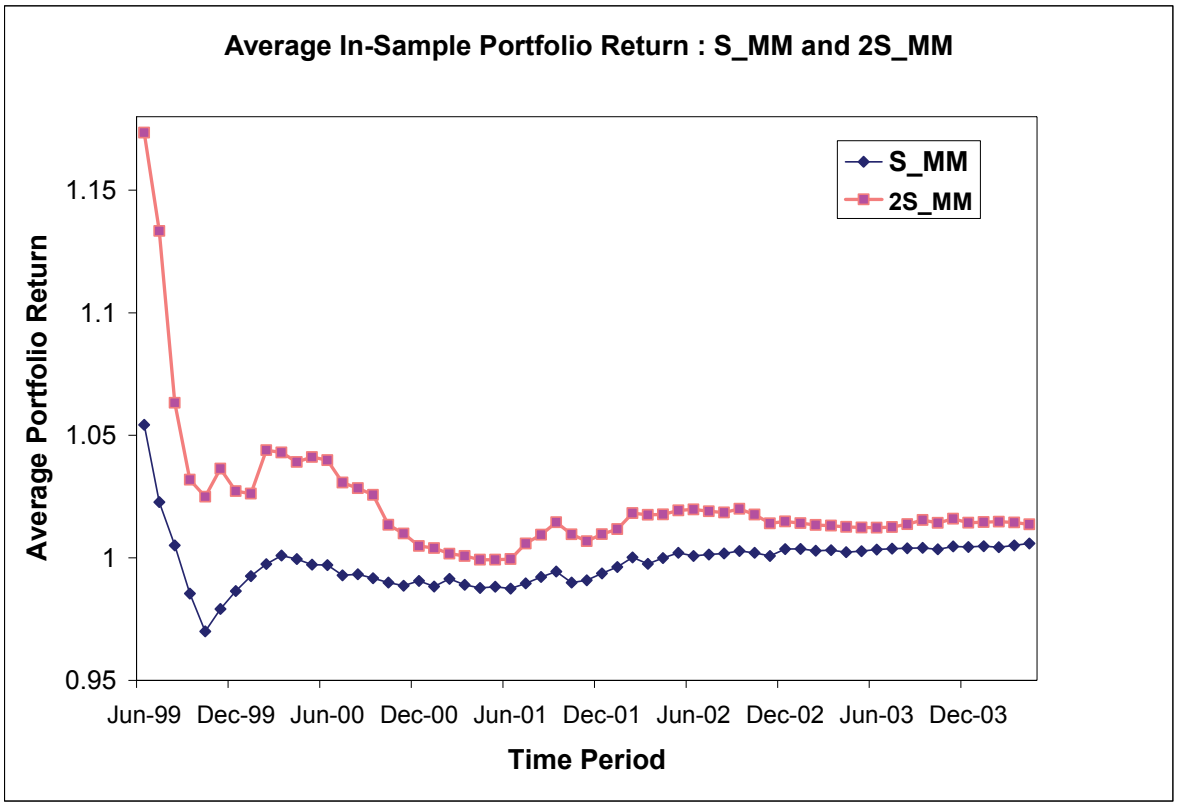
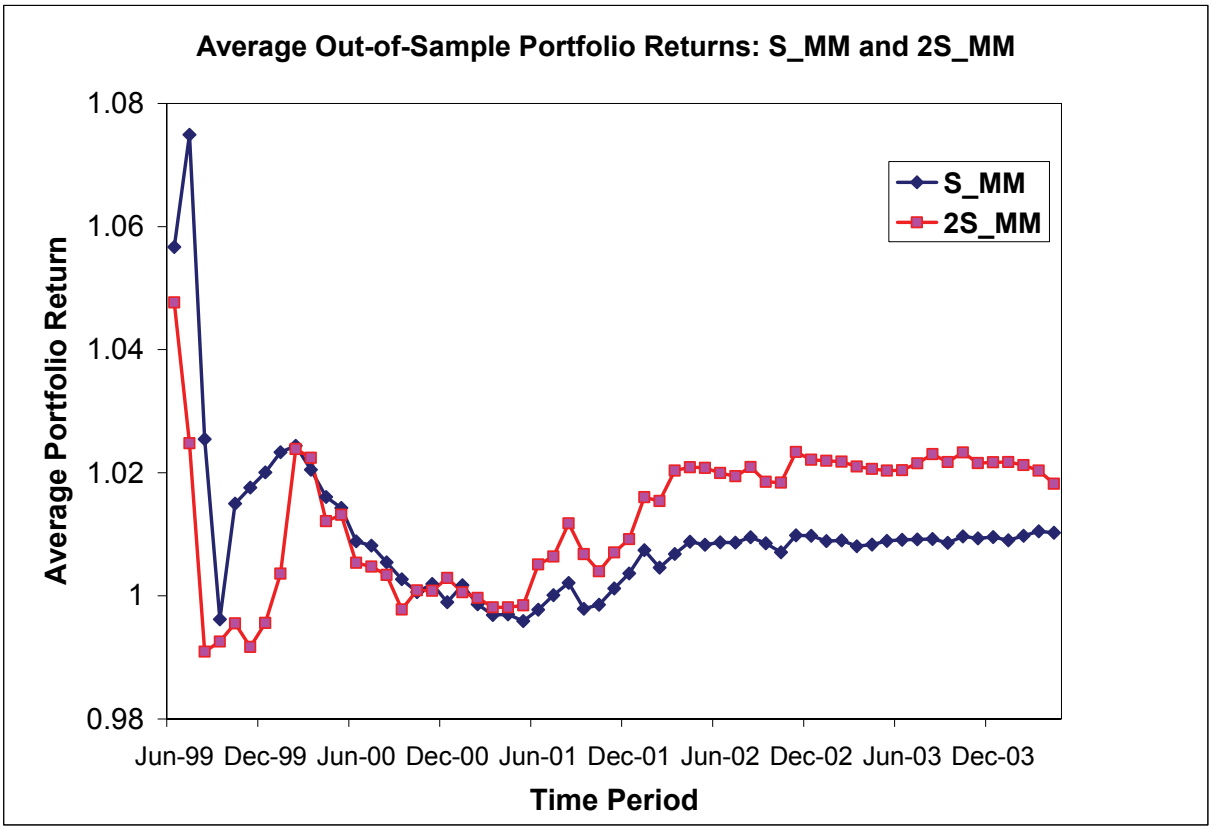


Figure 3: Comparison of Out-Of-Sample Analysis between Single Stage S_MM and Two Stage 2S_MM Models



Conclusion

A portfolio selection of stocks with maximum downside semi deviation measure is modeled as single stage and two stage stochastic programming models in this article. The single stage model and the two stage model incorporate uncertainty and at the same consider rebalancing the portfolio composition at the end of investment period. The comparison of the optimal portfolio returns, the in-sample portfolio returns and the out-of-sample portfolio returns show that the performance of the two stage model is better than that of the single stage model. Historical data was used for scenarios of future returns. Future research should generate scenarios of future asset returns using an appropriate scenario generation method before applying models developed in this article.

Acknowledgements

The work funded by the FRGS (Fundamental Research Grant Scheme) of Ministry for Higher Education of Malaysia, Grant 203/PJJAUH/671128 Universiti Sains Malaysia.

References

Acerbi, C. (2002). Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking Finance*, *26*(*7*), 1487-1503.

Beale, E. M. L. (1955). On minimizing a convex function subject to linear inequalities. *Journal of the Royal Statistical Society*, *Series B*, *17*, 173-184.

Carino, D. R., Myers, D. H., & Ziemba, W. T. (1998). Concepts, technical issues and use of the Russel-Yasuda Kasai financial planning model. *Operations Research*, *46*, 450-463.

Chang, K. W., Chen, H. J., & Liu, C. Y. (2002). A stochastic programming model for portfolio selection. *Journal of Chinese Institute of Industrial Engineers*, *19*(*3*), 31-41.

Crane, D. B. (1971). A stochastic programming model for commercial bank bond portfolio management. *Journal of Financial and Quantitative Analysis*, *6*, 955-976.

Dantzig, G. B. (1955). Linear programming under uncertainty. *Management Science*, *1*, 197-206.

Konno, H., & Yamazaki, H. (1991). Mean-absolute deviation portfolio optimization model and its application to Tokyo stock market. *Management Science*, *7*, 519-531.

Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance*, *8*, 77-91.

Markowitz,H. M. (1959). *Portfolio selection: Efficient diversification of investmen*t. New York: John Wiley& Sons.

Rockafellar, R. T., Uryasev, S., & Zabarankin, M. (2002). Deviation measures in risk analysis and optimization. *Technical Report 2002-7*, ISE Dept., University of Florida.

Rockafellar, R. T., Uryasev, S., & Zabarankin, M. (2004). Generalized deviation in risk analysis. *Technical Report 2004-4*, ISE Dept., University of Florida.

Ruszczynski, A., & Shapiro, A. (2004). Optimization of convex risk functions. *Working paper*.

Sortino, F. A., & Forsey, H. J. (1996). On the use and misuse of downside risk. *Journal of Portfolio Management*, *Winter*, 35-42.

Young, M. R. (1998). A minimax portfolio selection rule with linear programming solution. *Management Science*, *44*, 673-683

Worzel,K. J., Vassiadou-Zeniou, C., & Zenios, S. A. (1994). Integrated simulation and optimization models for tracking fixed-income securities. *Operations Research*, *42*(*2*), 223-233.

Zenios, S. A, Holmer, M. R., McKendall, R., & Vassiadou-Zeniou, C. (1998). Dynamic models for fixedincome portfolio management under uncertainty. *Journal of Economic Dynamics and Control*, *22*, 1517-1541.

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 547-557

1538 – 9472/10/$95.00

On Bayesian Shrinkage Setup for Item Failure Data Under a Family of Life Testing Distribution

Gyan Prakash
S. N. Medical College, Agra, U. P., India

Properties of the Bayes shrinkage estimator for the parameter are studied of a family of probability density function when item failure data are available. The symmetric and asymmetric loss functions are considered for two different prior distributions. In addition, the Bayes estimates of reliability function and hazard rate are obtained and their properties are studied.

Key words: Bayes shrinkage estimator; squared error loss function (SELF); LINEX loss function (LLF); reliability function; hazard rate.

Introduction

The probability density function (pdf) of a random variable x with parameter θ and two known positive constants w and v for the proposed family of life testing distribution is given as

$$f(x;\theta,w,v)=\frac{v}{\Gamma w}\left(\frac{x^{wv-1}}{\theta^{w}}\right)\exp\left(-\frac{x^{v}}{\theta}\right);\quad x>0,\ \theta>0,\ w,v>0. \tag{1.1}$$

For the different values of w and v, the distributions are given as:

w	v	Distribution
1	1	Exponential
	1	Two parameter Gamma
+ve Integer	1	Erlang
1		Two parameter Weibull
1	2	Rayleigh
3/2	2	Maxwell

The use of SELF in the Bayes estimation may not be appropriate when positive and negative errors have different consequences. To overcome this difficulty, an asymmetric loss function (LLF) was proposed by Varian (1975) and its invariant form for any parameter θ is given by (see Singh, et al., 2007)

$$L(\Delta)=e^{a\Delta}-a\Delta-1;\ a\neq 0$$

and

$$\Delta=\left(\hat{\theta}-\theta\right)/\theta. \tag{1.2}$$

where $\hat{\theta}$ is any estimate of the parameter θ. The sign and magnitude of 'a' represents the direction and degree of asymmetry respectively. The positive (negative) value of 'a' is used when overestimation is more (less) serious than underestimation. For small value of $|a|$, LLF is not far from SELF.

In many situations, the experimenter has some prior information about parameter in the form of a point guess value. Thompson (1968), Pandey and Singh (1977), Prakash and Singh (2006), Prakash and Singh (2008, 09) and others have suggested shrinkage estimators utilizing the point guess value of the parameter and have shown that they performed better when the guess value is in the vicinity of the true value. The shrinkage procedure has been applied in numerous problems, including mean survival time in epidemiological studies, forecasting of the money supply, estimating mortality rates and improving estimation in sample surveys.

Gyan Prakash is in the Department of Community Medicine. Email him at: ggyanji@yahoo.com.

The shrinkage estimator of the parameter θ when a prior point guess value θ_0 is available, is given by

$$S = k\,\hat{\theta} + (1-k)\,\theta_0 . \qquad (1.3)$$

Here $k \in [0,1]$ is the shrinkage factor and the experimenter according to his belief in the guess value specifies the values of the shrinkage factor.

In the item censored situations where n items are put to test without replacement and the test terminates as soon as the r^{th} item fails $(r \le n)$. Let $x_1, x_2, ..., x_r$ be the observed failure items for the first r components, then the likelihood function of r failure items $\underline{x} = (x_1, x_2, ..., x_r)$ is

$$L(\underline{x};\theta) = h(\underline{x}) \frac{1}{\theta^{r w}} \exp\left(-\frac{T_r}{\theta}\right), \qquad (1.4)$$

where $T_r = \sum_{i=1}^{r} x_i^{v} + (n-r)\, x_{(r)}^{v}$ and the function $h(\underline{x})$ is independent with the parameter θ. The statistic T_r is sufficient for θ and the UMVU estimator is $U_r = T_r / r w$. The risk of U_r under the SELF and LLF are obtained as

$$R_{(S)}(U_r) = \frac{\theta^2}{r}$$

and

$$R_{(L)}(U_r) = e^{-a}\left(1 - \frac{a}{r}\right)^{-r} - 1.$$

Here the suffixes S and L respectively show the risk taken under the SELF and LLF.

The inverted Gamma distribution with parameters α and β have been considered as the conjugate prior density for the parameter θ with pdf is

$$g_1(\theta) = \frac{\beta^{\alpha}}{\Gamma\alpha}\left(\frac{1}{\theta}\right)^{\alpha+1} \exp\left(-\frac{\beta}{\theta}\right); \theta > 0, \alpha, \beta > 0 \qquad (1.5)$$

having the prior mean is

$$\frac{\beta}{\alpha - 1}; \alpha > 1$$

and the prior variance is

$$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}; \alpha > 2 .$$

For the situation where life researchers have no prior information about the parameter θ, the uniform, quasi or improper prior may be used. This study considered a class of quasi prior defined as

$$g_2(\theta) = \frac{1}{\theta^{d}} \exp\left(-\frac{p\,d}{\theta}\right); \theta > 0, p, d > 0 . \qquad (1.6)$$

If $d = 0$ then a diffuse prior results, and if $d = 1, p = 0$ then a non–informative prior results. For a set of values of d and p, that satisfies the equality $\Gamma(d-1) = (p\,d)^{d-1}$ makes $g_2(\theta)$ an proper prior. The prior mean and prior variance are given as

$$\frac{\Gamma(d-2)}{(p\,d)^{d-2}}; d \ge 2$$

and

$$\frac{\Gamma(d-3)}{(p\,d)^{2d-4}}\left((p\,d)^{d-1} - (d-3)\Gamma(d-2)\right); d > 3.$$

Some Bayes estimators and Bayes Shrinkage estimators are suggested for the parameter θ when other parameters are known. The properties of these estimators are studied in terms of relative efficiencies empirically and by numerical example. The Bayes estimator of reliability function and hazard rate are obtained and their properties are studied.

Methodology

Bayes Shrinkage Estimators and their Properties

The posterior density $Z_1(\theta)$ for the parameter θ corresponding to prior $g_1(\theta)$ is obtained as

$$Z_1(\theta) = \frac{(r U_r + \beta)^{r w + \alpha}}{\Gamma(r w + \alpha)} \frac{e^{-(r U_r + \beta)/\theta}}{\theta^{r w + \alpha + 1}}; \ \theta > 0. \tag{2.1}$$

which is an inverted Gamma distribution with parameters $(r w + \alpha)$ and $(r U_r + \beta)$. The Bayes estimator of θ under SELF is obtained as

$$\hat{\theta}_1 = E_p(\theta) = \varphi_1 (r U_r + \beta), \tag{2.2}$$

where $\varphi_1 = (r w + \alpha - 1)^{-1}$ and suffix p indicates, the expectation is considered under the posterior density.

To utilize the prior information about the parameter θ in the form of a point guess value θ_0, the values of prior parameter β are chosen (Shirke & Nalawade, 2003) such as

$$E(\hat{\theta}_1) = \theta_0 \Rightarrow \varphi_1 \beta = (1 - \varphi_1 r)\theta_0. \tag{2.3}$$

Using (2.3) in (2.2), the shrinkage estimator takes the form (1.3) and is named the Bayes shrinkage estimator:

$$\bar{\theta}_1 = \lambda_1 U_r + (1 - \lambda_1)\theta_0 \ ; \ \lambda_1 = r \varphi_1. \tag{2.4}$$

The Bayes estimator θ under the LLF is obtained by simplifying the given equality

$$E_p\left(\frac{1}{\theta} e^{a\hat{\theta}_2/\theta}\right) = e^a E_p\left(\frac{1}{\theta}\right) \Rightarrow \hat{\theta}_2 = \varphi_2 (r U_r + \beta) \tag{2.5}$$

Similarly, the Bayes shrinkage estimator under the LLF is

$$\bar{\theta}_2 = \lambda_2 U_r + (1 - \lambda_2)\theta_0, \tag{2.6}$$

where

$$\lambda_2 = r \varphi_2, \ \varphi_2 = \frac{1}{a}\left(1 - \exp\left(\frac{-a}{r w + \alpha + 1}\right)\right).$$

The risks under the SELF and LLF of these estimators are obtained as

$$R_{(S)}(\hat{\theta}_i) = \theta^2 (r(r+1)\varphi_i^2 - 2 r \varphi_i + 1) + 2\theta\beta\varphi_i (r\varphi_i - 1) + \beta^2 \varphi_i^2,$$

$$R_{(L)}(\hat{\theta}_i) = \frac{e^{C_i}}{(1 - a\varphi_i)^r} - (a \varphi_i r + C_i + 1), \quad C_i = a\left(\frac{\varphi_i \beta}{\theta} - 1\right),$$

$$R_{(S)}(\bar{\theta}_i) = \theta^2 \left\{\lambda_i^2 \left(\frac{r+1}{r} + \delta(\delta - 2)\right) + (1-\delta)^2 (1 - 2\lambda_i)\right\}$$

and

$$R_{(L)}(\bar{\theta}_i) = \frac{r^r e^{a(\delta(1-\lambda_i) - 1)}}{(r - a\lambda_i)^r} - 1 + a(1-\delta)(1-\lambda_i);$$

$$\delta = \frac{\theta_0}{\theta}, \ i = 1, 2.$$

Similarly, the Bayes risks of these estimators are

$$R_{(BS)}(\hat{\theta}_i) = \beta^2 \left(\frac{r(r+1)\varphi_i^2 - 2r\varphi_i + 1}{(\alpha-1)(\alpha-2)} + \frac{2\varphi_i(r\varphi_i - 1)}{(\alpha-1)} + \varphi_i^2\right),$$

$$R_{(BL)}(\hat{\theta}_i) = \frac{e^{-a}}{(1 - a\varphi_i)^{\alpha + r}} - a \varphi_i (\alpha + r) + a - 1,$$

$$R_{(BS)}(\bar{\theta}_i) = (1-\lambda_i)^2 \left(\theta_0^2 - \frac{2\beta\theta_0}{\alpha - 1} + \frac{\beta^2}{(\alpha-1)(\alpha-2)}\right) + \frac{\beta^2 \lambda_i^2}{r(\alpha-1)(\alpha-2)}$$

and

$$R_{(BL)}\left(\bar{\theta}_i\right)=\frac{r^r e^{-a}}{\left(r-a\,\lambda_i\right)^r}\left(1-\frac{a\left(1-\lambda_i\right)\theta_0}{\beta}\right)^{-\alpha}$$
$$-1+\frac{a\left(\beta-\alpha\theta_0\right)}{\beta\left(1-\lambda_i\right)^{-1}}.$$
$$-1+\frac{a\left(\beta-\alpha\theta_0\right)}{\beta\left(1-\lambda_i\right)^{-1}}.$$

The relative efficiency of $\bar{\theta}_i$ with respect to U_r under the SELF and LLF loss criterions are defined as

$$RE_{(S)}\left(\bar{\theta}_i, U_r\right)=\frac{R_{(S)}\left(U_r\right)}{R_{(S)}\left(\bar{\theta}_i\right)}$$

and

$$RE_{(L)}\left(\bar{\theta}_i, U_r\right)=\frac{R_{(L)}\left(U_r\right)}{R_{(L)}\left(\bar{\theta}_i\right)}; i=1,2.$$

The expressions of relative efficiencies are the functions of r, a, v, δ and α. For the selected values of $r=04(02)10$; $a=0.25,0.50,1.00,1.50$; $v=1.00,1.50$; $\delta=0.40,0.50(0.25)1.50,1.60$ and $\alpha=1.50,03,05,07,10,15$. The relative efficiencies have been calculated and are presented in Tables 1–4 for selected parametric values.

The positive values of 'a' are considered because overestimation in mean life is more serious than underestimation. To guard against the large risk, the large values of 'a' may be ignored and the smaller values of $a(\leq 2)$ are considered (see Singh, et al., 2007).

Results

Tables 1 and 2 show that the estimator $\bar{\theta}_1$ is more efficient than the UMVU estimator U_r under the SELF and LLF for all selected parametric set of values when $r \leq 06$. For large $r \geq 08$, the effective interval decreases with large $\alpha \geq 10$. The efficiency attains maximum at the point $\delta=1.00$ (except $\alpha=1.50$ when the loss criterion is LLF) and the efficiency decreases as r increases for all considered values of δ.

In addition, under LLF loss criterion the efficiency increases with 'a' increases for all considered values of δ for small $r \leq 06$, and for large r in the interval $\delta \leq 1.00$. The estimator $\bar{\theta}_2$ performs uniformly well with respect to U_r for all considered values of the parametric space when r is small and in the interval $0.50 \leq \delta \leq 1.25$ when $r \geq 06$ (under the SELF–criterion).

In addition, the effective interval decreases as r or α increases (Table 3). The increasing trend in efficiency is also observed when 'a' increases in the interval $\delta \in [0.75,1.25]$ for other fixed parametric values.

The estimator $\bar{\theta}_2$ performs uniformly well with respect to U_r under LLF loss criterion for all considered values of the parameter space (Table 4). The increasing trend in efficiency is observed when 'a' increases for all δ when r is small and in the range $\delta \leq 1.25$ when $r \geq 08$.

Using Tables 3 and 4, it may be concluded that the efficiency reaches its maximum at the point $\delta=1.00$. The efficiency decreases as r increases for all considered values of parametric space.

Further, as v increases, the gain in efficiency is recorded only in close vicinity of the guess value and true value of the parameter but the effective interval becomes smaller for both the Bayes shrinkage estimators.

Remark

Note that the posterior density with respect to the quasi prior $g_2(\theta)$ is

$$Z_2(\theta)=\frac{\left(r\,U_r+p\,d\right)^{r\,w+d-1}}{\Gamma\left(r\,w+d-1\right)}\frac{e^{-\left(r\,U_r+p\,d\right)/\theta}}{\theta^{(r\,w+d)}}.$$

The only changes in the posterior (2.1) are replacement α and β by $d-1$ and $p\,d$ respectively. Hence, all the results are valid by substitution of these two.

Table 1: Relative Efficiency for the Bayes Shrinkage Estimator $\bar{\theta}_1$ with Respect to U_r Under the SELF

r	δ	α					
		1.50	3.00	5.00	7.00	10.00	15.00
04	0.40	1.2378	1.6544	1.6393	1.4741	1.2741	1.0864
	0.50	1.2462	1.8000	2.0000	1.9231	1.7423	1.5283
	0.75	1.2607	2.1176	3.2000	4.0000	4.6621	4.9846
	1.00	1.2656	2.2500	4.0000	6.2500	10.562	20.250
	1.25	1.2607	2.1176	3.2000	4.0000	4.6621	4.9846
	1.50	1.2462	1.8000	2.0000	1.9231	1.7423	1.5283
	1.60	1.2378	1.6544	1.6393	1.4741	1.2741	1.0864
06	0.40	1.1563	1.4337	1.4172	1.2658	1.0866	1.0708
	0.50	1.1615	1.5238	1.6667	1.6000	1.4286	1.2121
	0.75	1.1706	1.7067	2.3810	2.9091	3.3898	3.6530
	1.00	1.1736	1.7778	2.7778	4.0000	6.2500	11.111
	1.25	1.1706	1.7067	2.3810	2.9091	3.3898	3.6530
	1.50	1.1615	1.5238	1.6667	1.6000	1.4286	1.2121
	1.60	1.1563	1.4337	1.4172	1.2658	1.0866	1.0708
10	0.40	1.1163	1.3242	1.3081	1.1689	1.0721	0.7701
	0.50	1.1202	1.3889	1.5000	1.4412	1.2788	1.0614
	0.75	1.1267	1.5152	2.0000	2.3902	2.7656	2.9877
	1.00	1.1289	1.5625	2.2500	3.0625	4.5156	7.5625
	1.25	1.1267	1.5152	2.0000	2.3902	2.7656	2.9877
	1.50	1.1202	1.3889	1.5000	1.4412	1.2788	1.0614
	1.60	1.1163	1.3242	1.3081	1.1689	1.0721	0.7701
15	0.40	1.0927	1.2587	1.2437	1.1150	1.0219	0.7150
	0.50	1.0957	1.3091	1.4000	1.3474	1.1934	0.9763
	0.75	1.1008	1.4049	1.7818	2.0898	2.3967	2.5888
	1.00	1.1025	1.4400	1.9600	2.5600	3.6100	5.7600
	1.25	1.1008	1.4049	1.7818	2.0898	2.3967	2.5888
	1.50	1.0957	1.3091	1.4000	1.3474	1.1934	0.9763
	1.60	1.0927	1.2587	1.2437	1.1150	1.0219	0.7150

Table 2: Relative Efficiency for the Bayes Shrinkage Estimator $\bar{\theta}_1$ with respect to U_r under the LLF

r = 04		α					
a	δ	1.50	3.00	5.00	7.00	10.00	15.00
0.25	0.40	1.2666	1.7672	1.7814	1.6073	1.3911	1.1874
	0.50	1.2715	1.9093	2.1609	2.0867	1.8925	1.6607
	0.75	1.2770	2.2010	3.3852	4.2651	4.9906	5.3386
	1.00	1.2729	2.2882	4.1003	6.4364	10.923	21.021
	1.25	1.2594	2.1145	3.2063	4.0280	4.7259	5.0860
	1.50	1.2307	1.7608	1.9872	1.9241	1.7524	1.5412
	1.60	1.2257	1.6068	1.6256	1.4709	1.2763	1.0894
0.50	0.40	1.2997	1.9024	1.9552	1.7709	1.5347	1.3115
	0.50	1.3007	2.0404	2.3580	2.2880	2.0777	1.8240
	0.75	1.2968	2.3032	3.6144	4.5947	5.4006	5.7815
	1.00	1.2835	2.3423	4.2400	6.6931	11.414	22.060
	1.25	1.2613	2.1256	3.2413	4.0963	4.8409	5.2460
	1.50	1.2311	1.7640	1.9921	1.9437	1.7805	1.5706
	1.60	1.2310	1.6108	1.6262	1.4818	1.2912	1.1036
1.00	0.40	1.3830	2.2709	2.4450	2.2347	1.9422	1.6636
	0.50	1.3757	2.3977	2.9135	2.8603	2.6054	2.2894
	0.75	1.3510	2.5174	4.2662	5.5404	6.5833	7.0633
	1.00	1.3182	2.5891	4.6862	7.5059	12.955	25.292
	1.25	1.2782	2.2041	3.4259	4.3950	5.2801	5.8092
	1.50	1.2321	1.7870	2.0697	2.0563	1.9090	1.6956
	1.60	1.2323	1.6214	1.6820	1.5583	1.3718	1.1766
1.50	0.40	1.5036	2.8418	3.2760	3.0281	2.6406	2.2668
	0.50	1.4861	2.8655	3.8543	3.8407	3.5120	3.0895
	0.75	1.4372	2.9731	5.3708	7.1652	8.6319	9.2915
	1.00	1.3822	3.0604	5.5108	9.0029	15.782	31.192
	1.25	1.3225	2.4017	3.8550	5.0453	6.1813	6.9176
	1.50	1.2594	1.9032	2.2881	2.3249	2.1938	1.9654
	1.60	1.2335	1.7157	1.8501	1.7503	1.5608	1.3455

Table 3: Relative Efficiency for the Bayes Shrinkage Estimator $\bar{\theta}_2$ with Respect to U_r Under the SELF

r = 04		α					
a	δ	1.50	3.00	5.00	7.00	10.00	15.00
0.25	0.40	1.6932	1.6299	1.4640	1.3237	1.1778	1.0374
	0.50	1.9175	1.9995	1.9156	1.7925	1.6368	1.4674
	0.75	2.4770	3.2597	4.0396	4.5231	4.8735	5.0000
	1.00	2.7438	4.1266	6.4079	9.1891	14.298	25.314
	1.25	2.4770	3.2597	4.0396	4.5231	4.8735	5.0000
	1.50	1.9175	1.9995	1.9156	1.7925	1.6368	1.4674
	1.60	1.6932	1.6299	1.4640	1.3237	1.1778	1.0374
0.50	0.40	1.6944	1.6199	1.4540	1.3161	1.1730	1.0348
	0.50	1.9341	1.9981	1.9079	1.7850	1.6312	1.4642
	0.75	2.5485	3.3188	4.0782	4.5455	4.8817	4.9998
	1.00	2.8504	4.2566	6.5691	9.3816	14.537	25.631
	1.25	2.5485	3.3188	4.0782	4.5455	4.8817	4.9998
	1.50	1.9341	1.9981	1.9079	1.7850	1.6312	1.4642
	1.60	1.6944	1.6199	1.4540	1.3161	1.1730	1.0348
1.00	0.40	1.6917	1.5990	1.4342	1.3013	1.1636	1.0298
	0.50	1.9611	1.9928	1.8922	1.7701	1.6203	1.4578
	0.75	2.6919	3.4349	4.1529	4.5883	4.8971	4.9992
	1.00	3.0737	4.5267	6.9016	9.7765	15.026	26.276
	1.25	2.6919	3.4349	4.1529	4.5883	4.8971	4.9992
	1.50	1.9611	1.9928	1.8922	1.7701	1.6203	1.4578
	1.60	1.6917	1.5990	1.4342	1.3013	1.1636	1.0298
1.50	0.40	1.6830	1.5770	1.4147	1.2868	1.1544	1.0249
	0.50	1.9806	1.9846	1.8760	1.7554	1.6096	1.4516
	0.75	2.8351	3.5477	4.2240	4.6284	4.9113	4.9983
	1.00	3.3113	4.8108	7.2479	10.185	15.528	26.934
	1.25	2.8351	3.5477	4.2240	4.6284	4.9113	4.9983
	1.50	1.9806	1.9846	1.8760	1.7554	1.6096	1.4516
	1.60	1.6830	1.5770	1.4147	1.2868	1.1544	1.0249

Table 4: Relative Efficiency for the Bayes Shrinkage Estimator $\bar{\theta}_2$ with respect to U_r under the LLF

r = 04		α					
a	δ	1.50	3.00	5.00	7.00	10.00	15.00
0.25	0.40	1.8240	1.7718	1.5965	1.4448	1.2866	1.1344
	0.50	2.0515	2.1615	2.0788	1.9469	1.7782	1.5948
	0.75	2.5926	3.4508	4.3086	4.8389	5.2200	5.3538
	1.00	2.7988	4.2316	6.6005	9.4934	14.815	26.304
	1.25	2.4751	3.2671	4.0691	4.5772	4.9570	5.1097
	1.50	1.8947	1.9878	1.9172	1.8010	1.6490	1.4802
	1.60	1.6636	1.6171	1.4613	1.3250	1.1808	1.0401
0.50	0.40	1.9892	1.9353	1.7475	1.5846	1.4144	1.2504
	0.50	2.2395	2.3606	2.2710	2.1283	1.9461	1.7481
	0.75	2.8215	3.7598	4.6905	5.2598	5.6626	5.7964
	1.00	2.9917	4.5192	7.0418	10.119	15.778	27.986
	1.25	2.5656	3.3663	4.1820	4.7057	5.1063	5.2799
	1.50	1.9049	1.9946	1.9310	1.8208	1.6731	1.5056
	1.60	1.6644	1.6175	1.4633	1.3322	1.1910	1.0510
1.00	0.40	2.4607	2.4012	2.1776	1.9825	1.7780	1.5802
	0.50	2.7745	2.9280	2.8197	2.6462	2.4254	2.1857
	0.75	3.4681	4.6392	5.7837	6.4691	6.9391	7.0775
	1.00	3.5344	5.3440	8.3263	11.959	18.629	32.999
	1.25	2.8427	3.7029	4.5902	5.1748	5.6420	5.8720
	1.50	1.9870	2.0828	2.0351	1.9345	1.7908	1.6201
	1.60	1.6975	1.6564	1.5234	1.3986	1.2587	1.1151
1.50	0.40	3.2712	3.2019	2.9164	2.6657	2.4019	2.1455
	0.50	3.6926	3.9030	3.7633	3.5375	3.2503	2.9387
	0.75	4.4534	6.1474	7.6678	8.5593	9.1522	9.3049
	1.00	4.5695	6.7637	10.564	15.186	23.661	41.896
	1.25	3.3496	4.3562	5.4076	6.1173	6.7074	7.0303
	1.50	2.2055	2.3275	2.3004	2.2058	2.0581	1.8726
	1.60	1.8476	1.8227	1.6995	1.5742	1.4268	1.2692

Example: Exponential Failure Model

Two hundred electronic tubes were tested under the exponential failure model with the parameter $\theta = 04$ and the test was terminated after the first six items failed. The failure times (in hours) were recorded as follows:

83.5, 221, 356, 478, 535, 632

The relative efficiencies and Bayes risks of the proposed estimators were obtained for $\alpha = 5.00, 8.50, 10;\ \beta = 20, 32, 50;\ a = 0.50$ $1.00, 1.50$ and are presented in Table 5 for $\alpha = 8.50, \beta = 32$ and $a = 0.50$.

It may be concluded that the relative efficiency attains maximum at point $\theta = \theta_0$ for all considered values. Under the LLF criterion the gain in efficiency is larger than the SELF–criterion when $\delta \le 1.25$. Further, the Bayes risks are nominal when the loss criterion is LLF. The risks have the tendency to be smaller when $\theta > \theta_0$ and attains minimum when $\theta = \theta_0$ and then increases. Further, both the risk and Bayes risk decreases (increases) when $'a'(\alpha)$ increases under both loss criterions when other parametric values are fixed.

The Bayes Estimator of Reliability Function and Hazard Rate

The Reliability function $\Psi(t)$ at time $t\,(>0)$ is defined as

$$\Psi(t) = \frac{1}{\Gamma(w)} \int_{t^v/\theta}^{\infty} e^{-S} S^{w-1}\, dS. \qquad (3.1)$$

Similarly, the Hazard rate at time $t\,(>0)$ is given by

$$\rho(t) = \frac{v\, t^{wv-1}}{\theta^w e^{t^v/\theta}} \left\{ \int_{t^v/\theta}^{\infty} e^{-S} S^{w-1}\, dS \right\}^{-1}. \qquad (3.2)$$

In particular, for the exponential distribution $(w = v = 1)$ the Reliability function and Hazard rate are given as

$$\Psi(t) = \exp(-t/\theta) \text{ and } \rho(t) = 1/\theta. \qquad (3.3)$$

The Bayes estimator of the reliability function and hazard rates under the SELF, corresponding to the posterior density $Z_1(\theta)$ are obtained as

$$\Psi_1 = G(0, \infty, J_1) \text{ and } \rho_1 = G(0, \infty, J_2); \qquad (3.4)$$

where

$$G(u_1, u_2, \xi) = \frac{(r U_r + \beta)^{r_1}}{\Gamma(r_1)} \int_{u_1}^{u_2} e^{-(r U_r + \beta) z} z^{r_1 - 1}(\xi)\, dz,$$

ξ be the function of z, $r_1 = r w + \alpha$, $J_1 = \int_{t^v z}^{\infty} \frac{e^{-S} S^{w-1}}{\Gamma(w)} dS$ and $J_2 = \frac{v z^w t^{wv-1}}{e^{t^v z}} \left\{ \int_{t^v z}^{\infty} e^{-S} S^{w-1} dS \right\}^{-1}$.

Similarly, the Bayes estimate of the reliability function Ψ_2, and the Hazard rate ρ_2 under the LLF for the given posterior $Z_1(\theta)$ are obtained by solving the given equality

$$G\left(0, \infty, \left(J_1^{-1} \exp\left(a\ J_1^{-1} \Psi_{(L)}(t)\right)\right)\right) = e^a\, G\left(0, \infty, J_1^{-1}\right)$$

and

$$G\left(0, \infty, \left(J_2^{-1} \exp\left(a\, J_2^{-1} \rho_{(L)}(t)\right)\right)\right) = e^a\, G\left(0, \infty, J_2^{-1}\right) \qquad (3.5)$$

The close form of the Bayes estimators $\Psi(t)$ and $\rho(t)$ under the LLF are nonexistent, therefore, the risk and Bayes risks do not exist in the closed form. For convenience, consider Varian's (1975) asymmetric loss function defined for any parameter θ as

$$L(\Delta') = e^{a\,\Delta'} - a\,\Delta' - 1;\ \Delta' = \hat{\theta} - \theta.$$

Hence, the Bayes estimators $\Psi(t)$ and $\rho(t)$ under the LLF are given by

$$\Psi_3 = -\frac{1}{a}\ln\left\{\frac{(r\,U_r+\beta)^{r_1}}{\Gamma(r_1)}\;G\left(0,\infty,e^{-a\,J_1}\right)\right\}$$

and

$$\rho_3 = -\frac{1}{a}\ln\left\{\frac{(r\,U_r+\beta)^{r_1}}{\Gamma(r_1)}\;G\left(0,\infty,e^{-a\,J_2}\right)\right\}. \tag{3.6}$$

The risk and Bayes risk of these Bayes estimators under the SELF and LLF do not exist in a closed form. However, the numerical findings of the risk and Bayes risk for these Bayes estimators under SELF and LLF $R_{(S)}(\Psi_i)$, $R_{(L)}(\Psi_i)$, $R_{(BS)}(\Psi_i)$, $R_{(BL)}(\Psi_i)$, $R_{(S)}(\rho_i)$, $R_{(L)}(\rho_i)$, $R_{(BS)}(\rho_i)$ and $R_{(BL)}(\rho_i)$; $i=1,3$ are obtained for a particular case when $v=w=1$.

Example: Risks and Bayes Risks

Consider the above example with $t=250h$ the Bayes estimates for the reliability function and hazard rate, risks and Bayes risks as obtained are presented in Table 6.

The risk of the estimator Ψ_1 increases as β increases when $\alpha \geq 8.50$ under the LLF. A similar trend is observed for the risk and Bayes risk of Ψ_3 as β increases when $\alpha \geq 8.50$ under the LLF.

Further, the Bayes risk of Ψ_1 and Ψ_3 increases when 'a' increases under both loss criterions. The risk and Bayes risk decreases when α increases (except $\beta \geq 50.00$) when other parametric values are fixed for Ψ_1 (LLF–criterion) and Ψ_3 (SELF and LLF criterions).

The risk and Bayes risk for the estimators ρ_1 and ρ_3 increases as β increases for all the considered values of α under the SELF and LLF (except $\alpha = 5.00$) when other parametric values are fixed. The risk and Bayes risk of ρ_1 and ρ_3 also increases (decreases) under both loss criterions when 'a'(α) increases.

References

Pandey, B. N., & Singh, J. (1977). Estimation of variance of Normal population using apriori information. *Journal of Indian Statistical Association*, *15*, 141-150.

Prakash, G., & Singh, D. C. (2006). Shrinkage testimators for the inverse dispersion of the inverse Gaussian distribution under the LINEX loss function. *Austrian Journal of Statistics*, *35*(*4*), 463-470.

Prakash, G., and Singh, D.C. (2008). Shrinkage Estimation in Exponential Type-II Censored Data under LINEX Loss. *Journal of the Korean Statistical Society, 37 (1),* 53-61.

Prakash, G., and Singh, D. C. (2009). A Bayesian Shrinkage Approach in Weibull Type-II Censored Data Using Prior Point Information. *REVSTAT – Statistical Journal, 7 (2),* 171-187.

Shirke, D. T., & Nalawade, K. T. (2003). Estimation of the parameter of binomial distribution in presence of prior point information. *Journal of the Indian Statistical Association*, *41*(*1*), 117-128.

Singh, D. C., Prakash, G., & Singh, P. (2007). Shrinkage testimators for the shape parameter of Pareto distribution using the LINEX loss function. *Communication in Statistics Theory and Methods*, *36*(*4*), 741-753.

Thompson, J. R. (1968). Some shrinkage techniques for estimating the mean. *Journal of the American Statistical Association*, *63*, 113-122.

Varian, H. R. (1975). A Bayesian approach to real estate assessment. In *Studies in Bayesian Econometrics and Statistics in Honor of L. J. Savage*, Eds. S. E. Feinberge, and A. Zellner, 195-208. Amsterdam: North Holland.

Table 5: Risk and Bayes Risk of the Bayes Shrinkage Estimators

$r=06 :: \theta=04 :: \alpha=8.50 :: \beta=32.00 :: a=0.50$							
δ	0.40	0.50	0.75	1.00	1.25	1.50	1.60
θ_0	1.60	2.00	3.00	4.00	5.00	6.00	6.40
$RE_{(S)}(\bar{\theta}_1, U_r)$	1.1571	1.5140	3.1921	5.0625	3.1921	1.5140	1.1571
$RE_{(L)}(\bar{\theta}_1, U_r)$	1.3339	1.7305	3.5411	5.2767	3.1873	1.4909	1.1333
$RE_{(S)}(\bar{\theta}_2, U_r)$	1.0278	1.3889	3.4623	6.8918	3.4623	1.3889	1.0278
$RE_{(L)}(\bar{\theta}_2, U_r)$	1.1872	1.5902	3.8596	7.2137	3.4783	1.3748	1.0101
$R_{(BS)}(\bar{\theta}_1)$	3.7507	3.1416	2.0511	1.5778	1.7219	2.4832	2.9606
$R_{(BL)}(\bar{\theta}_1)$	0.0164	0.0130	0.0103	0.0098	0.0180	0.0336	0.0421
$R_{(BS)}(\bar{\theta}_2)$	4.3068	3.5505	2.1963	1.6086	1.7875	2.7328	3.3256
$R_{(BL)}(\bar{\theta}_2)$	0.0184	0.0141	0.0183	0.0100	0.0202	0.0398	0.0505

Table 6: Risk and Bayes Risk of the Reliability Function and Hazard Rates

$r=06 :: \theta=04 :: t=250 :: \alpha=8.50 :: \beta=32.00$							
$a \rightarrow$	0.50	1.00	1.50	$a \rightarrow$	0.50	1.00	1.50
Ψ_1	0.0097	0.0097	0.0097	Ψ_3	97.860	48.930	32.620
$R_{(S)}(\Psi_1)$	16.000	16.000	16.000	$R_{(S)}(\Psi_3)$	7.1480	9.0530	10.810
$R_{(BS)}(\Psi_1)$	35.192	35.192	35.192	$R_{(BS)}(\Psi_3)$	37.160	38.380	42.950
$R_{(L)}(\Psi_1)$	1.1353	3.0183	5.0025	$R_{(L)}(\Psi_3)$	1.1550	1.9160	3.8300
$R_{(BL)}(\Psi_1)$	1.7744	4.1825	6.6077	$R_{(BL)}(\Psi_3)$	1.9700	4.3600	6.7790
$a \rightarrow$	0.50	1.00	1.50	$a \rightarrow$	0.50	1.00	1.50
ρ_1	0.0218	0.0218	0.0218	ρ_3	53.270	26.704	17.849
$R_{(S)}(\rho_1)$	13.941	13.941	13.941	$R_{(S)}(\rho_3)$	6.7758	8.2750	9.9006
$R_{(BS)}(\rho_1)$	33.257	33.257	33.257	$R_{(BS)}(\rho_3)$	35.803	37.005	41.536
$R_{(L)}(\rho_1)$	1.0214	2.7575	4.6040	$R_{(L)}(\rho_3)$	1.1513	1.7864	3.6102
$R_{(BL)}(\rho_1)$	1.7085	4.0468	6.4038	$R_{(BL)}(\rho_3)$	1.9298	4.2714	6.6433

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 558-567

1538 – 9472/10/$95.00

Empirical Characteristic Function Approach to Goodness of Fit Tests for the Logistic Distribution under SRS and RSS

M. T. Alodat S. A. Al-Subh Kamaruzaman Ibrahim Abdul Aziz Jemain
Yarmouk University, Irbid, Jordan
Kebangsaan University, Selangor, Malaysia

The integral of the squares modulus of the difference between the empirical characteristic function and the characteristic function of the hypothesized distribution is used by Wong and Sim (2000) to test for goodness of fit. A weighted version of Wong and Sim (2000) under ranked set sampling**,** a sampling technique introduced by McIntyre (1952), is examined. Simulations that show the ranked set sampling counterpart of Wong and Sim (2000) is more powerful.

Key words: Goodness of fit test, empirical distribution function, logistic distribution, ranked set sampling, simple random sampling.

Introduction

In any one-sample goodness of fit test problem where a random sample $X_1, X_2,...,X_r$ from an unknown distribution function $F(x)$ is given in order to test the hypothesis $H_o : F(x) = F_o(x)$ for all x against the hypothesis $H_1 : F(x) \neq F_o(x)$, where $F_o(x)$ is a known distribution function. Stephens (1974) provided a practical guide to goodness of fit tests using statistics based on the empirical distribution function (EDF). Green and Hegazy (1976) studied modified forms of the Kolmogorov-Smirnov D, Cramer-von Mises W^2 and the Anderson-Darling A^2 goodness of fit tests. Stephens (1979) gave goodness of fit tests for the logistic distribution. A comprehensive survey for goodness of fit tests can be found in the book of D'Agostino and Stephens (1986).

Gürtler and Henze (2000) used another approach to test for goodness of fit for the Cauchy distribution. They built their test based on the weighted distance between the empirical characteristic function of the sample and the characteristic function of the null distribution, that is, they considered the test statistic of the form:

$$T = r\int_{-\infty}^{\infty} \left|\Phi_r(t) - e^{-|t|}\right|^2 w(t)dt, \quad w(t) = e^{-\kappa|t|}, \quad \kappa > 0,$$
$$= \frac{2}{r}\sum_{j,\kappa=1}^{r} \frac{\kappa}{\kappa^2 + (y_j - y_\kappa)^2} - 4\sum_{j=1}^{r} \frac{1+\kappa}{(1+\kappa)^2 + y_j^2} + \frac{2r}{2+\kappa}, \quad (1)$$

where $y_j = (x_j - \hat{\alpha})/\hat{\beta}$, and

$$\Phi_r(t) = \frac{1}{r}\sum_{j=1}^{r} \exp(Ity_j)$$

is the empirical characteristic function of the sample. The function $w(t)$ is a weight function and $\hat{\alpha}, \hat{\beta}$ are the Maximum Likelihood

M. T. Alodat is an Associate Professor in the Department of Statistics, Yarmouk University, Jordan. Email: alodatmts@yahoo.com. Sameer Al-Subh is an Assistant Professor Mathematics Department, Jerash University, Jordan. Email: salsubh@yahoo.com. Kamaruzaman Ibrahim is a Professor of statistics in the School of Mathematical Sciences, Kebangsaan University, Malaysia. Email: kamarulz@pkrisc.cc.ukm.my. Abdul Aziz Jemain is a professor of statistics in the School of Mathematical Sciences, Malaysia. Email: azizj@pkrisc.cc.ukm.my.

Estimators (MLE) of α and β, the location and the scale parameters of the Cauchy distribution. Wong and Sim (2000) studied the test statistic T when $w(t) \equiv 1$, for different distributions. Matsui and Takemura (2005) also considered the problem of Gürtler and Henze (2000) but used a different research design. For more information about the application of the empirical characteristic function to goodness of fit test see Feuerverger and Mureika (1977), Meintnis (2004), Epps (2005) and Towhidi & Salmanpour (2007).

In various situations, visual ordering of sample units (with respect to the variable of interest) is less expensive against its quantification. For statistical populations with such a property, McIntyre (1952) was the first to employ the visual ranking of sampling units in order to select a sample that is more informative than a simple random sample. Later, his sampling technique was known as Ranked Set Sampling (RSS). Without any theoretical developments, he showed that the RSS is more efficient and cost effective method than the Simple Random Sampling (SRS) technique. An RSS sample can be obtained as follows:

1. Select m random samples from the population of interest each of size m.
2. From the i^{th} sample detect, using a visual inspection, the i^{th} order statistic and choose it for actual quantification, say, Y_i, $i = 1,...,m$.
3. RSS is the set of the order statistics $Y_1,...,Y_m$.
4. The technique could be repeated r times to obtain additional observations.

Takahasi and Wakimoto (1968) developed the theoretical framework for RSS.

Visual ranking is accomplished based on an experimenter's experience. Hence, two factors affect the efficiency of an RSS: the set size and the ranking errors. The larger the set size, the larger the efficiency of the RSS; however, the larger the set size, visual ranking is more difficult and the ranking error is larger (Al-Saleh & Al-Omari, 2002). For this, several authors have modified MacIntyre's RSS scheme to reduce the error in ranking and to make visual ranking easier for an experimenter. Samawi, et al. (1996) investigated Extreme Ranked Set Sample (ERSS), i.e. they quantified the smallest and the largest order statistics. Muttlak (1997) introduced Median Ranked Set Sampling (MRSS) which consists of quantifying only the median in each set. Bhoj (1997) proposed a modification to the RSS and called it new ranked set sampling (NRSS). Al-Odat and Al-Saleh (2001) introduced the concept of varied set size RSS, which is called later by moving extremes ranked set sampling (MERSS). For more details about these developments see Chen (2000).

Stockes and Sager (1988) were the first who proposed a Kolmogorov-Smirnov goodness of fit test based on the empirical distribution function of an RSS. In addition, they derived the null distribution of their proposed test. Al-Subh, et al. (2008) studied the Chi-square test for goodness of fit test under the RSS technique and its modifications. Their simulation showed that the Chi-square test for the null logistic distribution is more powerful than its counterpart under SRS technique. This article examines the power of the test given in equation (1) when sample is selected using one of the modifications of the RSS, specifically, the modification that chooses only the i^{th} order statistic for quantification.

Problem Formalization

It can be noted that testing the hypotheses:

$$H_o : F(x) = F_o(x), \quad \forall x$$

vs.

$$H_1 : F(x) \neq F_o(x)$$

is equivalent to testing the hypothesis

$$H_o^* : G_i(y) = G_{io}(y), \quad \forall y$$

vs.

$$H_1^* : G_i(y) \neq G_{io}(y)$$

for some i, where $G_i(y)$, $G_{io}(y)$ are the cdf's of the i^{th} order statistics of random samples of size $2m-1$ chosen from $F(x)$ and $F_o(x)$, respectively. The rationale behind choosing an odd set size - rather than an even one - is to simplify the comparison with the median RSS, because an even set size produces two middle values. Moreover, quantifying the two middle sample units is more expensive than quantifying one sampling unit. If $f(y)$ and $f_o(y)$ are the corresponding pdf's of $F(x)$ and $F_o(x)$, respectively, then according to Arnold, et al. (1992), $G_i(y)$ and $G_{io}(y)$ have the following representations:

$$G_i(y)=\sum_{j=i}^{2m-1}\binom{2m-1}{j}[F(y)]^j[1-F(y)]^{(2m-1)-j}$$

and

$$G_{io}(y)=\sum_{j=i}^{2m-1}\binom{2m-1}{j}[F_o(y)]^j[1-F_o(y)]^{(2m-1)-j},$$

respectively. The corresponding pdf's are

$$g_i(y)=\frac{(2m-1)!}{(i-1)!(2m-1-i)!}F(y)^{i-1}(1-F(y))^{2m-1-i}f(y)$$

and

$$g_{io}(y)=\frac{(2m-1)!}{(i-1)!(2m-1-i)!}F_o(y)^{i-1}(1-F_o(y))^{2m-1-i}f_o(y),$$

respectively. It can be shown that $G_i(y)=G_{i0}(y)$ if and only if $F(x)=F_o(x)$, which means this statistical testing problem is invariant.

If ranked set sampling is employed to collect the data using the i^{th} order statistic, then the resulting data is used to build a test based on the empirical characteristic function of these data as described in equation (1). The empirical characteristic function and the population characteristic function that should be used, respectively, are:

$$\Phi_{ri}(t)=\frac{1}{r}\sum_{j=1}^{r}\exp(ItY_j),$$

and

$$\Phi_i(t)=\int_{-\infty}^{\infty}\exp(Ity)dG_{io}(y).$$

Hence, a ranked set sample counterpart of the test T is given by

$$T_i^*=r\int_{-\infty}^{\infty}\left|\Phi_{ri}(t)-\Phi_i(t)\right|^2 w(t)dt, \quad (2)$$

where $w(t)$ is a suitable weight function. Using complex integration, it may be shown that:

$$\Phi_{2m-1}(t)=Beta(1-I\kappa t,\ 1-I\kappa t).$$

The test rejects H_o^* for large values of T_i^*. Attention is restricted to the case when $F_o(x)=(1+e^{-(x-\theta)/\sigma})^{-1}$, that is, for the logistic distribution. Even for logistic distribution, the test T_i^* has no closed form as in the Cauchy case; for this, a simulation study is conducted to study the power of the test T_i^* and its counterpart T. The two tests will be compared in terms of power based on samples of the same size. The power of the T_i^* test can be calculated according to the equation

$$\text{Power of } T_i^*(H)=P_H(T_i^*>d_\alpha), \quad (3)$$

where H is a cdf under the alternative hypothesis H_1^*. Here d_α is the 100α percentage point of the distribution of T_i^* under H_o. The efficiency of the test statistic T_i^* relative to T is calculated as a ratio of powers:

$$eff\ (T_i^*,\ T)=\frac{\text{power of } T_i^*}{\text{power of } T},$$

thus, T_i^* is more powerful than T if $eff(T_i^*, T) > 1$.

Algorithms for Power and Percentage Point

The following two algorithms approximate the power and the percentage of the tests T and T_i^*.

Percentage Point Algorithm:

1. Simulate $Y_1,...,Y_r$ from $G_{io}(y)$.
2. Find T_i^* according to equation (2).
3. Repeat steps (1)-(2) to obtain $T_{i1}^*, ..., T_{i10,000}^*$.
4. Approximate d_α, the percentage point of T_i^*.

Power Algorithm:

1. Simulate $Y_1,...,Y_r$ from H, a distribution under H_1^*.
2. Find T_i^* according to equation (2).
3. Repeat the steps (1)-(2) to obtain $T_{i1}^*, ..., T_{i10,000}^*$.
4. Approximate the power of T_i^* as:

$$\text{Power of } T_i^*(H) = \frac{1}{10,000}\sum_{t=1}^{10,000} I(T_{it}^* > d_\alpha),$$

where $I(.)$ stands for indicator function.

Results

To compare tests T and T_i^*, a Monte Carlo simulation study was conducted to approximate the power of each test based on 10,000 iterations according to the algorithms shown. Due to symmetry the first and the last order statistics produced the same power; therefore, simulation results for the largest order statistic are not presented. The powers of the two tests were compared for samples sizes $r = 10,\ 20,\ 30$, set sizes $m = 1,\ 2,\ 3,\ 4$ and alternative distributions *Normal* = *N*(0. 1), *Laplace* = *L*(0, 1), *Lognormal* = *LN*(0, 1), *Cauchy* = *C*(0, 1), *StudentT* = *S*(5), *Uniform* = *U*(0, 1), *Beta* (0, 1), *ChiSquare* (5) and *Gamma* (2, 1). In addition, the following weight functions were used in the simulation study:

$$w_1(t) = \text{Real Part of Beta}(1-\kappa tI, 1-\kappa tI),$$
$$w_2(t) = \exp(-\kappa|t|),$$
$$w_3(t) = \exp(-\kappa t^2),$$
$$w_4(t) = |\cos(t)|\, e^{-\kappa t^2},$$

and

$$w_5(t) = (\kappa + t^2)^{-1}.$$

Simulation results are presented in Tables (1)-(5).

Simulation results for the uniform distribution show that the powers of all test statistics equal one, for this reason these powers are not reported in Tables (1)-(5). The simulation also shows that the efficiencies are equal to one for the non-symmetric alternatives: *Lognormal* = *LN*(0, 1), *ChiSquare* (5), *Gamma* (2, 1) and *Beta* (0, 1), thus, these are not presented in the tables.

Conclusion

Based on data in the tables, the following conclusions regarding T_i^* are put forth:

1. The efficiencies are greater than one for all alternatives, weight functions and all values of m, r and κ, thus indicating that the test T_1^* is more powerful than the test T.
2. It is noted that, for each alternative, the efficiency is increasing in m.
3. No clear pattern is observed in the efficiency values and the weight function, but for $\kappa = 1.5$ and $m = 4$, the efficiency has the highest values.
4. The worst value of the efficiency occurs when $H = N(0, 1)$ and $r = 50$.

This article considered a counterpart goodness of fit test based on the empirical characteristic function under ranked set sampling. The null

distribution and the power of the new test have no closed forms; therefore they have been obtained using simulation. The simulation results show that the ranked set sampling counterpart is more powerful than the empirical characteristic function based on a simple random sample. In addition, it also possible to improve the power of the test statistic (1) (see introduction) under different ranked set sampling schemes, however, this discussion is avoided due to space limitations.

Table 1: Values of $eff\ (T_i^*,\ T)$ Using $w_1(t)$ for $r = 10,\ 20,\ 30,\ 50$, $m = 1,\ 2,\ 3,\ 4$ and $\alpha = 0.05$

	r = 10, $w_1(t)$ = Real Part of Beta$(1-\kappa tI, 1-\kappa tI)$											
	$\kappa = 0.5$				$\kappa = 1$				$\kappa = 1.5$			
H\m	1	2	3	4	1	2	3	4	1	2	3	4
N(0, 1)	1	1.31	2.13	3.25	1	1.77	4.2	6.94	1	4.25	12.26	23.52
L(0, 1)	1	1.16	1.38	1.67	1	1.60	2.82	3.76	1	3.33	6.89	10
C(0, 1)	1	1.52	2.89	4.14	1	1.85	3.4	5.12	1	1.51	2.71	3.729
S(5)	1	1.23	1.7	2.20	1	1.48	2.73	4.02	1	2.63	6.21	10.04
r = 20												
N(0, 1)	1	1.43	2.05	2.45	1	1.45	2.26	2.62	1	2.42	4.6	5.53
L(0, 1)	1	1.15	1.35	1.56	1	1.47	2.25	2.79	1	2.27	4.35	5.94
C(0, 1)	1	1.79	3.66	5.36	1	2.06	3.83	5.47	1	1.69	3.25	4.12
S(5)	1	1.22	1.69	2.22	1	1.45	2.31	3.13	1	1.92	4.45	6.14
r = 30												
N(0, 1)	1	1.25	1.51	1.57	1	1.26	1.49	1.53	1	1.52	1.91	1.97
L(0, 1)	1	1.12	1.29	1.42	1	2.32	3.01	3.75	1	2.3	3.93	5.25
C(0, 1)	1	2.07	3.77	5.48	1	1.35	2.56	3.42	1	1.77	2.98	3.67
S(5)	1	1.18	1.63	1.89	1	1.32	1.82	2.33	1	1.69	2.98	3.8
r = 50												
N(0, 1)	1	1.06	1.09	1.09	1	1.05	1.06	1.06	1	1.08	1.09	1.09
L(0, 1)	1	1.05	1.12	1.21	1	1.16	1.35	1.52	1	1.44	2.06	2.37
C(0, 1)	1	1.57	2.81	3.53	1	2.08	3.36	3.83	1	1.73	2.53	2.74
S(5)	1	1.18	1.38	1.5	1	1.13	1.39	1.51	1	1.26	1.73	1.9

Table 2: Values of $eff\ (T_i^*,\ T)$ Using $w_2(t)$ for $r = 10,\ 20,\ 30,\ 50$, $m = 1,\ 2,\ 3,\ 4$ and $\alpha = 0.05$

	r = 10, $w_2(t) = \exp(-\kappa\|t\|)$											
	$\kappa = 0.5$				$\kappa = 1$				$\kappa = 1.5$			
*H**m*	1	2	3	4	1	2	3	4	1	2	3	4
N(0, 1)	1	1.31	1.91	2.82	1	1.35	2.4	3.88	1	1.57	3	4.82
L(0, 1)	1	1.13	1.22	1.38	1	1.15	1.61	1.87	1	1.49	2.11	2.69
C(0, 1)	1	1.67	2.44	3.56	1	1.76	3.34	4.85	1	1.53	3.13	4.43
S(5)	1	1.25	1.55	2.06	1	1.29	1.91	2.6	1	1.47	2.11	3.35
r = 20												
N(0, 1)	1	1.38	2.07	2.57	1	1.44	2.11	2.48	1	1.5	2.28	2.68
L(0, 1)	1	1.12	1.28	1.45	1	1.16	1.5	1.7	1	1.26	1.68	2.06
C(0, 1)	1	1.63	2.7	4.04	1	2.08	3.7	5.51	1	1.92	3.63	5.24
S(5)	1	1.3	1.68	2.2	1	1.3	1.79	2.37	1	1.39	2	2.64
r = 30												
N(0, 1)	1	1.28	1.64	1.78	1	1.26	1.49	1.55	1	1.28	1.51	1.55
L(0, 1)	1	1.11	1.15	1.28	1	1.21	1.38	1.55	1	1.19	1.55	1.8
C(0, 1)	1	1.54	2.73	3.99	1	1.86	3.58	4.77	1	1.78	3.37	4.41
S(5)	1	1.17	1.57	1.9	1	1.22	1.64	1.94	1	1.29	1.77	2.15
r = 50												
N(0, 1)	1	1.11	1.16	1.16	1	1.07	1.09	1.09	1	1.05	1.07	1.07
L(0, 1)	1	1.06	1.1	1.17	1	1.08	1.15	1.25	1	1.13	1.26	1.37
C(0, 1)	1	1.5	2.65	3.53	1	1.77	2.97	3.55	1	1.85	2.92	3.36
S(5)	1	1.21	1.43	1.6	1	1.13	1.34	1.47	1	1.18	1.37	1.5

Table 3: Values of $eff\ (T_i^*,\ T)$ Using $w_3(t)$ for $r = 10,\ 20,\ 30,\ 50$, $m = 1,\ 2,\ 3,\ 4$ and $\alpha = 0.05$

	r = 10, $w_3 = e^{-\kappa t^2}$											
	$\kappa = 0.5$				$\kappa = 1$				$\kappa = 1.5$			
*H**m*	1	2	3	4	1	2	3	4	1	2	3	4
N(0, 1)	1	1.35	2.33	3.66	1	1.46	3.43	5.6	1	1.59	5.21	9.18
L(0, 1)	1	1.48	2	2.46	1	1.88	2.98	3.85	1	2.45	4.48	6.17
C(0, 1)	1	1.78	3.37	5.53	1	1.97	3.91	5.88	1	1.82	3.29	4.79
S(5)	1	1.15	1.8	2.45	1	1.37	2.55	3.97	1	1.6	3.67	6.5
r = 20												
N(0, 1)	1	1.45	2	2.33	1	1.41	2.27	2.63	1	1.41	2.46	2.86
L(0, 1)	1	1.12	1.44	1.72	1	1.36	1.99	2.39	1	1.47	2.54	3.14
C(0, 1)	1	1.91	3.85	6.03	1	2.17	4.71	6.72	1	2.35	3.79	5.87
S(5)	1	1.29	1.73	2.3	1	1.34	2.13	2.9	1	1.51	2.68	3.7
r = 30												
N(0, 1)	1	1.19	1.4	1.45	1	1.24	1.45	1.48	1	1.34	1.6	1.64
L(0, 1)	1	1.12	1.33	1.52	1	1.19	1.73	2	1	1.56	2.27	2.81
C(0, 1)	1	1.99	3.95	5.35	1	2.26	4.26	5.78	1	2.84	4.48	5.76
S(5)	1	1.22	1.59	1.95	1	1.27	1.8	2.19	1	1.25	2	2.51
r = 50												
N(0, 1)	1	1.06	1.07	1.07	1	1.04	1.05	1.05	1	1.05	1.06	1.06
L(0, 1)	1	1.08	1.15	1.24	1	1.12	1.34	1.52	1	1.13	1.51	1.76
C(0, 1)	1	2.04	3.29	4.12	1	2.32	3.86	4.52	1	2.22	3.47	3.94
S(5)	1	1.15	1.35	1.45	1	1.12	1.37	1.48	1	1.17	1.49	1.62

Table 4: Values of $eff\ (T_i^*,\ T)$ Using $w_4(t)$ for $r = 10,\ 20,\ 30,\ 50$, $m = 1,\ 2,\ 3,\ 4$ and $\alpha = 0.05$

	r = 10, $w_4(t) = \lvert\cos(t)\rvert e^{-\kappa t^2}$											
	$\kappa = 0.5$				$\kappa = 1$				$\kappa = 1.5$			
*H**m*	1	2	3	4	1	2	3	4	1	2	3	4
N(0, 1)	1	2.15	4.99	8.55	1	2.92	8.08	15.10	1	5.61	15.6	31.60
L(0, 1)	1	1.95	3.54	4.56	1	3.35	6.40	9.25	1	3.41	7	11
C(0, 1)	1	1.95	3.51	5.02	1	1.84	3.26	4.82	1	1.61	3.01	4.09
S(5)	1	1.63	3.21	4.96	1	2.14	5.52	8.86	1	2.17	5.38	8.76
r = 20												
N(0, 1)	1	1.45	2.36	2.78	1	1.75	3.06	3.6	1	1.97	3.95	4.83
L(0, 1)	1	1.39	2.08	2.61	1	1.64	2.71	3.73	1	2.04	3.87	5.37
C(0, 1)	1	2.37	4.51	6.53	1	2.28	4.22	5.93	1	1.95	3.51	4.76
S(5)	1	1.54	2.37	3.18	1	1.57	3.02	4.35	1	2.13	3.90	5.60
r = 30												
N(0, 1)	1	1.27	1.49	1.53	1	1.24	1.55	1.59	1	1.36	1.77	1.81
L(0, 1)	1	1.33	1.76	2.17	1	1.39	2.23	2.71	1	1.87	3.23	3.98
C(0, 1)	1	2.59	4.83	6.2	1	2.1	4.17	5.43	1	1.89	3.41	4.33
S(5)	1	1.33	1.9	2.28	1	1.45	2.36	3.03	1	1.71	2.69	3.48
r = 50												
N(0, 1)	1	1.04	1.05	1.05	1	1.06	1.07	1.07	1	1.05	1.06	1.06
L(0, 1)	1	1.19	1.41	1.6	1	1.27	1.7	1.94	1	1.38	2.04	2.38
C(0, 1)	1	2.33	3.98	4.54	1	2.03	3.3	3.7	1	2.03	3.14	3.44
S(5)	1	1.2	1.46	1.59	1	1.24	1.57	1.71	1	1.24	1.62	1.83

Table 5: Values of $eff\ (T_i^*,\ T)$ Using $w_5(t)$ for $r = 10,\ 20,\ 30,\ 50$, $m = 1,\ 2,\ 3,\ 4$ and $\alpha = 0.05$

	$\alpha = 0.05, \text{r} = 10, w_5 = (k + t^2)^{-1}$											
	$\kappa = 0.5$				$\kappa = 1$				$\kappa = 1.5$			
H\m	1	2	3	4	1	2	3	4	1	2	3	4
N(0, 1)	1	1.56	2.76	4.26	1	1.35	2.39	3.6	1	1.37	2.24	3.26
L(0, 1)	1	1.48	1.85	2.27	1	1.23	1.54	1.74	1	1.08	1.36	1.45
C(0, 1)	1	2.13	3.56	5.6	1	1.86	3.06	4.41	1	1.64	2.75	4.39
S(5)	1	1.2	2.09	3	1	1.18	1.7	2.44	1	1.15	1.69	2.21
r = 20												
N(0, 1)	1	1.35	2.17	2.59	1	1.31	1.92	2.31	1	1.38	2	2.43
L(0, 1)	1	1.3	1.53	1.85	1	1.17	1.4	1.66	1	1.11	1.25	1.45
C(0, 1)	1	1.95	3.47	4.98	1	1.84	3.51	5.21	1	1.77	3.33	5.13
S(5)	1	1.23	1.88	2.52	1	1.2	1.77	2.33	1	1.23	1.64	2.14
r = 30												
N(0, 1)	1	1.3	1.58	1.63	1	1.27	1.53	1.61	1	1.25	1.56	1.64
L(0, 1)	1	1.16	1.33	1.51	1	1.18	1.34	1.46	1	1.13	1.24	1.34
C(0, 1)	1	1.78	3.38	4.73	1	1.92	3.58	5.1	1	1.94	3.67	5.17
S(5)	1	1.22	1.69	2.05	1	1.21	1.64	1.98	1	1.25	1.58	1.94
r = 50												
N(0, 1)	1	1.05	1.07	1.07	1	1.07	1.09	1.09	1	1.08	1.11	1.11
L(0, 1)	1	1.08	1.2	1.28	1	1.07	1.16	1.24	1	1.07	1.12	1.17
C(0, 1)	1	1.74	2.77	3.22	1	1.63	2.82	3.44	1	1.73	2.9	3.69
S(5)	1	1.2	1.41	1.54	1	1.2	1.4	1.51	1	1.19	1.37	1.51

References

Al-Odat, M. T., & Al-Saleh, M. F. (2001). A variation of ranked set sampling. *Journal of Applied Statistical Science*, *10*, 137-146.

Al-Subh, S. A., Alodat, M. T., Kamarlzaman, I., & Jemain, A. (2010). Chi-square test for goodness of fit using ranked set sampling. Under review.

Al-Saleh, M. F., & Al-Omari, A. I. (2002). Multistage ranked set sampling. *Journal of Statistical planning and Inference*, *102*, 273-286.

Arnold, B. C., Balakrishnan, N., & Nagaraja, H. N. (1992). *A first course in order statistics*. New York: John Wiley and Sons.

Bhoj, D. S. (1997). Estimation of parameters of the extreme value distribution using ranked set sampling. *Communications in Statistical Theory and Methods*, *26*(*3*), 653-667.

Chen, Z., Bai, Z., & Sinha, B. K. (2000). *Ranked set sampling: Theory and applications.* New York: Springer-Verlag.

D'Agostino, R., & Stephens, M. (1986). *Goodness- of- fit techniques*. New York: Marcel Dekker.

Epps, T. W. (2005). Tests for location-scale families based on the empirical characteristic function. *Metrika*, *62*, 99-114.

Feuerverger, A., & Mureika, R. A. (1977). The empirical characteristic function and its applications. *Annals of Statistics*, *5*, 88-97.

Green, J. R., & Hegazy, Y. A. S. (1976). Powerful modified-EDF goodness-of-fit tests. *Journal of the American Statistical Association*, *71*, 204-209.

Gürler, N., & Henze, N. (2000). Goodness-of-fit tests for the Cauchy distribution based on the empirical characteristic function. *Annals of Institute of Statistical Mathematics*, *52*, 267-286.

Matsui, M., & Takemura, A. (2005). Empirical characteristic function approach to goodness-of-fit tests for the Cauchy distribution with parameters estimated by MLE or EISE. *Annals of Institute of Statistical Mathematics*, *52*, 183-199.

McIntyre, G. A. (1952). A method of unbiased selective sampling using ranked sets. *Australian Journal of Agricultural Research*, *3*, 385-390.

Meintanis, S. G. (2004). A class of omnibus tests for the Laplace distribution based on the characteristic function. *Communication in Statistics-Theory and Method*s, *33*(*4*), 925-948.

Muttlak, H. A. (1997). Median ranked set sampling. *Journal of Applied Statistics Science*, *6*, 245-255.

Samawi, H. M., Mohmmad, S., & Abu-Dayyeh, W. (1996). Estimation the population mean using extreme ranked set sampling. *Biometrical Journal*, *38*, 577-586.

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, *69*, 730-737.

Stephens, M. A. (1977). Goodness of fit for the extreme value distribution. *Journal of Biometrika*, *64*, 583-588.

Stephens, M. A. (1979). Tests of fit for the logistic distribution based on the empirical distribution function. *Journal of Biometrika*, *66*(*3*), 591-595.

Stockes, S. L., & Sager T. W. (1988). Characterization of a ranked -set sample with application to estimating distribution functions. *Journal of the American Statistical Association*, *83*(*402*), 374-381.

Takahasi, K., & Wakitmoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics*, *20*, 1-31.

Towhidi, M., & Salmanpour, M. (2007). A new goodness-of-fit test based on the empirical characteristic function. *Communication in Statistics-Theory and Methods*, *36*, 2777-2785.

Wong, W. K., & Sim, C. H. (2000). Goodness-of-fit test based on empirical characteristic function. *Journal of Statistical Computation and Simulation*, *65*, 243-269.

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 568-578

1538 – 9472/10/$95.00

Bayesian Analysis of Location-Scale Family of Distributions Using S-PLUS and R Software

Sheikh Parvaiz Ahmad Aquil Ahmed
University of Kashmir,
Srinagar, India

Athar Ali Khan
Aligarh Muslim University,
U.P, India

The Normal and Laplace's methods of approximation for posterior density based on the location-scale family of distributions in terms of the numerical and graphical simulation are examined using S-PLUS and R Software.

Key words: Bayesian analysis, location-scale family, logistic distribution, Newton-Raphson iteration, normal and Laplace's approximation, S-PLUS, R software.

Introduction

A parametric location-scale model for a random variable *y* on (-∞ ,∞) is distributed with pdf of the form

$$p(y;\mu,\sigma) = \frac{1}{\sigma} f\left(\frac{y-\mu}{\sigma}\right), \quad -\infty < y < \infty \tag{1.1}$$

where μ $(-\infty < \mu < \infty)$ is a location parameter and $\sigma > 0$ is a scale parameter (not necessarily mean and standard deviation). This family can also be written as

$$y = \mu + \sigma z \tag{1.2}$$

where $z = \frac{y-\mu}{\sigma}$ is the standardized variate with density *f(z)*, David (1981). A few important models, namely, normal, logistic and extreme value are some important members of the location-scale family.

Sheikh Parvaiz Ahmad is an Assistant Professor in the Department of Statistics. Email: sprvz@yahoo.com. Aquil Ahmed is a Professor in the Department of Statistics. Email: aquilstat@yahoo.co.in. Athar Ali Khan is a Professor in the Department of Statistics. Email: atharkhan1962@gmail.com.

Bogdanoff and Pierce (1973) analyzed an extreme value model treating non informative priors for location and scale parameters. Stavrakakis and Drakopoulos (1995) and Galanis, et al. (2002) deal with an extreme value model with Bayesian statistics. Sinha (1986) and Khan (1997) also cite several references for non-normal *f(z)*.

Bayesian Analysis when Both Parameters μ and σ Are Unknown

Suppose that n observations $y^T = (y_1, y_2, \ldots, y_n)$ can be regarded as a random sample from a location-scale family of models in (1.2), but both μ *and* σ are unknown; in terms of general notation $\theta^T = (\mu, \sigma)$, the likelihood function is given by

$$p(y \mid \mu, \sigma) = \prod_{i=1}^{n} p(y_i \mid \mu, \sigma)$$

The log-likelihood is defined as

$$\begin{aligned} l(\mu,\sigma) &= \log \prod_{i=1}^{n} p(y_i \mid \mu, \sigma) \\ &= \log \prod_{i=1}^{n} \sigma^{-1} f\left(\frac{y-\mu}{\sigma}\right) \\ &= \sum_{i=1}^{n} \log f(z_i) - n \log \sigma \end{aligned}$$

or equivalently

$$l(\mu,\sigma)=\sum_{i=1}^{n} l_i \qquad (2.1)$$

where $l_i=\log f(z_i)-\log\sigma$ and $z_i=\dfrac{y_i-\mu}{\sigma}$.

Following the standard approach of Box and Tiao (1973), assume that a priori μ and σ are approximately independent, so that

$$p(\mu,\sigma)\cong p(\mu)p(\sigma) \qquad (2.2)$$

where $p(\mu)$ and $p(\sigma)$ are priors for μ and σ, respectively. Using Bayes theorem, the posterior density of $p(\mu,\sigma\mid y)$ is given by

$$p(\mu,\sigma\mid y)\quad\alpha\quad\prod_{i=1}^{n} p(y_i\mid\mu,\sigma)\,p(\mu)\,p(\sigma)$$

or

$$p(\mu,\sigma\mid y)\quad\alpha\quad\left[\prod_{i=1}^{n}\sigma^{-1}f(z_i)\right]p(\mu)\,p(\sigma) \qquad (2.3)$$

The joint posterior density of μ and σ is assumed to contain all information required in the statistical analysis (e.g., Box & Tiao, 1973), therefore, the main job remains to study the different features of $p(\mu,\sigma\mid y)$. The posterior mode can be obtained by maximizing (2.3) with respect to μ and σ. To formalize this, define

$$l^*(\mu,\sigma)=\log p(\mu,\sigma\mid y)$$

thus,

$$l^*(\mu,\sigma)=l(\mu,\sigma)+\log p(\mu)+\log p(\sigma). \qquad (2.4)$$

The maximization of $p(\mu,\sigma\mid y)$ is equivalent to maximizing $l^*(\mu,\sigma)$ with respect to (μ,σ). To apply the Newton-Raphson technique, partial derivatives of $l^*(\mu,\sigma)$ are needed and some notations must be defined for simplification purposes. For example

$$l_\mu=\frac{\partial l}{\partial\mu},\quad l_\sigma=\frac{\partial l}{\partial\sigma},\quad l_{\mu\sigma}=\frac{\partial^2 l}{\partial\mu\partial\sigma},\quad l_{\sigma\mu}=\frac{\partial^2 l}{\partial\sigma\partial\mu},$$

$$l_{\mu\mu}=\frac{\partial^2 l}{\partial\mu^2},\ \textit{and}\quad l_{\sigma\sigma}=\frac{\partial^2 l}{\partial\sigma^2}.$$

Similarly, define $l^*_\mu=l_\mu+\dfrac{p'(\mu)}{p(\mu)}$,

$$l^*_\sigma=l_\sigma+\frac{p'(\sigma)}{p(\sigma)},\ l^*_{\mu\sigma}=l_{\mu\sigma},\ l^*_{\sigma\mu}=l_{\sigma\mu},$$

$$l^*_{\mu\mu}=l_{\mu\mu}+\left[\frac{p'(\mu)}{p(\mu)}\right]',$$

and

$$l^*_{\sigma\sigma}=l_{\sigma\sigma}+\left[\frac{p'(\sigma)}{p(\sigma)}\right]'$$

where $f'(x)=Df(x)$ and $[f'(x)]'=D^2f(x)$, D stands for differential operator. Consequently, the score vector of log-posterior

$$U(\mu,\sigma)=(l^*_\mu,l^*_\sigma)^T$$

and Hessian matrix of log-posterior, that is,

$$H(\mu,\sigma)=\begin{bmatrix} l^*_{\mu\mu} & l^*_{\mu\sigma}\\ l^*_{\sigma\mu} & l^*_{\sigma\sigma}\end{bmatrix}$$

thus, the posterior mode $(\hat{\mu},\hat{\sigma})$ can be obtained from iteration scheme

$$\begin{bmatrix}\hat{\mu}\\ \hat{\sigma}\end{bmatrix}=\begin{bmatrix}\mu_0\\ \sigma_0\end{bmatrix}-H^{-1}(\mu_0,\sigma_0)\begin{bmatrix}l^*_\mu\\ l^*_\sigma\end{bmatrix} \qquad (2.5)$$

Consequently, the modal variance Σ can be obtained as

$$I^{-1}(\hat{\mu},\hat{\sigma})=-H^{-1}(\hat{\mu},\hat{\sigma}).$$

For drawing an inference about μ and σ simultaneously, the joint posterior $p(\mu,\sigma \mid y)$ is used. It is preferable to use approximations to this posterior as given below:

Normal Approximations

A bivariate normal approximation of $p(\mu,\sigma \mid y)$, is

$$p(\mu,\sigma \mid y) \cong N_2\left((\hat{\mu},\hat{\sigma})^T, I^{-1}(\hat{\mu},\hat{\sigma})\right) \tag{2.6}$$

Similarly, the Bayesian analog of likelihood ratio criterion is

$$-2[l^*(\mu,\sigma) - l^*(\hat{\mu},\hat{\sigma})] \approx \chi_2^2 \tag{2.7}$$

where the symbol $\approx$ means approximately distributed as. Defining $W(\mu,\sigma) = -2[l^*(\mu,\sigma) - l^*(\hat{\mu},\hat{\sigma})]$ using $W(\mu,\sigma)$ as a test criterion in hypothesis testing and construction of the credible region (confidence interval in non-Bayesian terminology).

Laplace's Approximation

Laplace's approximation of $p(\mu,\sigma \mid y)$ can also be written as

$$p(\mu,\sigma \mid y) \cong (2\pi)^{-1} \mid I(\hat{\mu},\hat{\sigma}) \mid^{\frac{1}{2}} exp[l^*(\mu,\sigma) - l^*(\hat{\mu},\hat{\sigma})] \tag{2.8}$$

The Marginal Inference

The marginal Bayesian inference about μ and σ is based on marginal posterior densities of these parameters. The marginal posterior for μ can be obtained after integrating out $p(\mu,\sigma \mid y)$ with respect to σ, that is,

$$p(\mu \mid y) = \int_0^\infty p(\mu,\sigma \mid y)\, d\sigma$$

Similarly, marginal posterior of σ can be obtained as

$$p(\sigma \mid y) = \int_{-\infty}^{\infty} p(\mu,\sigma \mid y)\, d\mu.$$

For normal likelihood $p(\mu,\sigma \mid y)$ and non-informative prior $p(\mu,\sigma)\,\alpha\,\frac{1}{\sigma}$, it can be shown that $p(\sigma \mid y)$ is the inverted χ-distribution (Box & Tiao, 1973; Zellener, 1971). But if either assumption of normality is extended to other members of location scale family or the prior is changed then closed form expressions cannot be obtained and approximations must be relied upon (Khan, 1997). In practice, the Gauss-Hermite quadrature (Naylor & Smith, 1982) can be used to find accurate approximations of $p(\mu \mid y)$ and $p(\sigma \mid y)$, however, following simple approximations is recommended.

Normal Approximation

The normal approximation of marginal posterior $p(\mu \mid y)$ is:

$$p(\mu \mid y) = N_1(\hat{\mu}, I_{11}^{-1}) \tag{3.1}$$

In addition, the Bayesian analog of likelihood ratio criterion can also be defined as a test criterion based on (3.1) as

$$(\mu - \hat{\mu})^T I_{11}(\mu - \hat{\mu}) \approx \chi_1^2 \tag{3.2}$$

Laplace's Approximation

The marginal posterior density $p(\mu \mid y)$ can alternatively be approximated by

$$p(\mu \mid y) \cong \left[\frac{\mid I(\hat{\mu},\hat{\sigma}) \mid}{2\pi \mid I(\mu,\hat{\sigma}(\mu)) \mid}\right]^{\frac{1}{2}} exp[l^*(\mu,\hat{\sigma}(\mu)) - l^*(\hat{\mu},\hat{\sigma})] \tag{3.3}$$

Similarly, $p(\sigma \mid y)$ can be approximated and results corresponding to normal and Laplace's approximation can be written as

$$p(\sigma \mid y) = N_1(\hat{\sigma}, I_{22}^{-1}) \tag{3.4}$$

or equivalently,

$$(\sigma-\hat{\sigma})^T I_{22}(\sigma-\hat{\sigma}) \approx \chi_1^2 \qquad (3.5)$$

$$p(\sigma \mid y) \cong$$

$$\left[\frac{|I(\hat{\mu},\hat{\sigma})|}{2\pi|I(\hat{\mu}(\sigma),\sigma)|}\right]^{\frac{1}{2}} exp[l^*(\hat{\mu}(\sigma),\sigma)-l^*(\hat{\mu},\hat{\sigma})] \qquad (3.6)$$

Bayesian Analysis of Logistic Distribution

The pdf of the logistic distribution is given by

$$p(y;\mu,\sigma)=\frac{e^{\frac{(y-\mu)}{\sigma}}}{\sigma\left(1+e^{\frac{(y-\mu)}{\sigma}}\right)^2},$$

$$-\infty < y < \infty,$$

$$\sigma > 0$$

The likelihood function is given by

$$p(y \mid \mu,\sigma)=\prod_{i=1}^{n} p(y_i \mid \mu,\sigma)$$

And the log-likelihood is defined as

$$l(\mu,\sigma)=log\prod_{i=1}^{n} p(y_i \mid \mu,\sigma)$$

$$=\sum_{i=n}^{n}(z_i-2log(1+e^{z_i}))-n\log\sigma \qquad (4.1)$$

where $z_i=\frac{y_i-\mu}{\sigma}$.

Taking partial derivatives with respect to μ and σ

$$l_\mu=\frac{\partial l}{\partial\mu}$$

$$=\frac{1}{\sigma}\sum_{i=1}^{n}\left(\frac{e^{z_i}-1}{e^{z_i}+1}\right)$$

$$l_\sigma=\frac{\partial l}{\partial\sigma}$$

$$=\frac{1}{\sigma}\sum_{i=1}^{n} z_i\left(\frac{e^{z_i}-1}{e^{z_i}+1}\right)-\frac{n}{\sigma}$$

$$l_{\mu\sigma}=\frac{\partial^2 l}{\partial\mu\partial\sigma}$$

$$=-\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(\frac{e^{2z_i}+2z_i e^{z_i}-1}{(e^{z_i}+1)^2}\right)$$

$$l_{\sigma\mu}=\frac{\partial^2 l}{\partial\sigma\partial\mu}$$

$$=-\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(\frac{e^{2z_i}+2z_i e^{z_i}-1}{(e^{z_i}+1)^2}\right)$$

$$l_{\mu\mu}=\frac{\partial^2 l}{\partial\mu^2}$$

$$=-\frac{2}{\sigma^2}\sum_{i=1}^{n}\left(\frac{e^{z_i}}{(e^{z_i}+1)^2}\right)$$

$$l_{\sigma\sigma}=\frac{\partial^2 l}{\partial\sigma^2}$$

$$=-\frac{2}{\sigma^2}\sum_{i=1}^{n} z_i\left(\frac{e^{2z_i}+z_i e^{z_i}-1}{(e^{z_i}+1)^2}\right)+\frac{n}{\sigma^2}$$

Following the standard approach of Box and Tiao (1973), Gelman, et al. (1995), it is assumed that the prior μ and σ are approximately independent so that

$$p(\mu,\sigma) \cong p(\mu)p(\sigma) \qquad (4.2)$$

where $p(\mu)p(\sigma)$ and $p(\sigma)$ are priors for μ and σ. Using Bayes theorem, the posterior density $p(\mu,\sigma \mid y)$ is

$$p(\mu,\sigma \mid y) \quad \alpha \quad \prod_{i=1}^{n} p(y_i \mid \mu,\sigma)\, p(\mu)\, p(\sigma) \qquad (4.3)$$

and the log-posterior is given by

$$log\ p(\mu,\sigma \mid y) = log \prod_{i=1}^{n} p(y_i \mid \mu,\sigma) + log\ p(\mu) + log\ p(\sigma) \quad (4.4a)$$

or

$$l^*(\mu,\sigma) = l(\mu,\sigma) + \log p(\mu) + \log p(\sigma) \quad (4.4b)$$

For a prior $p(\mu,\sigma) \cong p(\mu)p(\sigma) = 1$, $l^*_\mu = l_\mu$, $l^*_\sigma = l_\sigma$, $l^*_{\mu\sigma} = l_{\mu\sigma}, l^*_{\sigma\mu} = l_{\sigma\mu}, l^*_{\mu\mu} = l_{\mu\mu}$ and $l^*_{\sigma\sigma} = l_{\sigma\sigma}$. The posterior mode is obtained by maximizing (4.4) with respect to μ and σ. The score vector of the log posterior is given by

$$U(\mu,\sigma) = (l^*_\mu, l^*_\sigma)^T$$

and the Hessian matrix of the log posterior is

$$H(\mu,\sigma) = \begin{bmatrix} l^*_{\mu\mu} & l^*_{\mu\sigma} \\ l^*_{\sigma\mu} & l^*_{\sigma\sigma} \end{bmatrix}$$

Posterior mode $(\hat{\mu},\hat{\sigma})$ can be obtained from iteration scheme

$$\begin{bmatrix} \hat{\mu} \\ \hat{\sigma} \end{bmatrix} = \begin{bmatrix} \mu_0 \\ \sigma_0 \end{bmatrix} - H^{-1}(\mu_0,\sigma_0) \begin{bmatrix} l^*_\mu \\ l^*_\sigma \end{bmatrix}$$

consequently, the modal variance Σ can be obtained as

$$I^{-1}(\hat{\mu},\hat{\sigma}) = -H^{-1}(\hat{\mu},\hat{\sigma}).$$

For drawing inferences about μ and σ simultaneously, the joint posterior (7.3) is used.

Using normal approximation, a bivariate normal approximation of (7.3) can be written as

$$p(\mu,\sigma \mid y) \cong N_2\left((\hat{\mu},\hat{\sigma})^T, I^{-1}(\hat{\mu},\hat{\sigma})\right)$$

Similarly, a Bayesian analog of likelihood ratio criterion is

$$-2\left[l^*(\mu,\sigma) - l^*(\hat{\mu},\hat{\sigma})\right] \approx \chi^2_2$$

Using Laplace's approximation, $p(\mu,\sigma \mid y)$ can be written as

$$p(\mu,\sigma \mid y) \cong (2\pi)^{-1} \mid I(\hat{\mu},\hat{\sigma}) \mid^{\frac{1}{2}} exp[l^*(\mu,\sigma) - l^*(\hat{\mu},\hat{\sigma})]$$

The marginal Bayesian inferences about μ and σ are based on the marginal posterior densities of these parameters, and the marginal posterior for μ can be obtained after integrating out $p(\mu,\sigma \mid y)$ with respect to σ, that is

$$p(\mu \mid y) = \int_0^\infty p(\mu,\sigma \mid y)\, d\sigma$$

Similarly, the marginal posterior of σ can be obtained as

$$p(\sigma \mid y) = \int_{-\infty}^{\infty} p(\mu,\sigma \mid y)\, d\mu,$$

thus, normal approximation of the marginal posterior $p(\mu \mid y)$ is

$$p(\mu \mid y) = N_1(\hat{\mu}, I_{11}^{-1}).$$

The Bayesian analog of likelihood ratio criterion can also be defined as a test criterion as

$$(\mu - \hat{\mu})^T I_{11} (\mu - \hat{\mu}) \approx \chi^2_1$$

and Laplace's approximation of marginal posterior density $p(\mu \mid y)$ can be given by

$$p(\mu \mid y) \cong \left[\frac{|I(\hat{\mu},\hat{\sigma})|}{2\pi |I(\mu,\hat{\sigma}(\mu))|}\right]^{\frac{1}{2}} exp[l^*(\mu,\hat{\sigma}(\mu)) - l^*(\hat{\mu},\hat{\sigma})]$$

$$p(\sigma \mid y) \cong \left[\frac{|I(\hat{\mu},\hat{\sigma})|}{2\pi |I(\hat{\mu}(\sigma),\sigma)|}\right]^{\frac{1}{2}} exp[l^*(\hat{\mu}(\sigma),\sigma) - l^*(\hat{\mu},\hat{\sigma})]$$

Similarly, $p(\sigma \mid y)$ can be approximated with results corresponding to normal and Laplace's approximation can be written as

$$p(\sigma \mid y) = N_1(\hat{\sigma}, I_{22}^{-1})$$

or equivalently,

$$(\sigma - \hat{\sigma})^T I_{22} (\sigma - \hat{\sigma}) \approx \chi_1^2$$

Numerical and Graphical Illustrations

Numerical and graphical illustrations are implemented using both S-PLUS and R software for Logistic distribution. These illustrations are intended for the purpose of showing the strength of Bayesian methods in practical situations. The posterior mode and standard errors of parameters μ and σ of logistic distribution are presented in Table 4. A graphical display for comparing the posterior of μ using the Normal and Laplace approximations are shown in Figures 1 to 3 and a comparison for the posterior of σ is displayed in Figures 4 to 6. The graph shows that the two approximations are in close agreement.

Table 1: A Summary of Derivatives of Log Likelihoods

	Distributions		
Derivatives	Normal	Extreme-Value	Logistic
l_{μ}	$\frac{1}{\sigma}\sum_{i=1}^{n} z_i$	$-\frac{1}{\sigma}\sum_{i=1}^{n}(1-e^{z_i})$	$\frac{1}{\sigma}\sum_{i=1}^{n}\left(\frac{e^{z_i}-1}{e^{z_i}+1}\right)$
l_{σ}	$\frac{1}{\sigma}\sum_{i=1}^{n} z_i^{\,2} - \frac{n}{\sigma}$	$-\frac{1}{\sigma}\sum_{i=1}^{n} z_i(1-e^{z_i}) - \frac{n}{\sigma}$	$\frac{1}{\sigma}\sum_{i=1}^{n} z_i\left(\frac{e^{z_i}-1}{e^{z_i}+1}\right) - \frac{n}{\sigma}$
$l_{\mu\sigma}$	$-\frac{2}{\sigma^2}\sum_{i=1}^{n} z_i$	$-\frac{1}{\sigma^2}\sum_{i=1}^{n}(z_i e^{z_i} + e^{z_i} - 1)$	$-\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(\frac{e^{2z_i} + 2z_i e^{z_i} - 1}{(e^{z_i}+1)^2}\right)$
$l_{\sigma\mu}$	$-\frac{2}{\sigma^2}\sum_{i=1}^{n} z_i$	$-\frac{1}{\sigma^2}\sum_{i=1}^{n}(z_i e^{z_i} + e^{z_i} - 1)$	$-\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(\frac{e^{2z_i} + 2z_i e^{z_i} - 1}{(e^{z_i}+1)^2}\right)$
$l_{\mu\mu}$	$-\frac{n}{\sigma^2}$	$-\frac{1}{\sigma^2}\sum_{i=1}^{n} e^{z_i}$	$-\frac{2}{\sigma^2}\sum_{i=1}^{n}\left(\frac{e^{z_i}}{(e^{z_i}+1)^2}\right)$
$l_{\sigma\sigma}$	$-\frac{3}{\sigma^2}\sum_{i=1}^{n} z_i^{\,2} + \frac{n}{\sigma^2}$	$-\frac{1}{\sigma^2}\sum_{i=1}^{n}(z_i^{\,2} e^{z_i} + 2z_i e^{z_i} - 2z_i) + \frac{n}{\sigma^2}$	$-\frac{2}{\sigma^2}\sum_{i=1}^{n} z_i\left(\frac{e^{2z_i} + z_i e^{z_i} - 1}{(e^{z_i}+1)^2}\right) + \frac{n}{\sigma^2}$

where $z_i = \frac{y_i - \mu}{\sigma}$, $i = 1,2,...,n$.

Table 2: A Summary of Prior Densities for Location Parameter μ

Name of Density	$p(\mu)$	$\frac{p'(\mu)}{p(\mu)}$	$\left[\frac{p'(\mu)}{p(\mu)}\right]'$
Non-Informative	Constant	Zero	Zero
Normal	$c\sigma_0^{-1}\exp\left(-\frac{D^2}{2}\right)$	$-D\sigma^{-1}$	$-\sigma^{-2}$
Logistic	$c\exp(D)[1+\exp(D)]^{-2}$	$\sigma_0^{-1}[1-2F(D)]$	$\frac{2}{\sigma_0^2}F(D)[1-F(D)]$
Extreme-Value	$c\exp[D-\exp(D)]$	$\sigma_0^{-1}[1-F(D)]$	$-\sigma_0^{-2}\exp(D)$

where $D=\frac{\mu-\mu_0}{\sigma_0}$, $F(D)=\frac{e^D}{1+e^D}$, and c is the normalizing constant.

Table 3: A Summary of Prior Densities for σ

Name of Prior	$p(\sigma)$	$\frac{p'(\sigma)}{p(\sigma)}$	$\left[\frac{p'(\sigma)}{p(\sigma)}\right]'$
Non-Informative	$\frac{1}{\sigma}$	$-\frac{1}{\sigma}$	$\frac{1}{\sigma^2}$
Inverted Gamma	$c\sigma^{-(\alpha_0+1)}\exp\left(-\frac{1}{\sigma\beta_0}\right)$	$\frac{1}{\beta_0\sigma^2}-\frac{\alpha_0+1}{\sigma}$	$\frac{\alpha_0+1}{\sigma^2}-\frac{2}{\beta_0\sigma^3}$
Lognormal	$c\sigma_0^{-1}\exp\left(-\frac{D^2}{2}\right)$	$-\frac{D}{\sigma_0\sigma}-\frac{1}{\sigma}$	$\frac{1}{\sigma^2}-\frac{1}{(\sigma_0\sigma)^2}-\frac{D}{\sigma_0\sigma^4}$
Gamma	$c\sigma^{(\alpha_0-1)}e^{-\sigma\beta_0^{-1}}$	$\frac{1}{\beta_0}+\frac{\alpha_0-1}{\sigma}$	$-\frac{\alpha_0-1}{\sigma^2}$
Weibull	$c\sigma^{(\beta_0-1)}e^{-(\sigma_0\sigma)^{\beta_0}}$	$\frac{\beta_0-1}{\sigma}-\alpha_0\beta_0(\alpha_0\sigma)^{\beta_0-1}$	$-\frac{\beta_0-1}{\sigma^2}-\alpha_0{}^{\beta_0}\beta_0(\beta_0-1)\sigma^{(\beta_0-2)}$

where c is the normalizing constant and $D=\frac{\log\sigma-\mu_0}{\sigma_0}$.

Table 4: Posterior Mode and Posterior Standard Error of Parameters of Logistic Distribution with Different Priors

Prior	Posterior Mode μ	Posterior Standard Error μ	Posterior Mode σ	Posterior Standard Error σ
1	168.63355	2.679672	58.65997	1.320980
1/sigma	168.62814	2.678635	58.63024	1.319912
1/(mu*sigma)	168.58558	2.678692	58.62837	1.319845
1/(mu*sigma)^2	168.53766	2.677714	58.59681	1.318714

Figures 1-3: Comparing Normal and Laplace's Approximation for μ of Logistic Distribution for Various Priors in S-PLUS and R

Figure 1: Comparison between Normal and Laplace Approximations

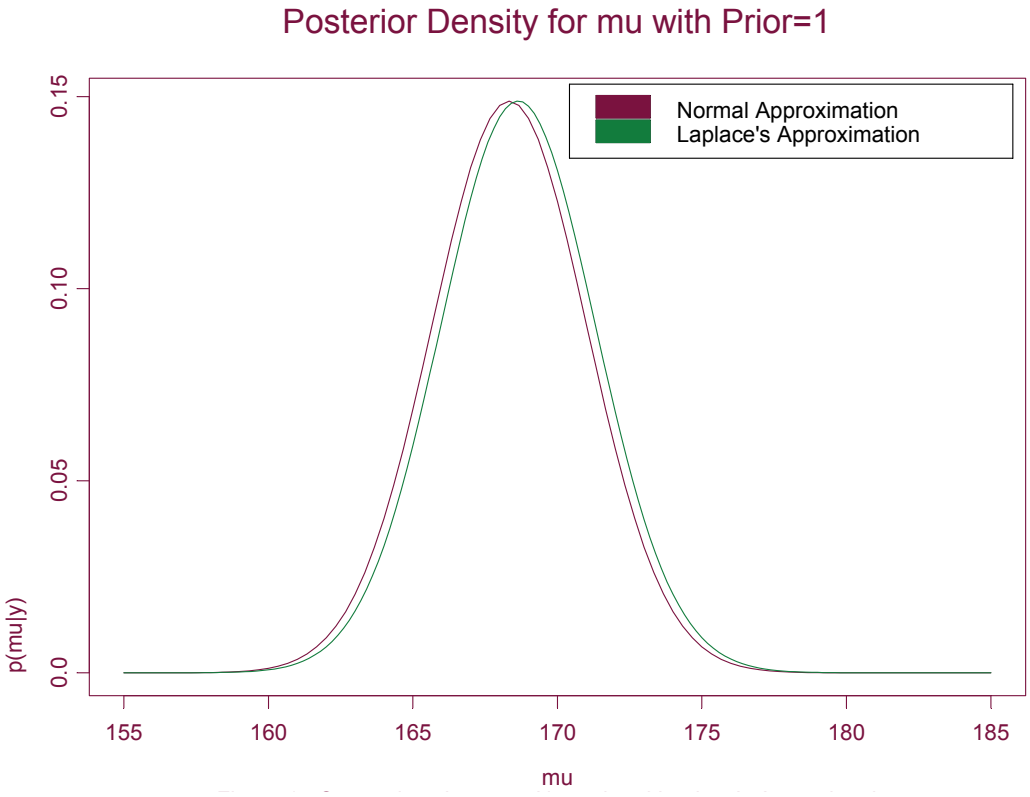


Figure 2: Comparison between Normal and Laplace Approximations

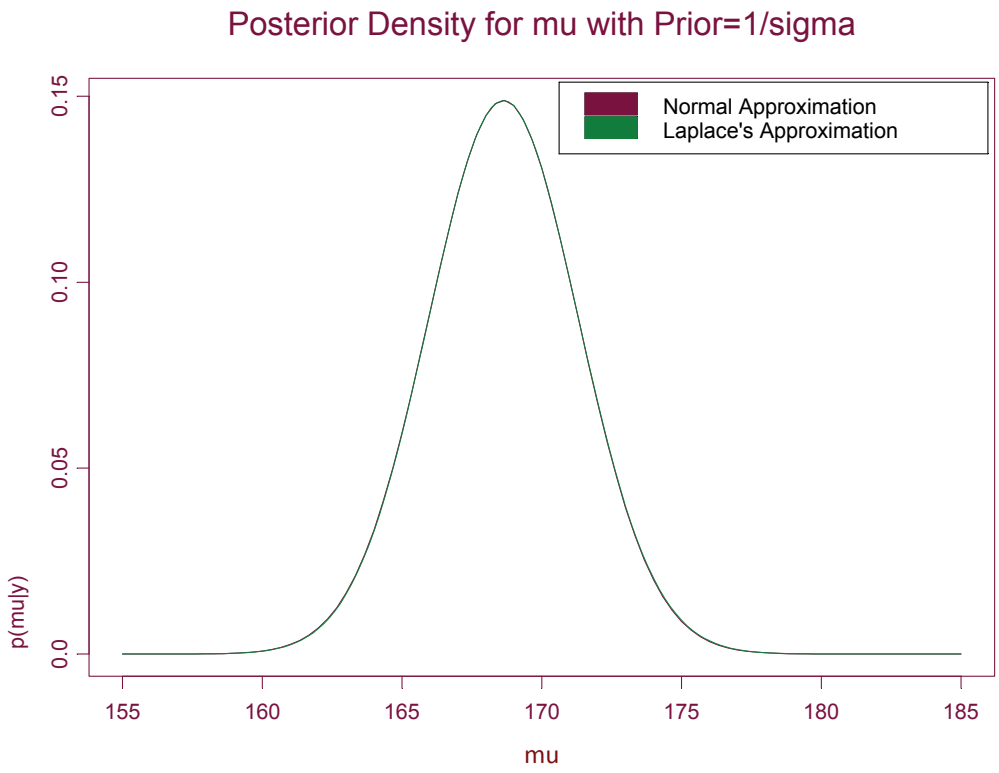
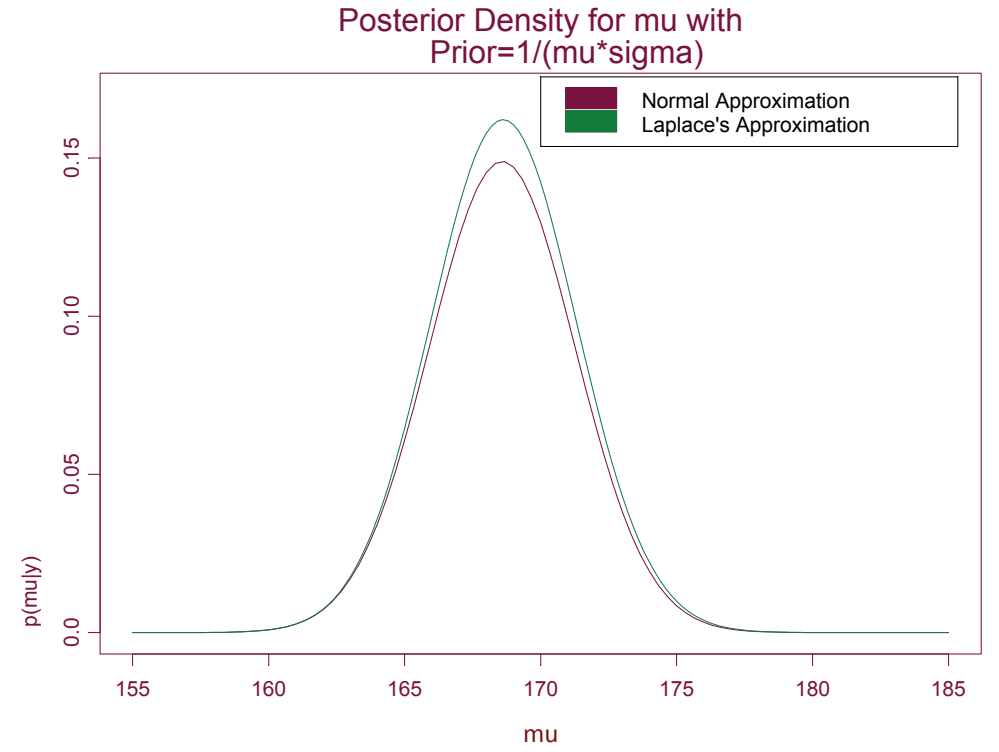


Figure 3: Comparison between Normal and Laplace Approximations



Figures 4-6: Comparing Normal and Laplace's Approximation for σ of Logistic Distribution for Various Priors in S-PLUS and R

Figure 4: Comparison between Normal and Laplace Approximations

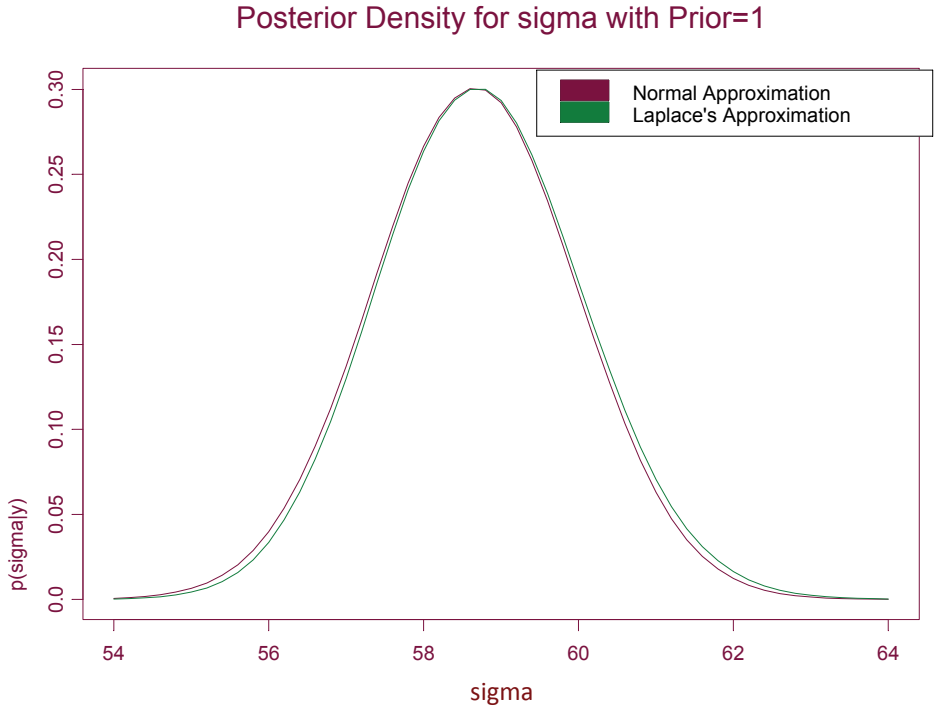


Figure 5: Comparison between Normal and Laplace Approximation

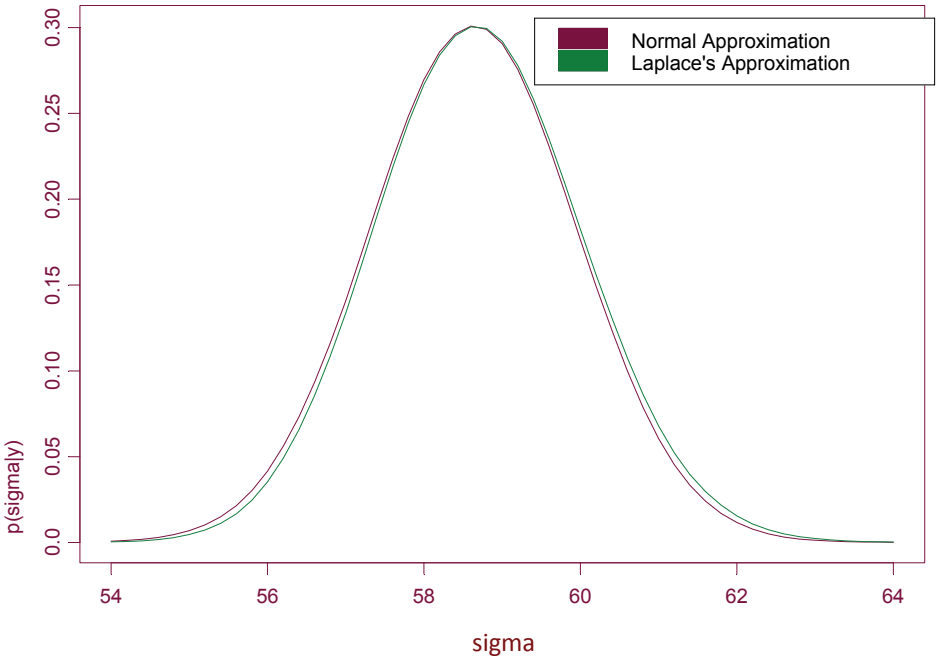
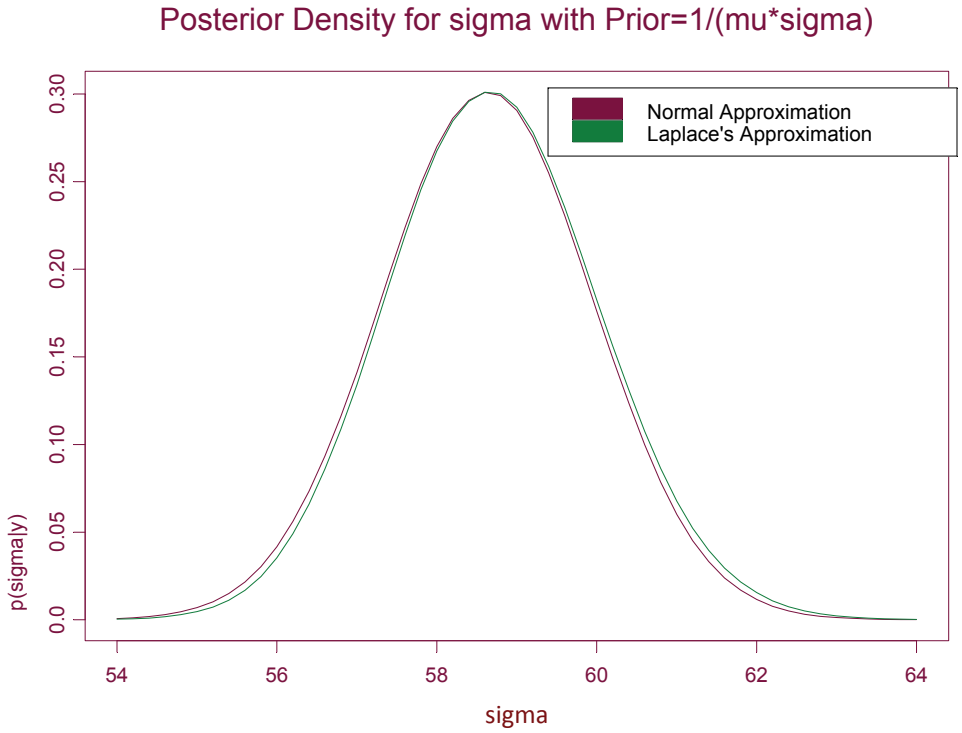


Figure 6: Comparison between Normal and Laplace Approximation



References

Bogdanoff, D., & Pierce, D. A. (1973). Bayes-fiducial inference for the Weibull distribution. *Journal of the American Statistical Association*, *68*, 659-664.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Reading, PA: Addison-Wesley.

David, H. A. (1981). *Order Statistics*. New York: Wiley.

Galanis, O. C., Tsapanos, T. M., Papadopoulos, G. A., & Kiratzi, A. A. (2002). Bayesian extreme values distribution for seismicity parameters in South America. *Journal of the Balkan Geophysical Society*, *5*, 77-86.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.

Khan, A. A. (1997). *Asymptotic Bayesian Analysis in Location-Scale Models*. Ph.D. Thesis submitted to the Dept. of Mathematics & Statistics, HAU, Hisar.

Khan, A. A., Puri, P. D., & Yaqub, M. (1996). Approximate Bayesian inference in location-scale models. *Proceedings of National Seminar on Bayesian Statistics and Applications*, *April 6-8, 1996*, 89-101. Department of Statistics BHU, Varanasi.

Naylor, J. C., & Smith, A. F. M. (1982). Applications of a method for efficient computation of posterior distributions. *Applied Statistics*, *31*, 214-225.

Sinha, S. K. (1986). *Reliability and life testing*. New Delhi: Wiley Eastern Limited.

Stavrakakis, G. N., & Drakopoulos, J. (1995). Bayesian probabilities of earthquake occurrences in Greece and surrounding areas. *Pageoph*, *144*, 307-319.

Zellener, A. (1971). *An introduction to Bayesian inference in econometrics*. New York: Wiley.

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 579-583

1538 – 9472/10/$95.00

ANN Forecasting Models for ISE National-100 Index

Ozer Ozdemir Atilla Aslanargun Senay Asma
Anadolu University,
Eskisehir, Turkey

Prediction of the outputs of real world systems with accuracy and high speed is crucial in financial analysis due to its effects on worldwide economics. Because the inputs of the financial systems are time-varying functions, the development of algorithms and methods for modeling such systems cannot be neglected. The most appropriate forecasting model for the ISE national-100 index was investigated. Box-Jenkins autoregressive integrated moving average (ARIMA) and artificial neural networks (ANN) are considered by using several evaluations. Results showed that the ANN model with linear architecture better fits the candidate data.

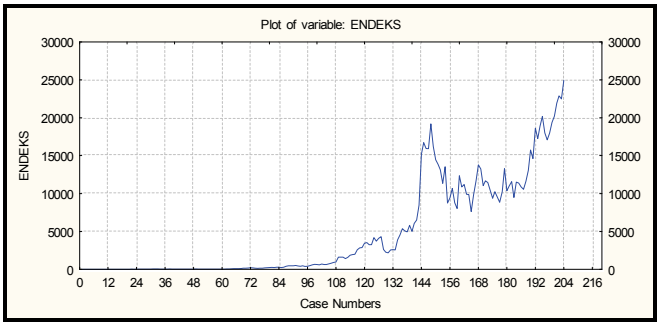
Key words: ISE stock market, time series modeling, artificial neural network, forecasting.

Introduction

Statistical modeling via artificial neural networks (ANN) has recently been widely used for time series forecasting for economy, hydrology, electricity, tourism, etc. Many studies in the literature support that ANN time series models show better results than Box-Jenkins (BJ) models (Aslanargun, 2007; Ansuj, et al., 1996; Chin & Arthur, 1996; Hill, et al., 1996; Kohzadi, et al., 1996; Maier & Dandy, 1996).

In this article, ANN forecasting models are used to study an economical forecasting problem, the case selected is from the Turkey Istanbul Stock Market. The performance of the determined model is demonstrated by comparing the models via mean square error. The aim of this study is to construct ANN forecasting models for the Istanbul Stock Exchange (ISE) national-100 index which has effects on much economic behavior. The time series taken at hand is shown in Figure 1.

Figure 1: The Fluctuation of Time Series in ISE Stock Market



Due to the nonlinear structure of the corresponding time series, ANN is the appropriate tool for accurate modeling. The

Senay Asma is an Assistant Professor in the Department of Statistics at Anadolu University, Turkey, where she obtained her B.Sc., M.Sc. and Ph.D. degrees in statistics. Her research interests include simulation, entropy and theory of information, coding theory, mixture models, theory of hypothesis tests. Email: senayyolacan@anadolu.edu.tr. Ozer Ozdemir is a Lecturer in the Department of Statistics at Anadolu University, Turkey, where he obtained his B.Sc., M.Sc. degrees in statistics. Currently he is doing his Ph.D. on neural network. His research interests include simulation, neural network, fuzzy logic and time series. Email: ozerozdemir@anadolu.edu.tr. Atilla Aslanargun is an Assistant Professor in the Department of Statistics. His research interests are statistical inference, stochastic process and time series analysis. Email: aaslanar@anadolu.edu.tr.

designer has to decide about the number of inputs and outputs, the activation functions, the algorithm for obtaining the weights of the net, the number of hidden layers and the number of neurons inside the hidden layers. Because all combinations of choices result in different ANN models the analysis becomes complicated, therefore, simulation of the various types of neural nets is crucial. Hence, to address such decisions in this study, the intelligent problem solver (IPS) module of the STATISTICA 7.0 was used. The corresponding program allows a researcher to construct one million ANN models at a time and select the best of them; thus, 100,000 ANN models were run to obtain the appropriate one for ISE Stock Market. Additionally, standard ARIMA models were also constructed.

Methodology

Over all the forecasting methods, the artificial neural network (ANN) is the most popular method. For different tasks such as classification, clustering, regression, etc. (Bishop, 1995) different types of neural networks are available, such as feed forward, radial basis function (RBF), Kohonen self-organizing and Bayesian. Training of the neural network is accomplished based on a specific cost function as sum of the square errors. Different types of available algorithms for training the network and include, among others, back propagation, conjugate gradient, quasi-Newton and steepest-descent. The weights of the network, also called parameters of the model, can be found by taking the derivative of the cost function subject to network parameters and updating those parameters until those which minimize the cost function are identified.

Network Overview

The following five networks are indicated as the best potential networks for the data in this study (Bishop, 1995; Haykin, 1999).

Linear

Linear networks have only two layers: an input and an output layer. This type of network is best trained using a Pseudo-Inverse technique.

Multi Layer Percepteron (MLP)

MLP networks are constructed of multiple layers of computational units. Each neuron in one layer is directly connected to the neurons of the subsequent hidden layer. The frequently used activation function is the sigmoid function. Multi-layer networks use a variety of learning techniques, the most popular being back-propagation.

Radial Basis Function (RBF)

The RBF network consists of an input layer, a hidden layer of radial units and an output layer of linear units. Typically, the radial layer has exponential activation functions and the output layer has linear activation functions.

Generalized Regression Neural Networks (GRNN)

The GRNN network is a type of Bayesian network. GRNN has exactly four layers: input, a layer of radial centers, a layer of regression units, and output. This network must be trained by a clustering algorithm.

Results

ANN and ARIMA Forecasting Model Analysis

ARIMA models are analyzed by the Time Series module of STATISTICA 7.0, and ANN models are obtained by using the IPS module. First, stationary of variance is considered for analyzing the time series aspect of ARIMA models. Because this time series is not stationary, natural log transformation for this time series is applied. Moreover, different transformations are applied due to the trend effect. Later, the ARIMA $(0,1,1)(0,1,0)_{12}$ model was found to be the best because it has the less mean square error (MSE) compared to the alternatives: the summary of this model, which is significant, is shown in Table 1.

Table 1: Summary of ARIMA $(0,1,1)(0,1,0)_{12}$ Model

Transformations: ln(x),D(1),D(1) Model:(0,1,1)(0,1,0) MS Residual= 0,02903				
Parameter	Par. Value	P	Lower	Upper
$q(1)$	0.890715	0.00	0.819855	0.961575

Forecasted values calculated for the period between January 2005 and June 2005 by using ARIMA $(0,1,1)(0,1,0)_{12}$ model are shown in Table 2 along with observed values of this period.

Table 2: Forecasting Values for ARIMA $(0,1,1)(0,1,0)_{12}$ Model

Period	Month-Year	Forecasting Values $(\hat{y})$	Observed Values (y)
205	January-2005	25946.05	27330.35
206	February-2005	26958.43	28396.17
207	March-2005	28010.31	25557.76
208	April-2005	29103.24	23591.64
209	May-2005	30238.82	25236.48
210	June-2005	31418.70	26957.32

The MSE is calculated by using forecasted and observed values in Table 2 as follows:

$$MSE = \frac{1}{6}\sum_{i=1}^{6}\left(y_i - \hat{y}_i\right)^2 = 14217239.31. \quad (1)$$

After calculating the MSE for the ARIMA model, the design of the ANN time series is prepared. The time series is in month periods, hence 12 inputs are taken and regarded for the months of a year and one output neuron is taken as the design of the neural net. The values of the first input neuron called X_1 is taken for the period January 1988 – December 2003, the second input neuron is one month delayed and so its period runs from February 1988 – January 2004. The remaining input neurons are constructed in similarly. The output neuron, Y, is the values of the period from January 1989 – December 2004.

Statistical modeling using ANN is analyzed by IPS. IPS provides the opportunity to conduct various experiments through different combinations of algorithms and designs. In this research study, the IPS was ordered to choose the 5 best models among 1,000 various neural nets. The minimum input number was 1 and the maximum input number was specified as 12. The models obtained from the IPS and their performance measures are shown in Tables 3 and 4.

Table 3: Summary of 5 Best Models for Forecasting

Profile	Train Error
Linear 2:2-1:1	0.045627
MLP s3 3:9-1-1:1	0.050194
GRNN 5:5-92-2-1:1	0.000030
GRNN 7:7-92-2-1:1	0.000026
RBF s6 12:72-16-1:1	0.000027

Table 4: Summary of 5 Best Models for Forecasting (continued)

Profile	Test Error	Training/ Members
Linear 2:2-1:1	0.058281	PI
MLP s3 3:9-1-1:1	0.077745	BP100, CG20, CG11b
GRNN 5:5-92-2-1:1	0.000043	SS
GRNN 7:7-92-2-1:1	0.000041	SS
RBF s6 12:72-16-1:1	0.000037	KM, KN, PI

The Generalized Regression neural network (GRNN), Multilayer Percepteron (MLP), Radial Basis Neural Network (RBF) and Linear neural networks performed well and produced the best results among all the predicted ANN time series models. In order to obtain a more accurate forecasting model, each model was used to calculate the forecast values by running the net again for the remaining test data and the MSE was calculated for each model respectively; results are shown in Table 5.

As apparent in Table 4, the ANN forecasting model consisting of linear neural net performs the forecasting with less error. It is concluded that, for fluctuation in the ISE stock market, the usage of Linear Neural Net Time Series provides more accurate results than the other variations of the ANN time series models. The weights (parameters) of this Linear 2:2-1:1 Neural Network are shown in Table 6.

Table 5: Mean Square Errors of the Best 5 Models

Model	MSE
Linear 2:2-1:1	3468672
MLP s3 3:9-1-1:1	14388771
GRNN 5:5-92-2-1:1	4124962
GRNN 7:7-92-2-1:1	4021692
RBF s6 12:72-16-1:1	15886195

Table 6: Weights of the Linear 2:2-1:1 Neural Network

	2.1
Thresh	-0,002914
1.1	0,147587
1.2	0,762368

For modeling, monthly index values have been taken from final quotations of the Istanbul Stock Exchange National-100 index between January 1988 and December 2004. Forecasting was done for a period between January 2005 and June 2005. The best models were determined by using the Box-Jenkins method and artificial neural networks for a time series which consisted of Istanbul Stock Exchange National-100 index values. Forecasting values for considered models of both methods are provided in Table 7.

Table 7: Forecasting Values for ANN and ARIMA

Period	Month-Year	Forecasting Values (ANN)	Forecasting Values (B.J.)
205	January-2005	24837.82	25946.05
206	February-2005	27235.23	26958.43
207	March-2005	25504.51	28010.31
208	April-2005	27133.13	29103.24
209	May-2005	26053.71	30238.82
210	June-2005	27154.03	31418.70

MSE values of the models considered in Table 7 are given in Table 8.

Table 8: Mean Square Errors for ANN and ARIMA

Model	MSE
Linear 2:2-1:1	3468672
ARIMA $(0,1,1)(0,1,0)_{12}$	14217239

Table 7 shows that MSE value which belongs to the Linear 2:2-1:1 model is smaller than the MSE value of the ARIMA $(0,1,1)(0,1,0)_{12}$ model which was found using BJ method.

Conclusion

This study presented ANN forecasting models that can be used as tools for predicting unexpected booms in the economy. The corresponding analyses were conducted by using IPS because it gives an opportunity to compare various types of ANN models together. Experimental studies were performed across 1,000 neural nets and the best 5 ANN models based on mean square error were evaluated. Additionally, ARIMA models were considered in order to evaluate the effectiveness of the presented ANN models. Finally, it was expressed that an ANN model conducted with linear architecture had better forecasting performance compared to the ARIMA model.

References

Aslanargun, A., Mammadov, M., Yazici, B., & Yolacan, S. (2007). Comparison of ARIMA, neural networks and hybrid models in time series: tourist arrival forecasting. *Journal of Statistical Computation and Simulation*, *77*(*1*), 29-53.

Ansuj, A. P., Camargo, M. F., Radharamanan, R., & Petry, D. G. (1996). Sales forecasting using time series and neural networks. *Computational and Industrial Engineering*, *31*(*1*), 421-424.

Chin, K., & Arthur, R. (1996). Neural network vs. conventional methods of forecasting. *Journal of Business Forecasting*, *14*(*4*), 17-22.

Hill, T., O'Connor, M., & Remus, W. (1996). Neural network models for time series forecasts. *Management Science*, *42*(*7*), 1082-1092.

Kohzadi, N., Boyd, M. S., Kermanshahi, B., & Kaastra, I. (1996). A comparison of artificial neural network and time series model for forecasting commodity prices. *Neurocomputing*, *10*(*2*), 169-181.

Maier, H. R., & Dandy, G. C. (1996). Neural network models for forecasting univariate time series. *Neural Networks World*, *6*(*5*), 747-772.

Zhu, X. T. (2008). Predicting stock index increments by neural networks: The role of trading volume under different horizons. *Expert Systems With Applications*, *34*(*4*), 3043-3054.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. New York: Prentice Hall.

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 584-595

1538 – 9472/10/$95.00

Markov Chain Analysis and Student Academic Progress: An Empirical Comparative Study

Shafiqah Alawadhi Mokhtar Konsowa
Kuwait University

An application of Markov Chain Analysis of student flow at Kuwait University is presented based on a random sample of 1,100 students from the academic years 1996-1997 to 2004-2005. Results were obtained for each college and in total which allows for a comparative study. The students' mean lifetimes in different levels of study in the colleges as well as the percentage of dropping out of the system are estimated.

Key words: Absorbing Markov chains, transition probabilities; absorbing state.

Introduction

The realization of the importance of education has increased among the public in Kuwait, and as a result, the traditional formal education has changed in recent years. To compete in the world market, nations are giving priority to higher education for the purpose of preparing students to be capable to bear the responsibility. Higher educational institutions in the state of Kuwait have encountered challenges in recent years. The economic and social needs associated with higher educational institutions have attracted the attention of many segments of the Kuwaiti public to observe the performance of these organizations with renewed and increased interest. The financial burden on the government to support higher educational institutions in Kuwait increases the responsibility of the institutions to maintain their efficiency. Kuwait University (KU) consists of 13 colleges which follow a regular semester system, with the exceptions of the Medical Sciences and Law colleges, which follow an annual system. Each semester consists of approximately 16 weeks and each year includes two semesters: Fall (September - January) and Spring (February - June). KU also offers a summer semester that is not compulsory; however, about 65% of students take summer courses. A typical student takes about 4 years to complete the required credit hours; students of Engineering and the Medical Sciences colleges are exceptions and they take slightly longer to complete the credit hours. The required numbers of credit hours to graduate from the different colleges are:

1. Allied Health, Social Sciences, Science and Business Administration: 126 to 130 credit hours.
2. Arts and Education: 132 credit hours.
3. Sharia & Islamic Studies: 142 credit hours.
4. Engineering and Petroleum: 144 credit hours.

Shafiqah Alawadhi is an Associate Professor in the Department of Statistics & Operations Research, Kuwait University. She received her Ph. D. from the Department of Mathematics, Aberdeen University, Scotland, UK. Her research interests include General Statistical Analysis, Bayesian Statistics especially Subjective Probability Assessment and Environmental Statistics. Email her at: alawadhidodo@yahoo.com. Mokhtar Konsowa is an Associate Professor in the Department of Statistics & Operations Research, Kuwait University. He received his Ph. D. from the Department of Mathematics, University of Cincinnati, USA. His research interests are in Probability and Stochastic Processes and focus on the area of random walks on graphs. Email him at: mokhtar@stat.kuniv.edu.

5. Medical Sciences (Pharmacy - 5 yrs; Dentistry - 6 yrs; Medicine - 7 yrs; annual system).
6. Law (4 yrs; annual system).

The regular registered credit hours per semester are between 12 and 19 but it is not allowed to be less than 12 credits, however, in summer semester it ranges between 3 and 9 credit hours. Depending on the completed credit hours the students are classified into 8 levels:

1. F: Freshman, a student who successfully completed less than 31 credit hours.
2. So: Sophomore, a student who successfully completed between 31 and 60 credit hours.
3. J: Junior, a student who successfully completed between 61 and 90 credit hours.
4. Se: Senior, a student who successfully completed more than 90 credit hours.
5. NR: Not registered, perhaps to take care of personal problems, but eventually will return to the system
6. G: Graduated from KU.
7. D: Dropped out or academically dismissed from KU.
8. T: Transferred to another college in KU.

The outcomes (graduates) of KU are considerably less than the incomes (freshmen). This is more visible in the scientific colleges. The attrition occurs from KU and in particular from scientific colleges at a significant percentage. It also happens that the transfer from scientific colleges to art colleges also occurs at a high percentage.

A high proportion of students may stay longer in their course of study for several reasons, especially in the freshman and senior stages. As such, it is necessary to study the average time that a student spends in each level, as well as the probability that a student who has been admitted will graduate or withdraw from each college specifically and generally from KU. A comparative study was conducted between the different colleges and between KU and each of these colleges in order to determine which college is closest to the normal average time. Finally the factors that cause a student to spend 0 increasing numbers of semesters in each level of study were investigated.

Markov analysis is used to investigate the flow process of students in KU. It has been employed in several flow processes (see Wainwright, 2007; Nichols, 2008; Al-Awadhi & Konsowa, 2007; Bessent & Bessent, 1980; Kolesat, 1970; Kwak, et al., 1985; Merddith, 1976; McNamara, 1974). Al-Awadhi and Konsowa (2007) studied student flow in the College of Science at KU. Bessent and Bessent (1980) studied the progression process of doctoral students in a university department to avoid undesirable future dissertation overload for supervising professors. Kwak, et al. (1985) were interested in forecasting student enrollment variations for an academic institution. Bessent and Bessent (1980) proposed an enrollment retention model using a Markov process to analyze enrollment rates for overlooked segments of the student population as well as the retention rate for specifying degree programs, rather than just the retention rates for aggregate incoming freshman. Reynolds and Porath (2008) studied absorbing Markov chains to model the academic progress of students attending the University of Wisconsin-Eau Claire over a specific time period.

Methodology

A random sample of 1,100 students was selected from the office of the Deanship of Admission and Registration. The data pertains to a period of 9 years from the academic years 1996-1997 to 2004-2005. Stratified random sampling was used for the sample collection. The sample size for each college was determined in proportion to the total number of students in each college and the sample from each college was divided into 8 groups proportionally to the number of students in each level defined as: freshmen, sophomore, junior, senior, non-register, graduate, drop out and transfer. Each of these groups was subdivided proportionally according to the departments in the college. Finally a random sample of each of the sub-groups was selected from each college.

The Conceptual Framework: Markov Chains

Some background concerning the Markov chains is presented, for additional detail; see Kwak, et al. (1985) or Resnick (1994). Consider a finite discrete time homogeneous stochastic process with index set $\mathbb{Z}^+ = \{0, 1, 2,...\}$; that is, a sequence $\{X_n: n \in \mathbb{Z}^+\}$ of random variables. As usual the subscript n in X_n stands for the time and X_n denotes the state of the process at time n. If $X_n \in S$, then S is called the state space of the stochastic process. The stochastic processes considered here satisfy the Markov property. Given the present state, the future of the process is independent of the past. That is, for i, j, x_0, ..., $x_{n-1}1 \in S$,

$$P(X_{n+1} = j \mid X_0 = x_0, ..., X_{n-1} = x_{n-1}, X_n = i)$$

$$= P(X_{n+1} = j \mid X_n = i) = P_{ij}$$

A stochastic process with this property is called a homogenous Markov chain. The quantity P_{ij} stands for the probability of moving from state i to state j in just one transition and all these quantities define the matrix of one-step transition probabilities P:

$$p = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{bmatrix} = (p_{ij})$$

where the finite set $I_k = \{1, 2, ..., k\}$ is the state space of the Markov chain. The entries P_{ij} of the matrix P must satisfy: (1) $P_{ij} \geq 0$, (2) $\Sigma_j P_{ij} = 1, i, j \in I_k$. The recurrent state i is called the absorbing state if $P_{ii} = 1$. The transient matrix P in its canonical form is $p = \left(\frac{I \mid 0}{R \mid Q}\right)$, where I is the identity matrix corresponding to the absorbing states, 0 is zero matrix, Q is the restriction of P to the transient states and $R = (P_{kl}$: k is a transient state and l is an absorbing state. The fundamental matrix N is defined to be $(I - Q)^{-1}$ where it is known that the matrix (I - Q) is invertible. Let N_{ij} be the ij^{th} entry of N, then N_{ij} is the mean number of visits to state j having started at state i. If M_i refers to the mean absorption time starting at state i, then $M_i = \sum_{j \in Q} N_{ij}$; that is $M = N\xi$, where ξ is a vector all of its components are 1's. Another quantity of interest is the probability U_{ij} that the chain starting in a transient i will end up in the absorbing state j. If U stands for the matrix with entries U_{ij}, then $U = NR$.

Data Analysis and Transition Probabilities

The frequency table of the university as a whole is compared with the frequency table of the university excluding the data of the Medical Sciences colleges. Because the study time for the Medical Sciences colleges is much longer than other colleges, it was thought that this may affect the analysis and the results may not be accurate. However, it was found that when analyzing the data with and without these colleges, the results are very close to each other (due to shortage of space, the comparison analysis is not shown here). This is most likely due to the sample size of the medical sciences colleges, which is small (about 45 cases) compared with the samples of other colleges.

Results

The data from each college was studied and the corresponding analysis is conducted. Table 1 displays a (7x7) frequency matrix for KU from which the transition probabilities are estimated. The matrix which represents the transition probabilities of remaining in or progressing to another state is referred as P in its canonical form and is presented in Table 2. The table shows that states G and O are considered to be absorbing states while the other states are transient states. Note that the transfer state is considered here as a transient state because the transfer is defined as transferred to

another college within KU, however when each college is analyzed separately the transfer state is considered as an absorbing state.

Note in Table 2 the probability that a freshman student remains in the state itself is 0.62 and the probability of progressing to sophomore is 0.33. For a sophomore student, the probability of progressing to the junior state is 0.41. The probability that a senior student remains in the same state is high (0.73). This may be attributed to the fact that the normal lifetime for some colleges such as the College of Engineering is more than 4 years and also the fact that the courses at the senior levels are tougher than those of the other levels. The probability of remaining a not registered student for another semester is 0.25 and to move to junior level is 0.22 while moving to other transient states varies between 0.12 and 0.16.

Proceeding as described, the diagonal elements of the matrix N represent the average life times that correspond to the transient states of the Markov chain. They are obtained for the levels F, So, J, Se, and NR and are found to be 2.752, 2.333, 2.041, 3.709, and 1.446 respectively. The average life time for a student at each level does not exceed 2.752, which is normal except for the senior level, which is found to be 3.7 semesters. This may be attributed to two reasons: (1) many students repeat courses to raise their GPA, and (2) Medical Sciences and Engineering students require more than 4 years to graduate so their senior states would be longer than the other colleges.

Table 1: Frequency Transition Data

	F	So	J	Se	NR	G	O
F	1,425	754	0	0	61	0	45
So	0	900	672	0	34	0	21
J	0	0	591	577	26	1	10
Se	0	0	0	1,260	7	456	11
NR	20	21	29	16	33	0	11
G	0	0	0	0	0	457	0
O	0	0	0	0	0	0	100

Vectors M are calculated and represent the average number of semesters needed to reach an absorbing state (graduate or dropout) starting at any given level. The components of M are: 10.181, 7.977, 5.829, 3.773 and 7.525 corresponding to the states F, So, J, Se and NR respectively. The probabilities of graduating or dropping out starting at a given level are displayed in Table 3 as matrix U.

The probabilities of graduating for freshmen, sophomore, junior, senior and non register levels are 0.86, 0.92, 0.95, 0.98 and 0.82 respectively. It is observed that the

Table 2: Transition Probability Matrix P

$P =$

	G	O	F	So	J	Se	NR
G	1	0	0	0	0	0	0
O	0	1	0	0	0	0	0
F	0	0.02	0.624	0.330	0	0	0.026
So	0	0.013	0	0.553	0.413	0	0.021
J	0	0.009	0	0	0.490	0.479	0.022
Se	0.263	0.006	0	0	0	0.727	0.004
NR	0	0.085	0.153	0.162	0.223	0.123	0.254

Table 3: Probability Matrix U

$$U = \begin{bmatrix} & G & O \\ F & 0.862 & 0.138 \\ So & 0.918 & 0.082 \\ J & 0.952 & 0.048 \\ Se & 0.975 & 0.025 \\ NR & 0.821 & 0.179 \end{bmatrix}$$

probability of progressing to graduation increases as a student moves to advanced levels and, as such, decreases for dropping out.

Data from Different Colleges

The data from each college was analyzed separately. The transition probability matrix and the fundamental matrix were obtained from the corresponding frequency tables. The columns in Table 4 represent the diagonal elements of the probability transition matrices of the corresponding colleges and these diagonal elements provide the probability that a student remains in the same state. It is noted that the probabilities to remain in the same state are generally close to 0.50.

The columns of Table 5 represent the diagonal elements of N matrices of the corresponding colleges. As noted, the diagonal elements of matrix N represent the average times a student spends in a level of study. For many levels, a student spends approximately 2 to 3 semesters in each state before passing to a higher level of study except for senior level, where they may remain more than 3 semesters for some colleges. A freshman student spends on average about 2.5 semesters, except those in the college of Science where about 3.9 semesters are required to make a transition. As the student advances to the senior level, the mean time increases with longest mean time in the colleges of Medical Sciences.

The M vectors for the university colleges represent the average times required to reach an absorbing state (graduate G, drop out D, transfer T) starting at a transient state (see Table 6). To reach an absorbing state in many colleges a freshman student takes more than 8 semesters, whereas a senior student takes commonly more than 3 semesters.

Table 4: Probabilities to Remain in the Transient States of Each College

	Arts	Soc	Law	Bus	Edn	Shar	Sci	Eng	Med	Allied
F	0.56	0.59	0.61	0.55	0.55	0.53	0.71	0.59	0.50	0.60
So	0.56	0.55	0.63	0.48	0.51	0.42	0.51	0.56	0.36	0.56
J	0.50	0.52	0.56	0.46	0.44	0.43	0.47	0.49	0.54	0.44
Se	0.66	0.65	0.61	0.63	0.74	0.65	0.73	0.78	0.90	0.71
NR	0.16	0.43	0.43	0.15	0.18	0.17	0.31	0.29	0.33	0.00

Table 5: Average Lifetimes of Students in Each Transient State

	Arts	Soc	Law	Bus	Edn	Shar	Sci	Eng	Med	Allied
F	2.3	2.5	2.6	2.3	2.3	2.2	3.9	2.5	2.0	2.3
So	2.4	2.3	2.7	2.1	2.1	1.8	2.2	2.4	1.6	2.3
J	2.1	2.2	2.4	1.9	1.8	1.9	1.9	2.1	2.2	1.9
Se	2.9	2.9	2.6	2.8	3.9	3.0	3.8	4.6	10.4	3.5
NR	1.3	1.8	1.8	1.3	1.3	1.3	1.7	1.5	1.6	1.1

On average, students in the colleges of Medical Sciences, Science and Engineering spend more time in the system to reach an absorbing state. This is due to the facts that the courses in these colleges are difficult and also because the number of credits for the colleges of Engineering and Medical Sciences is large compared to other colleges. In addition, it could be due to the repetition of the courses by the students in acquiring better GPAs, as in the case of the college of Science. A student with a low GPA finds it difficult to pass the core courses offered in the college. Thus many students repeat these courses to improve their GPA. At the senior level, a student from the college of Medical Sciences spends comparatively more time in practical classes (almost 5 years are needed to qualify them for professional careers).

At this stage, the U-matrices for the 10 colleges are calculated. The entry u_{ij} of a U matrix represents the probability of absorption at state j having started at state i. For the purpose of comparison, the entries of U-matrices are classified into three matrices corresponding to the three absorbing states: graduate, drop out, and transfer. (See Tables 7, 8 and 9.) For example, the first column of Table 7 stands for the probability of graduation of an arts' student enrolled in the respective levels F, So, J, Se and NR.

Table 6: Average Time Needed to Reach an Absorbing State

	Arts	Soc	Law	Bus	Edn	Shar	Sci	Eng	Med	Allied
F	7.66	9.45	10.37	8.08	9.46	7.69	8.20	10.1	15.30	8.15
So	7.10	7.32	7.94	6.53	7.77	6.49	5.70	8.70	13.60	6.73
J	5.07	5.15	5.20	4.74	5.73	5.03	5.43	6.70	12.60	5.44
Se	3.04	2.88	2.72	2.83	3.87	3.24	3.77	4.70	10.43	3.46
NR	5.91	6.31	5.71	6.72	8.48	5.48	7.61	6.80	15.90	6.27

Table 7: Probabilities of Reaching the Graduate State *G* for the 10 Colleges at KU

	Arts	Soc	Law	Bus	Edn	Shar	Sci	Eng	Med	Allied
F	0.60	0.91	0.98	0.80	0.91	0.77	0.37	0.77	0.93	0.69
So	0.79	0.95	1.00	0.91	0.99	0.89	0.49	0.88	0.95	0.81
J	0.85	0.98	1.00	0.95	1.00	0.96	0.78	0.94	1.00	0.99
Se	0.90	1.00	1.00	0.98	1.00	0.96	0.94	0.99	1.00	1.00
NR	0.72	0.98	1.00	0.81	0.97	0.85	0.45	0.62	0.94	0.95

Table 8: Probabilities of Reaching the Drop -out State *O* for the 10 Colleges at KU

	Arts	Soc	Law	Bus	Edn	Shar	Sci	Eng	Med	Allied
F	0.19	0.05	0.02	0.07	0.01	0.08	0.26	0.20	0.07	0.11
So	0.18	0.05	0	0.03	0	0.04	0.17	0.10	0.05	0.05
J	0.15	0.02	0	0.03	0	0.04	0.09	0.04	0	0
Se	0.09	0	0	0.02	0	0.04	0.03	0.01	0	0
NR	0.26	0.02	0	0.04	0	0.12	0.29	0.37	0.06	0.01

Table 9: Probability of Reaching the Transfer State T for the 10 Colleges at KU

	Arts	Soc	Law	Bus	Edn	Shar	Sci	Eng	Med	Allied
F	0.21	0.05	0	0.13	0.08	0.15	0.34	0.03	0	0.19
So	0.04	0	0	0.07	0.01	0.06	0.34	0.02	0	0.15
J	0	0	0	0.02	0	0	0.14	0.02	0	0
Se	0	0	0	0	0	0	0.03	0	0	0
NR	0.02	0	0	0.15	0.02	0.01	0.27	0.01	0	0.04

Table 7 shows that a student at a junior state attains the absorbing state G with an average probability above 0.94. The probability of graduation increases as the student moves to the senior level. For the colleges of Law, Social Science, Education, Allied Health and Medical Sciences, it is assumed that the senior students strive hard to graduate from KU.

The probability of reaching the graduating state from the freshman level is comparatively low for the college of Science (0.37); this may be due to the difficult courses in the college of Science and also to the fact that most of the students admitted to the college of Science have the lowest high school grades and are usually competitive students. Also, it is noted that the probability of graduation of students in the NR state is very low for the college of Science, thus, it is estimated that only 45% of the students who are non-registered will re-enter the system and complete their course of study.

The chances of dropping out of the system from the first levels of the colleges of Science and Arts are comparatively high. This may be due to probations. KU regulations do not allow a student to continue her/his study when she/he reaches 4 probations. Similarly, the probability that a student drops out after reaching the non registered state is considerably high for Arts, Sciences and Engineering students.

The percentage of students transferring from Science to other colleges is relatively high for freshman, sophomore, non register and junior states. Similarly the percentage of students transferring from the freshman and sophomore states in the Allied states. The probability of transferring is zero for students from the Law and Medical Sciences colleges and this can be attributed to the fact that once a student is enrolled in these annual systems, she/he cannot transfer to other colleges where all the credits she/he passed cannot be transferred to other semester system colleges.

If the transient states freshman, sophomore, junior, senior, non registered are respectively assigned the numbers 4, 5, 6, 7 and 8, then the mean number of semesters the chain remains in each of them after it is entered (including the entering step) are calculated by

$$E_i(r_i) = \sum_{k=0}^{\infty} P_{ii}^k = \frac{1}{1 - p_{ii}}; i = 4,5,6,7,8,$$

(see Table 10). It is noted that $E_i(r_i)$ refers to the mean continuous stay in state i once it is entered, whereas, N_{ii} stands for the mean number of visits to state i regardless of any departures from i. The slight differences between the entries of Tables 5 and 10 are explained by the small transition probabilities to the state NR in the transition matrix of Table 2.

The mean number of changes of a state in an absorbing chain can be calculated by setting $P_{ii} = 0$ for every transient state in the transition matrix P and then dividing each row by its row sum to obtain P^*. The i^{th} component of the new vector M^* gives the mean number of changes of the state i for the original process. The i^{th} component in the two vectors M and M^* may differ slightly if the repetitions of the states on the path from i to absorption is rare; otherwise, the two vectors

may differ significantly. From the matrix P^* of the different colleges, the vectors $M^* = N^*\xi$ are calculated and are shown in Table 11.

As noted, the significant differences between some of the components of M and M^* are interpreted by repeating courses and staying longer in some levels. The small values corresponding to the freshmen and sophomore levels in the college of Science depicts the occurrence of dropping out and transferring in these two levels.

Comparison between Scientific and Art Colleges

To compare the scientific and arts colleges, the colleges were divided into 2 groups: (1) Scientific colleges: Science, Engineering and Allied Health, and (2) Art colleges: Arts, Social Sciences and Sharia & Islamic Studies. The College of Medical Science which has a lengthier course schedule, the College of Law which is a professional course and the Colleges of Education and Business administration which have a mixture of science and art students have all been excluded from the groupings. Mean lifetimes in each state for the scientific and art colleges are shown in Table 12. It is clear that the lifetimes of students in scientific colleges are longer than that of art colleges, especially at the senior levels. The vector M as represented in Table 13 gives the average time spent by a student in the system before reaching one of the absorbing states. It is clear that the scientific student needs more time to reach one of the observing states (G, O, T). This may be due to the difficulty of the study in their colleges compared to art colleges. Going through the details of these absorbing states, U matrices are calculated.

Table 10: Average Number of Semesters the Chain Remains in Transient States

	Arts	Soc	Law	Bus	Edn	Shar	Sci	Eng	Med	Allied
$E_4(r_4)$	2.28	2.47	2.59	2.24	2.24	2.14	3.52	2.41	2.00	2.5
$E_5(r_5)$	2.28	2.22	2.73	1.93	2.05	1.73	2.04	2.25	1.6	2.3
$E_6(r_6)$	2.02	2.12	2.32	1.83	1.79	1.77	1.87	1.98	2.17	1.81
$E_7(r_7)$	2.96	2.88	2.57	2.71	3.87	2.83	3.77	4.6	10.38	3.45
$E_8(r_8)$	1.19	1.75	1.75	1.18	1.2	1.2	1.44	1.4	1.5	1

Table 11: M^* Vectors: Average Number of State-Changes Before Absorption

	Arts	Soc	Law	Bus	Edn	Shar	Sci	Eng	Med	Allied
F	3.29	3.95	4.1	3.79	3.84	3.73	2.98	3.8	3.9	3.38
So	3.02	3.09	3.16	3.08	3.05	3.17	2.40	3.1	3.0	2.76
J	2.09	2.09	2.15	2.13	2.04	2.27	2.04	2.15	2.00	2.13
Se	1.03	1.00	1.07	1.06	1.00	1.22	1.00	1.10	1.00	1.00
NR	2.96	2.79	2.61	3.55	3.77	2.91	3.29	2.8	4.5	3.00

Table 12: Mean Lifetime in Each State for Scientific and Art Colleges

	Science	Art
F	3.20	2.31
So	2.25	2.22
J	1.99	2.09
Se	4.20	2.98
NR	1.53	1.36

Table 13: Average Time Required Reaching One of the Absorbing States

	Science	Art
F	8.8	8.1
So	6.9	7.0
J	6.0	5.1
Se	4.3	3.1
NR	7.1	5.9

The U matrices display the probabilities of reaching the absorbing states G, O and T are as follows

$$U_{Sci} = \begin{bmatrix} & G & O & T \\ F & 0.554 & 0.221 & 0.225 \\ So & 0.680 & 0.130 & 1.190 \\ J & 0.879 & 0.055 & 0.066 \\ Se & 0.972 & 0.015 & 0.013 \\ NR & 0.589 & 0.294 & 0117 \end{bmatrix};$$

$$U_{Art} = \begin{bmatrix} & G & O & T \\ F & 0.701 & 0.137 & 0.162 \\ So & 0.849 & 0.120 & 0.032 \\ J & 0.901 & 0.098 & 0.001 \\ Se & 0.937 & 0.062 & 0.000 \\ NR & 0.803 & 0.185 & 0012 \end{bmatrix}$$

It is clear that - except for the senior state - the probability of graduation is higher for the art students than for the scientific students. The students in the NR state show a tendency to drop out rather than to transfer and the probabilities are higher in both cases for the Science students than for the Art students.

Graduation Process

An interesting use of the conditional probability for absorbing Markov chains is the following: Assume that for an absorbing chain, we start in a non-absorbing state and computing all the probabilities relative to the hypothesis that the process ends in a given absorbing state, for example, s_1. Then it is possible to obtain a new absorbing chain with a single absorbing state s_1. The non-absorbing states will be as before, except that we have new transition probabilities. Computing these probabilities is as follows: Let A be the event that the chain is absorbed in the state s_1. If s_i is a non-absorbing state then the transition probabilities of the new process are

$$p_i(X_1 = s_j \mid A) = \frac{p_i(A \mid X_1 = s_j)p_i(X_1 = s_j)}{p_i(A)} = \frac{p_j(A)p_{ij}}{p_i(A)},$$

where $p_j(A)$ is the occurrence probability of the event A starting at state s_j and p_{ij} is the one step transition probability from state s_i to state s_j. Using traditional notation, the equation above can be written as $\hat{p}_{ij} = \frac{u_{j1}p_{ij}}{u_{i1}}$, where $\hat{p}_{ij}$ stands for the elements of the new transition matrix. The canonical form of the new transition matrix $\hat{P}$ is $\hat{P} = \begin{pmatrix} \hat{I} & \hat{0} \\ \hat{R} & \hat{Q} \end{pmatrix}$.

The elements of submatrix $\hat{R}$ for a transient state s_i are given by $\hat{p}_{i1} = \frac{p_{i1}}{u_{i1}}$. That is $\hat{R} = \left[\frac{p_{i1}}{u_{i1}} \right]$. If U_0 is the diagonal matrix with diagonal elements $u_{j1}'s$ for the non-absorbing states $s_j's$, then $\hat{Q} = U_0^{-1}QU_0$, from which $\hat{Q}^n = U_0^{-1}Q^nU_0$. As such

$$\hat{N} = I + \hat{Q} + \hat{Q}^2 + \ldots = U_0^{-1}(\sum_{n=0}^{\infty} Q^n)U_0 = U_0^{-1}NU_0,$$

and $\hat{M} = \hat{N}\eta$ (Kemeny & Snell, 1970). Using the original transition matrix in its canonical form in Table 2, the matrix $N = (I - Q)^{-1}$ can be computed. Also, the first column of the U -Matrix (Table 3) constitutes the diagonal elements of the diagonal matrix U_0. In which case, the matrix $\hat{N} = U_0^{-1}NU_0$ would be:

$$\hat{N} = \begin{bmatrix} & F & So & J & Se & NR \\ F & 2.508 & 2.121 & 1.971 & 3.693 & 0.210 \\ So & 0.043 & 2.204 & 1.996 & 3.688 & 0.120 \\ J & 0.024 & 0.044 & 2.015 & 3.692 & 0.066 \\ Se & 0.006 & 0.011 & 0.018 & 3.692 & 0.016 \\ NR & 0.511 & 0.961 & 1.592 & 3.695 & 1.423 \end{bmatrix}$$

From which the vector $\hat{M}$ that determines the average number of semesters required for graduation starting from the different levels of study becomes

$$\hat{M} = \begin{bmatrix} F & 10.503 \\ So & 8.052 \\ J & 5.842 \\ Se & 3.744 \\ NR & 8.182 \end{bmatrix}$$

Conclusion

The lifetimes of a student in different levels varies depending on the colleges. A freshman student remains a freshman for an average of 2 to 3 semesters except in the College of Science where it is about 4 semesters. For the second state, sophomore, the life times in different colleges again varies from about 2 to 3 semesters. Junior students spend less time, less than 2 semesters for several colleges. At senior levels where the courses are comparatively more difficult, the life times vary significantly.

For students in the colleges of Medical Sciences, Engineering, Education and Science, the life times are comparatively high. The most common reasons for the lengthy life times are:

1) Most of the students take preliminary non-credit courses in English, Mathematics and Chemistry (these courses are compulsory if a student did not pass the university aptitude test once she/he is enrolled in scientific colleges). After a student registers in these intensive courses, she/he is not allowed to register in the same semester more than 3 credit hours.

2) The regular credit hours a student should be registered for should be between 15 and 19 credit hours but the actual average registered credit hours for a student in KU is calculated to be about 13 credit hours. This low registered number of credit hours delays the student graduation and prolongs the time of the study.

3) The minimum GPA for graduation is 2.00 and many students cannot achieve this GPA to graduate so they repeat some courses with a grade less than C to increase their overall GPA.

4) Being a Non Register student for some semesters delays graduation.

5) Each college has its own rules and bylaws for transferring, such as taking certain courses and requiring a certain GPA. When a student is not satisfied or not interested in the college she/he is admitted to, she/he may decide to transfer to another college and, as such, her/his case should meet the bylaws of that college. This increases the time period of the study.

The overall analysis of the results can be summarized as follows. Three states: graduation, dropping out, and transferring are classified in this analysis as absorbing states. Among the freshman students in the colleges of Arts, Sharia (Islamic studies), Business Administration, Engineering, Allied Health and Science, the graduation percentages range between 60% and 80%, while that of Social Sciences, Law, Education, and Medical Sciences range between 90% and 98%. For KU as a whole, the graduation percentage is about 77%. Similarly the probability of graduation of a student at the sophomore level ranges between 0.49 and 1 while the overall percentile is 86.6% for KU. Thus, an increased probability of graduation is observed as the student moves to the higher levels. For students who are in the junior and senior states, the probabilities of graduation are 0.95 and 0.98, respectively. The most remarkable point is that, for the sophomore, junior and senior students of the Law College, the probability of graduation is 1 which ensures that once the student reaches the sophomore state, he/she is certain of completing successfully her/his course of study.

Considering the students in the nonregistered state in Table 7, the overall graduation percentage is only 83%, and the colleges percentages vary between 45% (for the Science College) to 100% (for Law

College). It was found that the rate for the dropout state is very small. This is obvious when the graduation rate is high in each state. Thus, even though the students spend more time in the system, above 85% of them will reach graduation.

Comparing between colleges on the basis of the student lifetime was also investigated. For the Colleges of Sharia & Islamic study, Education and Allied health, students in the states of freshmen, sophomore and junior have approximately 2 semesters as their lifetimes which increase to approximately 3 semesters at the senior states. Students who belong to the colleges of Law, Arts, Social Sciences and Business Administration have lifetimes of about 3 semesters. For the College of Medical Sciences, the freshmen and junior students have normal lifetimes like other colleges of about 2 semesters and are decreasing in sophomore level, whereas the senior students have much longer period. In the Engineering College a student spends about 4.6 semesters in the senior state. For the college of Science, the mean lifetime varies from 2 to 4 semesters in the different stages of the study.

In comparing the mean lifetimes in the system for science and art colleges, it was found that a student in the scientific colleges stays on average 3.2 semesters in the freshmen state while in the arts colleges a student stays on average only 2.3 semesters. This lengthy lifetime in the scientific colleges was observed for each of the states. This implies that the average time that a student spends in the system is more than that in the art colleges (see Tables 2 and 3). This result is expected and agreeable because the core courses of the scientific colleges are more difficult than those of the art colleges.

It was also found that a freshmen student in the scientific colleges stays on average about 8.8 semesters before reaching an absorbing state. Whereas in the art colleges a freshmen student stays on average about 8.1 semesters. This confirms the above inference. Comparing the probabilities of attaining the stable absorbing states, it was found that the students of scientific colleges have less probability of graduating than the students in the art colleges. As such, the probability of dropping out is higher for the scientific students than for the art students.

To avoid such situations, KU should consider necessary actions to reduce drop out from the scientific colleges. Options may include raising the minimum high school grades for university admission or giving a counseling course for students at the time of admission to help students reduce strain in their studies. The average number of registered credit hours for a student in each semester should be between 15 and 19. At the same time students must avoid the conflict in exams schedule as well as the timing problem between lectures offered at different campus locations. Sections must be opened according to the number of students in each level. Levels should not be classified according to the number of credits but according to the kind of courses.

References

Al-Awadhi, S. & Konsowa, M. (2007). An Application of Absorbing Markov Analysis to the Student Flow in Kuwait University. *Kuwait Journal of Science and Engineering*, *34*(*2A*), 77-89.

Bessent, E.W., & Bessent, A. W. (1980). Student flow in a university department: Results of Markovian analysis. *Interfaces*, 37-43.

Kolesat, P. (1970). A Markovian model for hospital admission scheduling. *Management Science*, *16*, 384-396.

Kemeny, J. G., & Snell, J. L. (1970). *Finite Markov chains*. Princeton, NJ: D.Ban Mostrand Company, Inc.

Kwak, N. K., Brown, R. & Schniederjans, M. J. (1985). A Markov analysis of estimating student enrollment transition in a trimester institution. *Socio-Economic Planning Science*, *20*(*5*), 311-318.

Merddith, J. (1976). Selecting optimal training programs in a hospital for the mentally retarded. *Operations Research*, *24*, 899-915.

Mc Namara, J. F. (1974). Markov chain theory and technological forecasting. In *Futurism in Education*, S. P. Hendey & J. R.Yates (Eds.). Berkeley, CA: Mc Cutchan.

Nicholls, M. (2008). Short term prediction of student numbers in the Victorian secondary education system. *Australian and New Zealand Journal of Statistics*, *24*(*2*), 179-190.

Resnick, S. I. (1994). *Adventures in stochastic processes*. Boston, MA: Birkhauser.

Reynolds, D., & Porath, J. (2008). *Markov chains and student academic progress*. Unpublished manuscript, Department of Mathematics, University of Wisconsin-Eau Claire, USA.

Wainwright, P. (2007). *An enrollment retention study using a Markov model for a regional state university campus in transition*. Master thesis, Department of Mathematical sciences, Indiana University, USA, 2007.

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 596-598

1538 – 9472/10/$95.00

BRIEF REPORT

Bayesian Analysis for Component Manufacturing Processes

L. V. Nandakishore
Dr. M. G. R. University,
Chennai

In manufacturing processes various machines are used to produce the same product. Based on the age, make, etc., of the machines the output may not always follow the same distribution. An attempt is made to introduce Bayesian techniques for a two machine problem. Two cases are presented in this article.

Key words: Stochastic models, Bayesian Analysis, MVUE, Posterior distribution.

Introduction

Stochastic models can be better understood through the application of parametric, Bayesian and interval estimations. In this article, Bayesian Analysis of two machines producing the same component is attempted. If the first machine follows a distribution D_1 and the second machine follows distribution D_2, and λ_1 and λ_2 are the proportions of production for the two machines, then the total production equals $\lambda_1 + \lambda_2 = 1$.

In the final lot, a mixture of components from both the machines pooled together will have a distribution given by a linear combination of the two distributions as D = $\lambda_1 D_1 + \lambda_2 D_2$.

Case I

Assumptions:

1. The two machines produce components where the rate of production is not i.i.d.
2. The total lot collected has an observable distribution with an unknown parameter.
3. The number of components observed at sampled points in time is a discrete NB (N.p) distribution.
4. The log normal prior distribution of p is given by

$$\frac{1}{\beta p\sqrt{2\pi}}\, e^{-\frac{1}{2}((\frac{Logp-\alpha}{\beta})^2}$$

and is denoted by $\Lambda(\alpha,\beta^2)$.

If the number of components produced at sampled points of time $(t_1, t_2, ...t_n)$ is $(c_1, c_2, ...c_n)$ then D follows a negative binomial distribution given by

$$p_x = \binom{x+N-1}{N-1} p^N q^x$$

where

x = 0, 1, 2, 3…,

and

p + q = 1. (1.1)

Based on (1.1) the likelihood function of the number of components is given by

$$L(p/c_1, c_2, ...c_n) = \prod_{i=1}^{n} \binom{x_i+N-1}{N-1} p^{nN} q^X,$$

where

$$X = \sum_{i=1}^{n} x_i \quad (1.2)$$

for L to be the maximum likelihood estimator

L. V. Nandakishore is an Assistant Professor in the Department of Mathematics. Dr. M. G. R. University, Chennai 600095.
Email: arunalellapalli@yahoo.com.

$$\frac{\partial Log\, L}{\partial p} = 0 .$$

Hence,

$$\hat{p} = \frac{N}{(N + \bar{X})},$$

where

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} .$$

The sum of independent variables with a negative binomial distribution follows a negative binomial distribution (nN.p) with a probability mass function

$$f(x, p) = p(X = x) = \binom{x + nN - 1}{nN - 1} p^{nN} q^{X}$$

$$x = 1, 2, 3, \ldots \qquad (1.3)$$

where

$$E(X) = \frac{nN(1 - p)}{p} \qquad (1.4a)$$

and

$$Var\,(X) = \frac{nN(1 - p)}{p^2} \qquad (1.4b)$$

For large values of n, $E(\hat{p}) = p$ variance tends to 0, hence, the MVUE of p is $\hat{p}$.

Posterior Distribution

If the prior density of p is a log normal distribution given by

$$\tau(p/\alpha, \beta) = \frac{1}{\beta p \sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\left(\frac{Logp - \alpha}{\beta}\right)^2\right)}, \qquad (2.1)$$

where α is real, $\beta \succ 0$, and 0 < x < ∞ with mean $\alpha + \frac{\beta^2}{2}$ denoted by $\Lambda(\alpha, \beta^2)$, the marginal pdf of X is

$$f_X(x) = \int_0^1 f(x, p)\tau(p/\alpha, \beta)\, dp$$

$$= \int_0^1 f(x.p) \frac{1}{\beta p \sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\left(\frac{Logp - \alpha}{\beta}\right)^2\right)} dp \qquad (2.2)$$

Therefore, the posterior distribution of p given by

$$\frac{\text{Likelihood function * Prior}}{\text{Normalising constant}}$$

is

$$q(p/c_1, c_2, \ldots, c_n) = \frac{f(x, p)\tau(p/\alpha, \beta)\, dp}{\int_0^1 f(x, p)\tau(p/\alpha, \beta)\, dp} \qquad (2.3)$$

Case II

The numbers of components produced at discrete points of time in a given interval are observed if autoregressive processes of order n are considered. The initial observations preceding the sampled data must be determined first, which may not be possible in practical cases. If a first order AR model defined by $X_i = cX_{i-1} + g_i$ is considered where c is the parameter to be estimated, i = 1, 2, 3…, and g_i is the Gaussian noise, i.i.d. of normal variates with $N(0, \sigma^2)$ and stationary for c <1, then the backward shift operator defined by $\mathbf{B}\, X_i = X_{i-1}$ results in $X_i = (1 - cB)^{-1} g_i$. The product of n observations has a multivariate normal distribution with mean zero and variance matrix.

$\sigma^2 =$

$$\begin{pmatrix} (1-c^2)^{-1} & c(1-c^2)^{-1} & ..... & c^{n-1}(1-c^2)^{-1} \\ 0 & (1-c^2)^{-1} & c(1-c^2)^{-1}.... & c^{n-2}(1-c^2)^{-1} \\ 0 & 0 & (1-c^2)^{-1}... & c^{n-3}(1-c^2)^{-1} \\ 0 & 0 & 0 & (1-c^2)^{-1} \end{pmatrix}$$

Let

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ -c & 1 & 0 & 0 & 0 \\ 0 & -c & 0 & 0 & 0 \\ & & & & \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

then

$$P\sum P^i =$$

$$\sigma^2 \begin{pmatrix} (1-c^2)^{-2} & 0 & 0 & 0 .... & 0 \\ 0 & 1 & 0 & 0 .... & 0 \\ 0 & 0 & 1 & 0 ..... & 0 \\ & & & & \\ 0 & 0 & ,,,,,,,,,,,,,,,...,,1 & & \end{pmatrix} = \sigma^2 D,$$

which is the covariance matrix of Y=PX which has a multivariate normal dist with zero mean. Because $X=\prod_{i=1}^{n} X_i$ the joint pdf of its components is

$$\frac{1}{(\sigma\sqrt{2\pi})^n} exp(\frac{1}{2\sigma^2}(1-c^2)X_1^2 - \sum_{i=2}^{n}(X_i X_{i-1})^n.$$

Acknowledgements

I dedicate this paper to the memory of my late parents Mrs. & Professor Dr. L. V. K. V. Sarma, I. I. T. Madras, and also to Ms. Aruna for spontaneous support.

References

Alexander, Mood, & Boes. (1974). *Theory of Statistics*. New Delhi, India: Tata McGraw Hill.

Broemiling, L. D. (1993). *Bayesian analysis of linear models*. New York: Marcel Dekker Inc.

Radhakrishna, R. C. (2001). *Linear statistical inference and its applications* (*2nd Ed.*). New York: Wiley-Interscience.

Zellner, A. (1996). *An introduction to Bayesian inference to ecometrics*. New York: John Wiley.

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 599-603

1538 – 9472/10/$95.00

Emerging Scholars

Estimating the Non-Existent Mean and Variance of the F-Distribution by Simulation

Hamid Reza Kamali
Private Scholar

Parisa Shahnazari-Shahrezaei
Firoozkooh Branch Islamic Azad University,
Firoozkooh, Iran

In theory, all moments of some probability distributions do not necessarily exist. In the other words, they may be infinite or undefined. One of these distributions is the F-distribution whose mean and variance have not been defined for the second degree of freedom less than 3 and 5, respectively. In some cases, a large statistical population having an F-distribution may exist and the aim is to obtain its mean and variance which are an estimation of the non-existent mean and variance of F-distribution. This article considers a large sample F-distribution to estimate its non-existent mean and variance using Simul8 simulation software.

Key words: Probability distribution, F distribution, simulation.

Introduction

In practice, it is often necessary to calculate some properties of a statistical population. For this purpose, a random sample is taken from the population, its properties are calculated and these properties are developed to a population. A subset of individuals is selected from within a population to yield some knowledge about the whole population by sampling. The developed properties are approximate and are typically expressed as a confidence interval. Conversely, researchers may seek to calculate some properties of a random sample with the aim of estimating these properties to a population. In this case, the considered properties are calculated for the whole population once and then are used for drawn random samples. The most important statistical properties of a population and random sample are the mean and variance which are calculated from the probability distribution function of a population. If it is not possible to obtain these properties using a probability distribution function, all individuals of the population should be examined and used to obtain the considered properties. Although this work is easy for small and finite populations, it is not possible for infinite populations and a fairly large random sample is needed for approximate calculations.

Hamid Reza Kamali is a Ph.D. Student in the Department of Industrial Engineering. Email: hrkamali@gmail.com. Parisa Shahnazari-Shahrezaei is a Ph.D. Student in the Department of Industrial Engineering. Email: parisa_shahnazari@iaufb.ac.ir.

F-distribution

Probabilistic behavior of some random variables can be defined as a mathematical function of the value of the considered random variable which is called the probability distribution function. The most important probability distribution functions considered in this article are the Normal, Chi-square, and F distributions.

Equation (1) shows the Normal distribution function (Walpole & Myers, 1993) which includes two parameters $\mu \in R$ and $\sigma \in R^+$. Equations (2) and (3) express its mean and variance, respectively. The values of the Normal random variable, which belong to a real numbers set, have a specific mean and variance.

$$f_N(n) = \frac{1}{\sigma\sqrt{2\pi}} \text{EXP}(-\frac{(n-\mu)^2}{2\sigma^2}); \quad -\infty < n < +\infty \tag{1}$$

$$E(N) = \mu \tag{2}$$

$$Var(N) = \sigma^2 \tag{3}$$

In the case of $\mu = 0$ and $\sigma = 1$, the Normal distribution is called the Standard Normal distribution and its probability distribution function will be according to Equation (4):

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \text{EXP}(-\frac{z^2}{2}); \quad -\infty < z < +\infty \tag{4}$$

The Chi-square distribution (Walpole & Myers, 1993) has a parameter $\nu > 0$ which is called degree of freedom. Equations (5) to (7) display its probability distribution function, mean and variance, respectively. The values of a Chi-square random variable which belong to a positive real numbers set also have a specific mean and variance.

$$f_{\chi^2}(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}; \quad x > 0 \tag{5}$$

$$E(\chi^2) = \nu \tag{6}$$

$$Var(\chi^2) = 2\nu \tag{7}$$

The F-distribution (Walpole & Myers, 1993) has two parameters $\nu_1 > 0$ and $\nu_2 > 0$ which are called the first and second degree of freedom. Its probability distribution function, mean and variance are illustrated by Equations (8) to (10), respectively. The values of an F random variable belong to a positive real numbers set.

$$f_F(w) = \frac{\Gamma[\frac{(\nu_1+\nu_2)}{2}](\frac{\nu_1}{\nu_2})^{\nu_1/2} w^{(\nu_1/2-1)}}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})} (1+\frac{\nu_1}{\nu_2})^{-(\nu_1+\nu_2)/2}; \quad w > 0 \tag{8}$$

$$E(F) = \frac{\nu_2}{\nu_2 - 2}; \quad \nu_2 > 2 \tag{9}$$

$$Var(F) = \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}; \quad \nu_2 > 4 \tag{10}$$

As shown, the F-distribution's mean and variance have not been defined for the second degree of freedom less than 3 and 5, respectively. While confronting these circumstances, the approximate mean and variance of population can be obtained using a fairly large random sample and simulation. The Simul8 software is used in this research to conduct simulations. Because the F-distribution has not been defined in this software, simulation is performed using the relation between F and the Normal distribution. Equation (11) shows that the Chi-square distribution with ν degrees of freedom is the distribution of a sum of the squares of ν independent Standard Normal random variables (Walpole & Myers, 1993):

$$Z \approx N(0,1); \quad X = \sum_{i=1}^{\nu} Z_i^2 \Rightarrow X \approx \chi^2(\nu) \tag{11}$$

Also, the F-distribution with ν_1 and ν_2 degrees of freedom arises from the ratio of two independent Chi-Square random variables that have been divided on their degrees of freedom according to Equation (12) (Walpole & Myers, 1993):

$$X_1 \approx \chi^2(\nu_1);$$
$$X_2 \approx \chi^2(\nu_2); \quad (12)$$
$$W = \frac{X_1/\nu_1}{X_2/\nu_2} \Rightarrow W \approx F(\nu_1, \nu_2)$$

Simulation by the Simul8 software

In practice, many problems have probabilistic behavior. One of the most common cases is a queuing problem in which customer arrival rate, service time, and the like are not exact. Queuing theory can be used to solve these types of problems. Sometimes, a complex combination of several queues in exact and/or probabilistic manners with limitations, calculations and different conditions are observed. Examination of such a complex problem by queuing theory is possible theoretically, but nearly impossible practically. Hence, computer and simulation sciences are needed for calculations. Using simulation science, a complex system can be run virtually and its behavior can be forecasted and examined within a reasonable time. The Simul8 software used in this research has a user-friendly graphical aspect and includes the following elements (Simul8 Software, Version 2000):

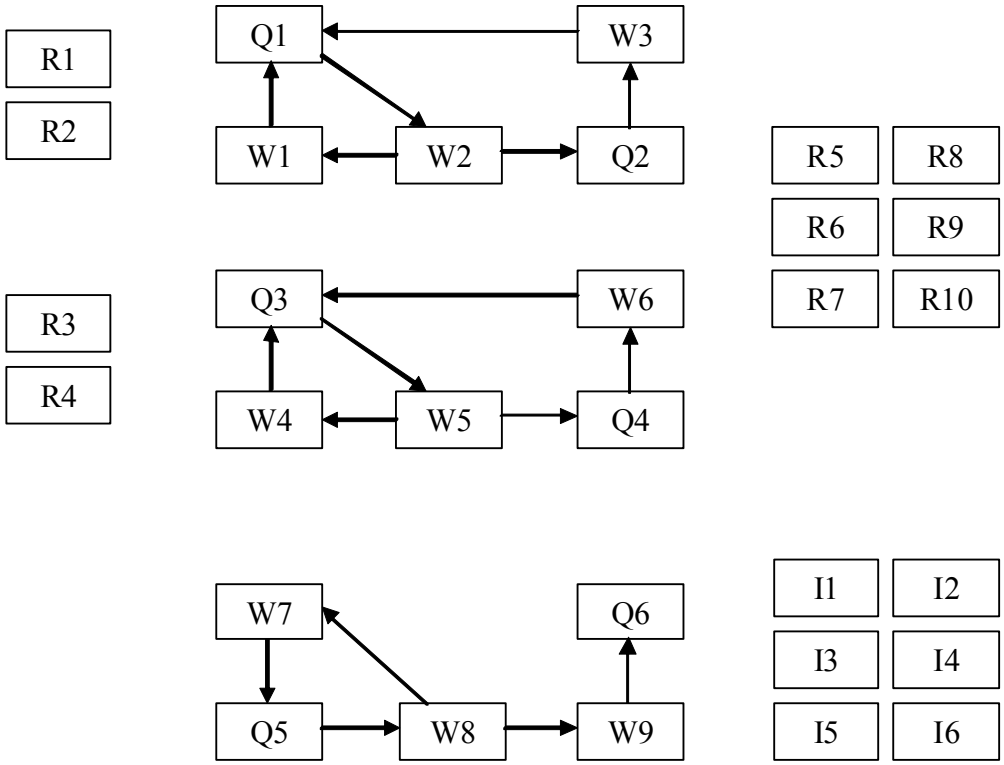
1) Object. An object is like a customer that moves in a special path to get some services in service centers or work stations and waits in queues until the conditions for service in the next service center are ready. This object may exist in a queue at first or may enter the system through an enter point after the system begins running. The object can have a label. A label is a variable that stores a property of an object.
2) Enter point. New objects come into the system through these points. The object's entry time probability distribution and values of its labels can be nominated.
3) Work station. In this element, the considered object gets service. Service time that can be probabilistic, number of servers, capacity and resources can be assigned. Also, calculations on the model's variables and object's labels can be done in this element.
4) Queue. Objects of the model wait in queues until the conditions to enter the next work station are ready. For example, if the element after the queue is a work station and it needs a resource to work, the object waits in the queue until that resource is ready. The queue capacity can be infinite or bounded.
5) Resource. This element is a variable that obtains the conditions and number of services. The value of each resource can be assigned at first or can be assigned in running time. After the service, the resource value can change to itself or another resource.
6) Exit point. Objects of the model go out of the model from an exit point, for example, the customers that have finished their service.
7) Information store. This element is a variable that stores properties of the model.

Proposed Model

The objective of this article is to obtain the mean of an F-distribution with $1 \le \nu_1 \le 30$ and $\nu_2 < 3$ and also the variance of an F-distribution with $1 \le \nu_1 \le 30$ and $\nu_2 < 5$. For this purpose, a random sample of size 500 is drawn from a statistical population having an F-distribution. In order to achieve more accurate results, the proposed model is run 100 times and the average of the obtained means and variances are recorded as final results. It should be noted that the proposed model generates Standard Normal random numbers at first and converts them to Chi-square random numbers. Afterwards, Chi-square random numbers are converted to F random numbers for calculations. In the proposed model, queues, work stations, resources and information stores are displayed with Q, W, R and I, respectively. As shown in Figure 1, the proposed model includes three parts. The first and second parts of the model generate the first and second Chi-square random numbers, respectively. Because the performance of both parts is similar, only the first part is described.

In the first part, there is one object in Q1 at first. This object enters W2 while running the model. The service time of W2 is exactly equal to zero. In the other words, W2 is a virtual work

Figure 1: Simulated Model by the Simul8 Software



station. In W2, a Standard Normal random number is assigned to the object's label. Afterwards, the object can enter W1 or Q2. In the proposed model, entrance priority belongs to W1, which is also a virtual work station. W1 uses one unit of R1 and then changes it to R2. When R1 finishes, the object cannot go to W1 from W2 and should inevitably enter Q2. In W1, the value of the object's label is squared and added to I1 and it then returns to Q1. The loop Q1-W2-W1 is repeated R1 times and a Chi-square random number with R1 degrees of freedom is generated and stored in I1. The object waits after entering Q2 until the conditions to enter W3 are ready. W3 requires R6 and changes it to R7 after using. The initial value of R6 is zero but after running the model, a value is assigned to it by the third part; R6 is an intermediate between the first and third parts. W3 which is a virtual work station changes the value of R1, that is now equal to zero, to its initial value and also changes the value of R2 and the object's label to zero. In the other words, the first part of model returns to its initial condition.

In the third part, there is only one object in Q5 at first, the object then enters W8 which is the only real work station of the model and has an exact service time. Because the first and second parts of the model have service times equal to zero they generate Chi-square random numbers I1 with R1 degrees of freedom and I2 with R3 degrees of freedom, respectively, when the object is in W8. Afterwards, the object enters W7 which is a virtual work station. In W7, an F random number is generated by the ratio of I1 and I2 which have been divided on their degrees of freedom (R1 and R3). The generated value and its square are respectively added to I3 and I4 to obtain the sum and sum of squares of values of the F random variable, and then W7 uses one unit of R5 and R8 and changes them to R6 and R9, in that order. Thus, W7 lets the first and second parts of the model generate new Chi-square random numbers. Calculation of sum and sum of squares of values of the F random

variable is repeated R5 times (which is equal to R8 times) in the loop Q5-W8-W7. R5 is considered equal to 500 in the proposed model. After finishing R5 and R8, the object enters W9 from W8. In W9, the values of the sample mean (I5) and sample variance (I6) are calculated using I3 and I4, which are respectively the sum and sum of squares of values of F random variable, according to Equations (13) and (14):

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{13}$$

$$S^2 = \frac{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}{n(n-1)} \tag{14}$$

By setting resources R1 and R3 equal to the first and second degree of freedom of an F-distribution respectively, mean and variance of a random sample of size 500 (I5 & I6) are calculated. To obtain more accurate results, the proposed model is run 100 times and the average of obtained means and variances are recorded. Table 1 shows the obtained results.

References

Walpole Ronald E. & Raymond H. Myers, Probability and Statistics for Engineers and Scientists, Fifth Edition, Macmillan, 1993.

Manual.pdf Help File, Simul8 Software Version 2000.

Table 1: Calculated Mean and Variance for the F-distribution

F Distribution	Mean		Variance			
V_1	V_2=1	V_2=2	V_2=1	V_2=2	V_2=3	V_2=4
1	11130.21	10.05	8.52E+11	108031.28	3793.61	41.65
2	33169.64	13.75	1.38E+13	541607.53	991.75	1174.13
3	37594.24	14.30	2.38E+13	458742.85	686.37	34.18
4	15012.24	10.74	1.54E+12	112955.47	3717.85	45.92
5	26761.31	12.81	7.69E+12	334859.40	1994.39	74.23
6	23044.24	13.13	4.51E+12	337649.74	5092.49	355.42
7	18210.40	11.22	2.36E+12	165931.92	1215.85	163.34
8	23826.42	13.79	4.96E+12	397856.63	2340.31	155.83
9	22405.18	11.67	3.29E+12	135107.50	3087.51	144.57
10	30296.88	12.90	9.19E+12	220879.02	6534.82	388.49
11	29119.63	11.62	8.05E+12	120507.71	1082.33	251.6
12	28509.47	11.65	6.61E+12	142528.84	2494.88	126.27
13	25922.59	11.60	5.29E+12	130915.62	1814.47	192.06
14	28315.80	12.34	1.07E+13	172780.09	3946.75	93.11
15	21680.74	13.58	3.33E+12	368276.62	1641.08	463.65
16	27846.46	11.80	6.59E+12	146747.04	1627.26	76.08
17	27757.44	13.02	8.41E+12	337102.72	2068.54	248.91
18	25305.75	11.32	4.68E+12	122134.97	634.55	113.93
19	29520.33	13.02	6.81E+12	338522.99	1357.71	171.95
20	24484.70	12.30	5.05E+12	218104.74	1434.49	251.87
21	23815.50	12.92	3.61E+12	245563.95	2027.46	121.14
22	24091.21	12.26	3.81E+12	220336.07	1567	121.76
23	24122.25	13.19	4.21E+12	295272.78	1133.46	150.74
24	24058.93	13.47	3.98E+12	359914.76	1061.62	189.02
25	22481.36	12.01	3.37E+12	191010.65	2112.89	287.59
26	25595.00	12.07	5.59E+12	196676.92	2368.16	157.14
27	27161.81	13.25	5.77E+12	319466.56	2136.64	122.85
28	26124.24	12.42	5.92E+12	230289.00	1710.13	245.72
29	26067.87	12.99	6.83E+12	237958.74	940.05	213.54
30	26485.85	13.58	5.27E+12	370863.43	2165.15	167.76

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 332-603

1538 – 9472/10/$95.00

Instructions For Authors

Follow these guidelines when submitting a manuscript:

1. *JMASM* uses a modified American Psychological Association style guideline.
2. Submissions are accepted via e-mail only. Send them to the Editorial Assistant at ea_jmasm@wayne.edu. Provide name, affiliation, address, e-mail address, and 30 word biographical statements for all authors in the body of the email message.
3. There should be no material identifying authorship except on the title page. A statement should be included in the body of the e-mail that, where applicable, indicating proper human subjects protocols were followed, including informed consent. A statement should be included in the body of the e-mail indicating the manuscript is not under consideration at another journal.
4. Provide the manuscript as an external e-mail attachment in MS Word for the PC format only. (Wordperfect and .rtf formats may be acceptable - please inquire.) Please note that Tex (in its various versions), Exp, and Adobe .pdf formats are designed to produce the final presentation of text. They are not amenable to the editing process, and are **NOT** acceptable for manuscript submission.
5. The text maximum is 20 pages double spaced, not including tables, figures, graphs, and references. Use 11 point Times Roman font.
6. Create tables without boxes or vertical lines. Place tables, figures, and graphs "in-line", not at the end of the manuscript. Figures may be in .jpg, .tif, .png, and other formats readable by Adobe Photoshop.
7. The manuscript should contain an Abstract with a 50 word maximum, following by a list of key words or phrases. Major headings are Introduction, Methodology, Results, Conclusion, and References. Center headings. Subheadings are left justified; capitalize only the first letter of each word. Sub-subheadings are left-justified, indent optional. Do not number headings or subheadings.
8. Number all formulas, tables, figures, and graphs, but do not use italics, bold, or underline.
9. Do not use underlining in the manuscript. Do not use bold, except for (a) matrices, or (b) emphasis within a table, figure, or graph. Do not number references. Do not use footnotes or endnotes.
10. In the References section, do not put quotation marks around titles of articles or books. Capitalize only the first letter of books. Italicize journal or book titles, and volume numbers. Use "&" instead of "and" in multiple author listings.
11. *Suggestions for style*: Instead of "I drew a sample of 40" write "A sample of 40 was selected". Use "although" instead of "while", unless the meaning is "at the same time". Use "because" instead of "since", unless the meaning is "after". Instead of "Smith (1990) notes" write "Smith (1990) noted". Do not strike spacebar twice after a period.

Print Subscriptions

Print subscriptions including postage for professionals are US $95 per year; for graduate students are US $47.50 per year; and for libraries, universities and corporations are US $195 per year. Subscribers outside of the US and Canada pay a US $10 surcharge for additional postage. Online access is currently free at http://www.jmasm.com/. Mail subscription requests with remittances to JMASM, P. O. Box 48023, Oak Park, MI, 48237. Email journal correspondence, other than manuscript submissions, to jmasm@wayne.edu.

Notice To Advertisers

Send requests for advertising information to jmasm@ wayne.edu.