1-1-1980

An empirical investigation of the consequences of error model misspecification.

David A. Wagstaff
University of Massachusetts Amherst

AN EMPIRICAL INVESTIGATION OF THE CONSEQUENCES
OF
ERROR MODEL MISSPECIFICATION

A Dissertation Presented

By

DAVID A. WAGSTAFF

Submitted to the Graduate School of the
University of Massachusetts in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 1980

PSYCHOLOGY

AN EMPIRICAL INVESTIGATION OF THE CONSEQUENCES

OF

ERROR MODEL MISSPECIFICATION

A Dissertation Presented

By

DAVID A. WAGSTAFF

Approved as to style and content by:

Jerome L. Myers

Jerome L. Myers, Chairperson

David W Hosmer Jr.

David W. Hosmer, Member

Ronnie Janoff-Bulman

Ronnie Janoff-Bulman, Member

Anne M. Stoddard

Anne M. Stoddard, Member

Charles E. Clifton, Jr.,
Acting Department Head,
Psychology

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to the members of my dissertation committee: Jerry Myers, Dave Hosmer, Ronnie Janoff-Bulman, and Anne Stoddard. While this dissertation benefited greatly from their many helpful comments, suggestions, and criticisms, I would be remiss if I did not acknowledge the important role that each member played in terms of my development as a statistician and data analyst.

Jerry introduced me to graduate-level statistics and, in doing so, made the learning of statistics an enjoyable enterprise. Dave indirectly sent me the first student that I would tutor in statistics (my wife, Bettie). Through Dave, I learned that teaching statistics could be a very rewarding experience. Ronnie supported my exploration in statistics. She served on the committees for my master's degree and my comprehensive examination. Anne completed the cycle. She taught my final graduate-level statistics course. (Anne also served on my comprehensive examination committee.)

I would also like to express my gratitude to Janet D. Elashoff. Without her timely hints and encouraging letters, I certainly would have been less eager to adopt an analytic approach.

Finally, I would like to thank Pat Milne of Wordstream Processing. Pat, as usual, did an outstanding job typing this dissertation.

ABSTRACT

A.B., 1969, Lafayette College, Easton, Pennsylvania

M.S., 1977, University of Massachusetts, Amherst, Massachusetts

Ph.D., 1980, University of Massachusetts, Amherst, Massachusetts

Directed by: Professor Jerome L. Myers

Four Monte Carlo simulations were undertaken in order to determine the consequences of analyzing data from a mixed sample as if it had come from a single, homogeneous population. More specifically, the study examined the effect that misspecification of the error model had on Ordinary Least Squares estimation and its associated hypothesis testing procedure for no slope (i.e., the t-test).

The obtained data and supporting analytic arguments suggest that misspecification of the error model will not seriously affect parameter estimation when the contaminating fraction is small and variables are measured on short, ordinal scales (e.g., the Likert scale). The obtained data also suggest that misspecification of the error model will not seriously affect the Type II error rate (i.e., the probability of accepting a false null hypothesis) providing that the postulated causal model does not include variables not found in (or exclude variables found in) the regression models for the separate components.

TABLE OF CONTENTS

CHAPTER

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

CHAPTER I

INTRODUCTION

A Hypothetical Example

Consider the following hypothetical study. A social psychologist is interested in investigating the extent to which the attribution of responsibility (Y) is influenced by the perceived similarity (X_1) between observer and actor, and by the seriousness (X_2) of any injuries or losses resulting from a harmful act. To be more specific, the social psychologist would like to determine the extent to which an observer would allow his judgment concerning the actor's responsibility for having caused the harmful act to be influenced by (a) the extent to which the observer perceives the actor to be "like" him, and (b) the seriousness of any injuries or losses which result from the harmful act.

To provide potential respondents with a relevant and engaging experimental task, the social psychologist decides to consider the situation where the observer (i.e., the respondent) is a "witness" to a traffic accident; the actor is the driver of the moving vehicle; and the accident results in the death of a jaywalker. Given this context, the social psychologist advances two hypotheses. First, he suggests that the extent to which an observer holds a driver responsible for a jaywalker's death is inversely related to the extent to which the observer perceives the driver to be "like" him. Second, he suggests that the extent to which an observer holds a driver responsible for a jaywalker's death is directly related to the "losses" the observer assigns to this kind of situation. In effect, the social psychologist is implicitly assuming that the observer is influenced more by the social identity of the jaywalker than by the actual physical nature of the injury or loss.

Fall semester begins and the social psychologist conducts the study. Unfortunately, he is unaware that the obtained data are described best as a mixture of two regression equations. That is, given his sample of N respondents, N_1 respondents provide data that support the hypothesized causal model while N_2 respondents provide data that challenge the hypothesized causal model. (Note, $N = N_1 + N_2$.)

Now, this mixture may have occurred for a number of reasons. For example, the two populations may reflect different social norms. The N_1 respondents may come from communities that require the driver to yield the right of way to all pedestrians (even jaywalkers). As a result, they are cognizant of the need to sanction a driver who fails to obey the law. More importantly, they are privy to the tacit knowledge as to the "type" of driver who should be imprisoned or fined for a particular kind of motor vehicle offense. In contrast, the N_2 respondents may come from communities that refuse to take any action as it is understood that pedestrians jaywalk at their own risk. Consequently, they may have little knowledge of, or experience with, the cognitive or social mechanisms which would assist them in assessing the seriousness of the loss or in assigning responsibility for the jaywalker's death.

On the other hand, the two populations may reflect different definitions of personal responsibility. Here, the N_1 respondents may use a pragmatic definition of personal responsibility that requires them to balance the consequences of sanctioning the driver against the losses incurred by the jaywalker and his immediate relations. As a result, they may feel justified in considering the social identities of both the driver and the jaywalker when assessing the seriousness of the loss and attributing responsibility for the death. In contrast, the N_2 respondents may use a strict moral definition of personal responsbility that requires them to ignore the social identities of driver and jaywalker. Because their available information is limited to a description of the participants and the act, they may base their judgments solely on the latter.

Then, the mixture may have occurred because one population incorrectly (or correctly) guessed the experimental hypotheses. For example, the N_2 respondents may have incorrectly concluded from the differences in social status between the driver and jaywalker that the study deals with social stereotypes. As a result, they may have intentionally biased their answers so that they would appear to be the kind of individuals who would not be unduly influenced by "extra-legal" considerations.

The point behind the preceding (and admittedly short) list of examples is that in any given study a researcher may unknowingly treat two or more heterogeneous groups as if they were one homogeneous population. By "treat," it is meant that the researcher analyzes his data on the erroneous assumption that all N observations come from the same population. It is clear from the hypothetical attribution study that an immediate consequence of "misspecifying the true error model" is the failure to identify correctly the population for which the postulated causal model would be most appropriate. What is not clear and, thus, needs to be demonstrated is whether or not an additional consequence of misspecifying the error model is the eventual acceptance of a causal model which is inappropriate for any population.

For the sake of completeness, let's define the term "appropriate" as it is used to describe a causal model. A causal model is appropriate if it correctly identifies the causally relevant variables. If the attribution of responsibility is indeed determined by perceived similarity and seriousness, then $Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$ may be an appropriate causal model. Now a causal model is to be distinguished from a mathematical model. Two different mathematical models are $Y_i = 2.54\ X_{1i} + 2.36\ X_{2i} + \varepsilon_i$ and $Y_i = 1.36 + 5\ X_{1i} + 1.00\ X_{2i} + \varepsilon_i$. However, as both cite the same variables as being causally relevant, both reflect the same causal model. Finally, while one mathematical model may be better than the other (where "better" is judged in terms of mean squared prediction error), both may be appropriate causal models.

Problem Statement

In order to develop causal theory from observational data, the researcher will often posit a simple linear model. That is, the researcher will express Y, the outcome variable, as a simple linear function of one or more explanatory variables, X_i ($i = 1, 2, \ldots, k$). Then, to determine whether the data are consistent with the postulated causal relationship between a given explanatory variable and Y, the researcher may use a test of statistical significance (e.g., the t-test for H_o: $\beta = 0$). If the test result suggests that the explanatory variable accounts for an appreciable proportion of the observed variation in Y, the researcher will attempt to use the estimated regression weight, $\hat{\beta}_i$, to say something about the magnitude of the causal effect of X_i on Y.

The selection of the most efficient estimator of $\underline{\beta}$, the vector of regression weights, depends on the distribution that is assumed for the errors of measurement. If the errors are adequately described by the Normal distribution, the Method of Ordinary Least Squares will provide the most efficient estimator of $\underline{\beta}$ (Harter, 1975). On the other hand, if the errors are more appropriately described as a mixture of two or more Normal distributions, Least Squares may provide a grossly inefficient estimator for both $\underline{\beta}$ and $V(\underline{\beta})$, the variance-covariance matrix for $\underline{\beta}$ (Mosteller and Tukey, 1977; Wainer and Thissen, 1976).

Many studies consider the statistical behavior of the various estimators of $\underline{\beta}$ when the errors of measurement are approximated best by the mixture of two Normal distributions. However, few studies actually address the problems that such mixtures may pose for the initial statement and subsequent refinement of causal theory. To be more specific, few studies ask how the development of a causal theory--the simple linear causal model--may be affected when the researcher fails to detect that the errors reflect a mixture of two (or more) distributions.

Then, few studies consider how ordinal measurement may affect a researcher's ability to detect mixed distributions. This apparent lack of interest on the part of social researchers is puzzling in that many of the variables that provide the basis for theory in the social sciences are measured on ordinal scales. Measurement on the short, ordinal scale (e.g., a 5-point Likert scale) is likely to be problematic in that it places a restriction on the extent to which marginal and conditional means of heterogeneous populations may differ. When dependent variables are restricted to a few discrete values, misspecification of the error model is (a) more likely to occur and (b) more difficult to detect.

Given the scarcity of research in this area, the present study seeks to determine the conditions under which misspecification of the error model and ordinal measurement lead to the statement of a completely misleading (as opposed to an incomplete) causal theory.

An incomplete causal theory is said to develop when the researcher fails to match the obtained causal model with the proper component population. A completely misleading causal theory is said to develop when the researcher accepts as viable a causal model which is not appropriate for any of the component populations.

Literature Review

Real data (in contrast to computer simulated data) may manifest non-normal error distributions for several reasons. First, the data may be generated by processes which are represented best by skewed distributions. For example, the length of time (in months) from a prisoner's release until reincarceration is a non-negative, positively skewed variable (see Witte and Schmidt, 1977). Second, the data may be generated by processes which are represented best by "long-tailed" distributions. Such distributions are more dense in the tails than the correspondening Normal distribution. For example, the distribution of income in the U.S. is described better by the Pareto distribution than by the less dense Normal distribution (see Hauseman and Wise, 1977). Finally, the data may be generated by processes which are represented best by "short-tailed" distributions. These distributions are less dense in the tails than the corresponding Normal distribution. For example, a respondent may be asked to use a 5-point bipolar scale to indicate the extent to which he agrees with a questionnaire item. Now the data obtained in this manner may be described better by a short-tailed distribution than by the more dense Normal distribution.

To generate data sets that approximate the kinds of non-normal distributions observed in real data, statisticians have frequently employed a mixture of two Normal distributions. For example, Elashoff (1972) has modeled a skewed error distribution with

$$\varepsilon_i \sim (1 - \pi)\ N\ (0,\ \sigma^2) + \pi\ N(\lambda(x),\ \sigma^2) \quad ,$$

where π is the mixing proportion and $\lambda(x)$ is either a constant or some function of the data values. When the errors are appropriately described by the above model, Least Squares estimators are biased and inefficient (Elashoff, 1972).[1]

The improvement in efficiency offered by alternative or "robust" estimators varies in a complex way with the mixing proportion, π; the degree of separation between the two populations, λ; the sample size, N; and the method of estimation. Because these four factors interact with one another, a straightforward interpretation of the various Monte Carlo simulations is inappropriate. However, it it would appear that Least Squares estimators fare poorly when the sample size is small, and the contaminating data points are both numerous and distant from the mean of the remaining data points.

[1] When the mixture results in a symmetric error distribution (say, one with a common mean, but unequal variances for the individual components), the Least Squares estimators are simply inefficient (see Andrews _et al._, 1972; Wainer and Thissen, 1976; Mosteller and Tukey).

If the outcome variable, Y, can be expressed as a simple linear function of one or more explanatory variables and the errors of measurement are appropriately modeled by a mixed or "compound" error model, the sample data are described best as a mixture of two regressions (see Elashoff, 1972; Quandt, 1972; Hosmer, 1974; Quandt and Ramsey, 1978; Kiefer, 1978). That is, data are collected from N observational units with respect to Y and $\underline{X}$, an N x p matrix with p explanatory variables. Then, with a probability equal to $(1 - \pi)$, the regression

$$Y_i = \alpha + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i$$

occurs; and with a probability equal to π, the regression

$$Y_i = \alpha^* + \sum_{j=1}^{p} \beta^*_j X_{ij} + \varepsilon^*_i$$

occurs (where one or more of the parameters in the second regression differs in value from its counterpart in the first regression).

The mixture may arise in one of two ways. First, it may occur as a result of some physical process. For some reason data are not collected on the variable or variables that would allow the researcher to assign each of the observational units to its respective population. As a result, the obtained sample contains two or more different statistical populations. Second, the mixture may occur as the result of a "structural change." That is, at some level of a known or unknown factor, a slight or radical alteration occurs in the nature of the relationship between the outcome variable and the explanatory variables.

Now, the literature on mixed regressions is particularly relevant in that it examines the factors which influence the researcher's ability to detect a misspecified error model. Specifically, this literature suggests that the researcher's ability to detect mixed regressions varies with the mixing proportion, π; the degree of separation between the mixed regression lines, λ; the sample size, N; and the method of estimation. Again, as these factors interact, a straightforward interpretation of the various Monte Carlo simulations is not feasible. However, it would appear that misspecification of the error model is most likely to go undetected when the regression lines are not well separated, and the sample is small in size and relatively free of contaminating data points.

In summary, the following points are noted. First, the factors which influence the researcher's ability to detect a misspecified error model are identical to the factors which enable a misspecified error model to produce unreliable and/or grossly misleading estimates. These factors are: the mixing proportion, π; the degree of separation between the marginal or conditional means, λ; and the sample size, N. Second, the more difficult it becomes to detect a misspecified error model, the less likely error model misspecification is to result in biased and grossly inefficient estimators. For example, a misspecified error model is difficult to detect when the two regressions (say) are not well separated and the obtained sample is relatively free of contaminating data points. On the other hand, a misspecified error model poses a serious threat to Least Squares estimation only when the two regressions are quite distant from one another and the obtained sample is relatively "noisy."

Finally, in that measurement on the short, ordinal scale restricts the extent to which heterogeneous groups may differ, the author suggests the following. First, as detection of mixed regressions depends on the degree to which the regression lines are separated, it is suggested that the researcher's ability to detect mixed regressions decreases with decreasing scale length, L. Second, as the realization of reliable estimates also depends on the separation between the different regression lines, it is suggested that the researcher's ability to obtain reliable estimates may also be influenced by scale length.

Study Objectives

Through a series of Monte Carlo simulations, the author seeks to determine the consequences of misspecifying the error model for stated combinations of the mixing proportion, sample size, scale length, and degree of separation between the mixed regressions. By "consequences," the author refers to any problem misspecification may pose for parameter estimation and statistical inference procedures. In addition, the author seeks to determine if a decrease in scale length can in fact affect a researcher's ability to detect mixed regressions (when the values of the mixing proportion, sample size, and degree of separation are fixed).

CHAPTER II

METHODOLOGY

Monte Carlo Simulation: General Remarks

The present study employs a series of Monte Carlo simulations to investigate the consequences of misspecifying the error model. In particular, it seeks to determine how estimation and inference are affected when all observations are incorrectly assumed to come from the same statistical population.

In general, a Monte Carlo simulation represents an attempt to have the computer generate values which behave as if they were the result of a random process (Chambers, 1977). For many statistical investigations, the approximate answer achieved through a Monte Carlo simulation provides a useful complement to the more rigorous analytic solution. However, for the statistical investigations in which an analytic solution is not readily forthcoming, the answer achieved through a Monte Carlo simulation may be the only obtainable one.

Whereas the analytic approach may suffer from the use of questionable simplifying mathematical assumptions, a Monte Carlo simulation faces the additional problem of having the computer generate values in accordance with the analyst's intentions and/or programmer's instructions. In each of the four simulations undertaken in this study, the computer is asked to generate data which approximate the kind that one would observe if a sample of N observations contained a mixture of two simple linear regressions. Specifically, the computer is asked to:

STEP

0 enter the necessary input data (e.g., X_1, X_2, . . . , X_N) and set the initial values of the design parameters (the mixing proportion, π; the sample size, N; the degree of separation between the two regressions, λ; and the scale length, L);

1 generate 2N random uniform deviates;

2 use the uniform deviates to generate N values of Y, an outcome variable that can assume any integer value from 1 to L, such that each Y_i has the "compound" probability density function $f_3(Y_i) = 1 - \pi)f_1\ (Y_i) + \pi\ f_2(Y_i)$, where $f_K(Y_i) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp - \frac{1}{2}\{[Y_i - E(Y_i|X_i)]\ \sigma\}^2$ for K = 1, 2;

3 calculate the Least Squares regression estimates (a, b, and $\hat{\sigma}^2_{y|x}$) and t-statistic testing H_o: $\beta = 0$;

4 repeat steps (1) through (3) until the number of replications equals 1000;

5 calculate summary statistics and moments for the empirical sampling distributions of the regression estimates and t-statistic;

6 change the values of specified design parameters and repeat Steps (1) through (5).

Estimation of the regression parameter is said to be affected if the mean of the 1000 Monte Carlo estimates differs from the correct value of the regression parameter. Inference is said to be affected if the probability of accepting a false null hypothesis (H_o: $\beta = 0$) is 0.10 or greater. Or to be more specific, inference is said to be affected if 100 or more of the computed t-statistics,

$$|t| = \frac{\hat{\beta} - 0}{\sqrt{\hat{\sigma}^2_\beta}}$$

where $\hat{\sigma}^2_\beta$ is the estimated variance of β, are less than $t_{N-2,0.975}$, the value of the t-distribution corresponding to the 97.5th percentile.

Simulation I utilizes a 2 x 4 factorial design with a single control group ($\pi = 0.00$). The degree of separation and mixing proportion are varied while the sample size and scale length are fixed. Further, X_1, X_2, . . . , and X_N are arrayed in a rectangular distribution to reflect the classic regression situation.

Simulation II utilizes a 2 x 3 x 5 factorial design. The sample size, scale length, and mixing proportion are varied while the separation factor is held constant. Here, X_1, X_2, . . . , and X_N are arrayed in a symmetrical distribution.

For Simulations III and IV, the method used to generate Y values is changed. This change is dictated in part by the fact that the continued use of the "old" FORTRAN code would have resulted in a subroutine of undesirable length (say, when L = 15). To ascertain the possible effects of this change in procedure, Simulation III replicates Simulation I.

Finally, to investigate the consequences of misspecifying both the structural model and the error model, Simulation IV utilizes a one-factor design.[1] The mixing proportion is varied while the sample size and scale length are fixed. Table 1 provides an overview of each simulation.

The Generation of Random Uniform Deviates: Subroutine Super

For a statistical investigation, the value of a Monte Carlo simulation is ultimately and directly related to the computer's ability to generate a sequence of random numbers. At present, it is not possible to program the computer to generate a sequence of numbers that exhibits true randomness. However, algorithms do exist which will result in a reasonable approximation to the desired random sequence.

The generation of values in accordance with a known statistical law typically begins with an attempt to generate random uniform deviates, u_1, u_2, . . . , u_m, $0 < u_i < 1$. The algorithms most frequently used to generate these deviates usually combine

[1] As Deegan (1976) notes, misspecification of the structural model is said to occur when the proposed model incorrectly includes an irrelevant causal variable or incorrectly excludes a relevant causal variable.

TABLE 1. Overview of Simulations I through IV.

	Simulation			
Parameter	I	II	III	IV
λ Separation	1, 3	1	1, 3	Undefined
π Mixing Proportion	0.00, 0.07, 0.14 0.28, 0.49	0.00, 0.07, 0.14 0.28, 0.49, 1.00	0.00, 0.07, 0.14 0.28, 0.49, 1.00	0.00, 0.07, 0.14 0.28, 0.49, 1.00
N Sample Size	35	35, 150	35	150
L Scale Length	7	5, 7, 9	7	15
$\underline{X}$ Data Values	Uniform Distribution	Symmetrical Distribution	Uniform Distribution	Skewed Distribution

two simpler or basic generators. These basic generators are combined in order to compensate for or eliminate the known patterns observed in the basic generator.

Subroutine Super is a FORTRAN version of the random uniform generator available at UCLA's Health Sciences Computer Facility. It combines a full period mixed multiplicative congruential generator,

$$r_i = (r_{i-1}\ (2^{\theta} + \mu)\ ),$$

with a 32-bit shift register generator,

$$r'_* = \mathrm{XOR}(\ r'_{i-1},\ \mathrm{SHIFT}(\ r'_{i-1},\ -17)\ )$$

$$r'_i = \mathrm{XOR}(\ r'_*\ ,\ \mathrm{SHIFT}(\ r'_*\ ,\ 15)\ )\ ,$$

in order to generate uniform deviates,

$$u_i = r'' / 281474976710655\ ,$$

where

$$r'' = \mathrm{XOR}(\ r_i\ ,\ r'_i\ )\ .$$ [1]

[1] Both XOR and SHIFT are FORTRAN string bit manipulation operations. For a more complete definition, the reader is referred to Chambers (1977).

Note, r_0 is the seed for the mixed congruential generator and r_0' is the seed for the shift register generator. (A seed is a six to nine digit odd integer used to start the generator.) To enhance Super's performance, r_0 and r_0' are changed following the generation of every 2N x 100 deviates.

A FORTRAN listing of Super (as modified for use in this study) is contained within appendix I. Appendix II provides a listing of the "original" version of Super. In addition, appendix II provides an evaluation of Super's performance as a random uniform generator.

The Generation of Discrete Integers: Subroutine Transform

Given the availability of a random uniform generator, there are many ways of generating a discrete outcome variable, Y_i, such that with a probability of $(1 - \pi)$

$$\mu_{y|x} = \alpha + \beta X_i \quad \text{and} \quad \sigma^2_{y|x} = \sigma^2_y (1 - \rho^2),$$

and with a probability of π

$$\mu_{y|x} = (\alpha + \lambda) + \beta X_i \quad \text{and} \quad \sigma^2_{y|x} = \sigma^2_y (1 - \rho^2),$$

where $\mu_{y|x}$ is the conditional mean; $\sigma^2_{y|x}$ is the conditional variance; σ^2_y is the marginal variance; ρ is the product moment correlation coefficient; α and β are the regression weights; and λ is the degree of separation between the mixed regressions. The present study adopts two different approaches for the generation of such integers. Again, the second transformation is adopted as it results in a considerably shorter FORTRAN code.

In the first approach (hereafter referred to as Transformation I) the 2N x 1 vector of uniform deviates generated by Super is reconfigured as an N x 2 matrix with elements p_{ij} (i = 1, 2, ... , N; j = 1, 2). Each p_{i1} or first column element is then used to identify the sampling population. Specifically, if $p_{ij} > \pi$, Y_i is drawn from Population I where $\mu_{y|x} = \alpha + \beta X_i$; if $p_{i1} \leq \pi$, Y_i is drawn from Population II. Next, each p_{i2} or second column element is used to determine an integer value for Y_i. Following a suggestion offered by Newman and Odell (1971), Y_i is varied from 1 to L, where L is the scale length, and is assigned that integer value which satisfies the inequality

$$F\left\{\frac{(Y_i - 1) - \mu_{y|x}}{\sigma_{y|x}}\right\} < p_{i2} < F\left\{\frac{Y_i - \mu_{y|x}}{\sigma_{y|x}}\right\}$$

where F( ) is the Standard Normal cumulative distribution function. (Note, the mean and variance are as defined in paragraph one, this section.)

To be more specific, an interval of unit length is partitioned into L sub-intervals

$$[a_o = 0, a_1] \ , \quad (a_1, a_2] \quad , \ . \ . \ . \ . \ , \quad (a_{L-1}, a_L = 1]$$

for each of the L possible values of X, and an integer ranging in value from 1 to L is associated with each sub-interval. As the partition boundaries are determined by F( ), the length of a given sub-interval at a given value of X is proportional to the probability of observing a particular value of Y at a particular value of X. When p_{i2} falls in a given sub-interval, Y_i is assigned the integer value associated with that sub-interval.

In the second approach (hereafter referred to as Transformation II) the 2N x 1 vector of uniform deviates is again reconfigured as an N x 2 matrix with elements p_{ij} (i = 1, 2, . . . , N; j = 1, 2). Again, each p_{i1} or first column element is used to identify the sampling population, while each p_{i2} or second column element is used to determine an integer value of Y_i. Specifically,

$$p_{i1} > \pi, \quad Y_i = \mathrm{INT}(\alpha + \beta X_i + kZ_i)$$

if

$$p_{i1} \leq \pi, \quad Y_i = \mathrm{INT}(\alpha + \lambda + \beta X_i + kZ_i) \quad ,$$

where $\mathrm{INT}(\mu_{y|x} + kZ_i)$ is the largest integer less than $\mu_{y|x} + kZ_i$; $k = \sigma_{y|x}$; and Z_i is a standard normal deviate satisfying the equality, $p_{i2} = F(Z_i)k$. By defining p_{i2} in terms of the cumulative Normal distribution function, Y_i is ensured of being sampled from a

Normal population.

A FORTRAN listing of Transformation I (for $\alpha = 1.67$; $\lambda = 3$; $\beta = 0.333$; $\rho = 0.667$; $\sigma_y^2 = 1$; and $L = 7$) is contained within appendix I. Appendix III provides a FORTRAN listing of Transformation II. In addition, it provides test data on the performance of each transformation.

The Computation and Display of Regression Statistics: Subroutine Regress, Subroutine Sort, Subroutine Stem

Subroutine Regress calculates the Least Squares regression estimates using the following equations:

the intercept, $\hat{\alpha} = \frac{1}{N}(\Sigma\ Y_i - \hat{\beta}\ \Sigma\ X_i)$;

the regression coefficient, $\hat{\beta} = \Sigma(X_i - \overline{X})(Y_i - \overline{Y})/\Sigma(X_i - \overline{X})^2$;

the correlation coefficient, $\hat{\rho} = \Sigma(X_i - \overline{X})^2\ \Sigma(Y_i - \overline{Y})^2$;

the conditional variance, $\hat{\sigma}^2_{y|x} = (1 - \hat{\rho}^2)\ \Sigma\ (Y_i - \overline{Y})^2/(N - 1)$;

the variance of the regression coefficient, $\hat{\sigma}^2_{\beta} = \hat{\sigma}^2_{y|x}\ /\Sigma(X_i - \overline{X})^2$;

the t-statistic for the null hypothesis, H_o: $\beta = 0$.

All summations are from 1 to N. To accomplish these calculations, the N values of X are entered into the program through a FORTRAN DATA statement on Step (0).

Subroutine Stem is a modification of an algorithm published in McNeil (1977). It provides a graphical display (specifically, a Stem-and-Leaf display) of and summary statistics for each of the generated empirical sampling distributions. (A Stem-and-Leaf display provides a two-dimensional representation of a batch of numbers. It is a histogram which uses digits instead of the usual "x" marks to note which values occurred and how often they occurred.)[1]

Sort is an IMSL (International Mathematical and Statistical Library) called subroutine that arranges a vector of unsorted numbers in ascending order. This subroutine--or another like it--is required for the execution of Subroutine Stem. All three subroutines are contained within appendix I.

Program Compute

Program Compute combines the previously cited FORTRAN subroutines in the manner shown in Figure 1.

[1] The Stem-and-Leaf displays are extremely useful in representing the degree of spread, symmetry, and peakedness exhibited by a batch of numbers. Unfortunately, as they are difficult and costly to reproduce by typewriter (when N = 1000), they are not included in this paper.

Figure 1. A Flowchart of Program Compute.

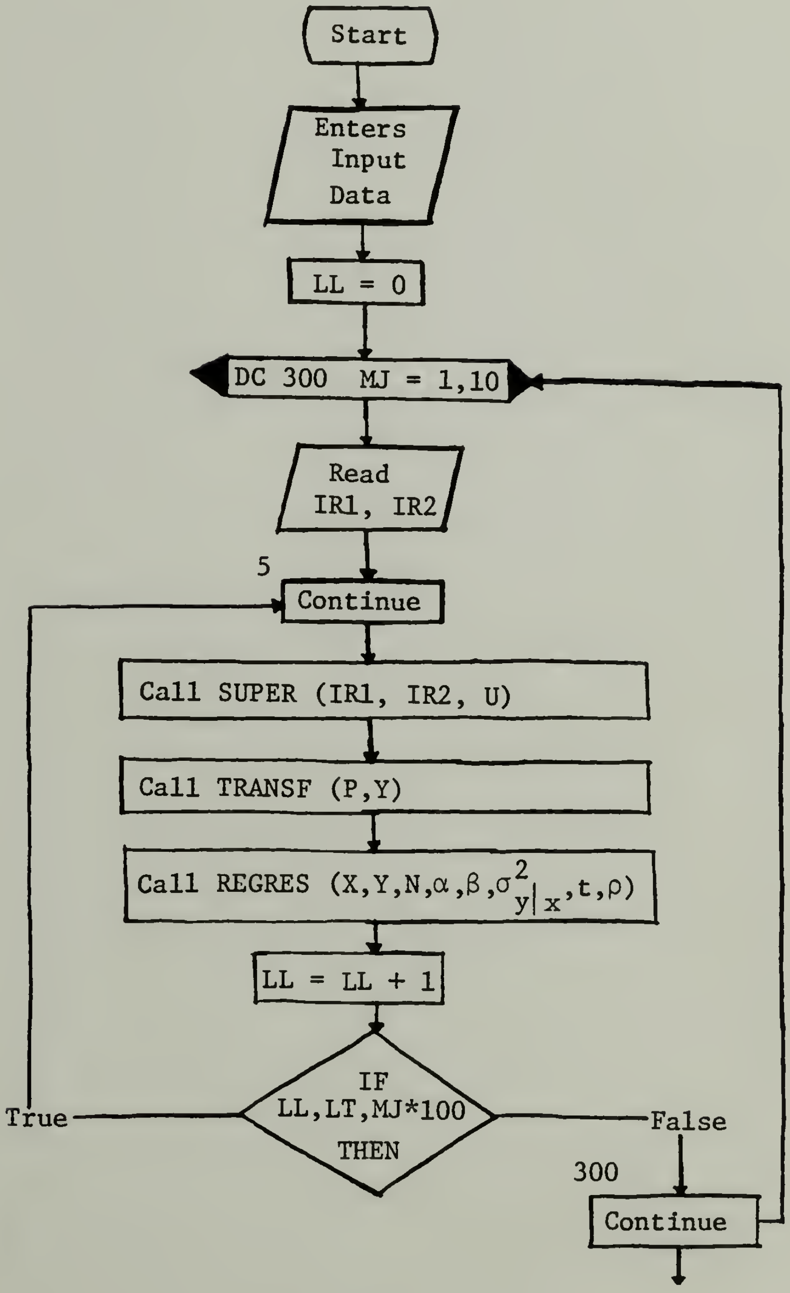
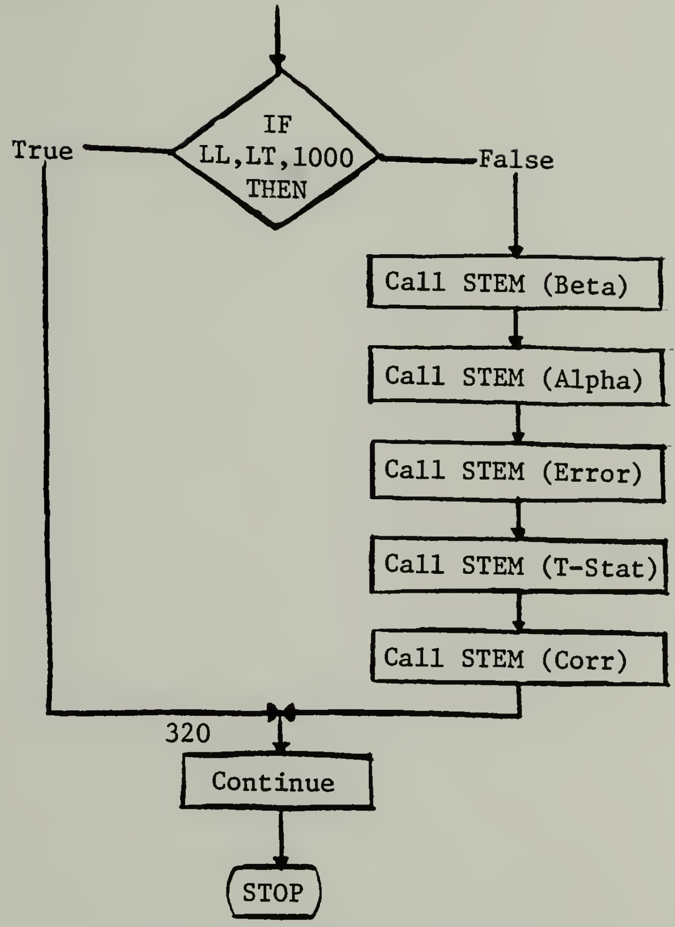


Figure 1 (continued).



(Note, IR1 is the seed for the multiplicative generator and IR2 is the seed for the shift register generator. In addition, note that Subroutine SORT is called by Subroutine STEM.)

CHAPTER III

RESULTS

Simulation I

In Simulation I, the mixing proportion, π, and the separation factor, λ, are varied for a fixed sample size (N = 35) and scale length (L = 7). Specifically, five values of Y, the outcome variable, are generated for each of the seven possible values of X, the explanatory variable. Then, with a probability of $(1 - \pi)$,

$$(1) \qquad Y_i = 1.67 + 0.333 X_i \quad ;$$

and with a probability of π,

$$(2) \qquad Y_i = (1.67 + \lambda) + 0.333 X_i \; .$$

Crossing the values for π (0.00, 0.07, 0.14, 0.28, 0.49) and λ (1, 3) results in a 2 x 4 factorial design with a single control group ($\pi = 0.00$).

A Note on Estimation and Notation

A researcher frequently undertakes a regression analysis assuming that the errors of measurement are normally distributed with common mean and constant variance. That is,

$$(3) \qquad \varepsilon_i \sim N(0, \sigma^2)$$

When this assumption is valid, the researcher may use the computationally simple Method of Ordinary Least Squares to derive estimators for the regression parameters. On the other hand, if the researcher knows that the errors are distributed as follows,

$$\varepsilon_i \sim (1 - \pi)\, N(0, \sigma^2) + \pi\, N(\lambda , \sigma^2) \quad , \tag{4}$$

he will generally use the Method of Maximum Likelihood to "fit" the postulated model to the data.

Because the data generated for this study are distributed in accordance with equation (4), rather extensive use is made of Maximum Likelihood estimators. Specifically, the formulas for the Maximum Likelihood estimators are used to pinpoint sources of bias (if any) in the Least Squares estimators derived under the misspecified error model (equation (3)).

As such, this study does not seek to compare Maximum Likelihood estimators with Least Squares estimators. (Indeed, when the errors are distributed as in equation (3), the Least Squares estimators are equivalent to the Maximum Likelihood estimators.) Instead, it seeks to compare estimators derived under two different error models, one which is true and the other which is false.

To minimize problems in notation, the present study uses Greek letters for the regression parameters (e.g., β), "circumflexed" Greek letters for the Maximum Likelihood estimators (e.g., $\hat{\beta}$), and lower case letters for the Least Squares estimators (e.g., b). Maximum Likelihood estimation under the mixed error model is described in appendix IV.

Estimation Under the Mixed Error Model

If the Monte Carlo samples are generated according to the design parameters for α, β, and λ, the Maximum likelihood estimators of the regression coefficient and residual variance are:

$$\hat{\beta} = \frac{\Sigma\,(Y_i - \hat{\alpha} - \lambda\, w_{2i})\, X_i}{\Sigma\, X_i^2}$$

(5) $$= \frac{\Sigma\,(X_i - \overline{X})(Y_i - \overline{Y})}{\Sigma\,(X_i - \overline{X})^2}$$

and

(6) $$\hat{\sigma}^2 = \frac{1}{N}\,\Sigma\,(Y_i - \hat{\alpha} - \hat{\beta}\, X_i)^2 - \frac{1}{N}\,\lambda^2\,\Sigma\, w_{2i} \quad ,$$

where

$$w_{2i} = \pi\, f_2(Y_i)/f_3(Y_i) \quad ,$$

for $f_3(Y_i)$, the weighted compound normal density function with components $f_1(y_i)$ and $f_2(Y_i)$,

and

(7) $$\hat{\alpha} = \overline{Y} - \hat{\beta}\,\overline{X} - \frac{1}{N}\,\lambda\,\sigma\, w_{2i} \quad .$$

From equation (5), it is clear that the (Maximum likelihood) estimator of the regression coefficient derived under the mixed error model is identical to the (Least Squares) estimator that is derived under the misspecified error model. Consequently, it follows that point estimation of the regression coefficient will not be affected by incorrectly asuming that all N observations come from the same statistical population.

However, it is clear from equation (6) that the (Maximum Likelihood) estimator of the residual variance derived under the mixed error model is smaller than the (Least Squares) estimator that is derived under the misspecified error model.[1] Then, from the known relationship between the residual variance and the variance of the regression coefficient, it follows that the (Maximum Likelihood) estimator of σ_β^2 that is derived under the mixed error model is smaller than the (Least Squares) estimator that is derived under the misspecified error model. Indeed, using a result from Elashoff (1972: eqn. 4.4), it can be argued that the (Least

[1] Equation (6) may be rewritten as follows:

$$\sigma_{ML}^2 = \sigma_{LS}^2 + c(c - \lambda) \qquad ,$$

where

$$c = \frac{1}{N} \lambda \Sigma w_{2i} \qquad .$$

Now, $\quad \sigma_{ML}^2 < \sigma_{LS}^2$

if $\quad c(c - \lambda) < 0 \qquad ,$

i.e., if $\quad c < \lambda \qquad ,$

or $\quad \frac{1}{N} \Sigma w_{2i} < 1 \qquad .$

As it is unlikely that $w_{2i} = 1$ for all N observations (where w_{2i} is the posterior probability that the "i"th observation comes from Population 2 given date vector $\underline{X}$), it is clear that

$$\sigma_{ML}^2 < \sigma_{LS}^2 \quad .$$

Squares) estimator of σ_β^2 is biased by a factor proportional to $\lambda^2\pi(1 - \pi)$. As a result, interval estimation of the regression coefficient and tests of significance will be affected by misspecification of the error model.

Estimation Under the Misspecified Error Model

Table 2 presents the basic results vis-à-vis the empirical sampling distribution of the Least Squares estimates of the regression coefficient. Note, the mean and variance of the control condition ($\pi = 0.00$) agree quite nicely with the intended values of 0.333 and 0.004205.[1] In addition, note that the means of the low separation condition ($\lambda = 1$) are less variable and closer to 0.333 than are the means of the high separation condition ($\lambda = 3$).

This finding appears to contradict the conclusion drawn from the mathematical analysis as it suggests that λ affects the point estimation of β. However, an analysis of variance (table 3) confirms the fact that effects due to π and/or λ cannot be used to explain the variation observed among the means of the empirical sampling distributions. Indeed, even when they are combined, these effects account for less than 2.5% of the observed variation (100% x (0.0041 + 0.0154 + 0.0053) = 2.48%).

[1] The value 0.004205 is obtained by substituting into the formula for estimated sample variance of b the values $\sigma_y^2 = 1$, $\sigma_x^2 = 4$, $\rho = 0.667$, and N = 35.

TABLE 2. The Mean (and Variance) of the Empirical Sampling Distribution of the Regression Coefficient by Degree of Separation and Mixing Proportion.

		Mixing Proportion				
		0.00	0.07	0.14	0.28	0.49
Separation	1	0.3334	0.3311	0.3337	0.3346	0.3338
		(0.4307)*	(0.4728)	(0.5849)	(0.5699)	(0.6355)
	3		0.3234	0.3206	0.3006	0.2918
			(0.7792)	(1.0710)	(1.5523)	(1.7416)

Notes:
*Multiply all variances by 10^{-2}.
Each distribution is based on 1000 replications.

TABLE 3. The Analysis of Variance of the Regression Coefficient Data.

Source of Variation	D.F.	Sum of Squares	ω^2
Mixing Proportion	3	0.3113	0.0041
Separation	1	1.1718	0.0154
Interaction	3	0.4029	0.0053
Residual	7992	73.9973	
Total	7999	75.8833	

Note: ω^2 is the sum of squares attributed to a given effect/the total sum of squares.

Although it not evident from the data presented thus far, the mean of the high separation condition would have been much closer to 0.333 had the Y values generated for the contaminating equation,

$$(8) \qquad Y_i = (\alpha + \lambda) + \beta X_i ,$$

not been subject to "ceiling" effects. To illustrate this point, Figure 2 provides an example of a scattergram of the contaminating equation used in the high separation condition ($\lambda = 3$).[1]

Note, the outcome variable is restricted to three values: 5, 6, and 7. While Y values less than 5 can occur, it is important to recall that the probability associated with such occurrences is quite small. In fact, when X_i equals 4, the probability of obtaining a Y value less than or equal to 5 is only 0.09. Then, as X_i increases in value, the value of this probability decreases. In effect, when X_i is greater than or equal to 4, the outcome variable is restricted to two values: 6 and 7.

By way of contrast, Figure 3 provides a scattergram of the generating equation used in the control condition ($\lambda = 0$). Note, the range of the outcome variable is much larger than that observed in Figure 2. Then, upon careful inspection of both figures, one should note that the number of conditional distributions which are

[1] For the cited values of α and β, the maximum value that λ can achieve is 3. Specifically, $\lambda_{max} = Y_{max} (\alpha + \beta X_{max})$.

Figure 2. Scattergram of 35 X_iY_i pairs - (λ=3)

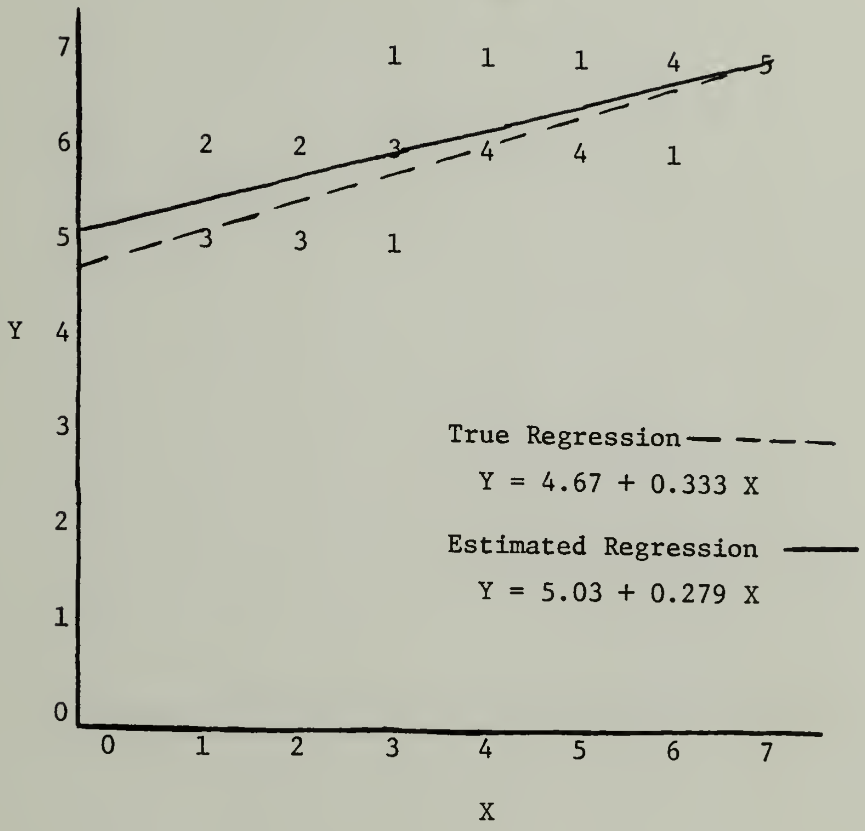
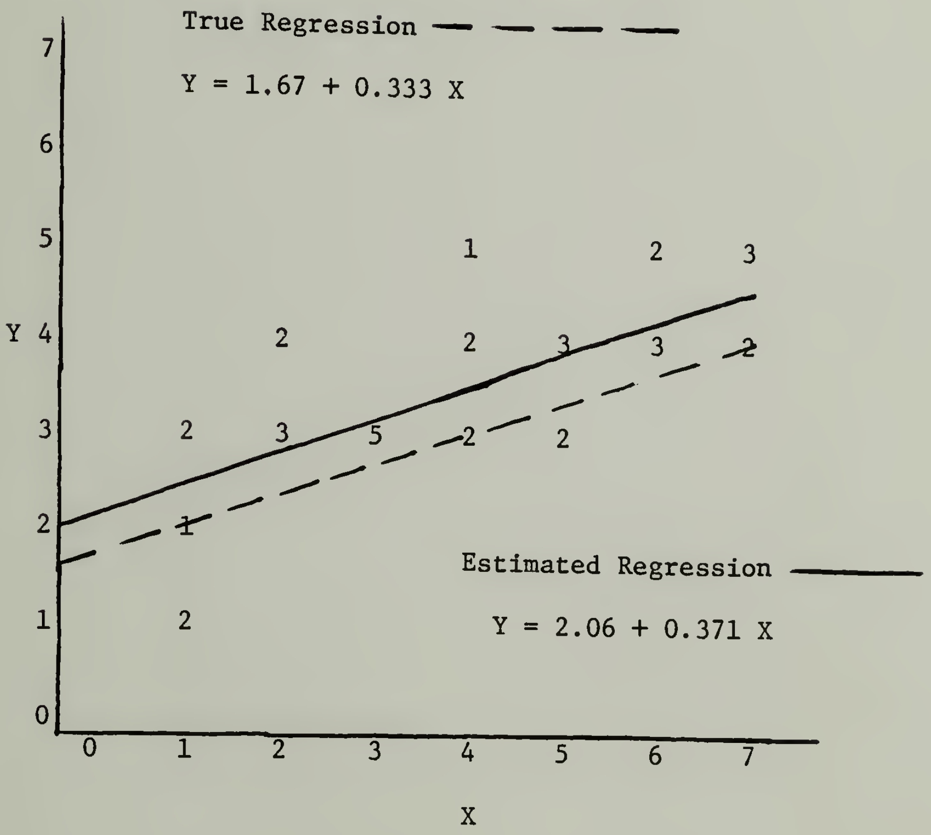


Figure 3. Scattergram of 35 X_iY_i Pairs - (λ=0)



truncated as well as the extent to which these distributions are truncated is less when $\lambda = 0$ than when $\lambda = 3$.

Whenever the range of the outcome variable is restricted, one or more of the conditional distributions will be truncated. As a result, the Least Squares estimators will be biased (see Hauseman and Wise, 1977; Takeshi, 1973). Table 4 presents the bias observed in the mean of the Least Squares estimates of α and β for specified values of λ. (Each mean is based on the 1000 estimates generated when π is either 0 or 1.) Note, when $\lambda = 0$, "floor" effects lead to a positive bias in the Least Squares estimator of α. As the value of λ increases, "floor" and "ceiling" effects combine to produce positive bias in the Least Squares estimator of α and negative bias in the Least Squares estimator of β. When λ attains its maximum value of 3, "ceiling" effects predominate, and the bias in the Least Squares estimators of α and β attain their maximum value.

In effect, equation (8) does not describe the regression which occurs with a probability of π. Instead, the contaminating regression is

$$(9) \qquad Y_i = (\alpha + \lambda) + (\beta + \theta) X_i + \varepsilon_i .$$

When "ceiling" and "floor" effects are non-existent, θ equals zero. (Note, the term corresponding to the "floor" effects observed in the Least Squares estimator of α has been absorbed by λ.) Estimates of θ may be obtained by subtracting β from $\overline{b}$, where $\overline{b}$ is the mean

TABLE 4. The Estimated Bias in the Least Squares Estimates by Degree of Separation.

	Degree of Separation			
	0.00	1.00	2.33	3.00
Bias in the Intercept	0.498	0.510	0.571	0.688
Observed Intercept	2.168	3.180	4.571	5.358
Bias in the Slope	0.000	-0.003	-0.029	-0.094
Observed Slope	0.000	0.330	0.304	0.239

of the 1000 sample regression coefficients generated when π equals zero or one. For example, when $\lambda = 3$,

$$
\begin{aligned}
(10) \qquad \hat{\theta} &= \bar{b} - \beta \\
&= 0.239 - 0.333 \\
&= -0.094 \quad .
\end{aligned}
$$

Returning to Table 2, one can easily verify that $\beta + \pi_j\theta$ - the expected value of the Least Squares estimator of β at the "j"th level of π - provides an accurate (2 decimal place) description of the means of the high separation condition. In fact, -0.094 is the Least Squares solution for θ when the high separation condition means are regressed on $\beta + \pi_j\theta$. Thus, one may conclude that the means of the high separation would have been much closer to β had it not been for the "ceiling" effects observed in the contaminating regression.

Before we consider how inference is affected by misspecification of the error model, we should note that the expected value of "b", the Least Squares estimator of β in the mixed sample, is a weighted sum. Specifically, if $\hat{b}^*$ and $\hat{b}^{**}$ are the Least Squares estimators of the regression coefficient for that portion of the sample from Populations I and II, respectively, then

$$
(11) \qquad E(\hat{b}) = (1 - \pi)\, E(\hat{b}^*) + \pi\, (\hat{b}^{**}).
$$

Moreover, the mean regression of the mixed sample is the weighted sum of the mean regressions for the individual populations.

that is,

$$(12) \qquad E\begin{Bmatrix} \hat{a} \\ \hat{b} \end{Bmatrix} = (1 - \pi)\ E\begin{Bmatrix} \hat{a}^* \\ \hat{b}^* \end{Bmatrix} + \pi\ E\begin{Bmatrix} \hat{a}^{**} \\ \hat{b}^{**} \end{Bmatrix} .[1]$$

To support this argument, Table 5 presents the observed and estimated mean values of the Least Squares estimates of the intercept and regression coefficient for the eight combinations of π and λ. The estimated mean value is the weighted sum of (a) the mean estimate observed for $\pi = 0$ and (b) the mean estimate observed for $\pi = 1$. While the estimated mean values are reasonably close to the observed mean values, the fit is better when $\lambda = 1$ than when $\lambda = 3$.

[1] An informal proof in terms of the mixed sample regression coefficient is as follows.

If, for Population I,

$$Y_i = \alpha + \beta X_i + \varepsilon_i \text{ with } E(\hat{b}^*) = \beta$$

and, for Population II,

$$Y_i = (\alpha + \lambda) + (\beta + \theta)X_i + \varepsilon_i \text{ with } E(\hat{b}^{**}) = \beta + \theta$$

then, for the mixed sample,

$$Y_i = (\alpha + \pi\lambda) + (\beta + \pi\theta)X_i + \varepsilon_i \text{ with } E(\hat{b}) = \beta + \pi\theta.$$

Since $\beta + \pi\theta = (1 - \pi)\ \beta + \pi\ (\beta + \theta)$, it follows that
$E(\hat{b}) = (1 - \pi)\ E(\hat{b}^*) + \pi\ E(\hat{b}^{**})$.

TABLE 5. Observed and Estimated Mean Values of the Intercept and Regression Coefficient by Degree of Separation and Mixing Proportion.

Separation	Mixing Proportion	Observed Mean	Estimated Mean	Difference^2
3	0.49	3.713	3.732	361*
	0.28	3.084	3.062	484
	0.14	2.620	2.615	25
	0.07	2.406	2.392	196
1	0.49	2.660	2.664	16
	0.28	2.455	2.452	9
	0.14	2.305	2.310	25
	0.07	2.247	2.240	49
3	0.49	0.292	0.287	25
	0.28	0.301	0.307	36
	0.14	0.321	0.320	1
	0.07	0.323	0.327	16
1	0.49	0.334	0.332	4
	0.28	0.335	0.333	4
	0.14	0.334	0.333	1
	0.07	0.331	0.333	4

Notes:

* Multiply column values by 10^{-6}.

When the Separation Factor is 3, the observed mixed regressions are:

$f_1(Y) = 2.168 + 0.334\ X$ and $f_2(Y) = 5.358 + 0.2393\ X$.

When the Separation Factor is 1, the observed mixed regressions are:

$f_1(Y) = 2.168 + 0.334\ X$ and $f_2(Y) = 3.180 + 0.3302\ X$.

Inference Under the Misspecified Error Model

It is clear from equation (6) that the Least Squares estimator of the residual variance is positively biased. Then, as the variance of the regression coefficient is proportional to the residual variance, it follows that the Least Squares estimator of the former will be positively biased.[1] Because of this bias, the actual distribution of the mixed sample t-statistic may be quite different from the t-distribution. As a result, the actual risk of accepting a false null hypothesis (i.e., the Type II error rate) may be quite different from some assumed nominal risk.

Table 6 presents the first four moments of the empirical sampling distribution of the t-statistic for the control condition ($\pi = 0.00$) and the eight experimental conditions. In addition, it presents an estimate of the Type II error rate associated with each condition. The latter is the proportion of computed t-statistics less than $t_{N-2,0.9875}$. (Note, the null hypothesis is H_o: $\beta = 0$. The alternative hypothesis is H_a: $\beta \neq 0$.)

Given table 6, it is clear that the value of $\bar{t}$, the mean of the computed t-statistics, decreases as the value of λ and/or π increases. When $\lambda = 1$, the positive bias in the Least Squares estimator of σ^2_β leads to a value to $\bar{t}$ which is slightly smaller

[1] One can verify this bias by regressing the variances of the sampling distributions of the Least Squares estimator of β (table 2) on $\sigma^2_\beta = \gamma Z_i$, where $Z_i = \lambda^2 \pi (1 - \pi)$. For both levels of λ, the fit accounts for approximately 98% of the observed variation.

TABLE 6. Moments of the Empirical Sampling Distribution of the t-statistic by Degree of Separation and Mixing Proportion.

Separation Factor	Mixing Proportion	Mean	Variance	Skewness	Kurtosis	Error Rate
3	0.49	2.214	1.1262	0.3061	0.3999	0.445
	0.28	2.467	1.2341	0.3328	0.6788	0.360
	0.14	3.138	1.3339	0.3436	0.0723	0.166
	0.07	3.764	1.5702	0.4574	0.3988	0.067
1	0.49	4.280	1.3823	0.4005	0.4876	0.020
	0.28	4.442	1.4093	0.4696	0.7682	0.011
	0.14	4.620	1.4774	0.3227	0.1251	0.008
	0.07	4.784	1.4648	0.4736	0.4178	0.004
	0.00	5.076	1.4651	0.4804	0.7408	0.002
		5.068	1.4171	0.3293	0.2371	0.003
		5.043	1.4380	0.3709	0.1931	0.001
		4.999	1.4290	0.6749	1.4225	0.001

When the Separation Factor is 3, the mixed regressions are:

$f_1(Y) = 2.168 + 0.333\ X$ and $f_2(Y) = 5.358 + 0.239\ X$.

When the Separation Factor is 1, the mixed regressions are:

$f_1(Y) = 2.168 + 0.333\ X$ and $f_2(Y) = 3.180 + 0.330\ X$.

than that obtained in the control condition. When $\lambda = 3$, the positive bias in the Least Squares estimator of σ_{β}^{2} combines with the negative bias in the Least Squares estimate of β and the value of $\bar{t}$ is reduced further. However, even when λ attains its maximum value and $\pi = 0.49$, the mean value for the t-statistic for H_o: $\rho = 0$ is the tabled criterion value, $t_{N-2,0.975}$.

Table 6 also suggests that an increase in λ or π leads to an increase in the probability of accepting a false null hypothesis. When $\lambda = 1$, the probability of accepting the null hypothesis is negligible. When $\lambda = 3$ and $\pi = 0.49$, the probability of accepting the null hypothesis is roughly 0.50.

Finally, as is shown in table 7, when the contaminating equation is

$$(13) \qquad Y_i = (\alpha + \lambda) - \beta X_i + \varepsilon_i,$$

$\bar{t}$ may be much less than $t_{N-2,0.075}$. Specifically, the expected value of the test statistic for H_o: $\beta = 0$ may be much less than the tabled criterion value when there is a modest degree of contamination ($\pi \geq 0.20$). For regressions with slopes of opposite sign, it is clear that the researcher stands an even chance of accepting a false null hypothesis when the value of π is much less than 0.50.

TABLE 7. Moments of the Empirical Sampling Distribution of the t-statistic by the Mixing Proportion.

Mixing Proportion	Mean	Variance	Skewness	Kurtosis	Error Rate
0.49	0.112	1.4262	-0.0509	0.3693	0.903
0.28	1.891	1.6372	0.3393	0.4693	0.578
0.14	3.231	1.7675	0.3892	0.7731	0.173
0.07	4.404	1.7822	0.5132	0.9522	0.056

Note: The mixed regressions are 2.168 + 0.333 X and 5.146 - 0.313 X.

Simulation II

In Simulation II the sample size, scale length, and mixing proportion are varied while the separation factor is held constant ($\lambda = 1$). Specifically, n_j ($\Sigma\, n_j = N$; $j = 1, 2, \ldots, L$) values of Y are generated for each of the L possible values of X. With a probability of $(1 - \pi)$,

(14) $$Y_i = \alpha + \beta X_i + \varepsilon_i,$$

and with a probability of π,

(15) $$Y_i = (\alpha + 1) + \beta X_i + \varepsilon_i.$$

Crossing N (35, 150), L (5, 7, 9) and π (0.00, 0.07, 0.14, 0.28, 1.00) results in a 2 x 3 x 5 factorial design. For L = 5, 7, and 9, the respective values for α are 1.00, 1.33, and 1.67.[1] For each combination of N and L, $\rho = 0.667$ and $\sigma_y^2 = \sigma_x^2$. (Hence, $\sigma_{y|x}^2 = \sigma_y^2\,(1 - 0.667^2) = 0.551\, \sigma_y^2$.)

The marginal distribution of X used in each N-L combination is shown in table 8. Note, the N values of X are symmetrically distributed with mean $\mu_x = ½(L + 1)$ and variance σ_x^2.

[1] The values for α are changed in order to ensure that $\beta = \rho$ at each combination of N and L.

TABLE 8. The Marginal Distribution of X by Scale Length and Sample Size.

Scale Length	Sample Size	Marginal Distribution								
		X=1	X=2	X=3	X=4	X=5	X=6	X=7	X=8	X=9
5	35	2	10	11	10	2				
	150	9	42	48	42	9				
9	35	1	3	8	11	8	3	1		
	150	5	14	33	46	33	14	5		
9	35	1	1	5	6	9	6	5	1	1
	150	4	6	17	28	40	28	17	6	4

Finally, the mean regressions observed for the six combinations of sample size and scale length are displayed in table 9. As noted earlier, each mean regression is the average of the 1000 regressions generated when $\pi = 0$ or $\pi = 1$.

As in Simulation I, the mean of the Least Squares estimates of α is consistently larger than the parameter's true value. Further, when there are marked "ceiling" effects, the mean of the estimates of β is much smaller than that parameter's true value. In effect, the Least Square estimator of α is positively biased because $\overline{Y}$ is a positively biased estimator of its population parameter. Now, $\overline{Y}$ is biased because (at least) one of the conditional distributions at $X = x_i$ is truncated. Then, as the Least Squares estimates of the intercept and regression coefficient are correlated, the latter must underestimate β as the bias in $\overline{Y}$ causes the former to overestimate α.

Estimation Under the Misspecified Error Model

Table 10 presents the means of the empirical sampling distributions of the regression and correlation coefficients. As in Simulation I, the mean of the generated mixed sample regression coefficients is a weighted average of the regression coefficients of the individual components (see equation (11)).

If there were no "floor" or "ceiling" effects, the following relationships would hold:

$$(17) \qquad E(\hat{b}) = E(\hat{b}^{*}) = E(\hat{b}^{**}).$$

TABLE 9. The Observed Regressions for Each Population by Scale Length and Sample Size.

Scale Length	Sample Size	Observed Regression Population I	 Population II
5	35	Y = 1.580 + 0.634 X	Y = 2.901 + 0.471 X
	150	Y = 1.588 + 0.630 X	Y = 2.897 + 0.472 X
7	35	Y = 1.903 + 0.645 X	Y = 3.359 + 0.535 X
	150	Y = 1.912 + 0.643 X	Y = 3.345 + 0.535 X
9	35	Y = 2.184 + 0.662 X	Y = 3.311 + 0.629 X
	150	Y = 2.203 + 0.659 X	Y = 3.355 + 0.619 X

Note: The intended value for the regression coefficient in each instance is 0.667. The intended values for the respective intercepts (α_1, α_2) are: (1.00, 2.00), (1.33, 2.33), and (1.67, 2.67).

TABLE 10. The Mean of the Empirical Sampling Distribution of the Regression and Correlation Coefficients by Scale Length, Mixing Proportion, and Sample Size.

		Regression Coefficient		Correlation Coefficient	
		Sample Size		Sample Size	
Scale Length	Mixing Proportion	n = 35	n = 150	n = 35	n = 150
5	1.00	0.471	0.472	0.608	0.602
	0.28	0.584	0.585	0.588	0.589
	0.14	0.601	0.610	0.607	0.614
	0.07	0.616	0.621	0.624	0.629
	0.00	0.634	0.630	0.652	0.646
7	1.00	0.535	0.535	0.605	0.611
	0.28	0.617	0.615	0.582	0.611
	0.14	0.630	0.630	0.597	0.607
	0.07	0.638	0.638	0.607	0.620
	0.00	0.645	0.643	0.624	0.630
9	1.00	0.629	0.619	0.658	0.641
	0.28	0.650	0.647	0.640	0.628
	0.14	0.653	0.652	0.645	0.636
	0.07	0.658	0.657	0.653	0.644
	0.00	0.662	0.659	0.662	0.651

However, the above data suggest quite strongly that both $\hat{\beta}^*$ and $\hat{\beta}^{**}$--the Least Squares estimators of the regression coefficient in Populations I and II, respectively--are biased for all six combinations of sample size and scale length. The estimator $\hat{\beta}^*$ is biased because the conditional distributions near X_{min} are truncated on the left. In contrast, the estimator $\hat{\beta}^{**}$ is biased because the conditional distributions near X_{max} are truncated on the right. As scale length increases, the bias observed in both estimators decreases in (absolute) value. However, this result is not attributable solely to the increase in scale length. It occurs in part because the values of α, β, $\sigma^2_{y|x}$, Y_{max}, and Y_{min} change with each change in scale length.

It is also clear from Table 10 that estimation of the correlation coefficient is not seriously affected by misspecification of the error model. (Although, it should be noted that the present form of contamination is quite mild.) Furthermore, while the mean of the sampled correlation coefficients is generally less than the value of the population parameter, it is apparent that the Least Squares estimator performs better when L = 9 than when L = 7 or 5.

Inference Under the Misspecified Error Model

Table 11 presents the moments of the empirical sampling distribution of the mixed sample t-statistic by scale length, sample size, and mixing proportion. As in Simulation I, the null hypo-

TABLE 11. Moments of the Sampling Distribution of the t-statistic by Scale Length, Sample Size, and Mixing Proportion.

L	N	π	Mean	Variance	Skewness	Kurtosis	Error Rate
5	150	1.00	9.252	1.1949	0.0562	-0.0759	0.000
		0.28	8.933	1.2249	0.2706	0.1428	0.000
		0.14	9.545	1.2757	0.1189	-0.0969	0.000
		0.07	9.919	1.3703	0.1580	0.1513	0.000
		0.00	10.379	1.3649	0.2116	-0.1715	0.000
	35	1.00	4.534	1.2358	0.3342	0.0605	0.005
		0.28	4.353	1.4789	0.2987	0.1498	0.020
		0.14	4.554	1.3592	0.3050	0.1992	0.006
		0.07	4.763	1.4182	0.2945	0.0736	0.004
		0.00	5.120	1.4704	0.1991	0.0934	0.002

TABLE 11 (continued)

L	N	π	Mean	Variance	Skewness	Kurtosis	Error Rate
7	150	1.00	9.478	1.5226	0.1572	0.2588	0.000
		0.28	9.027	1.1833	0.1077	-0.1358	0.000
		0.14	9.374	1.3342	-0.0094	-0.0450	0.000
		0.07	9.685	1.2414	0.2455	0.1701	0.000
		0.00	9.950	1.2921	0.2920	0.2228	0.000
	35	1.00	4.534	1.3947	0.1814	0.0841	0.010
		0.28	4.262	1.2453	0.2676	0.1586	0.015
		0.14	4.437	1.3561	0.2415	0.3130	0.016
		0.07	4.549	1.4049	0.5067	0.6947	0.005
		0.00	4.755	1.4141	0.2992	-0.0296	0.005

TABLE 11 (continued).

L	N	π	Mean	Variance	Skewness	Kurtosis	Error Rate
9	150	1.00	10.247	1.3424	0.1974	0.1473	0.000
		0.28	9.900	1.3104	0.1311	-0.0617	0.000
		0.14	10.120	1.3721	0.1276	-0.0536	0.000
		0.07	10.328	1.3010	0.1653	0.1211	0.000
		0.00	10.515	1.4900	0.1680	0.1471	0.000
	35	1.00	5.196	1.4284	0.4247	0.3508	0.001
		0.28	4.948	1.3368	0.3027	0.3239	0.003
		0.14	5.040	1.5343	0.3507	0.1446	0.005
		0.07	5.142	1.5087	0.3151	0.1446	0.003
		0.00	5.256	1.4205	0.2897	0.2781	0.002

thesis is H_o: $\beta = 0$, while the alternative hypothesis is H_a: $\beta \neq 0$. Given the above table, it is clear that the probability of accepting a false null hypothesis is virtually non-existent when $N = 150$ and extremely low when $N = 35$. Of course, different parameter values would lead to different error rates. In fact, careful examination of the expected value of the mixed sample t-statistic,

$$E(\ t\ ) \approx \frac{\beta}{\sqrt{\dfrac{(1 - \rho^2)\ \sigma_y^2 + \lambda^2\pi(1 - \pi)}{\Sigma\ (X_i - \overline{X})^2}}} \quad ,$$

suggests that the error rate would increase if β and ρ were allowed to reach their minimum values while λ, π, and σ_y^2 were allowed to reach their maximum values.[1]

[1] The formula for the mixed sample t-statistic was derived as follows.

$$E(\ t\ ) = E \left\{ \frac{\hat{\beta}}{\hat{\sigma}_\beta} \right\}$$

$$= E\ (\hat{\beta}) \cdot E(1/\hat{\sigma}_\beta)$$

$$\approx E\ (\hat{\beta}) \cdot (1/E(\hat{\sigma}_\beta)) \quad .$$

However,

$$\sigma_\beta^2 = \Sigma\ (X_i - \overline{X})^2 \cdot \sigma_\varepsilon^2 \ / \ [\ \Sigma\ (X_i - \overline{X})^2\ ]^{\,2} \quad .$$

Substituting

$$\sigma_\varepsilon^2 = \sigma_{y|x}^2 + \lambda^2\ \pi(1 - \pi) = \sigma_y^2\ (1 - \rho^2) + \lambda^2\pi(1 - \pi)$$

into the expression for σ_β^2, and that in turn into the expression for the expectation, one obtains--assuming $E(\hat{\beta}) = \beta$ - equation 18.

Simulation III

Using Transformation II to generate Y values, Simulation III replicates Simulation I. As before, the degree of separation and mixing proportion are varied while the sample size (N = 35) and scale length (L = 7) are fixed. Five values of Y, the outcome variable, are generated for each of the seven possible values of X. With a probability of $(1 - \pi)$,

$$Y_i = 1.67 + 0.333 X_i + \varepsilon_i \quad , \tag{19}$$

and with a probability of π,

$$Y_i = (1.67 + \lambda) + 0.333 X_i + \varepsilon_i . \tag{20}$$

Crossing π (0.00, 0.07, 0.14, 0.28, 0.49) and λ (1, 3) results in a 2 x 4 factorial design with a single control group ($\pi = 0.00$).

Panel (a) of table 12 presents the results of Simulation III in terms of the empirical sampling distribution of the regression coefficient. For ready comparison, panel (b) displays the corresponding results from Simulation I. From the above table, it is clear that Transformation I generates a conditional variance which is approximately 1.5 times larger than that generated by Transformation II. Of course, the two transformations could be programmed to produce identical results. To achieve this objective, one need only multiply the random error term generated by Transformation II by a constant. This constant, if chosen carefully, ensures that both transformations generate the same conditional variance.

TABLE 12. The Mean (and Variance) of the Regression Coefficient's Empirical Sampling Distribution by Degree of Separation and Mixing Proportion.

a. Simulation III

		Mixing Proportion				
		0.000	0.07	0.14	0.28	0.49
Separation	1	0.335 (0.288)*	0.333 (0.336)	0.333 (0.351)	0.333 (0.434)	0.334 (0.457)
	3		0.326 (0.660)	0.321 (0.952)	0.319 (1.435)	0.301 (1.626)

b. Simulation I

		Mixing Proportion				
		0.00	0.07	0.14	0.28	0.49
Separation	1	0.333 (0.431)*	0.331 (0.473)	0.334 (0.585)	0.335 (0.570)	0.334 (0.636)
	3		0.323 (0.779)	0.321 (1.017)	0.301 (1.552)	0.292 (1.742)

*Note: Multiply all variances by 10^{-2}.

Simulation IV

Simulation IV differs from the previous simulations in that the proposed causal model

(21) $$Y_i = \alpha + \beta X_{1i} + \varepsilon_i$$

is structurally misspecified. More specifically, it omits a relevant causal variable, X_2. Given a fixed sample size (N = 150) and scale length (L = 15), the regression which occurs with probability (1 = π) is

(22) $$Y_i = 2.67 + 0.333 X_{1i} - 0.166 X_{2i} + \varepsilon_i \quad ,$$

and the regression which occurs with probability π is

(23) $$Y_i = 1.67 + 0.667 X_{1i} + \varepsilon_i \quad .$$

The marginal frequencies for X_1 are: 0, 0, 8, 6, 10, 16, 22, 26, 22, 16, 10, 4, 4, 4, and 2. The marginal frequencies for X_2 are: 0, 0, 9, 5, 11, 15, 20, 26, 24, 16, 10, 6, 5, 2, and 1. The Y values are generated by Transformation II.

Now, the expected value of the mixed sample regression coefficient is again a weighted average of the individual component regression coefficients (see equation (11)). When $\pi = 1$, the expected value of the Least Squares estimator of the regression coefficient for X_1 is 0.667. However, when $\pi = 0$, the expected value of the regression coefficient for X_1 is not 0.333. Instead, it is 0.167.[1]

Table 13 presents the means (and variances) of the empirical sampling distributions of the Least Squares estimators of the parameters of the proposed causal model. Note, the above results are consistent with the analytic argument. Within reasonable error, $E(\hat{b}) = (1 - \pi)(0.167) + \pi(0.667)$ and $E(\hat{a}) = (1 - \pi)(2.67) + \pi(1.67)$.

[1] The Least Squares estimator of β_1 in the proposed causal model is

$$\hat{b}_1 = \Sigma\, (Y_i - \hat{a})\, X_{1i}/\Sigma\, X_{1i}^2 \quad .$$

Upon substitution of the true model of Y_i and taking expectations, one obtains

$$\begin{aligned} E(\hat{b}_1) &= \beta_1 - \beta \left\{ \frac{\Sigma\, X_{2i}X_{1i}}{\Sigma\, X_{1i}^2} \right\} \\ &= 0.333 - (0.166)(2506/2508) \\ &= 0.167 \quad . \end{aligned}$$

TABLE 13. The Mean (and Variance) of the Empirical Sampling Distribution of the Intercept and Regression Coefficient by Mixing Proportion.

	Mixing Proportion				
	0.00	0.07	0.14	0.28	1.00
Intercept	2.929 (0.033)	2.853 (0.078)	2.791 (0.121)	2.681 (0.182)	1.966 (0.030)
Regression Coefficient	0.173 (0.045)*	0.208 (0.145)*	0.243 (0.227)*	0.308 (0.360)*	0.667 (0.041)*

* Multiply variance by 10^{-2}

CHAPTER IV

DISCUSSION

The Effect of Error Model Misspecification on Parameter Estimation

For each simulation the correct error model may be written as follows:

(24) $$\varepsilon_i \sim (1 - \pi)\ N(0, \sigma^2) + \pi\ N(\mu(X), \sigma^2)\ ,$$

where

(25) $$\mu(X) = \lambda + \theta\ X_{1i} + \gamma\ X_{2i}\ .$$

This error model would be appropriate if, for example, the sample contained a mixture of two regressions:

(26) $$f_1(Y) = \alpha + \beta\ X_{1i} + \varepsilon_i\ ,$$

and

(27) $$f_2(Y) = (\alpha + \lambda) + (\beta + \theta)\ X_{1i} + \gamma X_{2i} + \varepsilon_i\ .$$

Now, analysis suggests that the Least Squares regression of Y on X_1 will lead to positively biased estimates. Specifically,

(28) $$E(\hat{a}) = \alpha + \pi\lambda\ ,$$

$$(29) \qquad E(\hat{b}) = \beta + \pi \left\{ \theta + \gamma \frac{\Sigma X_{1i}X_{2i}}{\Sigma X_{1i}^2} \right\}$$

and

$$(30) \qquad E(\hat{\sigma}^2) = \Sigma\ (1 - v_i)\ \sigma_{\varepsilon}^2 \quad ,$$

where

$$(31) \qquad v_i = \{\underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\}_{ii} \quad .$$

If each X_{hi} $(h = 1, 2)$ is expressed in mean deviation form (i.e., $x_{hi} = X_{hi} - X_h$) ,

$$(32) \qquad V(\hat{a}) = \frac{\sigma}{N} + \pi(1 - \pi)\ \Sigma\ (\lambda + \theta x_{1i} + \gamma x_{2i})^2/N^2 \quad ,$$

and

$$(33) \qquad V(\hat{b}) = \frac{\sigma}{\Sigma x_{1i}^2} + \pi(1 - \pi)\ \Sigma\ (x_{1i}^2)(\lambda + \theta x_{1i} + \gamma x_{2i})^2/(\Sigma\ x_{1i}^2)^2 \quad .$$

In Simulations I and II the mixed regressions are parallel to one another. (For the moment, let's ignore the fact that non-parallel regressions are actually obtained in these simulations). Furthermore, as tentative causal models, both regressions are correctly specified. In terms of equation (27) this means that both θ and γ are zero. In terms of equations (28) - (33), zero values for θ and γ imply the following. First, the Least Squares estimator of β will be unbiased. Consequently, point estimation of β will not be affected in this instance by misspecification of the error model. Second, the

Least Squares estimator of the residual variance will be positively biased. Consequently, interval estimation of β will be adversely affected. However, it is clear upon substituting $\theta = \gamma = 0$ into equation (33) that interval estimation of β is not seriously affected when N is large (or N is of moderate size and π is near either one of its limiting values).

However, as noted earlier, the mixed regressions are not parallel in either Simulation I or II. Specifically, while γ is zero, θ is not. In terms of equations (28) - (33), a non-zero value of θ implies the following. First, the Least Squares estimator of β will be biased. From equation (29), it is clear that the bias is given by $\pi\theta$. Second, the Least Squares estimator of the residual variance will again be positively biased, and interval estimation of β will again be adversely affected. Given equations (29) and (33), however, it is clear that the non-parallelism resulting from θ not being equal to zero will not seriously affect point or interval estimation of β. In fact, if θ is small, the point and interval estimates which are obtained will not be noticeably different from those obtained when the mixed regressions are parallel.

In Simulation IV the mixed regressions are by design not parallel. In addition, the causal model suggested by regressing Y on X_1 is misspecified. In terms of equation (27) this means that γ has a non-zero value while θ equals zero. In terms of equations (28) - (33), the non-zero value for γ implies the following. First, the

Least Squares estimator of β will be biased (providing that X_1 and X_2 are not independently distributed). From equation (29), it is clear that the bias is given by $\pi\lambda$ $(\Sigma\ x_{1i}x_{2i}/\Sigma\ x_{1i}^2)$. Second, the Least Squares estimator of the residual variance will again be positively biased, and interval estimation of β will again be adversely affected. However, from equations (29) and (33), it is clear that one could choose values for γ, for the correlation between X_1 and X_2, and for the N values of X_1 and X_2 and obtain point and interval estimates of the regression parameters as biased as or less biased than the estimates obtained when the mixed regressions are correctly specified, but not parallel.

In considering (a) parallel mixed regressions; (b) non-parallel, but correctly specified mixed regressions; and (c) non-parallel, misspecified mixed regressions, this study has sought to examine situations characteristic of most social science research. Equations (28) through (33) summarize the results obtained in the various simulations. Taken collectively, the analytic arguments and Monte Carlo results suggest that point estimation will be least affected when the mixed regressions are either parallel or non-parallel, but correctly specified. When the mixed regressions are non-parallel and misspecified, both point and interval estimation can be seriously affected.

The Effect of Error Misspecification on Statistical Inference

Because of the potential bias in the Least Squares estimator of the regression coefficient and the inevitable bias in the Least Squares estimator of the residual variance, the researcher runs an increased risk of accepting a false null hypothesis when he incorrectly assumes that the data come from the same statistical population. As observed in Simulations I and II, the mean of the computed mixed sample t-statistics is generally greater than the value of the central t distribution which is most frequently selected as a criterion for rejecting the null hypothesis. As a result, the observed Type II error rate is generally low (i.e., less than 10% when the mixing proportion is less than or equal to 0.10).

This finding may be of marginal utility as the error rates observed in Simulations I and II are determined in large part by the values assigned to the secondary parameters: N, σ_y^2/σ_x^2 , ρ_{yx}, and β. Specifically, the error rate observed at a given level of π would have been larger had N, ρ_{yx}, and β each been assigned a smaller value and σ_y^2/σ_x^2 been assigned a larger value. Using the results of Simulation II, it is easy to verify that an increase in the value of N leads to a decrease in the error rate. Then, using the results of Simulations I and II, it is easy to verify that an increase in the value of β and σ_y^2/σ_x^2 leads to a decrease in the error rate.

While the actual value of the error rate depends on such secondary parameters, certain conclusions (based on the expression for $\bar{t}$ and the results of Simulations II and III) are warranted. First, a researcher is least likely to erroneously conclude that $\beta = 0$ when π is low (say, less than 0.10). Second, the likelihood of making an incorrect inference is not significantly increased if the mixed regressions are correctly specified, parallel, and poorly separated. (If the mixed regressions are correctly specified and parallel, and yet quite distant from one another, there can be an appreciable increase in the error rate.) Third, although the mixed regressions may be correctly specified, the likelihood of making an incorrect inference will increase dramatically providing that the two regressions exhibit slopes of opposite sign. (If the mixed regressions have opposing slopes, the Least Squares estimator of the regression coefficient will exhibit a negative bias while the Least Squares estimator of the residual variance will exhibit a positive bias.) Fourth, and finally, a researcher is most likely to make an incorrect inference when he compounds misspecification of the error model with misspecification of the causal model. (Again, the Least Squares estimator of the regression coefficient will exhibit a negative bias while the Least Squares estimator of the residual variance will exhibit a positive bias.)

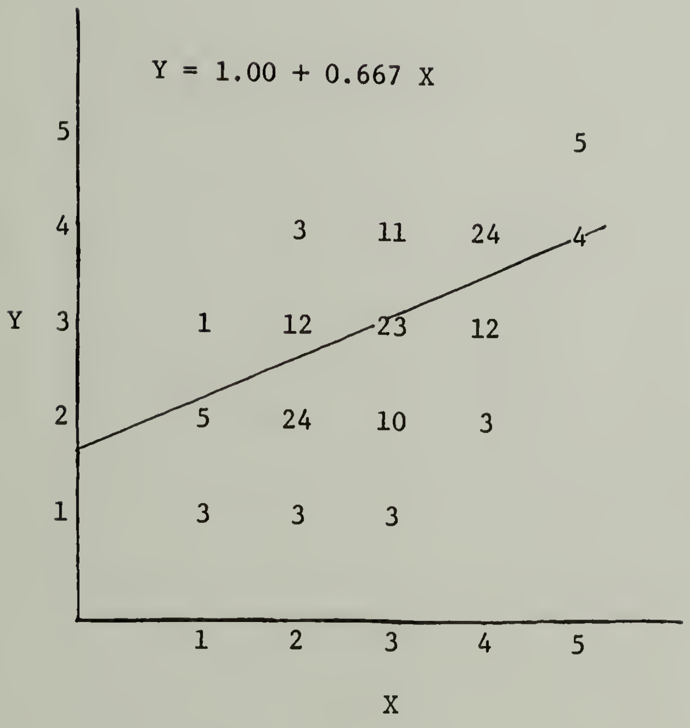
Scale Length and the Problem of Detecting Outliers

Error model misspecification would be of little consequence if outliers were easily detected. (Upon detection, one could in principle perform the necessary separate regressions.) However, outliers are seldom detected.

Short, ordinal scales complicate the process of detection because they restrict a variable's range. As a result, a researcher is less likely to observe any of the physical characteristics which mark mixed distributions (e.g., bimodality). Further, as short scales minimize the potential separation between the means of heterogeneous populations, a researcher is less likely to observe a Y_i that can be proven to be an outlier (e.g., shown to be more than two standard deviations from the center of the remaining observations).

Figure 4 illustrates these problems. It presents a scattergram of 150 X_iY_i pairs. As in Simulation II, $L = 5$; $\beta = \rho = 0.667$; and $\sigma_y^2 = 1$. Upon inspection, it is clear that an outlier will exert maximum influence on the Least Squares fit when $X = 1$ or $X = 5$. As the 150 pairs are presently configured, $\hat{\beta} = 0.667$. If a Y value at $X = 1$ is replaced by a 5, $\hat{\beta}$ becomes 0.615. Now, this slight reduction in value is not unexpected as π (1/150) is quite small. However--and, this is the important point--the "replaced" Y value would probably have to fall at a much greater distance from the remaining observations at $X = 1$ before the "eye" of the typically

Figure 4. Scattergram for 150 X-Y Pairs



untrained researcher would even begin to suspect its true nature. Then, visual detection of an outlier when X = 2, 3, or 4 is most unlikely even for the trained "eye."

Finally, if $Y_1 = 5$ were an outlier that dominated the Least Squares fit, $\hat{b}_{(1)}$--the regression weight calculated, omitting $Y_1 = 5$--would be quite different from $\hat{b}$. In this instance, $\hat{b}_{(1)} = 0.658$. Again, what is important is <u>not</u> that the Least Squares estimate of the regression weight is little changed by the omission of the suspect point. Rather, it is that there is little likelihood that the routine application of computerized statistical detection procedures will reveal outliers when variables are measured on short, ordinal scales.

Summary

The problems that misspecification of the error model pose for Least Squares estimation and its associated inference procedures are tolerable providing that the contaminating fraction is small and the hypothesized causal model is correctly specified. These problems become more severe if the hypothesized causal model is also misspecified. Finally, the inferential problems will be exacerbated further if variables are measured on short scales.

CHAPTER V

STUDY LIMITATIONS AND IMPLICATIONS

The present research investigates the effect which misspecification of the error model may have on the initial statement and subsequent refinement of causal theory. Now, implicit in the above are two study limitations. First, the present study only investigates the consequences of misspecifying the error model for the simplest mathematical representation of a causal theory: the single-stage, just-identified, fully recursive, linear model (see Land, 1969; Duncan, 1975). Second, the present study restricts its consideration of error model misspecification to the situation where the researcher fails to detect that the errors are described best as a mixture of two normal distributions. Other mixtures or non-normal distributions are not considered (e.g., Pollock, 1978; Hasseblad, 1969).

For an experimental discipline such as social psychology, the first limitation would not appear to be that great of a shortcoming. While it provides a questionable description of most social phenomena, the single-stage, linear recursive model nonetheless reflects quite well the level of complexity currently found in many social psychological theories.

The second limitation is potentially more serious. The appropriateness of using error models such as

$$\varepsilon_i \sim (1 - \pi)\ N(0,\sigma^2) + \pi\ N(\mu(x),\ \sigma^2)$$

to describe the kinds of non-normal error distributions observed in real data is already suspect (see Stigler, 1977).

Fortunately, the present study deals with issues which transcend the concerns one might have about the adequacy of the simulated data. In its less important role, the occurrence of (undetected) mixed regressions may simply reflect incomplete theorizing, faulty data screening procedures, or both. For example, upon noting that male and female respondents exhibit similar sample means and variances on all measured variables, a researcher may decide to "pool" the two samples and use the combined data to develop a causal model. Now, if the variance-covariance matrix for the male respondents is significantly different from that of the female respondents, the researcher would have erred in treating the two groups as if they came from the same population (see Sprecht and Warren, 1976). In this example, the resulting mixture of regressions could have been avoided had the researcher exercised more care when the data were screened. In other situations, however, the (undetected) mixture may be unavoidable. That is, the stated theory may be so incomplete that the researcher fails to obtain data on several key variables. As a result, the variance-covariance matrices based on the measured variables may indicate that the two groups come from the same population.

In its more important role, the occurrence of (undetected) mixed regressions suggests that there may be a limit to the extent to which a researcher can achieve an accurate description of any social process. All social processes are subject to change. Consequently, data that describe social processes may be a mixture of two or more statistical populations. If the mixture remains undetected, the researcher may never be in a position to provide an accurate description of the social process.

The present study addresses both of these issues. It documents the conditions for which improper or inadvertent "pooling" will seriously affect point and interval estimation. Moreover, it identifies some of the factors which contribute to the development of incomplete or erroneous causal theory. In accomplishing the latter, the present study seeks to redirect attention to the limitations inherent in ordinal measurement.

BIBLIOGRAPHY

Andrews, D., Bickel, P., Hampel, F., Huber, P., Rogers, W., and Tukey, J. Robust Estimation of Location: Survey and Advances, Princeton: Princeton University Press, 1972.

Chambers, J., Computational Methods for Data Analysis, New York: John Wiley, 1977.

Deegan, J., The consequences of model misspecification in regression analysis, Multivariate Behavioral Research, 1976, 11, 137-248.

Duncan, O., Introduction to Structural Equation Models, New York: Academic Press, 1975.

Elashoff, J., A model for quadratic outliers in linear regression, Journal of the American Statistical Association, 1972, 67, 478-485.

Harter, H., The method of least squares and some alternatives, International Statistical Review, 1975, 43, 269-278.

Hasselblad, V., Estimation of finite mixtures of distributions from the exponential family, Journal of the American Statistical Association, 1969, 64, 1459-1471.

Hastings, C., Approximations for Digital Computers, Princeton: Princeton University Press, 1955.

Hauseman, J., and Wise, D., Social experimentation, truncated distributions, and efficient estimation, Econometrica, 1977, 45, 919-938.

Hosmer, D., Maximum likelihood estimates of the parameters of a mixture of two regression lines, Communications in Statistics, 1974, 3, 995-1006.

Kiefer, N., Discrete parameter variation: Efficient estimation of a switching regression model, Econometrica, 1978, 46, 427-434.

Land, K., Principles of path analysis, In E. Borgatta (Ed.), Sociological Methodology 1969, San Francisco, CA: Jossey-Bass, 1969.

Maclaren, D., and Marsaglia, G., Uniform random number generators, Journal of the Association for Computing Machinery, 1965, 12, 83-89.

McNeil, D., Interactive Data Analysis, New York: John Wiley, 1977.

Mosteller, F., and Tukey, J., Data Analysis and Regression: A Second Course in Statistics, Reading, MA: Addison-Wesley, 1977.

Newman, T., and Odell, P., Generation of Random Variates, Griffin's Statistical Monographs, 1971.

Pollock, K., Inference robustness versus criterion robustness: An example, American Statistician, 1978, 32, 133-136.

Quandt, R., A new approach to estimating switching regressions, Journal of the American Statistical Association, 1972, 67, 306-311.

Quandt, R., and Ramsey, J., Estimating mixtures of normal distributions and switching regressions, Journal of the American Statistical Association, 1978, 73, 730-738.

Singleton, R., Algorithm 347: Sort, Communications of the Association for Computing Machinery, 1969, 12, 730-738.

Sprecht, D., and Warren, R., Comparing causal models, In D. Heise (Ed.), Sociological Methodology 1976, San Francisco, CA: Jossey-Bass, Inc., 1976, 46-82.

Stigler, S., Do robust estimators work with real data?, Annals of Statistics, 1977, 5, 1055-1098.

Takeshi, A., Regression analysis when the dependent variable is truncated normal, Econometrica, 1973, 41, 997-1016.

Tukey, J., A survey of sampling from contaminated distributions, In O. Olkin (Ed.), Contributions to Probability and Statistics, Stanford, CA: Stanford University Press, 1960, 448-485.

Wainer, H., and Thissen, D., Three steps toward robust regression, Psychometrica, 1976, 41, 9-34.

Witte, A., and Schmidt, P., An analysis of recidivism, using the truncated lognormal distribution, Applied Statistics, 1977, 26, 302-311.

APPENDIX I

PROGRAM COMPUTE, A FORTRAN LISTING

Program Compute is written in standard FORTRAN and runs on the CDC CYBER 175 computer under a NOS operating system at the University of Massachusetts (Amherst). It consists of a main program and five subroutines: SUPER, TRANS, REGRES, STEM, and SORT. For convenience, the call statements for SUPER, TRANS, and REGRES have been eliminated.

Subroutine SUPER is the segment of code which appears between FORTRAN statements 5 and 6. Again, it is a modified version of the random uniform generator available at UCLA's Health Sciences Computer Facility.

Subroutine TRANSF is the segment of code which appears between FORTRAN statements 10 and 225. Using the 2N random deviates generated by SUPER and Transformation I, it generates a mixed sample of N observations. As written in this appendix, the two conditional means are 1.67 + 0.333 X_i and 4.67 + 0.333 X_i; the common variance-covariance matrix is

$$\left\{\begin{array}{cc} 1.00 & 1.33 \\ 1.33 & 4.00 \end{array}\right\}$$

Subroutine REGRES is the segment of code which appears between FORTRAN statements 230 and 300. It takes the N observations generated by TRANSF and the N values of $\underline{X}$ entered earlier through a DATA statement and computes the regression estimates ($\hat{a}$, $\hat{b}$, $\hat{\sigma}^2$) and t-statistic for H_o: $\beta = 0$. It also contains the counter (LL) which ensures that 1000 samples are generated for every 10 seed pairs (IR1, IR2).

Subroutine STEM is a modification of an algorithm published in McNeil (1977). It provides the numerical summaries of the empirical sampling distributions of the regression estimates and t-statistic. In addition, it provides a graphical display (specifically, a Stem-and-Leaf display). It requires four inputs: (a) B, the vector of the Least Squares estimates; (b) N, the number of estimates; (c) Theta, the true value of the regression parameter or t-statistic; and (d) Print, a parameter which controls labeling. Its internal parameters include Iwidth, Atom, and Scale. Iwidth controls the number of characters printed on a single line. Atom prevents the impossible division by zero. Finally, Scale controls the depth of the display.

Subroutine SORT is a called IMSL (International Mathematical and Statistical Libraries) subroutine. The version of SORT printed in this appendix is Singleton's (1969) algorithm 347, the source cited for the IMSL subroutine. Its inputs include the unordered vector of Least Squares estimates (or t-statistics), and the numbers II = 1 and JJ = 1000.

```
      Program COMPUTE (Input, Output, Tape 5=Input, Tape 6=Output)
      Common Y(35)/Block/X(35)
      Dimension Beta(1000), Alpha(1000), Error(1000),
      Dimension T Stat(1000), Corr(1000)
      Dimension U(70), W(35,2)
      Equivalence(W(1,1), U(1)
C
C     This segment of code enters the necessary input data.
C
      DATA (X(k), k=1,35)/5*1.,5*2.,5*3.,5*4.,5*5.,5*6.,5*7./
      DATA PI/0.000/
      N = 35
      NN = 70
C
C     This segment of code reads in the seeds for the mixed
C     congruential generator and the shift register generator.
C
C
    1 Format(I6,I6)
      LL = 0
      DO 300 MJ = 1,10
      Read(5,1) IR1,IR2
C
C
      IM = IR1
      IT = IR2
C
C
C     This segment of code generates a vector of uniform
C     deviates.
C
    5 Continue
      M1 = 65539
      M2 = 4101
      M3 = 261
      DO 6 I=1,NN
         IM = M3*IM
         L = M1
         IF(IM.LT.0) L = M2
         IM = L*IM
         IF(IM.LT.0)IM = IM + 576460752303423487 + 1
         IB = IT
         IT = SHIFT(IT, -17)
         IB = XOR(IB,IT)
         IB = SHIFT(IB,15)
         IC = XOR(IB,IT)
         IT = IC
         IR = XOR(IM,IC)
```

```
               U(I) = IR
               U(I) = U(I)/281474976710655
               U(I) = ABS(U(I))
    6 Continue
   10 Continue
C
C     This segment of code transforms the uniform deviates into
C     discrete Y values.  The mixed regressions are Y = 1.67 +
C     0.333 X and Y = 4.67 + 0.333 X.
C
C     Computes Y values for X equals 1.0
C
      DO 45 k=1,5
         IF(W(k,1).GT.PI)15,30
C        Population One
   15    (Y(k) = 1.
         DO 25 MM = 1,1
            IF(W(k,2).GT.0.089) Y(k) = Y(k) + 1.
            IF(W(k,2).GT.0.499) Y(k) = Y(k) + 1.
            IF(W(k,2).GT.0.910) Y(k) = Y(k) + 1.
            IF(W(k,2).GT.0.996) Y(k) = Y(k) + 1.
   25    Continue
         GO TO 45
C        Population Two
   30    Y(k) = 3.
         DO 40 MM = 1,1
            IF(W(k,2).GT.0.004) Y(k) = Y(k) + 1.
            IF(W(k,2).GT.0.089) Y(k) = Y(k) + 1.
            IF(W(k,2).GT.0.499) Y(k) = Y(k) + 1.
            IF(W(k,2).GT.0.910) Y(k) = Y(k) + 1.
   40    Continue
   45 Continue
C
C     Computes Y values for X equals 2.0
C
      DO 75 k=6,10
         IF(W(k,1).GT.PI)55,65
C        Population One
   55    Y(k) = 1.
         DO 60 MM = 1,1
            IF(W(k,2).GT.0.037) Y(k) = Y(k) + 1.
            IF(W(k,2).GT.0.326) Y(k) = Y(k) + 1.
            IF(W(k,2).GT.0.813) Y(k) = Y(k) + 1.
            IF(W(k,2).GT.0.987) Y(k) = Y(k) + 1.
   60    Continue
         GO TO 75
C        Population Two
   65    Y(k) = 3.
         DO 70 MM = 1,1
            IF(W(k,2).GT.0.001) Y(k) = Y(k) + 1.
            IF(W(k,2).GT.0.037) Y(k) = Y(k) + 1.
```

```
                 IF(W(k,2).GT.0.326) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.813) Y(k) = Y(k) + 1.
      70      Continue
      75   Continue
C
C          Computes Y values for X equals 3.0
C
           DO 105 k=11,15
              IF(W(k,1).GT.PI)85,95
C             Population One
      85      Y(k) = 1.
              DO 90 MM = 1,1
                 IF(W(k,2).GT.0.013) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.185) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.671) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.963) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.999) Y(k) = Y(k) + 1.
      90      Continue
              GO TO 105
C             Population Two
      95      Y(k) = 4.
              DO 100 MM = 1,1
                 IF(W(k,2).GT.0.013) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.185) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.671) Y(k) = Y(k) + 1.
     100      Continue
     105   Continue
C
C          Computes Y values for X equals 4.0
C
           DO 135 k=16,20
              IF(W(k,1).GT.PI)115,125
C             Population One
     115      Y(k) = 1.
              DO 120 MM = 1,1
                 IF(W(k,2).GT.0.004) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.089) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.499) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.910) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.996) Y(k) = Y(k) + 1.
     120      Continue
              GO TO 135
C             Population Two
     125      Y(k) = 4.
              DO 130 MM = 1,1
                 IF(W(k,2).GT.0.004) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.089) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.499) Y(k) = Y(k) + 1.
```

```
    130       Continue
    135    Continue
C
C          Computes Y values for X equals 5.0
C
           DO 165 k=16,25
              IF(W(k,1).GT.PI)145,155
C             Population One
    145       Y(k) = 1.
              DO 150 MM = 1,1
                 IF(W(k,2).GT.0.001) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.037) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.327) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.814) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.987) Y(k) = Y(k) + 1.
    150       Continue
              GO TO 165
C             Population Two
    155       Y(k) = 4.
              DO 160 MM = 1,1
                 IF(W(k,2).GT.0.001) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.037) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.327) Y(k) = Y(k) + 1.
    160       Continue
    165    Continue
C
C          Computes Y values for X equals 6.0
C
           DO 195 k=26,30
              IF(W(k,1).GT.PI)175,185
C             Population One
    175       Y(k) = 2.
              DO 180 MM = 1,1
                 IF(W(k,2).GT.0.013) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.185) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.672) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.963) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.999) Y(k) = Y(k) + 1.
    180       Continue
              GO TO 195
C             Population Two
    185       Y(k) = 5.
              DO 190 MM = 1,1
                 IF(W(k,2).GT.0.013) Y(k) = Y(k) + 1.
                 IF(W(k,2).GT.0.185) Y(k) = Y(k) + 1.
    190       Continue
    195    Continue
C
C             Computes Y values for X equals 7.0
```

```
C
             DO 225 k=31,35
                IF(W(k,1).GT.PI)205,215
C               Population One
   205          Y(k) = 2.
                DO 210 MM = 1,1
                   IF(W(k,2).GT.0.004) Y(k) = Y(k) + 1.
                   IF(W(k,2).GT.0.089) Y(k) = Y(k) + 1.
                   IF(W(k,2).GT.0.500) Y(k) = Y(k) + 1.
                   IF(W(k,2).GT.0.910) Y(k) = Y(k) + 1.
                   IF(W(k,2).GT.0.996) Y(k) = Y(k) + 1.
   210          Continue
                GO TO 225
C               Population Two
   215          Y(k) = 5.
                DO 220 MM = 1,1
                   IF(W(k,2).GT.0.004) Y(k) = Y(k) + 1.
                   IF(W(k,2).GT.0.089) Y(k) = Y(k) + 1.
   220          Continue
   225       Continue
   230       Continue
C
C         This segment of code computes the Least Squares estimates
C         for the simple linear regression of Y on X.
C
C
          Sum  Y = 0.0
          Sum YY = 0.0
          Sum  X = 0.0
          Sum XX = 0.0
          Sum XY = 0.0
C
          DO 240 k=1,N
             Sum  X = Sum  X + X(k)
             Sum XX = Sum XX + X(k)**2.
             Sum  Y = Sum  Y + Y(k)
             Sum YY = Sum YY + Y(k)**2.
             Sum XY = Sum XY + X(k)*Y(k)
  240     Continue
          Var X = N*Sum XX - Sum X**2.
          Var Y = N*Sum YY - Sum Y**2.
          Covxy = N*Sum XY - Sum X*Sum Y
          DDD   = Covxy/((Var X*Var Y)**0.5)
          BBB   = Covxy/Var X
          AAA   = (Sum Y - BBB*Sum X)/N
          EEE   = (Var Y*(1. - DDD**2.))/((N-2)*N)
          TTT   = BBB/((N*EEE/Var X)**0.5)
C
```

```
C
          LL = LL + 1
          Alpha (LL) = AAA
          Beta(LL)   = BBB
          Error(LL)  = EEE
          Corr(LL)   = DDD
          T Stat(LL) = TTT
C
C
          IF(LL.LT.MJ*100) GO TO 5
C
   300    Continue
C
          IF(LL.LT.1000) GO TO 320
C
C
          Call STEM(Alpha, 1000,1.67,1)
          Call STEM(Beta,1000,0.333,2)
          Call STEM(Error,1000,0.555,3)
          Call STEM(T Stat,1000,0.0,4)
          Call STEM(Corr,1000,0.667,5)
C
C
   320    Continue
          Stop
          End
```

```
C
      Subroutine STEM (B, N, Theta, Print)
C
      Real Min, Loq, Med, Upq, Max, Mean, Mse B, Kurt
      Integer Print
      Dimension LF(1000), IA(20), M(4)
      Dimension B(n)
      DATA IA/"0,", "1", "2", "3", "4", "5", "6", "7", "8", "9",
      2"0", "1", "2", "3", "4", "5", "6", "7", "8", "9"/
      DATA M/"  -", " ","      ", "+"/
      DATA Iwidth/120/
      DATA Atom/0.001/
      Scale = 5.
C
C
      Sum B  = 0.0
      Sum B2 = 0.0
      Sum B3 = 0.0
      Sum B4 = 0.0
C
      DO 10 k = 1,N
         Sum B  = Sum B  + B(k)
         Sum B2 = Sum B2 + B(k)**2.
         Sum B3 = Sum B3 + B(k)**3.
         Sum B4 = Sum B4 + B(k)**4.
  10  Continue
C
C     Calculates the various summary measures
C
      Mean  =  Sum B/N
      Var B =  (Sum B2 - N*Mean**2.)/(N-1)
      SD    =  Var B**0.5
      Bias  =  Mean - Theta
      MSE B =  Var B + (N/(N-1))*Bias**2.
      T     =  Mean
      R     =  N
      F     =  ((R-1)/R)**0.5
      Skew  =  (Sum B3 - 3.*Sum B2*T + 3.*Sum B*T**2. - R*T**3.)/R
      Skew  =  Skew/(SD*F)**3.
      Kurt  =  (Sum B4 - 4.*Sum B3*T + 6.*Sum B2*T**2.
     2 - 4.*Sum B*T**3.+N*T**4.)
      Kurt  =  Kurt/(((N-1)*Var B)/N)**1.
      Kurt  =  (Kurt/N) - 3.
C
C     Sets up the Stemleaf Display
C
```

```
      IF(Print.EQ.1) Write(6,100)
      IF(Print.EQ.2) Write(6,110)
      IF(Print.EQ.3) Write(6,120)
      IF(Print.EQ.4) Write(6,130)
      IF(Print.EQ.5) Write(6,135)
C
      Call Sort(B,1000)
C
      Min = B(1)
      Loq = (B(250) + B(251))/2.
      Med = (B(500) + B(501))/2.
      Upq = (B(750) + B(751))/2.
      Max = B(1000)
      Spd = Upq = Loq
      Write(6,140) Min, Loq, Med, Upq, Max, Spd, N
C
C     This segment initiates McNeil's (1977) algorithm.
C
      R = (Atom + (B(N) - B(1)))/Scale
      C = 10.**(11-INT(ALOG10(R) + 10))
      MM = MINØ(2,MAXØ(INT(R*C/25.),0))
      K = 3*MM + 2 - 150/(N + 50)
      IF((K-1)*(K-2)*(K-5).EQ.0) C = C*10
      MU = 10
      IF(K*(K-4)*(K-8).EQ.0)MU = 5
      IF((K-1)*(K-5)*(K-6).EQ.0) MU = 20
      I = 1
      IF(B(1).GE.0) I = 2
      II = 1
      D = MU*(INT(B(II)*C/MU) + I-2)/10.
C
C
C
 20   DO 30 k = 1, IWIDTH
         LF(k) = M(2)
 30   Continue
      IF(I.EQ.2.OR.D.LE.0) GO TO 40
      I = 2
      D = D - MU/10.0
 40   J = 0
 50   J = J + 1
      IX = INT(0.5 + ABS(B(II)*C-10*INT(D)))
      IF((B(II)*C-10*D).GE.0.5+(MU-1)*(I-1)) GO TO 60
      IF(J.LE.IWIDTH) LF(J) = IA(1 + IX)
      II = II + 1
      IF(II.GT.N) GO TO 60
      GO TO 50
```

```
   60     ID = MOD(IABS(INT(D)),100)
          K1 = 1 + ID/10
          K2 = 1 + ID - 10*(K1 - 1)
          IF(J.LE.IWIDTH + 1) GO TO 70
          LF(IWIDTH-2) = M(4)
          LF(IWIDTH-1) = IA(1 + (J-IWIDTH + 2)/10)
          LF(IWIDTH) = IA(J - IWIDTH + 3 - 10*((J - IWIDTH + 2)/10))
   70     K = MINØ(IWIDTH,J)
          Write(6,80) M(I), IA(K1), IA(K2), M(3), (LF(J),J=1,K)
          IF(II.GT.N) GO TO 90
          D = D + MU/10.0
          GO TO 20
   90     Continue
C
C
C
          Write(6,150)
          Write(6,155) Mean, Mse B, Var B, Bias
          Write(6,160) Skew, Kurt
C
          IF(Print.EQ.1) Write(6,180) (B(k),k=1,N)
          IF(Print.EQ.2) Write(6,185) (B(k),k=1,N)
          IF(Print.EQ.3) Write(6,185) (B(k),k=1,n)
          IF(Print.EQ.4) Write(6,180) (B(k),k=1,N)
          IF(Print.EQ.5) Write(6,185) (B(k),k=1,N)
C
          Write(6,195)
C
C
C
   80     Format(6X,120A1)
  100     Format(44X,"Stemleaf Display - Alpha"////)
  110     Format(45X,"Stemleaf Display - Beta"////)
  120     Format(44X,"Stemleaf Display - Error Term"////)
  130     Format(43X,"Stemleaf Display - T Stat"////)
  135     Format(45X,"Stemleaf Display - Corr"////)
  140     Format("ØMIN = ",F8.3,3X,"Loq = ",F8.3,3X,"Med =",F8.3,3X,
        2 "Upq =",2F8.3,3X,"Max =",F8.3/"ØSpd =",F8.3,3X,"N =", I6//)
  150     Format(4X,"Mean",16X,"Mse",17X,"var",17X,"Bias"/)
  155     Format(F8.5,3E20.5//)
  160     Format("ØSkewness =",E20.5,6X,"Kurtosis =",E20.5//)
  180     Format(10F8.3)
  185     Format(10F8.4)
  195     Format("           "//)
C
C
C
          Return
          End
```

```
C
      Subroutine Sort(B, II, JJ)
C
      Dimension B(1), IU(16), IL(16)
      Integer B, T, TT
C
C
      M = 1
      I = II
      J = JJ
    5 IF(I.GE.J) GO TO 70
   10 K = I
      IJ = (J+I)/2
      T = B(IJ)
      IF(B(I).LE.T) GO TO 20
      B(IJ) = B(I)
      B(I) = T
      T = B(IJ)
   20 L = J
      IF(B(IJ).GE.T) GO TO 40
      B(IJ) = B(J)
      B(J) = T
      T = B(IJ)
      IF(B(I).LE.T) GO TO 40
      B(IJ) = B(I)
      B(I) = T
      T = B(IJ)
      GO TO 40
   30 B(L) = B(K)
      B(K) = TT
   40 L = L-1
      IF(B(L).GT.T) GO TO 40
      TT = B(L)
   50 K = K+1
      IF(B(K).LT.T) GO TO 50
      IF(K.LE.L) GO TO 30
      IF(L-I.LE.J-K) GO TO 60
      IL(M) = I
      IU(M) = L
      I = K
      M = M+1
      GO TO 80
   60 IL(M) = k
      IU(M) = J
      J = L
      M = M+1
      GO TO 80
```

```
 70     M = M-1
        IF(M.EQ.0) Return
        I = IL(M)
        J = IU(M)
 80     IF(J-I.GE.II) GO TO 10
        IF(I.EQ.II) GO TO 5
        I = I-1
 90     I = I+1
        IF(I.EQ.J) GO TO 70
        T = B(I+1)
        IF(B(I).LE.T) GO TO 90
        K = I
100     B(K+1) = B(K)
        K = K-1
        IF(T.LT.B(K)) GO TO 100
        B(K+1) = T
        GO TO 90
        End
```

APPENDIX II

TESTS OF SUBROUTINE SUPER

Prior to the initiation of this study, the author was shown the results of a previous evaluation of SUPER.[1,2] Viewed individually and collectively, the test results suggest quite strongly that there are no serious deficiencies in SUPER's performance as a pseudo-random number generator. A FORTRAN listing of SUPER is provided below.

[1] The testing program was written in FORTRAN by Alan Van Hull. Both testing program and test results were graciously shown to the author by Mr. Robert Gonter, Associate Director, UMASS Computer Center.

[2] The specific tests were suggested by Maclaren and Marsaglia (1965) as a means of examining a generator's ability to produce random points which are uniformly distributed in a k-dimensional space. This property, as Chambers (1977) notes, is important in terms of this study as $k = 2$ uniform variates are needed to generate each Y_i value.

```
      Subroutine SUPER (IR1, IR2, IC, N, X)
      Dimension X(1)
C
C     IR1 is the seed for the mixed multiplicative generator.
C     IR2 is the seed for the shift register generator.
C     IC is the start constant with values of either 0 or 1.
C     N is the number of uniform deviates generated.
C     X is the returned array of random uniform deviates.
C
C
      IF(IC)5, 5, 10
    5 IM = IR1
      IT = IR2
   10 Continue
      M1 = 65539
      M2 = 4101
      M3 = 261
C
C     M1 = (2**16) + 3
C     M2 = (2**12) + 5
C     M3 = (2**8) + 5
C
      DO 15 I = 1, N
          IM = M3*IM
          L  = M1
          IF(IM.LT.0) L = M2
          IM = L*IM
          IF(IM.LT.0) IM = IM + 576460752303423487 + 1
          IB = IT
          IT = SHIFT(IT, -17)
          IB = XOR(IB, IT)
          IB = SHIFT(IB, 15)
          IC = XOR(IB, IT)
          IT = IC
          IR = XOR(IM, IC)
          X(I) = IR
          X(I) = X(I)/281474976718655
          X(I) = ABS(X(I))
   15 Continue
      Return
      End
```

APPENDIX III

TESTS OF SUBROUTINE TRANSFORM

Each transformation was evaluated in terms of its ability to generate N = 35 $(X_i Y_i)$ pairs which behaved as if they had been randomly sampled from a population with conditional mean, 1.67 + 0.333 X_i, and variance-covariance matrix,

$$\begin{Bmatrix} \sigma_Y^2 & \sigma_{XY}^2 \\ \sigma_{XY}^2 & \sigma_X^2 \end{Bmatrix} = \begin{Bmatrix} 1.00 & 1.33 \\ 1.33 & 4.00 \end{Bmatrix} .$$

The X_i values were arrayed as in Simulation I. Following 3000 replications, Transformation I yielded a conditional mean and covariance matrix of

$$2.168 + 0.3334\ X_i \quad \text{and} \quad \begin{Bmatrix} 1.069 & 1.334 \\ 1.334 & 4.000 \end{Bmatrix} ,$$

while Transformation II yielded a conditional mean and covariance matrix of

$$1.979 + 0.331\ X_i \quad \text{and} \quad \begin{Bmatrix} 0.844 & 1.332 \\ 1.332 & 4.000 \end{Bmatrix} .$$

Note, Transformation I overestimates the marginal variance, σ_y^2, while Transformation II underestimates it. The implications of the differing values for σ_y^2 are clear in that:

$$\rho = \frac{\sigma_{XY}^2}{\sqrt{\sigma_X^2 \sigma_Y^2}} ,$$

$$\sigma_{Y|X}^2 = \sigma_Y^2 (1 - \rho^2) ,$$

$$\sigma_\beta^2 = \frac{\sigma_{Y|X}^2}{\Sigma (X_i - \overline{X})^2} ,$$

$$t = \frac{\beta - \beta_o}{\sqrt{\sigma_\beta^2}} ,$$

and, finally, floor and ceiling effects are proportional to $\sigma_{Y|X}^2$. (As a result, the estimate of α is smaller when Transformation II is used than when Transformation I is used.)

The code for Transformation I is found in appendix I. The code for Transformation II is given below. Note, it is both simpler and more flexible than the code for Transformation I. More importantly, it readily lends itself to any necessary correction of the realized σ_y^2 value. A deficiency in σ_y^2 can be corrected by multiplying the random error term, Distur(k), by a constant such that $\sigma_{Y|X}^2$ has the intended value. (Distur(k) is a random normal deviate generated by subroutine TRF. The latter is based on an algorithm published in Hastings (1955).)

```
C
C       This segment of code transforms uniform deviates into
C       discrete values for the equations shown below.
C
        DO 20 k = 1,N
           IF(W(k,1).GT.PI)10,15
   10      Continue
           Call TRF(k, W(k,2), Distur(k) )
           Y(k) = 1.67 + 0.333*X(k) + Distur(k)* Const
           Y(K) = AINT(Y(k))
           IF(Y(k).GT.ULIMIT) Y(k) = ULIMIT
           IF(Y(k).LT.1.) Y(k) = 1.
           GO TO 20
   15      Continue
           Call TRF(k, W(k,2), Distur(k) )
           Y(k) = 4.67 + 0.333*X(k) + Distur(k)*Const
           Y(k) = AINT(Y(k))
           IF(Y(k).GT. ULIMIT) Y(k) = ULIMIT
           IF(Y(k).LT.1.) Y(k) = 1.
   20      Continue
```

* * * * * * * * *

```
           Subroutine TRF(k, Pr, Z)
           DATA A1,A2,A3/2.515517, 0.802853, 0.010328/
           DATA B1,B2,B3/1.432788, 0.189269, 0.001308/
C
           IF(Pr.GT.0.5) PR = 1. - Pr
C
           T2 = ALOG(2.0/Pr**2.)
           T1 = T2**0.5
           AA = A1 + A2*T1 + B2*T2 + B3*T1*T2
           Z  = T1 - (AA/BB)
C
           IF(PR.GT.0.5) Z = -Z
C
           Return
           End
```

APPENDIX IV

MAXIMUM LIKELIHOOD ESTIMATION FOR MIXED REGRESSIONS

Maximum Likelihood estimation of the parameters associated with mixed regressions is described by Elashoff (1972), Hosmer (1974), and Kiefer (1978). Basically, the problem is as follows. Data are collected from "N" individuals with respect to a response variable, Y_i (i = 1, . . . , N), and "p" explanatory variables. With a probability equal to $(1 - \pi)$, the regression

$$Y_i = \alpha + \sum_{h=1}^{p} \beta_h X_{hi} + \varepsilon_i$$

occurs; and with a probability equal to π, the regression

$$Y_i = \alpha^* + \sum_{h=1}^{p} \beta_h^* X_{hi} + \varepsilon_i'$$

occurs where one or more of the parameters in the second regression differs in value from its counterpart in the first regression.

Given this mixture of regressions, the density function for Y_i is

$$(1) \qquad f_3(Y_i) = (1 - \pi)\, f_1(Y_i) + \pi\, f_2(Y_i) \quad ,$$

where $f_1(Y_i)$ and $f_2(Y_i)$ are the density functions corresponding to the first and second regressions. Estimation of $(\alpha, \beta_1, \beta_2, \ldots ,$

β_p, σ^2, α^*, β_1^*, β_2^*, . . . , β_p^*, and σ^{*2}) by Maximum Likelihood requires that the Log-Likelihood Function (LLF) be differentiated with respect to the parameters of interest.

In general, estimation proceeds as follows. First, the Likelihood Function (LF) is defined for the "N" independent observations on Y:

(2) $$LF = \Pi f_3(Y_i) \quad .$$

Second, the Log-Likelihood Function (LLF) is derived by taking the natural logarithm of the Likelihood Function:

(3) $$LLF = \Sigma \ln f_3(Y_i) \quad .$$

Third, the Log-Likelihood Function is differentiated with respect to θ, the parameter of interest:

(4) $$\frac{\partial LLF}{\partial\theta} = \Sigma \frac{\partial}{\partial\theta} \ln f_3(Y_i)$$

$$= \left\{ \Sigma \ (1 - \pi) \frac{\partial}{\partial\theta} f_1(Y_i) + \pi \frac{\partial}{\partial\theta} f_2(Y_i) \right\} / f_3(Y_i) \quad .$$

since,

(5) $$f_j(Y_i) = (2\pi\sigma_j^2)^{-\frac{1}{2}} \exp -\tfrac{1}{2} (Y_i - \lambda_j(\theta))^2/\sigma_j^2 \ , \quad j = 1, 2$$

it follows that

(6) $$\frac{\partial}{\partial\theta} f_j(Y_i) = (2\pi\sigma_j^2)^{-\frac{1}{2}} \frac{\partial}{\partial\theta} \exp -\tfrac{1}{2}(Y_i - \lambda_j(\theta))^2/\sigma_j^2 + \left\{ \exp -\tfrac{1}{2} (Y_i - \lambda_j(\theta))^2/\sigma_j^2 \right\} \frac{\partial}{\partial\theta} (2\pi\sigma_j^2)^{-\frac{1}{2}} \ ,$$

where $\lambda_j(\theta)$ is the "j"th equation for the regression of Y_i on the "p" explanatory variables. Finally, to obtain the Maximum Likelihood estimator for θ, $\frac{\partial LLF}{\partial \theta}$ is set equal to zero and the resulting equation is solved for θ.

For the specific case investigated in this study, i.e.,

$$f_3(Y_i) = (1 - \pi)\, f_1(Y_i) + \pi\, f_2(Y_i) \quad , \tag{7}$$

where

$$f_1(Y_i) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp -\tfrac{1}{2}\, (Y_i - \alpha - \beta X_i)^2/\sigma^2 \tag{8}$$

and

$$f_2(Y_i) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp -\tfrac{1}{2}\, (Y_i - \alpha - \lambda - \beta X_i)^2 / \sigma^2 \tag{9}$$

implementation of the cited procedure yields:

$$\begin{aligned} \frac{\partial LLF}{\partial \alpha} &= \Sigma \left\{ w_{1i}(Y_i - \alpha - \beta X_i) + w_{2i}(Y_i - \alpha - \lambda - \beta Z_i) \right\} / \sigma^2 , \\ &= \Sigma\, (Y_i - \alpha - \beta X_i) - \lambda\, \Sigma\, w_{2i} \quad ; \end{aligned} \tag{10}$$

$$\begin{aligned} \frac{\partial LLF}{\partial \beta} &= \Sigma \left\{ w_{1i}(Y_i - \alpha - \beta X_i)(-X_i) + w_{2i}(Y_i - \alpha - \lambda - \beta X_i)(-X_i) \right\} / \sigma^{2i} \\ &= \Sigma(Y_i - \alpha - \beta - \lambda\, w_{2i})X_i \quad ; \end{aligned} \tag{11}$$

$$\begin{aligned} \frac{\partial LLF}{\partial \sigma^2} &= \tfrac{1}{2}\, \Sigma \left\{ w_{1i}(Y_i - \alpha - \beta X_i)^2 - w_{1i}\, \sigma^2 \right\} / \sigma^4 \\ &\quad + \tfrac{1}{2}\, \Sigma \left\{ w_{2i}(Y_i - \alpha - \lambda - \beta X_i)^2 - w_{2i}\sigma^2 \right\} / \sigma^4 \quad ; \end{aligned} \tag{12}$$

where

$$w_{1i} = (1 - \pi) f_1(Y_i) \;/\; f_3(Y_i) \quad , \tag{13}$$

$$w_{2i} = \pi\; f_2(Y_i) \;/\; f_3(Y_i) \quad , \tag{14}$$

and

$$w_{1i} + w_{2i} = 1 \text{ for } i = 1, 2, . . . , N.$$

Upon setting expressions (10) - (12) equal to zero and solving for the parameter, one obtains the following:

(15) $$\hat{\alpha} = \overline{Y} - \hat{\beta}\ \overline{X} - \frac{1}{N} \Sigma\, \lambda\, w_{2i} \quad ,$$

(16) $$\hat{\beta} = \Sigma\, (Y_i - \hat{\alpha} - \lambda\, w_{2i})X_i \ /\ \Sigma\, X_i^2 \quad , \text{ and}$$

(17) $$\hat{\sigma}^2 = \frac{1}{N} \Sigma\, (Y_i - \hat{\alpha} - \hat{\beta}\, X_i)^2 - \lambda^2\, \Sigma\, w_{2i} \quad .$$

Finally, upon substitution of the expression for $\hat{\alpha}$ into that for $\hat{\beta}$, one can show that

(18) $$\hat{\beta} = \Sigma\, (X_i - \overline{X})(Y_i - \overline{Y}) \ /\ \Sigma\, (X_i - \overline{X})^2 \quad .$$