Rita Cristina Pinto de Sousa

Mestre em Estatística

Parameter Estimation in the Presence of Auxiliary Information

Dissertação para obtenção do Grau de Doutora em
Estatística e Gestão de Risco, Especialidade em Estatística

Orientador : Sat Gupta, Professor of Statistics,
University of North Carolina at Greensboro, USA

Co-orientador : Pedro Corte Real, Professor Auxiliar,
Universidade Nova de Lisboa, Portugal

Júri:

Presidente: Prof. Doutora Maria Luísa Dias de Carvalho de Sousa Leonardo

Arguentes: Prof. Doutor Dinis Duarte Ferreira Pestana
Prof. Doutora Maria Teresa Themido da Silva Pereira

Vogais: Prof. Doutor João Tiago Praça Nunes Mexia
Prof. Doutora Célia Maria Pinto Nunes
Prof. Doutor Sat Gupta
Prof. Doutor Pedro Alexandre da Rosa Corte Real

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Novembro, 2013

Rita Cristina Pinto de Sousa

Mestre em Estatística

Parameter Estimation in the Presence of Auxiliary Information

Dissertação para obtenção do Grau de Doutora em
Estatística e Gestão de Risco, Especialidade em Estatística

Orientador : Sat Gupta, Professor of Statistics,
University of North Carolina at Greensboro, USA

Co-orientador : Pedro Corte Real, Professor Auxiliar,
Universidade Nova de Lisboa, Portugal

Júri:

Presidente: Prof. Doutora Maria Luísa Dias de Carvalho de Sousa Leonardo

Arguentes: Prof. Doutor Dinis Duarte Ferreira Pestana
Prof. Doutora Maria Teresa Themido da Silva Pereira

Vogais: Prof. Doutor João Tiago Praça Nunes Mexia
Prof. Doutora Célia Maria Pinto Nunes
Prof. Doutor Sat Gupta
Prof. Doutor Pedro Alexandre da Rosa Corte Real

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Novembro, 2013

Parameter Estimation
in the Presence of Auxiliary Information

Aos meus Pais e Irmã

Acknowledgements

The last four years were a great challenge for me. I have to thank many people who made this interesting journey possible. Their scientific or emotional support was crucial to reach the end of my thesis.

- First of all I would like to thank my Mentors, the advisor Sat Gupta and the co-adviser Pedro Corte Real for sharing their great knowledge and experience with me. I have to point out their tremendous support and their orientation ability that made me easily forget and overcome the physical distance.
  I thank Professor Pedro Corte Real especially for having encouraged me to invest in my school graduation which contributed to my growth as a researcher. His friendly receptivity, the sharing of his knowledge and his great ideas were fundamental for my motivation and for a better outcome. His experience and friendship are for me an important and true reference.
  I thank Professor Sat Gupta especially for sharing with me his pleasure, knowledge and motivation in working on this theme. The chance to work with him allowed me to meet many people, to learn a lot and take a broad view of estimation with auxiliary information. He proved to be a very friendly person. I consider him an excellent reference as person, as a mentor and as a researcher.
  I am deeply grateful to them. Their support greatly contributed for my development as research and as person.

- I would like to thank the Professors and people from the university staff who support the post-graduate students. A special thanks to the coordinator Manuel Esquível for his support and his incentive to carry on, always understanding my constraints as working student.

- I would like to thank all my Co-Authors who helped me write the chapters of this thesis. Their collaboration improved and expanded my knowledge on several topics.
  A special thanks to Javid for sharing with me his great theoretical knowledge and

experience and to Nursel for being a good reference as a young and active researcher.

- I also would like to thank the anonymous referees for making very constructive suggestions which resulted in a significant improvement over the original version of the papers to be found in the chapters of this thesis.

- Thanks to my work colleagues for their support and encouragement.
  A special reference to Pedro and São, my classmates at Statistics Portugal, for their advice and for having always a friendly word encouraging me and giving me strength to carry on.
  A special message of affection to my colleagues from afrolatINE for having always supported me and for having contributed to my well being and happiness. Who dances is happier.

- Thanks to my parents and my sister for their affection and unconditional support.
  From the beginning my parents always supported me, encouraging me in my decision and doing everything they could to allow me to complete successfully this journey. They are my true inspiration.
  My sister is a true example of a blood sister and especially a sister by heart. Thank you for being as you are.

- I would also like to thank my friends for all their support, for being always there and for their comprehension for my very often absence. They know who they are.
  I can not forget to mention Graça, Joana, Magda, Marinha, Marta, Mónica, Rosário and Silvia. To them I owe my deep friendship and a lot of my well being as a person.
  A special word for Rui who accompanied me in the final phase of this thesis. Thanks for your encouragement, for your patience, for your fellowship and affection.

Abstract

In survey research, there are many situations when the primary variable of interest is sensitive. The sensitivity of some queries can give rise to a refusal to answer or to false answers given intentionally. Survey can be conducted in a variety of settings, in part dictated by the mode of data collection, and these settings can differ in how much privacy they offer the respondent. The estimates obtained from a direct survey on sensitive questions would be subject to high bias. A variety of techniques have been used to improve reporting by increasing the privacy of the respondents.

The Randomized Response Technique (RRT), introduced by Warner in 1965, develops a random relation between the individual's response and the question. This technique provides confidentiality to respondents and still allows the interviewers to estimate the characteristic of interest at an aggregate level.

In this thesis we propose some estimators to improve the mean estimation of a sensitive variable based on a RRT by making use of available non-sensitive auxiliary information. In the first part of this thesis we present the ratio and the regression estimators as well as some generalizations in order to study the gain in the estimation over the ordinary RRT mean estimator. In chapters 4 and 5 we study the performance of some exponential type estimators, also based on a RRT. The final part of the thesis illustrates an approach to mean estimation in stratified sampling. This study confirms some previous results for a different sample design. An extensive simulation study and an application to a real dataset are done for all the study estimators to evaluate their performance. In the last chapter we present a general discussion referring to the main results and conclusions as well as showing an application to a real dataset which compares the performance of study estimators.

Keywords: Auxiliary variable; Exponential estimator; Randomized response technique; Ratio estimator; Regression estimator; Sensitive variable.

Resumo

Em estudos de pesquisa por inquérito existem muitas situações em que a variável de interesse é sensível. A sensibilidade de algumas questões pode dar origem a recusas na resposta ou a falsas respostas dadas de forma intencional. Os inquéritos podem assumir diversas configurações, em parte relacionadas com o método de recolha e com o grau de privacidade que é oferecido aos respondentes. As estimativas obtidas por inquérito direto em questões sensíveis estariam sujeitas a erros elevados. Muitas técnicas têm sido utilizadas para melhorar as respostas através do aumento de privacidade dos inquiridos. A Técnica de Resposta Aleatorizada, introduzida por Warner em 1965, desenvolve uma relação aleatória entre as respostas individuais e a questão. Esta técnica providencia confidencialidade aos respondentes e ainda permite aos entrevistadores estimar a característica de interesse num nível mais agregado.

Nesta tese propõem-se alguns estimadores para melhorar a estimação da média de uma variável sensível baseada numa técnica de resposta aleatorizada com recurso a informação auxiliar disponível não sensível. Na primeira parte da tese apresentam-se os estimadores da razão e da regressão bem como algumas generalizações para estudar o ganho na estimação face ao estimador ordinário da média. Nos capítulos 4 e 5 estuda-se a performance de alguns estimadores do tipo exponencial, também baseados numa técnica de resposta aleatorizada. A parte final da tese ilustra uma aproximação à estimação da média com amostragem estratificada. Este estudo vem confirmar resultados anteriores com um novo desenho amostral. Um extenso estudo de simulação e uma aplicação a dados reais são feitos para avaliar a performance de todos os estimadores. No último capítulo apresenta-se uma discussão geral, bem como uma aplicação a dados reais onde se compara a performance dos estimadores em estudo.

Palavras-chave: Estimador da razão; Estimador da regressão; Estimador exponencial; Técnica de resposta aleatorizada; Variável auxiliar; Variável sensível.

Contents

List of Figures

List of Tables

Listings

List of Abbreviations

ARB - Absolute Relative Bias

Deff - Design Effect

ICT - Information and Communication Technologies

MES - Monthly Economic Survey

MSE - Mean Square Error

NACE - Statistical Classification of Economic Activities in the European Community

PRE - Percent Relative Efficiency

RRT - Randomized Response Technique

SRS - Simple Random Sample

SRSWOR - Simple Random Sampling Without Replacement

Str - Stratified Sample

1

General Introduction

One of the major problems in survey research involving sensitive questions is the social desirability response bias (Edwards, 1957). For various reasons individuals in a sample survey may prefer not to confide to the interviewer the correct answers to certain questions. In such cases the individuals may elect not to reply at all or to reply with incorrect answers. The resulting evasive answer bias is ordinarily difficult to assess. That bias is potentially removable through allowing the interviewer to maintain privacy using a randomization device (Warner, 1965).

Randomized response is a research method used in structured survey interview. It was first proposed by Warner in 1965 and later modified by Greenberg et al. in 1969. This technique allows respondents to respond to sensitive issues while maintaining confidentiality. It provides confidentiality to respondents through a random relation between the individual's response and the question. It still allows the interviewers to estimate the characteristic of interest at an aggregate level.

Gupta and Thornton (2002) showed that Randomized Response Technique (RRT) is effective in circumventing the social desirability response bias, and is more friendly and portable than other methods such as the method which uses a bogus pipeline (Jones and Sigall, 1971).

RRT models may be classified as Full RRT model, Partial RRT model or Optional RRT model depending on the level of scrambling. In the Full RRT model (Eichhorn and Hayre, 1983) all the respondents are asked to provide a scrambled response. When a predetermined proportion of randomly selected respondents are asked to provide a true response we have a Partial RRT model (Mangat and Singh, 1990). Gupta et al. (2002) proposed an Optional RRT model where the respondents are allowed to report a true response or a scrambled response depending on whether the respondents find the question sensitive or not.

In RRT work, generally the focus is on the estimation of the mean of a sensitive variable or the prevalence of a sensitive characteristic in the population. The mean can be estimated by using one of many RRT but we propose some estimators which improve the mean estimation considerably by using non-sensitive auxiliary information. In such cases, one will be able to observe an auxiliary variable directly but will have to rely on some RRT to collect information on the variable of interest, resulting from a sensitive issue. Given that our main aim is to evaluate the performance of the mean estimator in the presence of auxiliary information, we opt for using an additive Full RRT method to scramble the sensitive variable.

The main goal of this thesis is to improve the parameter estimation of a sensitive variable in the presence of auxiliary information. For that purpose we introduce some estimators for the population mean based on the additive Full RRT technique. Expressions are derived for the *Bias* and Mean Square Error (*MSE*) for all the proposed estimators. Furthermore, an extensive simulation study and an application to a real dataset are done for all the study estimators. All the applications are developed using the statistical software R [1].

This thesis is based on five papers to be found in chapters 2–6. Each chapter presents, at least, a new estimator and evaluates its performance comparing it to the other estimators previously proposed. The contents of this thesis are as follows:

- In **Chapter 2** we propose a ratio estimator for the mean of a sensitive variable using information from a non-sensitive auxiliary variable. We generalize the proposed estimator to the case of transformed ratio estimators. We show that there is hardly any difference in the first order and second order approximations for *MSE* even for small sample sizes. We also show that the proposed estimator does better than the ordinary RRT mean estimator which does not use the auxiliary information (Sousa et al., 2010).

- In **Chapter 3** we introduce a regression estimator which performs better than the ratio estimator even for modest correlation between the primary and the auxiliary variables. We consider a generalized regression-cum-ratio estimator that has even smaller *MSE*. It is shown that the proposed regression estimator performs better than the ratio estimator and the ordinary RRT mean estimator that does not utilize the auxiliary information (Gupta et al., 2012).

- In **Chapter 4** we propose exponential type estimators using one and two auxiliary variables to improve the efficiency of mean estimator based on a RRT. It is shown the proposed exponential type estimators are more efficient than the existing estimators described in Sousa et al. (2010) and Gupta et al. (2012)(Koyuncu et al., 2013).

- In **Chapter 5** we propose an improved exponential type estimator which is more efficient than the Koyuncu et al. (2013) estimator, which in turn was shown to be more efficient than the usual mean estimator, ratio estimator, regression estimator, and the Gupta et al. (2012) estimator. It is shown that the improved difference-cum-exponential estimator can produce further improvement relative to other estimators previously proposed (Gupta et al., 2013).

- In **Chapter 6** we extend the ratio and regression estimators to the stratified sampling setting. Although both the ratio and regression estimators perform better than the ordinary RRT mean estimator, the improvement is much larger with the regression estimator. The results agree with the findings of Sousa et al. (2010) and Gupta et al. (2012) in simple random sampling. We show that the advantage of using the RRT in the presence of auxiliary information still holds in the context of stratified sampling (Sousa et al., 2013).

- In **Chapter 7** we present a general discussion referring to the main results and conclusions. We present a study with a real dataset and we show the numerical results for the *Bias* and *MSE*, as well as graphic evidence which illustrates the performance of the main study estimators.

In the last part of each chapter we attach the R routines developed for the simulation studies and for the numerical examples.

References

EDWARDS, A. L. 1957. *The social desirability variable in personality assessment and research*, New York: Dryden, Praeger.

EICHHORN, B. H. & HAYRE, L. S. 1983. Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*, 7, 307-316.

GREENBERG, B., ABDUL-ELA, A., SIMMONS, W. & HORVITZ, D. 1969. The unrelated question randomized response model: theoretical framework. *Journal of American Statistical Association*, 520-539.

GUPTA, S. N., GUPTA, B. C. & SINGH, S. 2002. Estimation of sensitivity level of personal interview survey questions. *Journal of Statistical Planning and Inference*, 100, 239-247.

GUPTA, S., SHABBIR, J., SOUSA, R. & REAL, P. C. 2012. Estimation of the Mean of a

Sensitive Variable in the Presence of Auxiliary Information. *Communications in Statistics - Theory and Methods*, 41(13-14), 2394-2404.

GUPTA, S., SHABBIR, J., SOUSA, R. & REAL, P. C. 2013. Improved exponential type estimators of the mean of a sensitive variable in the presence of non-sensitive auxiliary information. (*submitted*)

GUPTA, S. & THORNTON, B. 2002. Circumventing social desirability response bias in personal interview surveys. *American Journal of Mathematical and Management Sciences*, 22, 369-383.

JONES, E. E. & SIGALL, H. 1971. The Bogus Pipeline: A New Paradigm for Measuring Affect and Attitude. *Psychological Bulletin*, 76, 349-364.

KOYUNCU, N., GUPTA, S. & SOUSA, R. 2013. Exponential type estimators of the mean of a sensitive variable in the presence of non-sensitive auxiliary information. *Communications in Statistics - Simulation and Computation*. (*accepted*).

MANGAT, N. S. & SINGH, R. 1990. An Alternative Randomized Response Procedure. *Biometrika*, 77, 439-442.

SOUSA, R., GUPTA, S., SHABBIR, J. & REAL, P. C. 2013. Improved Mean Estimation of a Sensitive Variable Using Auxiliary Information in Stratified Sampling. *Journal of Statistics and Management Systems*. (*submitted*).

SOUSA, R., SHABBIR, J., REAL, P. C. & GUPTA, S. 2010. Ratio estimation of the mean of a sensitive variable in the presence of auxiliary information. *Journal of Statistical Theory and Practice*, 4(3), 495-507.

WARNER, S. L. 1965. Randomized response: a survey technique for elimination evasive answer bias. *Journal of American Statistical Association*, 60, 63-69.

[1] The R Project for Statistical Computing: `www.r-project.org`.

2

Ratio Estimation of the Mean of a Sensitive Variable in the Presence of Auxiliary Information

Abstract

We propose a ratio estimator for the mean of sensitive variable utilizing information from a non-sensitive auxiliary variable. Expressions for the *Bias* and Mean Square Error (*MSE*) of the proposed estimator (correct up to first and second order approximations) are derived. We show that the proposed estimator does better than the ordinary Randomized Response Technique (RRT) mean estimator that does not utilize the auxiliary information. We also show that there is hardly any difference in the first order and second order approximations for *MSE* even for small sample sizes. We also generalize the proposed estimator to the case of transformed ratio estimators but these transformations do not result in any significant reduction in *MSE*. An extensive simulation study is presented to evaluate the performance of the proposed estimator. The procedure is also applied to some financial data (purchase orders (sensitive variable) and gross turnover (non-sensitive variable)) in 2009 for 5090 companies in Portugal from a survey on Information and Communication Technologies (ICT) usage.

Published as: SOUSA, R., SHABBIR, J. REAL, P. C. & GUPTA, S. (2010). Ratio estimation of the mean of a sensitive variable in the presence of auxiliary information. *Journal of Statistical Theory and Practice*, 4(3), 495-507.

2.1 Introduction

In survey research, there are many situations when the primary variable of interest (Y) is sensitive and direct observation on this variable may not be possible. However, we may be able to directly observe a highly correlated auxiliary variable (X). For example, Y may be the number of abortions a woman might have had in her life and X may be her age. Similarly Y may be the total purchase orders in a year for a company and X may be the total turn-over for that company in that year. In such cases, one will be able to observe X directly but will have to rely on some Randomized Response Technique (RRT) to collect information on Y. In such situations, mean of Y can be estimated by using one of many randomized response techniques but this estimator can be improved considerably by utilizing information from the auxiliary variable X. Many authors have presented ratio estimators when both Y and X are directly observable. These include Kadilar and Cingi (2006), Turgut and Cingi (2008), Singh and Vishwakarma (2008), Koyuncu and Kadilar (2009) and Shabbir and Gupta (2010).

Also, many authors have estimated the mean of a sensitive variable when the primary variable is sensitive and there is no auxiliary variable available. These include Eichhorn and Hayre (1983), Gupta and Shabbir (2004), Gupta et al. (2002), Saha (2008) and Gupta et al. (2010).

In this paper, we propose a ratio estimator where the RRT estimator of the mean of Y is further improved by using information on an auxiliary variable X. Expressions for the *Bias* and *MSE* for the proposed estimator are derived, correct up to both the first order and second order approximations. It is shown that the two approximations are very similar even for moderate sample size. We also observe that there is considerable reduction in *MSE* when auxiliary information is used, particularly when the correlation between the study variable and the auxiliary variable is high.

2.2 Terminology

Let Y be the study variable, a sensitive variable which cannot be observed directly. Let X be a non-sensitive auxiliary variable which is strongly correlated with Y. Let S be a scrambling variable independent of Y and X. The respondent is asked to report a scrambled response for Y given by $Z = Y + S$ but is asked to provide a true response for X. Let a random sample of size n be drawn without replacement from a finite population $U = (U_1, U_2, ..., U_N)$. For the i^{th} unit $(i = 1, 2, ..., N)$, let y_i and x_i respectively be the values of the study variable Y and auxiliary variable X. Moreover, let $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$, $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ and $\bar{z} = \frac{\sum_{i=1}^{n} z_i}{n}$ be the sample means and $\bar{Y} = E(Y)$, $\bar{X} = E(X)$ and $\bar{Z} = E(Z)$ be the population means for Y, X and Z, respectively. We assume that $\bar{X}$ is known and $\bar{S} = E(S) = 0$. Thus, $E(Z) = E(Y)$. Let us also define $\delta_z = \frac{\bar{z} - \bar{Z}}{\bar{Z}}$ and $\delta_x = \frac{\bar{x} - \bar{X}}{\bar{X}}$, such that

$E(\delta_i) = 0, i = z, x.$

If information on X is ignored, then an unbiased estimator of μ_Y is the ordinary sample mean $(\bar{z})$ given by (2.1) below

$$\hat{\mu}_Y = \bar{z}. \tag{2.1}$$

The mean square error (*MSE*) of $\hat{\mu}_Y$ is given by

$$MSE(\hat{\mu}_Y) = \frac{1-f}{n}\left(S_y^2 + S_s^2\right), \tag{2.2}$$

where
$f = n/N$, $S_y^2 = \frac{1}{N-1}\sum_{i=1}^N (y_i - \bar{Y})^2$, $S_x^2 = \frac{1}{N-1}\sum_{i=1}^N (x_i - \bar{X})^2$ and $S_s^2 = \frac{1}{N-1}\sum_{i=1}^N (s_i - \bar{S})^2$.

2.3 The Proposed Estimator

We propose the following ratio estimator for estimating the population mean of the study variable Y using the auxiliary variable X:

$$\begin{aligned} \hat{\mu}_R &= \bar{z}\left(\frac{\bar{X}}{\bar{x}}\right) \\ &= \bar{Z}\left(1+\delta_z\right)\left(1+\delta_x\right)^{-1}. \end{aligned} \tag{2.3}$$

Using Taylor's approximation and retaining terms of order up to 4, (2.3) can be rewritten as

$$\hat{\mu}_R - \bar{Z} \cong \bar{Z}\{\delta_z - \delta_x - \delta_z\delta_x + \delta_x^2 - \delta_x^3 + \delta_x^4 + \delta_z\delta_x^2 - \delta_z\delta_x^3\}. \tag{2.4}$$

Under the assumption of bivariate normality (see Sukhatme and Sukhatme, 1984), we have $E(\delta_z^2) = \frac{1-f}{n}C_z^2$, $E(\delta_x^2) = \frac{1-f}{n}C_x^2$, $E(\delta_x\delta_z) = \frac{1-f}{n}C_{zx}$, where $C_{zx} = \rho_{zx}C_zC_x$ and C_z and C_x are the coefficients of variation of Z and X, respectively. Also we have:

$$E(\delta_z\delta_x^3) = \left(\tfrac{1-f}{n}\right)^2 3\rho_{zx}C_zC_x^3, \quad E(\delta_z^2\delta_x^2) = \left(\tfrac{1-f}{n}\right)^2(1+2\rho_{zx}^2)C_z^2C_x^2,$$
$$E(\delta_x^4) = \left(\tfrac{1-f}{n}\right)^2 3C_x^4, E(\delta_z\delta_x^2) = E(\delta_z^2\delta_x) = E(\delta_x^3) = 0,$$

and

$$C_z^2 = C_y^2 + \frac{S_s^2}{\bar{Y}^2}, \rho_{zx} = \frac{\rho_{yx}}{\sqrt{1+\frac{S_s^2}{S_y^2}}}.$$

Recognizing that $\bar{Z} = \bar{Y}$ in Equation (2.4), we can get expressions for the *Bias* of $\hat{\mu}_R$, correct up to second order of approximation, as given by

$$Bias^{(2)}(\hat{\mu}_R) \cong Bias^{(1)}(\hat{\mu}_R) + 3\left(\frac{1-f}{n}\right)^2 \bar{Y}\left[C_x^4 - \rho_{yx}C_yC_x^3\right], \tag{2.5}$$

where

$$Bias^{(1)}(\hat{\mu}_R) = \left(\frac{1-f}{n}\right)\bar{Y}\left[C_x^2 - \rho_{yx}C_yC_x\right] \tag{2.6}$$

is the *Bias* corresponding to first order of approximation.

Similarly from (2.4), *MSE* of $\hat{\mu}_R$, correct up to second order of approximation, is given by

$$MSE^{(2)}(\hat{\mu}_R) = E(\hat{\mu}_R - \bar{Z})^2 \cong \bar{Z}^2 E\{\delta_z - \delta_x - \delta_z\delta_x + \delta_x^2 - \delta_x^3 + \delta_x^4 + \delta_z\delta_x^2 - \delta_z\delta_x^3\}^2$$

or

$$MSE^{(2)}(\hat{\mu}_R) \cong \bar{Z}^2 E\{\delta_z^2 + \delta_x^2 - 2\delta_z\delta_x + 3\delta_z^2\delta_x^2 + 3\delta_x^4 - 6\delta_z\delta_x^3 - 2\delta_z^2\delta_x + 4\delta_z\delta_x^2 - 2\delta_x^3\}.$$

Since $\bar{Z} = \bar{Y}$, we have

$$\begin{aligned} MSE^{(2)}(\hat{\mu}_R) \cong\ & MSE^{(1)}(\hat{\mu}_R) \\ & +3\bar{Y}^2\left(\tfrac{1-f}{n}\right)^2 C_x^2\left[(1+2\rho_{yx}^2)C_y^2 + 3C_x^2 - 6\rho_{yx}C_yC_x\right], \end{aligned} \tag{2.7}$$

where

$$MSE^{(1)}(\hat{\mu}_R) \cong \left(\frac{1-f}{n}\right)\bar{Y}^2\left(C_y^2 + C_x^2 - 2\rho_{yx}C_yC_x\right) \tag{2.8}$$

is the *MSE* corresponding to the first order approximation. The difference between the two approximations for *MSE* is given by

$$3\bar{Y}^2\left(\frac{1-f}{n}\right)^2 C_x^2\left[(1+2\rho_{yx}^2)C_y^2 + 3C_x^2 - 6\rho_{yx}C_yC_x\right],$$

and it converges to zero as $n \to N$. Our simulation results in Section 2.4 will also confirm this pattern.

According to the first order of approximation, $MSE^{(1)}(\hat{\mu}_R) < MSE(\hat{\mu}_Y)$ if

$$\left(\rho_{yx} - \frac{1}{2}\frac{C_x}{C_y}\right) > 0. \tag{2.9}$$

If second order approximation is used, we can easily see that $MSE^{(2)}(\hat{\mu}_R) < MSE(\hat{\mu}_Y)$ if

$$2\rho_{yx}\frac{C_x}{C_x} + 3\left(\frac{1-f}{n}\right)\left[6\rho_{yx}C_xC_y - 3C_x^2 - (1+2\rho_{yx}^2)C_y^2\right] > 1. \tag{2.10}$$

2.4 A Simulation Study

In this section, we conduct a simulation study with particular focus on the following two issues:

a. How does the ratio estimator $\hat{\mu}_R$ compare with $\hat{\mu}_R$ the RRT mean estimator $\hat{\mu}_Y$;

b. How do the *Bias* and *MSE* for the ratio estimator, correct up to second order of approximation, compare with the *Bias* and *MSE* expressions correct up to first order of approximation.

We considered 3 bivariate normal populations with different covariance matrices to represent the distribution of (Y, X). The scrambling variable *S* is taken to be a normal distribution with mean equal to zero and standard deviation equal to 10% of the standard deviation of *X*. The reported response is given by $Z = Y + S$.

All of the simulated populations have theoretical mean of $[Y, X]$ as $\mu = [2, 2]$ and covariance matrices as given below.

Population 1

$N = 1000$

$\Sigma = \begin{bmatrix} 9 & 1.9 \\ 1.9 & 4 \end{bmatrix}, \rho_{XY} = 0.3167.$

Population 2

$N = 1000$

$\Sigma = \begin{bmatrix} 10 & 3 \\ 3 & 2 \end{bmatrix}, \rho_{XY} = 0.6708.$

Population 3

$N = 1000$

$\Sigma = \begin{bmatrix} 6 & 3 \\ 3 & 2 \end{bmatrix}, \rho_{XY} = 0.8660.$

For each population we considered five sample sizes: $n = 20, 50, 100, 200$ and 300.

The Absolute Relative Bias (*ARB*) for the two estimators is given by $|Bias(\hat{\mu}_Y)/\bar{Y}|$ and $|Bias(\hat{\mu}_R)/\bar{Y}|$. We estimate the *ARB* using 5000 samples of size *n* selected from each population. The empirical *ARB* values for both estimators are given in Table 2.1. As expected, the *ARB* generally decreases as the sample size increases, with some exceptions due to random fluctuations.

The RRT mean estimator should generally perform better than the ratio estimator because this is an unbiased estimator. Nevertheless, the ratio estimator produces fairly good results.

Table 2.1: Empirical *ARB* for RRT mean estimator and ratio estimator (bold).

Population		**Empirical *ARB***				
N	ρ_{XY}	$n=20$	$n=50$	$n=100$	$n=200$	$n=300$
1000	0.3549	0.0021 **0.0223**	0.0011 **0.0071**	0.0010 **0.0057**	0.0018 **0.0006**	0.0016 **0.0009**
	0.6965	0.0010 **0.0193**	0.0021 **0.0061**	0.0014 **0.0015**	0.0014 **0.0029**	0.0011 **0.0021**
	0.8783	0.0012 **0.0181**	0.0013 **0.0063**	0.0008 **0.0023**	0.0013 **0.0026**	0.0011 **0.0020**

The theoretical *ARB* results for the ratio estimator, correct up to first and second degree of approximation, are presented in Table 2.2.

One can see that second order approximation as compared to first order approximation does not result in major difference in *ARB* even for modest sample size of $n = 20$ and 50.

Table 2.2: Theoretical *ARB* for ratio estimator based on 1^{st} and 2^{nd} order (bold) approximation.

Population		**Theoretical *ARB***				
N	ρ_{XY}	$n=20$	$n=50$	$n=100$	$n=200$	$n=300$
1000	0.3549	0.0224 **0.0258**	0.0087 **0.0092**	0.0041 **0.0042**	0.0018 **0.0019**	0.0011 **0.0011**
	0.6965	0.0155 **0.0167**	0.0060 **0.0062**	0.0029 **0.0029**	0.0013 **0.0013**	0.0007 **0.0007**
	0.8783	0.0142 **0.0153**	0.0055 **0.0057**	0.0026 **0.0026**	0.0012 **0.0012**	0.0007 **0.0007**

Table 2.3 below gives empirical and theoretical *MSE's* for the ratio estimator based on both the first order and second order approximations. As we see from the table, there is hardly a difference between the two approximations even for small samples. Hence the Percent Relative Efficiency (*PRE*) is calculated based on first order of approximation only. We use the following expression to find the *PRE* of ratio estimator as compared to the RRT mean estimator:

$$PRE = \frac{MSE(\hat{\mu}_Y)}{MSE(\hat{\mu}_R)} \times 100.$$

All the percent relative efficiencies are greater than 100 indicating that the ratio estimator is better than the RRT mean estimator.

There are small differences between MSE values based on first and second order approximation for smaller sample sizes (n=20 and 50) but the MSE values are very similar when the sample size is larger. We can also note that the ratio estimator gets more and more efficient as the coefficient of correlation between X and Y increases. We can further note that for small correlation values, the ratio estimator may not be better than the RRT mean estimator, particularly so if sample size is small.

Table 2.3: *MSE* correct up to 1^{st} and 2^{nd} order approximations and *PRE* for the ratio estimator relative to the RRT mean estimator.

Population			***MSE* Estimation**			***MSE* Condition**		***PRE***	
N	ρ_{XY}	n	Emprirical	1^{st}Order	2^{nd}Order	1^{st}Order[1]	2^{nd}Order[2]	1^{st}Order	2^{nd}Order
1000	0.3549	20	0.5782	0.4462	0.5249	0.0340	0.6947	104.86	89.15
		50	0.1837	0.1730	0.1848		0.9464		98.15
		100	0.0819	0.0820	0.0846		1.0304		101.57
		200	0.0358	0.0364	0.0370		1.0723		103.37
		500	0.0219	0.0219	0.0214		1.0863		103.398
	0.6965	20	0.3434	0.3036	0.3327	0.4785	2.9075	171.63	156.65
		50	0.1202	0.1177	0.1221		3.0887		165.49
		100	0.0548	0.0558	0.0568		3.1492		168.67
		200	0.0248	0.0248	0.0250		3.1794		170.30
		500	0.0152	0.0145	0.0145		3.1894		170.85
	0.8783	20	0.1178	0.1012	0.1139	0.5919	2.9424	309.31	274.89
		50	0.0406	0.0392	0.0412		3.0185		294.99
		100	0.0183	0.0186	0.0190		3.0439		302.36
		200	0.0083	0.0083	0.0083		3.0565		306.18
		500	0.0050	0.0048	0.0048		3.0608		307.48

[1] *MSE* comparison condition based on 1^{st} order approximation given in expression (2.9).

[2] *MSE* comparison condition based on 2^{nd} order approximation given in expression (2.10).

2.5 Numerical Example

We now compare the RRT mean estimator and the ratio estimator using a real data set. The data come from a sample from the survey on Information and Communication Technologies (ICT) usage in enterprises in 2009 with seat in Portugal (Smilhily and Storm, 2010). This survey intends to promote the development of the national statistical system in the information society and to contribute to a deeper knowledge about the usage of ICT by enterprises. The target population covers all industries with one and more persons employed in the sections of economic activity C (Manufacturing) to N (Administrative and support service activities) and S (Other service activities), from NACE[1] Rev. 2 (Eurostat, 2008). The data are essentially collected using Electronic Data Interchange, applying direct connection between information systems at the respondent and the National Statistics Institute. For some enterprises the paper questionnaire is still used. The

[1]NACE is derived from the French title "Nomenclature générale des Activités économiques dans les Communautés Européennes" (Statistical classification of economic activities in the European Communities).

questions in the structural business surveys mainly deal with characteristics that can be found in the organisations' annual reports and financial statements, such as employment, turnover and investment.

In our application the study variable Y is the purchase orders in 2009, collected by the ICT survey in that year. This is typically a confidential variable for enterprises, only known from business surveys. The auxiliary variable X is the turnover of each enterprise. This information can be easily obtained from enterprise records available in the public domain, as administrative information. In 2009 the population survey contained approximately 278000 enterprises and we know the value of X for all these enterprises. The purchase orders information was collected in the ICT survey and we have the values of Y for 5090 enterprises (which answered this question in the ICT survey in 2009). For this study, these 5090 enterprises are considered as our population. The scrambling variable S is taken to be a normal random variable with mean equal to zero and standard deviation equal to 10% of the satandard deviation of X, that is $\sigma_S = 0.1\sigma_X$. The reported response is given by $Z = Y + S$ (the purchase order value plus a random quantity).
The variables Y and X are strongly correlated so we can take advantage of this correlation by using the ratio estimator. In the next tables we present the results for the RRT mean estimator and for the ratio estimator for different sample sizes.

Population Characteristics:

$N = 5090, \rho_{XY} = 0.9832$
$\mu_X = 32.53, \mu_Y = 26.06, \sigma_X = 183.42, \sigma_Y = 67.07$ (in millions of Euros)
$\gamma_1^X = 31.54, \gamma_1^Y = 36.12, \gamma_2^X = 1481.08, \gamma_2^Y = 1839.13$

where γ_1 and γ_2 are the coefficients of *skewness* and *kurtosis*, respectively. We use the following samples sizes in our simulation study: $n = 100, 200, 300, 400, 500, 1000, 1500$ and 2000.

The empirical *ARB* values for both estimators, based on 5000 iterations, are given in Table 2.4. As expected, the bias decreases as the sample size increases, except for some random fluctuation. We expect the RRT mean estimator to perform better than the ratio estimator because this is an unbiased estimator, however, we don't see major differences between the two for larger samples.

Table 2.4: Empirical *ARB* for the RRT mean estimator and the ratio estimator (bold).

Population		**Empirical *ARB***							
N	ρ_{XY}	$n = 100$	$n = 200$	$n = 300$	$n = 400$	$n = 500$	$n = 1000$	$n = 1500$	$n = 2000$
5090	0.9832	0.0219 **0.0284**	0.0002 **0.0198**	0.0096 **0.0171**	0.0107 **0.0183**	0.0163 **0.0166**	0.0145 **0.0149**	0.0106 **0.0127**	0.0096 **0.0121**

The theoretical *ARB* results for the ratio estimator, correct up to first degree of approximation, are presented in Table 2.5. We use only the first order approximations from here on since the first and second order approximations are very similar, as we have seen earlier.

Table 2.5: Theoretical *ARB* for the RRT mean estimator and the ratio estimator.

Population		**Theoretica *ARB***							
N	ρ_{XY}	$n = 100$	$n = 200$	$n = 300$	$n = 400$	$n = 500$	$n = 1000$	$n = 1500$	$n = 2000$
5090	0.9832	0.0368	0.0180	0.0118	0.0086	0.0068	0.0030	0.0018	0.0011

Table 2.6 presents the results for the empirical *MSE* estimates, the theoretical estimates, correct up to first degree of approximation and the *PRE* of ratio estimator relative to the RRT mean estimator.

Table 2.6: *MSE*, corrected to 1^{st} order approximation, and *PRE* for the ratio estimator related to the RRT mean estimator.

Population			***MSE* Estimation**		***PRE***
N	ρ_{XY}	n	Empirical	Theoretical	
5090	0.9832	100	12.8924	15.2630	2286.36
		200	6.4608	7.4786	
		300	4.4498	4.8838	
		400	3.5279	3.5864	
		500	2.7380	2.8079	
		1000	1.4117	1.2510	
		1500	0.8805	0.7321	
		2000	0.6033	0.4726	

Clearly the ratio estimator performs better than the RRT mean estimator for the real data also. The effect of sample size on the *PRE* calculation is neutralized when first order approximation is used, as can be seen from Equations (2.2) and (2.8).

2.6 Transformed Ratio Estimators

Now consider the transformed ratio estimator:

$$\hat{\mu}_{TR} = \bar{z}\left(\frac{c\bar{X} + d}{c\bar{x} + d}\right), \tag{2.11}$$

where c and d are the unit-free parameters, which may be quantities such as the coefficient of *skewness* and coefficient of *kurtosis* for X. Many researchers have used transformed ratio estimators. These include Sisodia and Dwivedi (1981), Singh et al. (1973), Kulkarni (1977), Upadhyaya and Singh (1999), Upadhyaya et al. (2000) and Chandra and Singh (2005).

We can rewrite (2.11) using relative error terms in the form

$$\hat{\mu}_{TR} = \bar{z}\,(1+\delta_z)\,(1+\eta\delta_x)^{-1}\,, \tag{2.12}$$

where $\eta = \dfrac{c\bar{X}}{c\bar{X}+d}$.

Expanding (2.12), the *Bias*, correct up to first order of approximation, is given by

$$Bias^{(1)}(\hat{\mu}_{TR}) \cong \left(\frac{1-f}{n}\right)\bar{Y}\{\eta^2 C_x^2 - \eta\rho_{yx}C_yC_x\}. \tag{2.13}$$

By (2.6) and (2.13) $Bias^{(1)}(\hat{\mu}_{TR}) < Bias^{(1)}(\hat{\mu}_R)$ if

$$(\eta-1)\left\{\rho_{yx} - \frac{(\eta+1)C_x}{C_y}\right\} > 0. \tag{2.14}$$

Similarly *MSE* of $\hat{\mu}_{TR}$, to first order of approximation, is given by

$$MSE^{(1)}(\hat{\mu}_{TR}) \cong \left(\frac{1-f}{n}\right)\bar{Y}^2\left(C_y^2 + \eta^2C_x^2 - 2\eta\rho_{yx}C_yC_x\right). \tag{2.15}$$

By (2.8) and (2.15) $MSE^{(1)}(\hat{\mu}_{TR}) < MSE^{(1)}(\hat{\mu}_R)$ if

$$(\eta-1)\left\{\rho_{yx} - \frac{(\eta+1)C_x}{2C_y}\right\} > 0. \tag{2.16}$$

Now we conduct a simulation study with particular focus on the comparison between the ratio estimator $\hat{\mu}_R$ and the transformed ratio estimator $\hat{\mu}_{TR}$. We considered the same three bivariate normal populations as in the previous simulation study (Section 2.4).

The scrambling variable S is taken to be a normal random variable with mean equal to zero and the standard deviation equal to 10% of the standard deviation of X. The reported response is given by $Z = Y + S$. To compare these estimators, we present the results for the RRT mean estimator $(\hat{\mu}_Y)$, the ratio estimator $(\hat{\mu}_R)$ and for transformed ratio estimator $\hat{\mu}_{TRi}(i = 1, 2, 3, 4)$ with four different combinations of parameters c and d:

1. $\hat{\mu}_{TR1} = \bar{z}\left(\dfrac{c\bar{X}+d}{c\bar{x}+d}\right)$,
   where $c = 1$ and $d =$ coefficient of *skewness*;

2. $\hat{\mu}_{TR2} = \bar{z}\left(\dfrac{c\bar{X}+d}{c\bar{x}+d}\right)$,
   where $c = 1$ and $d =$ coefficient of *kurtosis*;

3. $\hat{\mu}_{TR3} = \bar{z}\left(\dfrac{c\bar{X}+d}{c\bar{x}+d}\right)$,
   where $c =$coefficient of *skewness* and $d =$ coefficient of *kurtosis*;

4. $\hat{\mu}_{TR4} = \bar{z}\left(\dfrac{c\bar{X} + d}{c\bar{x} + d}\right)$,
   where c =coefficient of *kurtosis* and $d =$ coefficient of *skewness*.

The empirical *ARB* values for these six estimators are given in Table 2.7.

Table 2.7: Empirical *ARB* for the RRT mean estimator, the ratio estimator and for the transformed ratio estimators.

Population			**Empirical *ARB***					
N	ρ_{XY}	n	$\hat{\mu}_Y$	$\hat{\mu}_R$	$\hat{\mu}_{TR1}$	$\hat{\mu}_{TR2}$	$\hat{\mu}_{TR3}$	$\hat{\mu}_{TR4}$
1000	0.3209	20	0.0002	0.0337	0.0435	0.0006	0.0026	0.0366
		50	0.0007	0.0118	0.0146	0.0002	0.0019	0.0126
		100	0.0003	0.0052	0.0065	0.0000	0.0009	0.0056
		150	0.0000	0.0032	0.0040	0.0001	0.0003	0.0035
		200	0.0012	0.0025	0.0030	0.0008	0.0016	0.0027
		300	0.0020	0.0041	0.0045	0.0023	0.0021	0.0043
	0.6746	20	0.0011	0.0122	0.0113	0.0111	0.0018	0.0119
		50	0.0004	0.0042	0.0038	0.0037	0.0015	0.0041
		100	0.0001	0.0022	0.0021	0.0021	0.0005	0.0022
		150	0.0005	0.0016	0.0015	0.0016	0.0002	0.0016
		200	0.0010	0.0005	0.0005	0.0001	0.0013	0.0005
		300	0.0015	0.0013	0.0014	0.0011	0.0016	0.0014
	0.8684	20	0.0006	0.0120	0.0115	0.0108	0.0013	0.0119
		50	0.0005	0.0041	0.0039	0.0036	0.0012	0.0040
		100	0.0001	0.0018	0.0017	0.0017	0.0005	0.0018
		150	0.0002	0.0010	0.0010	0.0012	0.0001	0.0010
		200	0.0009	0.0004	0.0004	0.0001	0.0011	0.0004
		300	0.0014	0.0010	0.0011	0.0010	0.0015	0.0010

The empirical *ARB* results in the Table 2.7 and the theoretical *ARB* results, to first degree of approximation, in the Table 2.8 indicate that the transformed ratio estimators do not produce major reductions in *ARB* as compared to the ratio estimator when sample size is large. Some reduction is observed for small sample size when using transformations where the additive parameter (d) is the *kurtosis*.

Table 2.8: Theoretical *ARB* to 1^{st} order approximation for the RRT mean estimator, the ratio estimator and for the transformed ratio estimators.

Population			**Theoretical *ARB* (1^{st} Order)**				
N	ρ_{XY}	n	$\hat{\mu}_R$	$\hat{\mu}_{TR1}$	$\hat{\mu}_{TR2}$	$\hat{\mu}_{TR3}$	$\hat{\mu}_{TR4}$
1000	0.3209	20	0.0248	0.0310	0.0017	0.0031	0.0267
		50	0.0096	0.0120	0.0006	0.0012	0.0103
		100	0.0046	0.0057	0.0003	0.0006	0.0049
		150	0.0029	0.0036	0.0002	0.0004	0.0031
		200	0.0020	0.0025	0.0001	0.0003	0.0022
		300	0.0012	0.0015	0.0001	0.0001	0.0013
	0.6746	20	0.0124	0.0116	0.0108	0.0031	0.0121
		50	0.0048	0.0045	0.0042	0.0012	0.0047
		100	0.0023	0.0021	0.0020	0.0006	0.0022
		150	0.0014	0.0013	0.0012	0.0004	0.0014
		200	0.0010	0.0009	0.0009	0.0003	0.0010
		300	0.0006	0.0006	0.0005	0.0001	0.0006
	0.8684	20	0.0123	0.0118	0.0108	0.0020	0.0121
		50	0.0048	0.0046	0.0042	0.0008	0.0047
		100	0.0023	0.0022	0.0020	0.0004	0.0022
		150	0.0014	0.0014	0.0013	0.0002	0.0014
		200	0.0010	0.0010	0.0009	0.0002	0.0010
		300	0.0006	0.0006	0.0005	0.0001	0.0006

Table 2.9 presents the results for the empirical *MSE* estimates and for the theoretical estimates, correct up to first order of approximation. Both results indicate that modest gains can be achieved by using transformations where the additive parameter (d) is the coefficient of *skewness*.

Table 2.9: Empirical *MSE* and theoretical (bold) *MSE* to 1^{st} order of approximation for the RRT mean estimator, the ratio estimator and for the transformed ratio estimators.

Population			***MSE***				
N	ρ_{XY}	n	$\hat{\mu}_R$	$\hat{\mu}_{TR1}$	$\hat{\mu}_{TR2}$	$\hat{\mu}_{TR3}$	$\hat{\mu}_{TR4}$
1000	0.3209	20	0.5799 **0.4496**	0.6584 **0.4672**	0.4097 **0.3994**	0.4686 **0.4663**	0.6010 **0.4548**
		50	0.1881 **0.1743**	0.2000 **0.1811**	0.1546 **0.1549**	0.1793 **0.1808**	0.1915 **0.1763**
		100	0.0872 **0.0826**	0.0914 **0.0858**	0.0750 **0.0734**	0.0875 **0.0857**	0.0884 **0.0835**
		150	0.0546 **0.0520**	0.0571 **0.0540**	0.0475 **0.0462**	0.0551 **0.0539**	0.0554 **0.0526**
		200	0.0395 **0.0367**	0.0412 **0.0381**	0.0343 **0.0326**	0.0394 **0.0381**	0.0400 **0.0371**
		300	0.0223 **0.0214**	0.0232 **0.0222**	0.0196 **0.0190**	0.0227 **0.0222**	0.0225 **0.0217**
	0.6746	20	0.3226 **0.2939**	0.3197 **0.2884**	0.3956 **0.3885**	0.5167 **0.5162**	0.3216 **0.2961**
		50	0.1165 **0.1140**	0.1148 **0.1118**	0.1497 **0.1507**	0.1979 **0.2002**	0.1159 **0.1132**
		100	0.0558 **0.0540**	0.0548 **0.0530**	0.0728 **0.0714**	0.0967 **0.0948**	0.0554 **0.0536**
		150	0.0352 **0.0340**	0.0346 **0.0333**	0.0460 **0.0449**	0.0608 **0.0597**	0.0350 **0.0338**
		200	0.0254 **0.0240**	0.0250 **0.0235**	0.0330 **0.0317**	0.0434 **0.0421**	0.0253 **0.0238**
		300	0.0145 **0.0140**	0.0142 **0.0137**	0.0189 **0.0185**	0.0250 **0.0246**	0.0144 **0.0139**
	0.8684	20	0.1117 **0.0984**	0.1083 **0.0947**	0.1973 **0.1920**	0.3119 **0.3113**	0.1106 **0.0971**
		50	0.0396 **0.0381**	0.0382 **0.0367**	0.0743 **0.0744**	0.1195 **0.1207**	0.0391 **0.0377**
		100	0.0188 **0.0181**	0.0181 **0.0174**	0.0361 **0.0353**	0.0584 **0.0572**	0.0186 **0.0178**
		150	0.0118 **0.0114**	0.0114 **0.0110**	0.0228 **0.0222**	0.0367 **0.0360**	0.0117 **0.0112**
		200	0.0086 **0.0080**	0.0083 **0.0077**	0.0164 **0.0157**	0.0263 **0.0254**	0.0085 **0.0079**
		300	0.0049 **0.0047**	0.0047 **0.0045**	0.0094 **0.0091**	0.0151 **0.0148**	0.0048 **0.0046**

Table 2.10 gives the *PRE* of various transformed ratio estimators relative to the ratio estimator based on first order approximation.

We can observe that the transformed ratio estimators that utilize the parameter d as coefficient of *skewness* result in higher *PRE* as compared to the ratio estimator when the correlation is larger. This was expected based on Condition (2.16) and Table 2.11 below.

Table 2.10: *PRE* for the transformed ratio estimator related to the ratio estimator based on 1^{st} order of approximation.

Population		***PRE*** **(1^{st} Order)**			
N	ρ_{XY}	$\hat{\mu}_{TR1}$	$\hat{\mu}_{TR2}$	$\hat{\mu}_{TR3}$	$\hat{\mu}_{TR4}$
	0.3209	96.24	112.56	96.41	98.86
1000	0.6746	101.93	75.65	56.94	100.64
	0.8684	103.87	51.24	31.60	101.28

Note that the transformed ratio estimator performs better than the ratio estimator when the condition in (2.16) is satisfied.

Table 2.11: Calculations for the expression in (2.16).

Population		Condition **(*MSE* - 1^{st} Order)**			
N	ρ_{XY}	$\hat{\mu}_{TR1}$	$\hat{\mu}_{TR2}$	$\hat{\mu}_{TR3}$	$\hat{\mu}_{TR4}$
	0.3209	-0.0299	0.0855	-0.0285	-0.0088
1000	0.6746	0.0127	-0.2160	-0.5074	0.0043
	0.8684	0.0108	-0.2751	-0.6259	0.0037

2.7 Conclusions

We can observe from this study that the estimation of the mean of a sensitive variable can be improved by using a non-sensitive auxiliary variable. The ratio estimators, in spite of being biased, can have much better *PRE* as compared to the RRT mean estimator. Our simulation study and the numerical example show that this improvement can be quite substantial if the correlation between the study variable and the auxiliary variable is high. We also note that there is hardly any difference in the *Bias* or *MSE* of the proposed estimator when using first or second order approximation. It is further noticed that the transformed ratio estimators produce very minimal gain over the ordinary ratio estimator.

References

CHANDRA, P. & SINGH, H. P. 2005. A family of estimators for population variance using knowledge of kurtosis of an auxiliary variable in sample surveys. *Statistics in Transition*, 7(1), 27-34.

EICHHORN, B. H. & HAYRE, L. S. 1983. Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*, 7, 307-316.

EUROSTAT. 2008. NACE Rev. 2 - Statistical classification of economic activities in the

European Community. *Official Publications of the European Communities*, 112-285 and 306-311.

GUPTA, S. N., GUPTA, B. C. & SINGH, S. 2002. Estimation of sensitivity level of personal interview survey questions. *Journal of Statistical Planning and Inference*, 100, 239-247.

GUPTA, S. & SHABBIR, J. 2004. Sensitivity estimation for personal interview survey questions. *Statistica*, 64, 643-653.

GUPTA, S., SHABBIR, J. & SEHRA, S. 2010. Mean and sensitivity estimation in optional randomized response models. *Journal of Statistical Planning and Inference*, 140(10), 2870-2874.

KADILAR, C. & CINGI, H. 2006. Improvement in estimating the population mean in simple random sampling. *Applied Mathematics Letters*, 19(1), 75-79.

KOYUNCU, N. & KADILAR, C. 2009. Efficient estimators for the population mean, *Hacettepe Journal of Mathematics and Statistics*, 38(2), 217-225.

KULKARNI, S. P. 1977. A note on modified ratio estimator using transformation, *Journal of the Indian Society of Agricultural Statistics*, 30(2), 125-128.

SAHA, A. 2008. A randomized response technique for quantitative data under unequal probability sampling. *Journal of statistical Theory and Practice*, 2(4), 589-596.

SHABBIR, J. & GUPTA, S. 2010. Estimation of the finite population mean in two-phase sampling when auxiliary variables are attribute, *Hacettepe Journal of Mathematics and Statistics*, 39(1), 121-129.

SINGH, J., PANDAY, B. N. & HIRANO, K. 1973. On the utilization of a known coefficient of kurtosis in the estimation procedure of variance. *Annals of the Institute of Statistical Mathematics*, 25, 51-55.

SINGH, H. P. & VISHWAKARMA, G. 2008. Some families of estimators of variance of stratified random sample mean using auxiliary information. *Journal of Statistical Theory and Practice*, 2(1), 21-43.

SISODIA, B. V. S. & DWIDEDI, V. K. 1981. A modified ratio estimator using coefficient of variation of auxiliary variable. *Journal of the Indian Society of Agricultural Statistics*, 33, 13-18.

SMILHILY, M. & STORM, H. 2010. ICT usage in enterprises - 2009, *Eurostat Publications*, Issue 1.

SUKHATME, P.V. & SUKHATME, B.V. 1984. *Sampling theory of surveys with applications*, 3^{rd} Ed., Ames, Iowa, Iowa State University Press.

TURGUT, Y. & CINGI, H. 2008. New generalized estimators for the population variance using auxiliary information. *Hacettepe Journal of Mathematics and Statistics*, 37(2), 177-184.

UPADHYAYA, L. N. & SINGH, H. P. 1999. Use of transformed auxiliary variable in estimating the finite population mean. *Biometrical Journal*, 41(5), 627-636.

UPADHYAYA L. N., SINGH, G. N. & SINGH, H. P. 2000. Use of transformed auxiliary variable in the estimation of population ratio in sample surveys. *Statistics in Transition*, 4(6), 1019-1027.

Appendix A - R Routines

Listing 2.1: R Code for Simulation Study of Proposed Estimator in Chapter 2

```

proj1 <- function(N,sigma,mu)
{
  set.seed(100)
  #Generation of a bivariate normal population
  data_yx <- mvrnorm(N, mu, sigma)

  #Study variable
  Y <- data_yx[,1]
  #Auxiliary variable, correlated with Y
  X <- data_yx[,2]

  #Coefficient of correlation between Y and X
  Ro_YX <- cor(Y,X)

  #Scrambling variable independent of Y and X, with mean=0
  S <- rnorm(N,mean=0,sd=0.1*sd(X))
  #Scrambled response
  Z <- Y+S

  #Coefficient of correlation between Z and X
  Ro_ZX <- Ro_YX/sqrt(1+(var(S)/var(Y)))

  #population
  univ <- data.frame(cbind(Y=Y,S=S,Z=Z,X=X,NRAND=runif(N)))
  univ <- univ[order(univ$NRAND),]

  #Mean of Y
  my <- mean(univ$Y)
  mz <- mean(univ$Z)
  mx <- mean(univ$X)
  ms <- mean(univ$S)

  #Sample dimension
  dim_samp <- c(20,50,100,200,300)

  res <- NULL
  for (i in 1:length(dim_samp))
  {
    #sample dimension
    n <- dim_samp[i]
    #sample
    samp <- univ[1:n,]
    #sampling rate
    f <- n/N

```

```
    #estimators
    est1 <- mean(samp$Z)
    est2 <- mean(samp$Z)*(mean(univ$X)/mean(samp$X))

    #Ratio
    R <- mean(univ$X)/mean(samp$X)

    #Mean Square Error of 1st estimator
    mse1 <- ((1-f)/n)*var(univ$Z)

    #Coefficient of variation
    c_x <- sd(univ$X)/mx
    c_y <- sd(univ$Y)/my
    c2_x <- c_x^2
    c2_y <- c_y^2
    c2_z <- c2_y+(var(univ$S)/(my^2))
    c_z <- sqrt(c2_z)

    #Bias of ratio estimator - 1st degree approximation
    bias2i <- ((1-f)/n)*my*(c2_x-Ro_ZX*c_z*c_x)
    #Bias of ratio estimator - 2nd degree approximation
    bias2ii <- bias2i*(1+((1-f)/n)*3*c2_x)

    #Mean Square Error of ratio estimator - 1st degree approximation
    mse2i <- ((1-f)/n)*(my^2)*(c2_z+c2_x-2*Ro_ZX*c_z*c_x)
    #Mean Square Error of ratio estimator - 2nd degree approximation
    mse2ii <- mse2i+3*(my^2)*(((1-f)/n)^2)*c2_x*((1+2*(Ro_ZX^2))*c2_z
        +3*c2_x-6*Ro_ZX*c_z*c_x)

    aux_bias <- (c_x-Ro_ZX*c_z)

    aux_mse1 <- (Ro_ZX-(1/2)*(c_x/c_z))
    aux_mse2 <- 2*Ro_ZX*(c_z/c_x)-3*((1-f)/n)*((1+2*(Ro_ZX^2))*c2_z
          +3*c2_x-6*Ro_ZX*c_z*c_x)

    emp <- NULL

    #Empirical results
    #Simulation of 5000 replicas of estimates
    ...

    #Results
    res <- rbind(res,c(N,n,Ro_YX,Ro_ZX,R,my,mz,ms,
        med_est1,med_est2,bias2i,bias2ii,emp_mse1,mse1,
        emp_mse2,mse2i,mse2ii,aux_bias,aux_mse1,aux_mse2))
  }
  colnames(res) <- c("N","n","RhoXY","RhoZX","R","mY","mZ","mS",
      "Est1","Est2","BIAS2I","BIAS2II","EMP_MSE1","MSE1",
      "EMP_MSE2","MSE2I","MSE2II","AUX_BIAS","AUX_MSE1","AUX_MSE2")
  return(res)
```

```
}

#Package for generation
require(MASS)
N<-1000
#Parameters
sigma1 <- matrix(c(9,1.9,1.9,4),2,2)
sigma2 <- matrix(c(10,3,3,2),2,2)
sigma3 <- matrix(c(6,3,3,2),2,2)
mu <- c(2,2)

res <- NULL
for (i in 1:length(N))
{
  res <- rbind(res,proj1(N[i],sigma1,mu))
  res <- rbind(res,proj1(N[i],sigma2,mu))
  res <- rbind(res,proj1(N[i],sigma3,mu))
}
write.table(res,"chapter2_ss_results1.txt",sep="\t",dec=",",row.names=FALSE)
```

Listing 2.2: R Code for Simulation Study of Transformed Ratio Estimators in Chapter 2

```

mykurtosis <- function(x)
{
  m4 <- mean((x-mean(x))^4)
  kurt <- m4/(sd(x)^4)
  return(kurt)
}
myskewness <- function(x)
{
  m3 <- mean((x-mean(x))^3)
  skew <- m3/(sd(x)^3)
  return(skew)
}
proj1_transf <- function(N,sigma,mu)
{

  #Generation of a bivariate normal population
  data_yx <- mvrnorm(N, mu, sigma)

  #Study variable
  Y <- data_yx[,1]
  #Auxiliary variable, correlated with Y
  X <- data_yx[,2]

  #Coefficient of correlation between Y and X
  Ro_YX <- cor(Y,X)

  #Scrambling variable independent of Y and X, with mean=0
  S <- rnorm(N,mean=0,sd=0.1*sd(X))
  #Scrambled response
  Z <- Y+S

  #Coefficient of correlation between Z and X
  Ro_ZX <- Ro_YX/sqrt(1+(var(S)/var(Y)))

  #population
  univ <- data.frame(cbind(Y=Y,S=S,Z=Z,X=X,NRAND=runif(N)))
  univ <- univ[order(univ$NRAND),]

  #Mean of Y
  my <- mean(univ$Y)
  mz <- mean(univ$Z)
  mx <- mean(univ$X)

  #Samples dimension
  dim_samp <- c(20,50,100,150,200,300)

```

```
res <- NULL
for (i in 1:length(dim_samp))
{
  #sample dimension
  n <- dim_samp[i]
  #sample
  samp <- univ[1:n,]
  #sampling rate
  f <- n/N

  #Ratio
  R <- mean(univ$X)/mean(samp$X)

  #Ordinary meam
  est1 <- mean(samp$Z)
  #Ratio estimator
  est2 <- mean(samp$Z)*(mean(univ$X)/mean(samp$X))

  #Coefficient of variation
  c_x <- sd(univ$X)/mx
  c_y <- sd(univ$Y)/my
  c2_x <- c_x^2
  c2_y <- c_y^2
  c2_z <- c2_y+(var(univ$S)/(my^2))
  c_z <- sqrt(c2_z)

  #Bias of ratio estimator - 1st degree approximation
  bias2i <- ((1-f)/n)*my*(c2_x-Ro_ZX*c_z*c_x)
  #Bias of ratio estimator - 2nd degree approximation
  bias2ii <- bias2i*(1+((1-f)/n)*3*c2_x)

  #Mean Square Error of 1st estimator (ordinal mean)
  mse1 <- ((1-f)/n)*(var(univ$Y)+var(univ$S))

  #Mean Square Error of ratio estimator - 1st degree approximation
  mse2i <- ((1-f)/n)*(my^2)*(c2_z+c2_x-2*Ro_ZX*c_z*c_x)
  #Mean Square Error of ratio estimator - 2nd degree approximation
  mse2ii <- mse2i+3*(my^2)*(((1-f)/n)^2)*c2_x*((1+2*(Ro_ZX^2))*c2_z
      +3*c2_x-6*Ro_ZX*c_z*c_x)

  nu <- 1
  aux_m <- c2_x-2*Ro_ZX*c_z*c_x

  s <- myskewness(univ$X)
  k <- mykurtosis(univ$X)

  vc <- c(1,1,1,s,k)
  vd <- c(s,k,Ro_YX,k,s)

  #Initialize the variables est3, mse3i, ...
```

```

    for (i in 1:length(vc))
    {
      nu <- (vc[i]*mean(univ$X))/(vc[i]*mean(univ$X)+vd[i])
      vnu <- c(vnu,nu)

      aux_bias1 <- (nu-1)*(Ro_ZX-(nu+1)*c_x/c_z)
      aux_mse1 <- (nu-1)*(Ro_ZX-(nu+1)*c_x/(2*c_z))

      vb1 <- c(vb1,aux_bias1)
      vm1 <- c(vm1,aux_mse1)

      #Transformed ratio estimator
      est3 <- c(est3,mean(samp$Z)*(vc[i]*mean(univ$X)+vd[i])
              /(vc[i]*mean(samp$X)+vd[i]))

      #Mean Square Error of transformed ratio estimator
      #1st degree approximation
      mse3i <- c(mse3i,((1-f)/n)*(my^2)*(c2_z+(nu^2)*c2_x-2*nu*Ro_ZX*c_z*c_x))

      #Mean Square Error of transformed ratio estimator
      #1st degree approximation
      mse3ii <- c(mse3ii,mse3i[i]+3*(my^2)
                 *(((1-f)/n)^2)*c2_x*((nu^2)*(1+2*(Ro_ZX^2))
                 *c2_z+3*(nu^4)*c2_x-6*(nu^3)*Ro_ZX*c_z*c_x))

      #Bias of transformated ratio estimator - 1st degree approximation
      bias3i <- c(bias3i,((1-f)/n)*my*((nu^2)*c2_x-nu*Ro_ZX*c_z*c_x))
      #Bias of transformated ratio estimator - 2nd degree approximation
      bias3ii <- c(bias3ii,bias3i[i]
                  +(((1-f)/n)^2)*3*my*((nu^4)*(c2_x^2)-(nu^3)
                  *Ro_ZX*c_z*(c_x^3)))
    }

    #Empirical results
    #Simulation of 5000 replicas of estimates
    ...

    #Results
    res <- rbind(res,c(N,n,Ro_YX,Ro_ZX,R,my,med_est1,med_est2,
                  med_est3,
                  bias2i,bias2ii,
                  bias3i,
                  bias3ii,
                  emp_mse1,mse1,emp_mse2,mse2i,mse2ii,
                  emp_mse3,
                  mse3i,
                  mse3ii,
                  vnu,
                  vb1,
```

```
                   vm1))
  }
  colnames(res) <- c("N","n","RhoXY","RhoZX","R","mY","Est1","Est2",
                     paste("Est3_",1:length(vc),sep=""),
                     "BIAS2I","BIAS2II",
                     paste("BIAS3I_",1:length(vc),sep=""),
                     paste("BIAS3II_",1:length(vc),sep=""),
                     "EMP_MSE1","MSE1","EMP_MSE2","MSE2I","MSE2II",
                     paste("EMP_MSE3_",1:length(vc),sep=""),
                     paste("MSE3I_",1:length(vc),sep=""),
                     paste("MSE3II_",1:length(vc),sep=""),
                     paste("NU_",1:length(vc),sep=""),
                     paste("AUX3_BIAS1_",1:length(vc),sep=""),
                     paste("AUX3_MSE1_",1:length(vc),sep=""))
  return(res)
}
#Package for generation
require(MASS)
N <- 1000

#Parameters
sigma1 <- matrix(c(9,1.9,1.9,4),2,2)
sigma2 <- matrix(c(10,3,3,2),2,2)
sigma3 <- matrix(c(6,3,3,2),2,2)
mu <- c(2,2)

res <- NULL
for (i in 1:length(N))
{
  res <- rbind(res,proj1_transf(N[i],sigma1,mu))
  res <- rbind(res,proj1_transf(N[i],sigma2,mu))
  res <- rbind(res,proj1_transf(N[i],sigma3,mu))
}
write.table(res,"chapter2_ss_results2.txt",sep="\t",dec=",",row.names=FALSE)
```

Listing 2.3: R Code for Numerical Example of Proposed Estimator in Chapter 2

```

proj1_real <- function(Y,X,N)
{

  #Coefficient of correlation between Y and X
  Ro_YX <- cor(Y,X)

  #Scrambling variable independent of Y and X, with mean=0
  S <- rnorm(N,mean=0,sd=sd(X)*0.1)
  #Scrambled response
  Z <- Y+S

  #Coefficient of correlation between Z and X
  Ro_ZX <- Ro_YX/sqrt(1+(var(S)/var(Y)))

  #population
  univ <- data.frame(cbind(Y=Y,S=S,Z=Z,X=X,NRAND=runif(N)))
  univ <- univ[order(univ$NRAND),]

  #Mean of Y
  my <- mean(univ$Y)
  mz <- mean(univ$Z)
  mx <- mean(univ$X)

  #Samples dimension
  dim_samp <- c(100,200,300,400,500,1000,1500,2000)

  res <- NULL
  for (i in 1:length(dim_samp))
  {
    #sample dimension
    n <- dim_samp[i]
    #sample
    samp <- univ[1:n,]
    #Sampling rate
    f <- n/N

    #estimators
    est1 <- mean(samp$Z)
    est2 <- mean(samp$Z)*(mean(univ$X)/mean(samp$X))

    #Ratio
    R <- mean(univ$X)/mean(samp$X)

    #Mean Square Error of 1st estimator
    mse1 <- ((1-f)/n)*(var(univ$Y)+var(univ$S))

```

```
    #Coefficient of variation
    c_x <- sd(univ$X)/mx
    c_y <- sd(univ$Y)/my
    c2_x <- c_x^2
    c2_y <- c_y^2
    c2_z <- c2_y+(var(univ$S)/(my^2))
    c_z <- sqrt(c2_z)

    #Bias of ratio estimator - 1st degree approximation
    bias2i <- ((1-f)/n)*my*(c2_x-Ro_ZX*c_z*c_x)
    #Bias of ratio estimator - 2nd degree approximation
    bias2ii <- bias2i*(1+((1-f)/n)*3*c2_x)

    #Mean Square Error of ratio estimator - 1st degree approximation
    mse2i <- ((1-f)/n)*(my^2)*(c2_z+c2_x-2*Ro_ZX*c_z*c_x)
    #Mean Square Error of ratio estimator - 2nd degree approximation
    mse2ii <- mse2i+3*(my^2)*(((1-f)/n)^2)*c2_x*((1+2*(Ro_ZX^2))*c2_z
        +3*c2_x-6*Ro_ZX*c_z*c_x)

    aux_bias <- (c_x-Ro_ZX*c_z)

    aux_mse1 <- (Ro_ZX-(1/2)*(c_x/c_z))
    aux_mse2 <- 2*Ro_ZX*(c_z/c_x)-3*((1-f)/n)*((1+2*(Ro_ZX^2))*c2_z
          +3*c2_x-6*Ro_ZX*c_z*c_x)

    #Empirical results
    #Simulation of 5000 replicas of estimates
    ...

    #Results
    res <- rbind(res,c(N,n,Ro_YX,Ro_ZX,R,my,med_est1,med_est2,
        bias2i,bias2ii,emp_mse1,mse1,emp_mse2,mse2i,mse2ii,
        aux_bias,aux_mse1,aux_mse2))
  }
  colnames(res) <- c("N","n","RhoXY","RhoZX","R","mY","Est1","Est2",
            "BIAS2I","BIAS2II","EMP_MSE1","MSE1","EMP_MSE2",
            "MSE2I","MSE2II","AUX_BIAS","AUX_MSE1","AUX_MSE2")
  return(res)
}

#Package for generation
require(MASS)

#Import data
data_yx <- read.table("IUTICE09.dat",sep="\t",dec=",",header = T)
#Study variable (purchase, millions of euros)
Y <- data_yx[,3]
#Auxiliary variable, correlated with Y (turnover, millions of euros)
X <- data_yx[,2]
```

```
#Data application
N <- dim(data_yx)[1]
res <- proj1_real(Y,X,N)

#Export data
write.table(res,"chapter2_ne_results",sep="\t",dec=",",row.names=FALSE)
```

3

Estimation of the Mean of a Sensitive Variable in the Presence of Auxiliary Information

Abstract

Sousa et al. (2010) introduced a ratio estimator for the mean of a sensitive variable and showed that this estimator performs better than the ordinary mean estimator based on a Randomized Response Technique (RRT). In this paper, we introduce a regression estimator that performs better than the ratio estimator even for modest correlation between the primary and the auxiliary variables. The underlying assumption is that the primary variable is sensitive in nature but a non-sensitive auxiliary variable exists that is positively correlated with the primary variable. Expressions for the *Bias* and Mean Square Error (*MSE*) are derived based on the first order of approximation. It is shown that the proposed regression estimator performs better than the ratio estimator and the ordinary RRT mean estimator (that does not utilize the auxiliary information). We also consider a generalized regression-cum-ratio estimator that has even smaller *MSE*. An extensive simulation study is presented to evaluate the performances of the proposed estimators in relation to other estimators in the study.

Published as: GUPTA, S., SHABBIR, J., SOUSA, R. & REAL, P.C. 2012. Estimation of the Mean of a Sensitive Variable in the Presence of Auxiliary Information. *Communications in Statistics - Theory and Methods*, 41(13-14), 2394-2404.

The procedure is also applied to some financial data: purchase orders (a sensitive variable) and gross turnover (a non-sensitive variable) in 2009 for a population of 5336 companies in Portugal from a survey on Information and Communication Technologies (ICT) usage.

3.1 Introduction

In survey research, direct reliable observation on the variable of interest (Y) is sometimes not possible because the variable may be sensitive in nature. In this paper we focus on estimating the mean of a sensitive variable Y using an auxiliary variable (X) that can be directly observed and that is correlated with the variable of the interest. For example, Y may be the total number of abortions a woman of child bearing age might have had and X may be her current age. Similarly, Y may be the total value of purchase orders in a year for a company and X may be the total turnover for that company in that year. In such situations, mean of Y can be estimated by using one of many randomized response techniques if the auxiliary information is to be ignored.

Many authors have estimated the mean of a sensitive variable when the primary variable is sensitive and there is no auxiliary variable available. These include Eichhorn and Hayre (1983), Gupta and Shabbir (2004), Gupta et al. (2002, 2010), Wu et al. (2008), Saha (2008) and Perri (2008). Also, many authors have presented ratio and regression estimators when both Y and X are directly observable. These include Kadilar and Cingi (2005), Kadilar et al. (2007), Shabbir and Gupta (2007, 2010) and Nangsue (2009).

In this paper, we propose a regression estimator where the RRT estimator of the mean of Y is further improved by using an auxiliary variable X. We also consider a generalized regression-cum-ratio estimator under the same conditions. Expressions for the *Bias* and *MSE* for the proposed estimators are derived, correct up to first order of approximation. We compare the performances of the proposed estimators with those of the ratio and the ordinary RRT mean estimators. We observe that there is considerable reduction in *MSE*, particularly when the correlation between the study variable and the auxiliary variable is high.

3.2 Terminology

Let Y be the study variable, a sensitive variable which cannot be observed directly due to respondent bias. Let X be a non-sensitive auxiliary variable which has a positive correlation with Y. Let S be a scrambling variable independent of Y and X. The respondent is asked to report a scrambled response for Y given by $Z = Y + S$ but is asked to provide a true response for X. Let a random sample of size n be drawn without replacement from a finite population $U = (U_1, U_2, ..., U_N)$. For the i^{th} unit $(i = 1, 2, ..., N)$, let y_i and x_i, respectively, be the values of the study variable Y and auxiliary variable X. Let $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$, $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ and $\bar{z} = \frac{\sum_{i=1}^{n} z_i}{n}$ be the sample means and $\bar{Y} = E(Y)$, $\bar{X} = E(X)$, and $\bar{Z} = E(Z)$ be the corresponding population means for Y, X and Z, respectively. We assume that $\bar{X}$ is known and $\bar{S} = E(S) = 0$. Thus, $E(Z) = E(Y)$ and $C_z^2 = C_y^2 + \frac{S_s^2}{\bar{Y}^2}$, where C_z and C_y are the coefficients of variation of z and y, respectively. If $e_0 = \frac{\bar{z}-\bar{Z}}{\bar{Z}}$, $e_1 = \frac{\bar{x}-\bar{X}}{\bar{X}}$, $e_2 = \frac{s_x^2 - S_x^2}{S_x^2}$, and $e_3 = \frac{s_{zx} - S_{zx}}{S_{zx}}$, then we have $E(e_i) = 0$, $i = 0, 1, 2, 3$.

If information on X is ignored, then an unbiased estimator of μ_Y is the ordinary sample mean $(\bar{z})$ given by (3.1) below.

$$\hat{\mu}_Y = \bar{z} \tag{3.1}$$

The mean square error (*MSE*) of $\hat{\mu}_Y$ is given by

$$MSE(\hat{\mu}_Y) = \frac{1-f}{N}\left(S_y^2 + S_s^2\right), \tag{3.2}$$

where $f = n/N$, $S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$ and $S_s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(s_i - \bar{S})^2$.

3.3 The Ratio Estimator

Sousa et al. (2010) proposed a ratio estimator for the mean of sensitive variable (Y) utilizing information from a non-sensitive auxiliary variable (X). This estimator is given by

$$\hat{\mu}_R = \bar{z}\left(\frac{\bar{X}}{\bar{x}}\right). \tag{3.3}$$

Bias and *MSE* of $\hat{\mu}_R$, correct up to first order of approximation, are given by

$$Bias(\hat{\mu}_R) \cong \left(\frac{1-f}{n}\right)\bar{Y}\left(C_x^2 - \rho_{zx}C_zC_x\right) \tag{3.4}$$

and

$$MSE(\hat{\mu}_R) \cong \left(\frac{1-f}{n}\right)\bar{Y}^2\left(C_z^2 + C_x^2 - 2\rho_{zx}C_zC_x\right). \tag{3.5}$$

It can be observed that $MSE(\hat{\mu}_R) < MSE(\hat{\mu}_Y)$ if

$$\rho_{yx} > \frac{1}{2}\frac{C_x}{C_y}\sqrt{1 + \frac{S_s^2}{S_y^2}}. \tag{3.6}$$

3.4 Ordinary Regression Estimator

Assuming linear relationship between Y and X, we propose the following regression estimator for the population mean of Y

$$\hat{\mu}_{Reg} = \bar{z} + \hat{\beta}_{zx}\left(\bar{X} - \bar{x}\right), \tag{3.7}$$

where $\hat{\beta}_{zx}$ is the sample regression coefficient between Z and X and $Z = Y + S$ is the scrambled response on Y. Using Taylor's approximation and retaining terms of order up to 2, (3.7) can be rewritten as

$$\hat{\mu}_{Reg} - \bar{Z} \cong \bar{Z}e_0 - \beta_{zx}\bar{X}\left[e_1 + e_1e_3 - e_1e_2\right]. \tag{3.8}$$

From Mukhopadhyay (1998, p. 123), we have:

$E(e_0^2)=\dfrac{1-f}{n}C_z^2$, $E(e_1^2)=\dfrac{1-f}{n}C_x^2$, $E(e_{12})=\dfrac{1-f}{n}\dfrac{1}{\bar{X}}\dfrac{\mu_{03}}{\mu_{02}}$, $E(e_{13})=\dfrac{1-f}{n}\dfrac{1}{\bar{X}}\dfrac{\mu_{12}}{\mu_{11}}$,
where $\mu_{rs} = \dfrac{1}{N-1}\sum_{i=1}^{N}(z_i - \bar{Z})^r(x_i - \bar{X})^s$ and C_x, C_z are the coefficients of variation of x and z, respectively.
Also we have: $\beta_{zx} = \dfrac{S_{zx}}{S_x^2} = \dfrac{S_{yx}}{S_x^2} = \rho_{yx}\dfrac{S_y}{S_x} = \beta_{yx}$, $\rho_{zx} = \dfrac{\rho_{yx}}{\sqrt{1+\frac{S_s^2}{S_y^2}}}$,
where ρ_{yx} and ρ_{zx} are the coefficients of correlation between y and x, and between z and x, respectively.

Recognizing that $\bar{Z} = \bar{Y}$ in Equation (3.8), the *Bias* and *MSE* of $\hat{\mu}_{Reg}$, to first order of approximation, are given by

$$Bias(\hat{\mu}_{Reg}) \cong -\beta_{zx}\left(\frac{1-f}{n}\right)\left\{\frac{\mu_{12}}{\mu_{11}} - \frac{\mu_{03}}{\mu_{02}}\right\} \tag{3.9}$$

and

$$MSE(\hat{\mu}_{Reg}) \cong \left(\frac{1-f}{n}\right)\bar{Y}^2C_z^2\left(1-\rho_{zx}^2\right) = \left(\frac{1-f}{n}\right)S_y^2\left\{\left(1+\frac{S_s^2}{S_y^2}\right) - \rho_{yx}^2\right\}. \tag{3.10}$$

It can be verified easily that

(i) $MSE(\hat{\mu}_{Reg}) < MSE(\hat{\mu}_Y)$ if

$$\rho_{yx}^2 > 0; \tag{3.11}$$

(ii) $MSE(\hat{\mu}_{Reg}) < MSE(\hat{\mu}_R)$ if

$$(C_x - C_z\rho_{zx})^2 > 0. \tag{3.12}$$

These conditions will always hold true indicating that up to first order of approximation, the regression estimator performs better than $\hat{\mu}_Y$ and $\hat{\mu}_R$.

3.5 Generalized Regression-cum-ratio Estimator

Many authors have used regression-cum-ratio estimators that combine both the regression estimator and ratio estimator. These include Ray and Singh (1981), Perri (2004), and Kadilar and Cingi (2004, 2006). We consider a similar hybrid estimator, as a generalized regression-cum-ratio estimator with constant coefficients whose values are to be determined later from optimality considerations. The main idea is to see if further gains can be achieved by using a generalized regression-cum-ratio estimator, as compared to regression estimator given by (3.7). This estimator is given by:

$$\hat{\mu}_{GRR} = \left[k_1 \bar{z} + k_2 \left(\bar{X} - \bar{x}\right)\right] \left(\frac{\bar{X}}{\bar{x}}\right), \tag{3.13}$$

where k_1 and k_2 are constants.

Solving (3.13) using Taylor's approximation and retaining terms of order up to 2, we have

$$\hat{\mu}_{GRR} - \bar{Y} \cong (k_1 - 1)\bar{Y} + k_1 \bar{Y}(e_0 - e_1 - e_0 e_1 + e_1^2) - k_2 \bar{X}(e_1 - e_1^2). \tag{3.14}$$

From (3.14), the *Bias* and *MSE* of $\hat{\mu}_{GRR}$ to first order of approximation are given by

$$Bias(\hat{\mu}_{GRR}) \cong (k_1 - 1)\bar{Y} + k_1 \bar{Y}\left(\tfrac{1-f}{n}\right)\left\{C_x^2 - \rho_{zx} C_z C_x\right\} + k_2 \bar{X}\left(\tfrac{1-f}{n}\right) C_x^2 \tag{3.15}$$

and

$$\begin{aligned} MSE(\hat{\mu}_{GRR}) \cong\ & (k_1 - 1)^2 \bar{Y}^2 + k_1^2 \bar{Y}^2 \left(\tfrac{1-f}{n}\right)\left\{C_z^2 + 3C_x^2 - 4\rho_{zx} C_z C_x\right\} \\ & + k_2^2 \bar{X}^2 \left(\tfrac{1-f}{n}\right) C_x^2 - 2k_1 \bar{Y}^2 \left(\tfrac{1-f}{n}\right)\left\{C_x^2 - \rho_{zx} C_z C_x\right\} \\ & - 2k_2 \bar{Y}\bar{X}\left(\tfrac{1-f}{n}\right) C_x^2 - 2k_1 k_2 \bar{Y}\bar{X}\left(\tfrac{1-f}{n}\right)\left\{\rho_{zx} C_z C_x - 2C_x^2\right\}. \end{aligned} \tag{3.16}$$

Differentiating (3.16) with respect to k_1 and k_2 we get the following optimum values:

$$k_{1(opt)} = \frac{1 - \left(\tfrac{1-f}{n}\right) C_x^2}{1 - \left(\tfrac{1-f}{n}\right)\left\{C_x^2 - C_z^2\left(1 - \rho_{zx}^2\right)\right\}} \tag{3.17}$$

and

$$k_{2(opt)} = \frac{\bar{Y}}{\bar{X}}\left\{1 + k_{1(opt)}\left(\frac{\rho_{zx} C_z}{C_x} - 2\right)\right\}. \tag{3.18}$$

which minimize the *MSE*.

Substituting the optimum values of k_1 and k_2 in (3.16), we get

$$MSE(\hat{\mu}_{GRR})_{min} \cong \frac{\bar{Y}^2 C_z^2\left(1-\rho_{zx}^2\right)\left(\frac{1-f}{n}\right)\left\{1-\left(\frac{1-f}{n}\right)C_x^2\right\}}{C_z^2\left(1-\rho_{zx}^2\right)\left(\frac{1-f}{n}\right)+\left\{1-\left(\frac{1-f}{n}\right)C_x^2\right\}}. \tag{3.19}$$

It can be verified that:

(i) $MSE(\hat{\mu}_{GRR})_{min} < MSE(\hat{\mu}_Y)$ if

$$\left(\frac{1-f}{n}\right)\left\{S_y^2+S_s^2\right\} > 0, \tag{3.20}$$

which is always true.

(ii) $MSE(\hat{\mu}_{GRR})_{min} < MSE(\hat{\mu}_R)$ if

$$\left(\frac{C_x}{C_z}-\rho_{zx}\right)^2+\frac{\left(\frac{1-f}{n}\right)C_z^2\left(1-\rho_{zx}^2\right)^2}{\left(\frac{1-f}{n}\right)C_z^2\left(1-\rho_{zx}^2\right)+\left(1-\left(\frac{1-f}{n}\right)C_x^2\right)} > 0. \tag{3.21}$$

(iii) $MSE(\hat{\mu}_{GRR})_{min} < MSE(\hat{\mu}_{Reg})$ if

$$\left(\frac{1-f}{n}\right)C_z^2\left(1-\rho_{zx}^2\right) > 0, \tag{3.22}$$

which is always true.

From these conditions we can conclude that the generalized estimator in (3.13) with optimal coefficients is always better than $\hat{\mu}_Y$, and $\hat{\mu}_{Reg}$. It is also better than the ratio and regression estimators if

$$1-\left(\frac{1-f}{n}\right)C_x^2 > 0$$

or

$$\left(\frac{1-f}{n}\right)C_x^2 < 1. \tag{3.23}$$

The last condition is very likely to hold true. So, with this generalized regression-cum-ratio estimator, we may be able to achieve further gain in terms of *MSE*, as can be observed from the simulation results in the next section.

3.6 The Simulation Study

In this section, we present results of a simulation study with particular focus on the performance for the regression estimator $\hat{\mu}_{Reg}$ and the proposed generalized regression-cum-ratio estimator $\hat{\mu}_{GRR}$ as compared to the RRT mean estimator $\hat{\mu}_Y$, and the ratio estimator $\hat{\mu}_R$.

We consider three finite sub-populations of size 1000 each from bivariate normal populations with different covariance matrices to represent the distribution of (Y, X). The scrambling variable S is taken to be a normal variate with mean equal to zero and standard deviation equal to 10% of the standard deviation of X. The reported response is given by $Z = Y + S$.

All of the simulated populations have theoretical mean of $[Y, X]$ as $\mu = [2, 2]$. The covariance matrices (Σ) are as given below.

Population 1

$$\Sigma = \begin{bmatrix} 9 & 1.9 \\ 1.9 & 4 \end{bmatrix}, \rho_{XY} = 0.3209.$$

Population 2

$$\Sigma = \begin{bmatrix} 10 & 3 \\ 3 & 2 \end{bmatrix}, \rho_{XY} = 0.6746.$$

Population 3

$$\Sigma = \begin{bmatrix} 6 & 3 \\ 3 & 2 \end{bmatrix}, \rho_{XY} = 0.8684.$$

For each population, we consider five sample sizes: $n = 50, 100, 200$ and 300.

Table 3.1 below gives empirical and theoretical *MSE*'s for various estimators based on the first order approximation. We estimate the empirical *MSE* using 5000 samples of various sizes selected from each population. We use the following expression to find the *PRE* of ratio, regression and generalized regression-cum-ratio estimators as compared to the RRT mean estimator:

$$PRE = \frac{MSE(\hat{\mu}_Y)}{MSE(\hat{\mu}_\alpha)} \times 100,$$

where $\alpha = R, Reg, GRR$.

Table 3.1: *MSE* correct up to 1^{st} order approximation and *PRE* for the ratio estimator ($\hat{\mu}_R$), the regression estimator ($\hat{\mu}_{Reg}$) and the generalized regression-cum-ratio estimator ($\hat{\mu}_{GRR}$) relative to the RRT mean estimator.

Population				***MSE* Estimation**			
N	ρ_{XY}	n	Estimator	Empirical	Theoretical	*PRE*	Condition[1]
1000	0.3209	50	$\hat{\mu}_R$	0.1885	0.1743	98.62	0.0186
			$\hat{\mu}_{Reg}$	0.1557	0.1543	111.43	
			$\hat{\mu}_{GRR}$	0.1571	0.1485	115.77	
		100	$\hat{\mu}_R$	0.0877	0.0826	98.62	0.0088
			$\hat{\mu}_{Reg}$	0.0733	0.0731	111.43	
			$\hat{\mu}_{GRR}$	0.0736	0.0718	113.46	
		200	$\hat{\mu}_R$	0.0382	0.0367	98.62	0.0039
			$\hat{\mu}_{Reg}$	0.0331	0.0325	111.43	
			$\hat{\mu}_{GRR}$	0.0331	0.0322	112.33	
		300	$\hat{\mu}_R$	0.0222	0.0214	98.62	0.0023
			$\hat{\mu}_{Reg}$	0.0194	0.0189	111.43	
			$\hat{\mu}_{GRR}$	0.0194	0.0189	111.95	
1000	0.6746	50	$\hat{\mu}_R$	0.1181	0.1140	167.23	0.0094
			$\hat{\mu}_{Reg}$	0.1041	0.1040	183.20	
			$\hat{\mu}_{GRR}$	0.1035	0.1014	187.97	
		100	$\hat{\mu}_R$	0.0548	0.0540	167.23	0.0044
			$\hat{\mu}_{Reg}$	0.0505	0.0493	183.20	
			$\hat{\mu}_{GRR}$	0.0502	0.0487	185.45	
		200	$\hat{\mu}_R$	0.0246	0.0240	167.23	0.0020
			$\hat{\mu}_{Reg}$	0.0224	0.0219	183.20	
			$\hat{\mu}_{GRR}$	0.0224	0.0218	184.20	
		300	$\hat{\mu}_R$	0.0143	0.0140	167.23	0.0012
			$\hat{\mu}_{Reg}$	0.0131	0.0128	183.20	
			$\hat{\mu}_{GRR}$	0.0131	0.0127	183.78	
1000	0.8684	50	$\hat{\mu}_R$	0.0399	0.0381	300.26	0.0094
			$\hat{\mu}_{Reg}$	0.0283	0.0284	402.62	
			$\hat{\mu}_{GRR}$	0.0288	0.0282	405.49	
		100	$\hat{\mu}_R$	0.0186	0.0181	300.26	0.0045
			$\hat{\mu}_{Reg}$	0.0141	0.0135	402.62	
			$\hat{\mu}_{GRR}$	0.0141	0.0134	403.97	
		200	$\hat{\mu}_R$	0.0083	0.0080	300.26	0.0020
			$\hat{\mu}_{Reg}$	0.0061	0.0060	402.62	
			$\hat{\mu}_{GRR}$	0.0061	0.0060	403.22	
		300	$\hat{\mu}_R$	0.0048	0.0047	300.26	0.0012
			$\hat{\mu}_{Reg}$	0.0036	0.0035	402.62	
			$\hat{\mu}_{GRR}$	0.0036	0.0035	402.97	

[1] *MSE* comparison base on 1^{st} order approximation given in expression 3.23.

For the regression and the generalized regression-cum-ratio estimators all the percent relative efficiencies are greater than 100 indicating that all these estimators are better than the RRT mean estimator. The same cannot be said about the ratio estimator because it is better than RRT mean estimator only for larger correlation values between X and Y. As expected, the generalized regression-cum-ratio estimator presents larger percent relative efficiencies although the improvement over the ordinary regression estimator is only modest.

3.7 Numerical Example

We now compare the performances of different estimators using a real data set. We focus on the ratio estimator, regression estimator and generalized regression-cum-ratio estimator. The sample data come from a very large survey on Information and Communication Technologies (ICT) usage in enterprises in 2009 with seat in Portugal (Smilhily and Storm, 2010). This survey intends to promote the development of the national statistical system in the information society and to contribute to a deeper knowledge about the usage of ICT by enterprises. The target population covers all industries with one and more persons employed in the sections of economic activity C (Manufacturing) to N (Administrative and support service activities) and S (Other service activities), from NACE[1] Rev. 2 (Eurostat, 2008). The data are collected mainly using Electronic Data Interchange, applying direct connection between information systems at the respondent and the National Statistics Institute. For some enterprises the paper questionnaire is still used. The questions in the structural business surveys mainly deal with characteristics that can be found in the organisations' annual reports and financial statements, such as employment, turnover and investment.

In our application the study variable Y is the purchase orders in 2010, collected by the ICT survey in that year. This is typically a confidential variable for enterprises, only known from business surveys. The auxiliary variable X is the turnover of each enterprise. This information can be easily obtained from enterprise records available in the public domain, as administrative information. In 2010, the population survey contained approximately 278000 enterprises and we know the value of X for all these enterprises. The purchase orders information was collected in the ICT survey and we have the values of Y for 5336 enterprises (which answered this question in the ICT survey in 2010). For this study, these 5336 enterprises are considered as our population so that its parameters are known. The scrambling variable S is taken to be a normal random variable with mean equal to zero and standard deviation equal to 10% of the standard deviation of X, that is $\sigma_S = 0.1\sigma_X$. The reported response is given by $Z = Y + S$ (the purchase order value plus a random quantity). The variables Y and X are strongly correlated so we can take advantage of this correlation by using the ratio and regression estimators, as well as a hybrid estimator that combines both. In Table 6.1 we present the results for the ratio estimator, the regression estimator and the generalized ratio-cum-regression estimator for different sample sizes.

[1]NACE is derived from the French title "Nomenclature générale des Activités économiques dans les Communautés Européennes" (Statistical classification of economic activities in the European Communities).

Population Characteristics:

$N = 5336, \rho_{XY} = 0.9632$
$\mu_X = 22.99, \mu_Y = 30.19, \sigma_X = 172.09, \sigma_Y = 138.65$ (in millions of Euros)
and $\beta_{YX} = 0.7763$

We use the following samples sizes in our simulation study: $n = 100, 300, 500, 1000$ and 2000.

Table 3.2 below presents the results for the empirical *MSE* estimates, the theoretical estimates, correct up to first degree of approximation, and the *PRE* of ratio, regression and generalized regression-cum-ratio estimators relative to the RRT mean estimator. We estimate the empirical *MSE* using 5000 samples of size n selected from the population.

Table 3.2: *MSE* correct up to 1^{st} order approximation and *PRE* for the ratio estimator ($\hat{\mu}_R$), the regression estimator ($\hat{\mu}_{Reg}$) and the generalized regression-cum-ratio estimator ($\hat{\mu}_{GRR}$) relative to the RRT mean estimator.

Population				***MSE* Estimation**			
N	ρ_{XY}	n	Estimator	Empirical	Theoretical	*PRE*	Condition[1]
5336	0.9636	100	$\hat{\mu}_R$	11.5741	16.4778	1162.46	0.3189
			$\hat{\mu}_{Reg}$	8.8601	16.4153	1166.88	
			$\hat{\mu}_{GRR}$	11.6905	15.3461	1248.18	
		300	$\hat{\mu}_R$	4.3423	5.2828	1162.46	0.1022
			$\hat{\mu}_{Reg}$	3.9360	5.2628	1166.88	
			$\hat{\mu}_{GRR}$	4.3858	5.0879	1206.99	
		500	$\hat{\mu}_R$	2.6995	3.0438	1162.46	0.0589
			$\hat{\mu}_{Reg}$	2.6596	3.0323	1166.88	
			$\hat{\mu}_{GRR}$	2.7166	2.9460	1201.03	
		1000	$\hat{\mu}_R$	1.4224	1.3645	1162.46	0.0264
			$\hat{\mu}_{Reg}$	1.4265	1.3594	1166.88	
			$\hat{\mu}_{GRR}$	1.4287	1.3253	1196.91	
		2000	$\hat{\mu}_R$	0.5817	0.5249	1162.46	0.0102
			$\hat{\mu}_{Reg}$	0.6100	0.5229	1166.88	
			$\hat{\mu}_{GRR}$	0.5869	0.5106	1194.95	

[1] *MSE* comparison base on 1^{st} order approximation given in expression 3.23.

According to the results in Table 3.2, all of the percent relative efficiencies are greater than 100, so all the estimators perform better than the RRT mean estimator for the real data also. The *PRE* of the generalized regression-cum-ratio estimator is better than the other estimators, particularly when the sample size is small.
Note that the sample size does not play a role in the *PRE* calculation for the ratio and regression estimators, as can be seen from Equations (3.2), (3.5) and (3.10).

3.8 Conclusions

We can observe from this study that the estimation of the mean of a sensitive variable can be improved by using a non-sensitive auxiliary variable. Although both the ratio and regression estimators perform better than the ordinary RRT mean estimator, the improvement is much larger with the regression estimator. Our simulation study shows that this improvement can be quite substantial for large sample sizes, particularly if the correlation between the study variable and the auxiliary variable is high. Further gains, although modest, can be achieved by using a generalized regression-cum-ratio estimator.

References

EICHHORN, B. H. & HAYRE, L. S. 1983. Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*, 7, 307-316.

EUROSTAT. 2008. NACE Rev. 2 - Statistical classification of economic activities in the European Community. *Official Publications of the European Communities*, 112-285 and 306-311.

GUPTA, S. N., GUPTA, B. C. & SINGH, S. 2002. Estimation of sensitivity level of personal interview survey questions. *Journal of Statistical Planning and Inference*, 100, 239-247.

GUPTA, S. & SHABBIR, J. 2004. Sensitivity estimation for personal interview survey questions. *Statistica*, 64, 643-653.

GUPTA, S., SHABBIR, J. & SEHRA, S. 2010. Mean and sensitivity estimation in optional randomized response models. *Journal of Statistical Planning and Inference*, 140(10), 2870-2874.

KADILAR, C., CANDAN, M. & CINGI, H. 2007. Ratio estimators using robust regression. *Hacettepe Journal of Mathematics and Statistics*, 36(2), 81-188.

KADILAR, C. & CINGI, H. 2004. Ratio estimators in simple random sampling. A*pplied Mathematics and Computation*, 151, 893-902.

KADILAR, C. & CINGI, H. 2005. A new estimator using two auxiliary variables. *Applied Mathematics and Computation*, 162, 901-908.

KADILAR, C. & CINGI, H. 2006. Improvement in estimating the population mean in simple random sampling. *Applied Mathematics Letters*, 19(1), 75-79.

MUKHOPADHYAY, P. 1998. *Theory and Methods of Survey Sampling*, New Delhi, Prentice-Hall of India.

NANGSUE, N. 2009. Adjusted Ratio and Regression Type Estimators for Estimation of Population Mean when some Observations are missing. *World Academy of Science, Engineering and Technology*, 53, 781-784.

PERRI, P. F. 2004. On the efficient use of regression-in-ratio estimator in simple random sampling. *Proceeding of the XLII Meeting of the Italian Statistical Society*, Bari (Italy), 537-540.

PERRI, P. F. 2008. Modified randomized devices for Simmons' model. *Model Assisted Statistics and Applications*, 3, 233-239.

RAY, S. K. & SINGH, R. K. 1981. Difference-cum-ratio type estimators. *Journal of Indian Statistical Association*, 19, 147-151.

SAHA, A. 2008. A randomized response technique for quantitative data under unequal probability sampling. *Journal of statistical Theory and Practice*, 2(4), 589-596.

SHABBIR, J. & GUPTA, S. 2007. On improvement in variance estimation using auxiliary information. *Communication in Statistics-Theory and Methods*, 36(12), 2177-2185.

SHABBIR, J. & GUPTA, S. 2010. Estimation of the finite population mean in two-phase sampling when auxiliary variables are attribute. *Hacettepe Journal of Mathematics and Statistics*, 39(1), 121-129.

SMILHILY, M. & STORM, H. 2010. ICT usage in enterprises - 2009. *Eurostat Publications*, Issue 1.

SOUSA, R., SHABBIR, J. REAL, P. C. & GUPTA, S. 2010. Ratio estimation of the mean of a sensitive variable in the presence of auxiliary information. *Journal of Statistical Theory and Practice*, 4(3), 495-507.

WU, J-W, TIAN, G-L & TANG, M-L. 2008. Two new models for survey sampling with sensitive characteristic: design and analysis. *Metrika*, 67, 251-263.

Appendix B - R Routines

Listing 3.1: R Code for Simulation Study of Proposed Estimator in Chapter 3

```

proj2_2nd_estimator <- function(N,sigma,mu)
{

  #Generation of a bivariate normal population
  data_yx <- mvrnorm(N, mu, sigma)

  #Study variable
  Y <- data_yx[,1]
  #Auxiliary variable, correlated with Y
  X <- data_yx[,2]

  #Coefficient of correlation between Y and X
  Ro_YX <- cor(Y,X)

  #Scrambling variable independent of Y and X, with mean=0
  S <- rnorm(N,mean=0,sd=0.1*sd(X))
  #Scrambled response
  Z <- Y+S

  #Coefficient of correlation between Z and X
  Ro_ZX <- Ro_YX/sqrt(1+(var(S)/var(Y)))

  #population
  univ <- data.frame(cbind(Y=Y,S=S,Z=Z,X=X,NRAND=runif(N)))
  univ <- univ[order(univ$NRAND),]

  #Mean of Y
  mz <- mean(univ$Z)
  mx <- mean(univ$X)
  my <- mean(univ$Y)

  mu11 <- sum((univ$Z-mz)*(univ$X-mx))/(N-1)
  mu12 <- sum((univ$Z-mz)*((univ$X-mx)^2))/(N-1)
  mu02 <- sum((univ$X-mx)^2)/(N-1)
  mu03 <- sum((univ$X-mx)^3)/(N-1)

  beta_zx <- Ro_YX*(sd(univ$Y)/sd(univ$X))

  #Samples dimension
  dim_samp <- c(50,100,150,200,300)

  #Initialize the variables...
  for (i in 1:length(dim_samp))
  {
    #sample dimension
```

```
    n <- dim_samp[i]
    #sample
    samp <- univ[1:n,]
    #Sampling rate
    f <- n/N

    #Ratio
    R <- mean(univ$X)/mean(samp$X)

    #Ordinary meam
    est1 <- mean(samp$Z)
    #Ratio estimator
    est2 <- mean(samp$Z)*(mx/mean(samp$X))
    #Regression estimator
    est3 <- mean(samp$Z)+beta_zx*(mx-mean(samp$X))

    #Coefficient of variation
    c_x <- sd(univ$X)/mx
    c_y <- sd(univ$Y)/my
    c2_x <- c_x^2
    c2_y <- c_y^2
    c2_z <- c2_y+(var(univ$S)/(my^2))
    c_z <- sqrt(c2_z)

    k1 <- (1-((1-f)*c2_x/n))/(1-((1-f)/n)*(c2_x-c2_z*(1-(Ro_ZX^2))))
    k2 <- (mz/mx)*(1+k1*((Ro_ZX*c_z/c_x)-2))
    #Generalized regression-cum-ratio estimator
    est5 <- (k1*mean(samp$Z)+k2*(mx-mean(samp$X)))*(mx/mean(samp$X))

    #Mean Square Error of 1st estimator (ordinal mean)
    mse1 <- ((1-f)/n)*(var(univ$Y)+var(univ$S))

    #Bias of ratio estimator - 1st degree approximation
    bias2i <- ((1-f)/n)*my*(c2_x-Ro_ZX*c_z*c_x)
    #Mean Square Error of ratio estimator - 1st degree approximation
    mse2i <- ((1-f)/n)*(my^2)*(c2_z+c2_x-2*Ro_ZX*c_z*c_x)

    #Bias of regression estimator - 1st degree approximation
    bias3i <- -beta_zx*((1-f)/n)*((mu12/mu11)-(mu03/mu02))
    #Mean Square Error of regression estimator - 1st degree approximation
    mse3i <- ((1-f)/n)*(my^2)*c2_z*(1-(Ro_ZX^2))

    #Bias of genetalized regression-cum-ratio estimator
    #1st degree approximation
    bias5i <- (k1-1)*my+k1*my*((1-f)/n)*(c2_x-Ro_ZX*c_z*c_x)
          +k2*mx*((1-f)/n)*c2_x
    #Mean Square Error of generalized regression-cum-ratio estimator
    #1st degree approximation
    mse5i <- ((k1-1)^2)*(my^2)+(k1^2)*(my^2)*((1-f)/n)*(c2_z
        +3*c2_x-4*Ro_ZX*c_z*c_x)+(k2^2)*(mx^2)*((1-f)/n)*c2_x
```

```
              -2*k1*(my^2)*((1-f)/n)*(c2_x-Ro_ZX*c_z*c_x)
              -2*k2*my*mx*((1-f)/n)*c2_x
              -2*k1*k2*my*mx*((1-f)/n)*(Ro_ZX*c_z*c_x-2*c2_x)

    cond1 <- ((1-f)/n)*c2_x

    #Empirical results
    #Simulation of 5000 replicas of estimates
    ...

    #Results
    res <- rbind(res,c(N,n,Ro_YX,Ro_ZX,R,
                    c_x,c_y,c_z,mx,my,mz,
                    med_est1,med_est2,med_est3,med_est5,
                    bias2i,bias3i,bias5i,
                    emp_mse1,mse1,emp_mse2,mse2i,
                    emp_mse3,mse3i,emp_mse5,mse5i,cond1))
  }
  colnames(res) <- c("N","n","RhoXY","RhoZX","R",
                    "Cx","Cy","Cz","mX","mY","mZ",
                    "Est1","Est2","Est3","Est5",
                    "BIAS2I","BIAS3I","BIAS5I",
                    "EMP_MSE1","MSE1","EMP_MSE2","MSE2I",
                    "EMP_MSE3","MSE3I","EMP_MSE5","MSE5I","COND1")
  return(res)
}

#Package for generation
require(MASS)
N <- 1000

#Parameters
sigma1 <- matrix(c(9,1.9,1.9,4),2,2)
sigma2 <- matrix(c(10,3,3,2),2,2)
sigma3 <- matrix(c(6,3,3,2),2,2)
mu <- c(2,2)

res <- NULL
for (i in 1:length(N))
{
  res <- rbind(res,proj2_2nd_estimator(N[i],sigma1,mu))
  res <- rbind(res,proj2_2nd_estimator(N[i],sigma2,mu))
  res <- rbind(res,proj2_2nd_estimator(N[i],sigma3,mu))
}
write.table(res,"chapter3_ss_results.txt",sep="\t",dec=",",row.names=FALSE)
```

Listing 3.2: R Code for Numerical Example of Proposed Estimator in Chapter 3

```

proj2_2nd_estimator_real <- function(Y,X,N)
{
  #Coefficient of correlation between Y and X
  Ro_YX <- cor(Y,X)

  #Scrambling variable independent of Y and X, with mean=0
  S <- rnorm(N,mean=0,sd=sd(X)*0.1)
  #Scrambled response
  Z <- Y+S

  #Coefficient of correlation between Z and X
  Ro_ZX <- Ro_YX/sqrt(1+(var(S)/var(Y)))

  #population
  univ <- data.frame(cbind(Y=Y,S=S,Z=Z,X=X,NRAND=runif(N)))
  univ <- univ[order(univ$NRAND),]

  #Mean of Y
  mz <- mean(univ$Z)
  mx <- mean(univ$X)
  my <- mean(univ$Y)
  ms <- mean(univ$S)

  mu11 <- sum((univ$Z-mz)*(univ$X-mx))/(N-1)
  mu12 <- sum((univ$Z-mz)*((univ$X-mx)^2))/(N-1)
  mu02 <- sum((univ$X-mx)^2)/(N-1)
  mu03 <- sum((univ$X-mx)^3)/(N-1)

  beta_zx <- Ro_YX*(sd(univ$Y)/sd(univ$X))

  #Samples dimension
  dim_samp <- c(100,300,500,1000,2000)

  #Initialize variables...
  for (i in 1:length(dim_samp))
  {
    #sample dimension
    n <- dim_samp[i]
    #sample
    samp <- univ[1:n,]
    #Sampling rate
    f <- n/N

    #Ratio
    R <- mean(univ$X)/mean(samp$X)

```

```
    #Ordinary meam
    est1 <- mean(samp$Z)
    #Ratio estimator
    est2 <- mean(samp$Z)*(mx/mean(samp$X))
    #Regression estimator
    est3 <- mean(samp$Z)+beta_zx*(mx-mean(samp$X))

    #Coefficient of variation
    c_x <- sd(univ$X)/mx
    c_y <- sd(univ$Y)/my
    c2_x <- c_x^2
    c2_y <- c_y^2
    c2_z <- c2_y+(var(univ$S)/(my^2))
    c_z <- sqrt(c2_z)

    k1 <- (1-((1-f)*c2_x/n))/(1-((1-f)/n)*(c2_x-c2_z*(1-(Ro_ZX^2))))
    k2 <- (mz/mx)*(1+k1*((Ro_ZX*c_z/c_x)-2))
    #Generalized regression-cum-ratio estimator
    est5 <- (k1*mean(samp$Z)+k2*(mx-mean(samp$X)))*(mx/mean(samp$X))

    #Mean Square Error of 1st estimator (ordinal mean)
    mse1 <- ((1-f)/n)*(var(univ$Y)+var(univ$S))

    #Bias of ratio estimator - 1st degree approximation
    bias2i <- ((1-f)/n)*my*(c2_x-Ro_ZX*c_z*c_x)
    #Mean Square Error of ratio estimator - 1st degree approximation
    mse2i <- ((1-f)/n)*(my^2)*(c2_z+c2_x-2*Ro_ZX*c_z*c_x)

    #Bias of regression estimator - 1st degree approximation
    bias3i <- -beta_zx*((1-f)/n)*((mu12/mu11)-(mu03/mu02))
    #Mean Square Error of regression estimator - 1st degree approximation
    mse3i <- ((1-f)/n)*(my^2)*c2_z*(1-(Ro_ZX^2))

    #Bias of genetalized regression-cum-ratio estimator
    #1st degree approximation
    bias5i <- (k1-1)*my+k1*my*((1-f)/n)*(c2_x-Ro_ZX*c_z*c_x)
          +k2*mx*((1-f)/n)*c2_x
    #Mean Square Error of generalized regression-cum-ratio estimator
    #1st degree approximation
    mse5i <- ((k1-1)^2)*(my^2)+(k1^2)*(my^2)*((1-f)/n)
    *(c2_z+3*c2_x-4*Ro_ZX*c_z*c_x)+(k2^2)*(mx^2)*((1-f)/n)
    *c2_x-2*k1*(my^2)*((1-f)/n)*(c2_x-Ro_ZX*c_z*c_x)
    -2*k2*my*mx*((1-f)/n)*c2_x-2*k1*k2*my*mx*((1-f)/n)*(Ro_ZX*c_z*c_x-2*c2_x)

    cond1 <- ((1-f)/n)*c2_x

    #Empirical results
    #Simulation of 5000 replicas of estimates
    ...

```

```
    #Results
    res <- rbind(res,c(N,n,Ro_YX,Ro_ZX,R,
                     c_x,c_y,c_z,mx,my,mz,ms,
                     med_est1,med_est2,med_est3,med_est5,
                     bias2i,bias3i,bias5i,
                     emp_mse1,mse1,emp_mse2,mse2i,
                     emp_mse3,mse3i,emp_mse5,mse5i,cond1))
  }
  colnames(res) <- c("N","n","RhoXY","RhoZX","R",
                     "Cx","Cy","Cz","mX","mY","mZ","ms",
                     "Est1","Est2","Est3","Est5",
                     "BIAS2I","BIAS3I","BIAS5I",
                     "EMP_MSE1","MSE1","EMP_MSE2","MSE2I",
                     "EMP_MSE3","MSE3I","EMP_MSE5","MSE5I","COND1")
  return(res)
}

#Package for generation
require(MASS)

#Import data
data_yx <- read.table("IUTICE10.txt",sep="\t",dec=",",header = T)
#Study variable (purchase, millions of euros)
Y <- data_yx[,3]
#Auxiliary variable, correlated with Y (turnover, millions of euros)
X <- data_yx[,2]

#Data application
N <- dim(data_yx)[1]
res <- proj2_2nd_estimator_real(Y,X,N)

#Export data
write.table(res_exp,"chapter3_ne_results.txt",sep="\t",dec=",",row.names=FALSE)
```

4

Exponential Type Estimators of the Mean of a Sensitive Variable in the Presence of Non Sensitive Auxiliary Information

Abstract

Sousa et al. (2010) and Gupta et al. (2012) suggested ratio and regression type estimators of the mean of a sensitive variable using non-sensitive auxiliary variable. This paper proposes exponential type estimators using one and two auxiliary variables to improve the efficiency of mean estimator based on a Randomized Response Technique (RRT). The expressions for the Mean Square Errors (*MSE*'s) and bias, up to fisrt order approximation, have been obtained. It is shown that the proposed exponential type estimators are more efficient than the existing estimators. The gain in efficiency over the existing estimators has also been shown with a simulation study and by using real data.

Accepted as: KOYUNCU, N., GUPTA, S., SOUSA, R. 2013. Exponential type estimators of the mean of a sensitive variable in the presence of non-sensitive auxiliary information. *Communications in Statistics - Simulation and Computation*.

4.1 Introduction

Randomized Response Technique (RRT) is used to estimate the proportion of people in a community bearing a stigmatizing characteristic like habitual tax evasion, reckless driving, indiscriminate gambling, abortion etc. In such situations we cannot expect to get a truthful direct response to a sensitive question. Eichhorn and Hayre (1983), Gupta and Shabbir (2004), Gupta et al. (2002, 2010), Wu et al. (2008), Perri (2008), and many others have estimated the mean of a sensitive variable when the study variable is sensitive and there is no auxiliary variable. Sousa et al. (2010) and Gupta et al. (2012) suggested mean estimators based on RRT models using an auxiliary variable that can be directly observed. In sampling literature, Bahl and Tuteja (1991), Shabbir and Gupta (2007), Grover (2010) and Koyuncu (2012) have studied exponential type estimators to get more efficient estimates. In this study we have proposed exponential type estimators of the mean of a sensitive variable using non-sensitive auxiliary information. We have discussed the cases when one or two non-sensitive auxiliary variables are available.

4.2 Terminology

Let Y be the study variable, a sensitive variable which cannot be observed directly. Let X_1 and X_2 be non-sensitive auxiliary variables which have a positive correlation with Y. Let S, be a scrambling variable, independent of Y, X_1 and X_2. The respondent is asked to report a scrambled response for Y given by $Z = Y + S$ but is asked to provide a true response for X_1 and X_2. Let a random sample of size n be drawn without replacement from a finite population $U = (U_1, U_2, ..., U_N)$. For the i^{th} unit $(i = 1, 2, ..., N)$, let y_i, x_{1i} and x_{2i} respectively be the values of the study variable Y and auxiliary variables X_1 and X_2. Let $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$, $\bar{x}_1 = \frac{\sum_{i=1}^{n} x_{1i}}{n}$, $\bar{x}_2 = \frac{\sum_{i=1}^{n} x_{2i}}{n}$ and $\bar{z} = \frac{\sum_{i=1}^{n} z_i}{n}$ be the sample means and $\bar{Y} = E(Y)$, $\bar{X}_1 = E(X_1)$, $\bar{X}_2 = E(X_2)$ and $\bar{Z} = E(Z)$ be the population means for Y, X_1, X_2 and Z respectively. We assume that $\bar{X}_1$, $\bar{X}_2$ are known and $\bar{S} = E(S) = 0$. Thus $E(Z) = E(Y)$ and $C_z^2 = C_y^2 + \left(S_s^2/\bar{Y}^2\right)$, where C_z and C_y are the coefficients of the variation of z and y respectively.

To obtain the bias and *MSE* expressions, let us define

$$e_0 = \frac{\bar{z} - \bar{Z}}{\bar{Z}},\ e_1 = \frac{\bar{x}_1 - \bar{X}_1}{\bar{X}_1},\ e_2 = \frac{\bar{x}_2 - \bar{X}_2}{\bar{X}_2},\ e_3 = \frac{s_{x1}^2 - S_{x1}^2}{S_{x1}^2} \text{ and } e_4 = \frac{s_{zx1}^2 - S_{zx1}^2}{S_{zx1}^2}.$$

Using these notations,

$$E(e_i) = 0,\ \ i = 0, 1, 2, 3, 4.$$

$$E(e_0^2) = \lambda C_z^2,\ E(e_1^2) = \lambda C_{x1}^2,\ E(e_2^2) = \lambda C_{x2}^2,\ E(e_0 e_1) = \lambda C_{zx1},$$

$E(e_0e_2) = \lambda C_{zx2}$, $E(e_1e_2) = \lambda C_{x1x2}$, $E(e_1e_3) = \lambda \dfrac{1}{\bar{X}_1}\dfrac{\mu_{03}}{\mu_{02}}$,

$E(e_1e_4) = \lambda \dfrac{1}{\bar{X}_1}\dfrac{\mu_{12}}{\mu_{11}}$,

where $\lambda = \dfrac{1-f}{n}$ and $\mu_{rs} = \dfrac{1}{N-1}\sum_{i=1}^{N}(z_i - \bar{Z})^r(x_{1i} - \bar{X}_1)^s$.

4.3 Estimators Review

We describe below some existing mean estimators and their bias and *MSE* formulas.

(i) Ordinary sample mean $(\bar{z})$ of scrambled responses:

$$\hat{\mu}_Y = \bar{z}. \tag{4.1}$$

$$MSE(\hat{\mu}_y) = \lambda\left(S_y^2 + S_s^2\right), \tag{4.2}$$

where
$S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$, $S_{x1}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{1i} - \bar{X}_1)^2$, $S_s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(s_i - \bar{S})^2$.

(ii) Sousa et al. (2010) ratio type estimator:

$$\hat{\mu}_R = \bar{z}\frac{\bar{X}_1}{\bar{x}_1}. \tag{4.3}$$

$$Bias(\hat{\mu}_R) \cong \bar{Y}\lambda\left(C_{x1}^2 - C_{x1z}\right), \tag{4.4}$$

where $C_z^2 = \left(C_y^2 + \dfrac{S_s^2}{\bar{Y}^2}\right)$, $\rho_{zx1} = \dfrac{\rho_{yx1}}{\sqrt{1 + \dfrac{S_s^2}{S_y^2}}}$, $\bar{Z} = \bar{Y}$.

$$MSE(\hat{\mu}_R) \cong \lambda\bar{Y}^2\left(C_z^2 + C_{x1}^2 - 2C_{x1z}\right). \tag{4.5}$$

(iii) Gupta et al. (2012) regression estimator:

$$\hat{\mu}_{Reg} = \bar{z} + \hat{\beta}_{zx1}\left(\bar{X}_1 - \bar{x}_1\right), \tag{4.6}$$

where $\hat{\beta}_{zx1} = \dfrac{S_{zx1}}{S_{x1}^2} = \dfrac{S_{yx1}}{S_{x1}^2}$, is the sample regression coefficient between Z and X_1.

$$Bias(\hat{\mu}_{Reg}) \cong -\beta_{zx1}\lambda\left(\frac{\mu_{12}}{\mu_{11}} - \frac{\mu_{03}}{\mu_{02}}\right), \tag{4.7}$$

where $\beta_{zx1} = \dfrac{S_{zx1}}{S_{x1}^2} = \dfrac{S_{yx1}}{S_{x1}^2} = \rho_{yx1}\dfrac{S_y}{S_{x1}} = \beta_{yx1}$ is the population regression coefficient between Z and X_1.

Recognizing $\bar{Z}=\bar{Y}$

$$MSE(\hat{\mu}_{Reg}) \cong \lambda \bar{Y}^2 C_z^2 \left[1 - \frac{S_{zx1}^2}{S_{x1}^2 S_z^2}\right] = \lambda \bar{Y}^2 C_z^2 \left[1 - \rho_{zx1}^2\right]$$

or

$$MSE(\hat{\mu}_{Reg}) \cong \lambda S_y^2 \left[\left(1 + \frac{S_s^2}{S_y^2}\right) - \rho_{yx1}^2\right]. \tag{4.8}$$

(iv) Gupta et al. (2012) generalized regression-cum-ratio estimator:

$$\hat{\mu}_{GRR} = \left[k_1 \bar{z} + k_2 \left(\bar{X} - \bar{x}\right)\right] \left(\frac{\bar{X}}{\bar{x}}\right), \tag{4.9}$$

where k_1 and k_2 are constants.

$$Bias(\hat{\mu}_{GRR}) \cong (k_1 - 1)\bar{Y} + k_1 \bar{Y} \lambda \left\{C_x^2 - \rho_{zx} C_z C_x\right\} + k_2 \bar{X} \lambda C_x^2. \tag{4.10}$$

$$\begin{aligned} MSE(\hat{\mu}_{GRR}) \cong \;& (k_1 - 1)^2 \bar{Y}^2 + k_1^2 \bar{Y}^2 \lambda \left\{C_z^2 + 3C_x^2 - 4\rho_{zx} C_z C_x\right\} \\ & + k_2^2 \bar{X}^2 \lambda C_x^2 - 2k_1 \bar{Y}^2 \lambda \left\{C_x^2 - \rho_{zx} C_z C_x\right\} \\ & - 2k_2 \bar{Y} \bar{X} \lambda C_x^2 - 2k_1 k_2 \bar{Y} \bar{X} \lambda \left\{\rho_{zx} C_z C_x - 2C_x^2\right\}. \end{aligned} \tag{4.11}$$

From Equation (4.11), the optimum values of k_1 and k_2 are given by

$$k_{1(opt)} = \frac{1 - \lambda C_x^2}{1 - \lambda \left\{C_x^2 - C_z^2 \left(1 - \rho_{zx}^2\right)\right\}} \tag{4.12}$$

and

$$k_{2(opt)} = \frac{\bar{Y}}{\bar{X}} \left\{1 + k_{1(opt)} \left(\frac{\rho_{zx} C_z}{C_x} - 2\right)\right\}, \tag{4.13}$$

the minimum *MSE* of $\hat{\mu}_{GRR}$ can be written as follows:

$$MSE(\hat{\mu}_{GRR})_{min} \cong \frac{\bar{Y}^2 C_z^2 \left(1 - \rho_{zx}^2\right) \lambda \left\{1 - \lambda C_x^2\right\}}{C_z^2 \left(1 - \rho_{zx}^2\right) \lambda + \left\{1 - \lambda C_x^2\right\}}. \tag{4.14}$$

4.4 Proposed Exponential Type Estimators

Our first proposed estimator, which we call "generalized regression-cum-exponential estimator" follows Grover (2010) and Shabbir and Gupta (2007), and is given by

$$\hat{\mu}_{exp1} = \left[w_1 \bar{z} + w_2 \left(\bar{X}_1 - \bar{x}_1\right)\right] exp\left(\frac{\bar{X}_1 - \bar{x}_1}{\bar{X}_1 + \bar{x}_1}\right), \tag{4.15}$$

where w_1 and w_2 are suitable weights. Expressing (4.15) in terms of e's (defined earlier) and retaining terms of e's up to second-order we have

$$\hat{\mu}_{exp1}-\bar{Z} \cong \left[w_1\bar{Z} + w_1\bar{Z}e_0 - w_2\bar{X}_1e_1 - \frac{1}{2}w_1\bar{Z}e_1 - \frac{1}{2}w_1\bar{Z}e_0e_1 + \frac{1}{2}w_2\bar{X}_1e_1^2 + \frac{3}{8}w_1\bar{Z}e_1^2 - \bar{Z}\right]. \tag{4.16}$$

The *Bias* and *MSE* of $\hat{\mu}_{exp1}$, to the first order of approximation, are given by

$$Bias\left(\hat{\mu}_{exp1}\right) \cong \left(w_1 - 1\right)\bar{Y} + \lambda\left\{\frac{1}{2}w_1\bar{Y}\left(\frac{3}{4}C_{x1}^2 - C_{zx1}\right) + \frac{1}{2}w_2\bar{X}_1C_{x1}^2\right\} \tag{4.17}$$

and

$$\begin{aligned} MSE\left(\hat{\mu}_{exp1}\right) \cong\ & \{\bar{Y}^2 + w_1^2\bar{Y}^2\left(1 + \lambda\left(C_z^2 + C_{x1}^2 - 2C_{zx1}\right)\right) \\ & + w_2^2\bar{X}^2\lambda C_{x1}^2 + w_1\bar{Y}^2\left(\lambda\left(C_{zx1} - \frac{3}{4}C_{x1}^2\right) - 2\right) \\ & - w_2\bar{Y}\bar{X}\lambda C_{x1}^2 + 2w_1w_2\bar{Y}\bar{X}\lambda\left(C_{x1}^2 - C_{zx1}\right)\}, \end{aligned} \tag{4.18}$$

and optimum values of w_1 and w_2 respectively are found as

$$w_1^* = \frac{1 - \frac{1}{8}\lambda C_{x1}^2}{1 + \lambda C_z^2\left(1 - \rho_{zx1}^2\right)} \tag{4.19}$$

and

$$w_2^* = \frac{\bar{Y}}{2\bar{X}_1}\frac{C_{x1}^2 - 2C_{x1}^2 + 2C_{zx1} + \lambda C_{x1}^2\left(C_z^2\left(1 - \rho_{zx1}^2\right) + \frac{1}{4}\left(C_{x1}^2 - C_{zx1}\right)\right)}{C_{x1}^2\left[1 + \lambda C_z^2\left(1 - \rho_{zx1}^2\right)\right]}. \tag{4.20}$$

Substituting these optimum values in (4.18), the minimum *MSE* of $\hat{\mu}_{exp1}$ can be written as follows:

$$MSE_{min}\left(\hat{\mu}_{exp1}\right) \cong \bar{Y}^2\left[1 - \frac{\lambda^2C_{x1}^2\left(\frac{1}{16}C_{x1}^2 + C_z^2\left(1 - \rho_{zx1}^2\right)\right) + 4}{4 + \left[1 + \lambda C_z^2\left(1 - \rho_{zx1}^2\right)\right]}\right] \tag{4.21}$$

or

$$MSE_{min}\left(\hat{\mu}_{exp1}\right) \cong \left[\frac{MSE\left(\hat{\mu}_{Reg}\right)}{\left[1 + \frac{MSE\left(\hat{\mu}_{Reg}\right)}{\bar{Y}^2}\right]} - \frac{\lambda C_{x1}^2\left(MSE\left(\hat{\mu}_{Reg}\right) + \lambda\frac{1}{16}C_{x1}^2\bar{Y}^2\right)}{4\left[1 + \frac{MSE\left(\hat{\mu}_{Reg}\right)}{\bar{Y}^2}\right]}\right]. \tag{4.22}$$

Note that the optimum choice of the constants w_1 and w_2 involve unknown parameters. These quantities can be guessed through a pilot sample survey or through experience gathered in due course of time, as mentioned Upadhyaya and Singh (2006), and Koyuncu and Kadilar (2009).

The estimator defined in (4.15) can be generalized to the case of multiple auxiliary variables. We consider below the case of two auxiliary non-sensitive variables. This estimator is given by

$$\hat{\mu}_{exp2} = \left[d_1\bar{z} + d_2\left(\bar{X}_1 - \bar{x}_1\right) + d_3\left(\bar{X}_2 - \bar{x}_2\right)\right] exp\left(\frac{\left(\bar{X}_1 - \bar{x}_1\right) + \left(\bar{X}_2 - \bar{x}_2\right)}{\left(\bar{X}_1 + \bar{x}_1\right) + \left(\bar{X}_2 + \bar{x}_2\right)}\right). \tag{4.23}$$

Expressing (4.23) in terms of e's and retaining up to second-order terms in e's we have

$$\begin{aligned}\hat{\mu}_{exp2} \cong\ & \left\{d_1\bar{Z}\left(1+e_0\right) - d_2\bar{X}_1e_1 - d_3\bar{X}_2e_2\right\} \\ & \left\{\begin{array}{l} 1 - \dfrac{\bar{X}_1}{2\left(\bar{X}_1+\bar{X}_2\right)}e_1 - \dfrac{\bar{X}_2}{2\left(\bar{X}_1+\bar{X}_2\right)}e_2 + \dfrac{3\bar{X}_1^2}{8\left(\bar{X}_1+\bar{X}_2\right)^2}e_1^2 \\ +\dfrac{6\bar{X}_1\bar{X}_2}{8\left(\bar{X}_1+\bar{X}_2\right)^2}e_1e_1 + \dfrac{3\bar{X}_2^2}{8\left(\bar{X}_1+\bar{X}_2\right)^2}e_2^2 \end{array}\right\}.\end{aligned} \tag{4.24}$$

The *Bias* and *MSE* of $\hat{\mu}_{exp2}$, to the first order of approximation, are given by

$$\begin{aligned}Bias\left(\hat{\mu}_{exp2}\right) \cong\ & \left(d_1-1\right)\bar{Z} \\ & +\frac{d_1\lambda\bar{Z}}{2\left(\bar{X}_1+\bar{X}_2\right)}\left(\begin{array}{l} -\bar{X}_1C_{zx1} - \bar{X}_2C_{zx2} + \dfrac{3\bar{X}_1^2}{4\left(\bar{X}_1+\bar{X}_2\right)}C_{x1}^2 \\ +\dfrac{3\bar{X}_2^2}{4\left(\bar{X}_1+\bar{X}_2\right)}C_{x2}^2 + \dfrac{3\bar{X}_1\bar{X}_2}{2\left(\bar{X}_1+\bar{X}_2\right)}C_{x1x2} \end{array}\right) \\ & +\frac{d_2\lambda\bar{X}_1}{2\left(\bar{X}_1+\bar{X}_2\right)}\left(\bar{X}_1C_{x1}^2 + \bar{X}_2C_{x1x2}\right) \\ & +\frac{d_3\lambda\bar{X}_2}{2\left(\bar{X}_1+\bar{X}_2\right)}\lambda\left(\bar{X}_1C_{x1x2} + \bar{X}_2C_{x2}^2\right)\end{aligned} \tag{4.25}$$

and

$$\begin{aligned}MSE\left(\hat{\mu}_{exp2}\right) \cong\ & \bar{Z}^2 + d_1A - d_2B - d_3C + d_1^2D + d_2^2\bar{X}_1^2\lambda C_{x1}^2 + d_3^2\bar{X}_2^2\lambda C_{x2}^2 \\ & +2d_1d_2F + 2d_1d_3G + 2d_2d_3\bar{X}_1\bar{X}_2\lambda C_{x1x2},\end{aligned} \tag{4.26}$$

where

$$\begin{aligned}A =\ & \bar{Z}^2\left(-2+\lambda\left\{\begin{array}{l} \dfrac{\bar{X}_1C_{zx1}}{\left(\bar{X}_1+\bar{X}_2\right)} + \dfrac{\bar{X}_2C_{zx2}}{\left(\bar{X}_1+\bar{X}_2\right)} - \dfrac{3\bar{X}_1^2C_{x1}^2}{4\left(\bar{X}_1+\bar{X}_2\right)^2} - \dfrac{6\bar{X}_1\bar{X}_2C_{x1x2}}{4\left(\bar{X}_1+\bar{X}_2\right)^2} \\ -\dfrac{3\bar{X}_2^2C_{x2}^2}{4\left(\bar{X}_1+\bar{X}_2\right)^2} \end{array}\right\}\right), \\ B =\ & \lambda\frac{\bar{Z}}{\left(\bar{X}_1+\bar{X}_2\right)}\left(\bar{X}_1^2C_{x1}^2 + \bar{X}_1\bar{X}_2C_{x1x2}\right), \\ C =\ & \lambda\frac{\bar{Z}}{\left(\bar{X}_1+\bar{X}_2\right)}\left(\bar{X}_2^2C_{x2}^2 + \bar{X}_1\bar{X}_2C_{x1x2}\right), \\ D =\ & \bar{Z}^2 + \lambda\left\{\begin{array}{l} \bar{Z}^2C_z^2 + \dfrac{\bar{X}_1^2\bar{Z}^2C_{x1}^2}{\left(\bar{X}_1+\bar{X}_2\right)^2} + \dfrac{\bar{X}_2^2\bar{Z}^2C_{x2}^2}{\left(\bar{X}_1+\bar{X}_2\right)^2} - 2\dfrac{\bar{X}_1\bar{Z}^2C_{zx1}}{\left(\bar{X}_1+\bar{X}_2\right)} - 2\dfrac{\bar{X}_2\bar{Z}^2C_{zx2}}{\left(\bar{X}_1+\bar{X}_2\right)} \\ +\dfrac{2\bar{X}_1\bar{X}_2\bar{Z}^2C_{x1x2}}{\left(\bar{X}_1+\bar{X}_2\right)^2} \end{array}\right\},\end{aligned}$$

$$F = \lambda \left(\frac{\bar{Z}\bar{X}_1^2}{(\bar{X}_1 + \bar{X}_2)} C_{x1}^2 + \frac{\bar{Z}\bar{X}_1\bar{X}_2}{(\bar{X}_1 + \bar{X}_2)} C_{x1x2} - \bar{Z}\bar{X}_1 C_{zx1} \right),$$
$$G = \lambda \left(\frac{\bar{Z}\bar{X}_2^2}{(\bar{X}_1 + \bar{X}_2)} C_{x2}^2 + \frac{\bar{Z}\bar{X}_1\bar{X}_2}{(\bar{X}_1 + \bar{X}_2)} C_{x1x2} - \bar{Z}\bar{X}_2 C_{zx2} \right),$$

and optimum values of d_1, d_2 and d_3 are respectively found as

$$d_1^* = \frac{1}{2D} \frac{\begin{bmatrix} A\left(D\lambda S_{x1x2} - FG\right)^2 + (BDG + CDF + 2AFG)\left(D\lambda S_{x1x2} - FG\right) \\ -G\,(CD + AG)\left(D\lambda S_{x1}^2 - F^2\right) - F\,(AF + BD)\left(D\lambda S_{x2}^2 - G^2\right) \\ -A\left(D\lambda S_{x1}^2 - F^2\right)\left(D\lambda S_{x2}^2 - G^2\right) \end{bmatrix}}{\left(D\lambda S_{x1}^2 - F^2\right)\left(D\lambda S_{x2}^2 - G^2\right) - \left(D\lambda S_{x1x2} - FG\right)^2}, \quad (4.27)$$

$$d_2^* = \frac{1}{2} \frac{(AF + BD)\left(D\lambda S_{x2}^2 - G^2\right) - (AG + CD)\left(D\lambda S_{x1x2} - FG\right)}{\left(D\lambda S_{x1}^2 - F^2\right)\left(D\lambda S_{x2}^2 - G^2\right) - \left(D\lambda S_{x1x2} - FG\right)^2}, \quad (4.28)$$

and

$$d_3^* = \frac{1}{2} \frac{(AG + CD)\left(D\lambda S_{x1}^2 - F^2\right) - (AF + BD)\left(D\lambda S_{x1x2} - FG\right)}{\left(D\lambda S_{x1}^2 - F^2\right)\left(D\lambda S_{x2}^2 - G^2\right) - \left(D\lambda S_{x1x2} - FG\right)^2}. \quad (4.29)$$

Substituting these optimum values in (4.26), the minimum MSE of $\hat{\mu}_{exp2}$ can be written as follows:

$$\begin{aligned} MSE\left(\hat{\mu}_{exp2}\right)_{min} = \; & \bar{Z}^2 - \frac{A^2}{4D} \\ & - \frac{1}{4D} \frac{\begin{bmatrix} (AG + CD)^2 \left(D\lambda S_{x1}^2 - F^2\right) \\ + (AF + BD)^2 \left(D\lambda S_{x2}^2 - G^2\right) \\ -2\,(AG + CD)(AF + BD)\left(D\lambda S_{x1x2} - FG\right) \end{bmatrix}}{\left(D\lambda S_{x1}^2 - F^2\right)\left(D\lambda S_{x2}^2 - G^2\right) - \left(D\lambda S_{x1x2} - FG\right)^2}. \end{aligned} \quad (4.30)$$

4.5 Comparison with Gupta et al. (2012) Estimators

First, we compare the proposed "generalized regression-cum-exponential estimator" with the Gupta et al. (2012) regression estimator. Note that

$MSE\left(\hat{\mu}_{exp1}\right) < MSE\left(\hat{\mu}_{Reg}\right)$ if

$$MSE\left(\hat{\mu}_{Reg}\right) - \frac{MSE\left(\hat{\mu}_{Reg}\right)}{\left[1 + \frac{MSE\left(\hat{\mu}_{Reg}\right)}{\bar{Y}^2}\right]} + \frac{\lambda C_x^2 \left(MSE\left(\hat{\mu}_{Reg}\right) + \lambda \frac{1}{16} C_x^2 \bar{Z}^2\right)}{4\left[1 + \frac{MSE\left(\hat{\mu}_{Reg}\right)}{\bar{Y}^2}\right]} > 0$$

or

$$\frac{\frac{(MSE\left(\hat{\mu}_{Reg}\right))^2}{\bar{Y}^2}}{\left[1+\frac{MSE\left(\hat{\mu}_{Reg}\right)}{\bar{Y}^2}\right]}+\frac{\lambda C_x^2\left(MSE\left(\hat{\mu}_{Reg}\right)+\lambda\frac{1}{16}C_x^2\bar{Z}^2\right)}{4\left[1+\frac{MSE\left(\hat{\mu}_{Reg}\right)}{\bar{Y}^2}\right]}>0. \tag{4.31}$$

From (4.31), we can see easily that proposed "generalized regression-cum-exponential estimator" is always more efficient than regression estimator of Gupta et al. (2012).

Secondly, we compare the proposed "generalized regression-cum-exponential estimator" with Gupta et al. (2012) generalized regression-cum-ratio estimator

$$MSE\left(\hat{\mu}_{exp1}\right)_{min}<MSE\left(\hat{\mu}_{GRR}\right)_{min} \text{ if}$$

$$\frac{\lambda^2C_x^2\left(\frac{1}{16}C_x^2+C_z^2\left(1-\rho_{zx}^2\right)\right)+4}{4\left[1+\lambda C_z^2\left(1-\rho_{zx}^2\right)\right]}+\frac{C_z^2\left(1-\rho_{zx}^2\right)\lambda\left\{1-\lambda C_x^2\right\}}{C_z^2\left(1-\rho_{zx}^2\right)\lambda+\left\{1-\lambda C_x^2\right\}}>1. \tag{4.32}$$

When the condition (4.32) is satisfied, we can infer that the suggested estimator is more efficient than Gupta et al. (2012) generalized regression-cum-ratio estimator.

4.6 Simulation Study

In this section, we investigate the efficiency of proposed exponential estimators to existing estimators. The simulation study is carried out to compare the *Bias* and *MSE* of the estimators both empirically and theoretically. In the simulation study, we consider two finite populations of size $N = 1000$ generated from a multivariate normal distribution with the same theoretical mean of $[Y, X_1, X_2]$ as $\mu = [5, 5, 5]$ and different covariance matrices as given below.

Population 1

$$\sigma^2=\begin{bmatrix}10 & 3 & 2.9\\ 3 & 2 & 1.1\\ 2.9 & 1.1 & 2\end{bmatrix},\ \rho_{X1Y}=0.6817,\ \rho_{X2Y}=0.6705.$$

Population 2

$$\sigma^2=\begin{bmatrix}6 & 3 & 2.9\\ 3 & 2 & 1.1\\ 2.9 & 1.1 & 2\end{bmatrix},\ \rho_{X1Y}=0.8706,\ \rho_{X2Y}=0.8706.$$

The scrambling variable S is taken to be a normal variable with mean equal to zero and standard deviation equal to 10% of the standard deviation of X_1. The reported response is given by $Z = Y + S$. For each population we considered four sample sizes:

$n = 50, 100, 200$ and 300. The percent relative efficiency (*PRE*) is calculated from following equations

$$PRE = \frac{MSE(\hat{\mu}_Y)}{MSE(\hat{\mu}_i)} \times 100,$$

where $i = R, Reg, GRR, exp1, exp2$.

The empirical *MSE*, theoretical *MSE* and Percent Relative Efficiency (*PRE*) values for all estimators are given in Table 4.1 and Table 4.2.

Table 4.1: Empirical *MSE*, theoretical MSE correct up to 1^{st} order approximation and *PRE* of all estimators.

Population				***MSE* Estimation**		
N	ρ_{X1Y} ρ_{X2Y}	n	Estimator	Empirical	Theoretical	*PRE*
1000	0.6817 0.6705	50	$\hat{\mu}_Y$	0.1193	0.1953	100.00
			$\hat{\mu}_R$	0.1193	0.1145	170.64
			$\hat{\mu}_{Reg}$	0.1083	0.1047	186.50
			$\hat{\mu}_{GRR}$	0.1094	0.1043	187.26
			$\hat{\mu}_{exp1}$	0.1089	0.1042	187.34
			$\hat{\mu}_{exp2}$	0.0857	0.0827	236.13
		100	$\hat{\mu}_Y$	0.0900	0.0925	100.00
			$\hat{\mu}_R$	0.0544	0.0542	170.64
			$\hat{\mu}_{Reg}$	0.0499	0.0496	186.50
			$\hat{\mu}_{GRR}$	0.0503	0.0495	186.86
			$\hat{\mu}_{exp1}$	0.0501	0.0495	186.90
			$\hat{\mu}_{exp2}$	0.0390	0.0393	235.69
		200	$\hat{\mu}_Y$	0.0404	0.0411	100.00
			$\hat{\mu}_R$	0.0240	0.0241	170.64
			$\hat{\mu}_{Reg}$	0.0220	0.0220	186.50
			$\hat{\mu}_{GRR}$	0.0221	0.0220	186.66
			$\hat{\mu}_{exp1}$	0.0220	0.0220	186.67
			$\hat{\mu}_{exp2}$	0.0172	0.0175	235.47
		300	$\hat{\mu}_Y$	0.0236	0.0240	100.00
			$\hat{\mu}_R$	0.0141	0.0141	170.64
			$\hat{\mu}_{Reg}$	0.0129	0.0129	186.50
			$\hat{\mu}_{GRR}$	0.0130	0.0129	186.59
			$\hat{\mu}_{exp1}$	0.0130	0.0129	186.60
			$\hat{\mu}_{exp2}$	0.0103	0.0102	235.40

From Table 4.1 and Table 4.2 we can confirm that suggested generalized regression-cum-exponential estimator is always more efficient than Gupta et al. (2012) regression estimator. Generalized regression-cum-exponential estimator is the most efficient estimator for using one auxiliary variable. Suggested exponential estimator with two auxiliary variables performs better than the estimator with one auxiliary variable, as expected.

Table 4.2: Table 4.1 Continued.

Population				*MSE* Estimation		
N	ρ_{X1Y} ρ_{X2Y}	n	Estimator	Empirical	Theoretical	*PRE*
1000	0.8706 0.8428	50	$\hat{\mu}_Y$	0.1198	0.1181	100.00
			$\hat{\mu}_R$	0.0400	0.0395	299.42
			$\hat{\mu}_{Reg}$	0.0287	0.0289	409.40
			$\hat{\mu}_{GRR}$	0.0291	0.0288	409.86
			$\hat{\mu}_{exp1}$	0.0289	0.0288	410.03
			$\hat{\mu}_{exp2}$	0.0078	0.0073	1626.30
		100	$\hat{\mu}_Y$	0.0547	0.0560	100.00
			$\hat{\mu}_R$	0.0188	0.0187	299.42
			$\hat{\mu}_{Reg}$	0.0138	0.0137	409.40
			$\hat{\mu}_{GRR}$	0.0140	0.0137	409.62
			$\hat{\mu}_{exp1}$	0.0139	0.0137	409.70
			$\hat{\mu}_{exp2}$	0.0037	0.0034	1625.73
		200	$\hat{\mu}_Y$	0.0246	0.0249	100.00
			$\hat{\mu}_R$	0.0086	0.0083	299.42
			$\hat{\mu}_{Reg}$	0.0063	0.0061	409.40
			$\hat{\mu}_{GRR}$	0.0063	0.0061	409.49
			$\hat{\mu}_{exp1}$	0.0063	0.0061	409.53
			$\hat{\mu}_{exp2}$	0.0017	0.0015	1625.44
		300	$\hat{\mu}_Y$	0.0143	0.0145	100.00
			$\hat{\mu}_R$	0.0049	0.0048	299.42
			$\hat{\mu}_{Reg}$	0.0037	0.0035	409.40
			$\hat{\mu}_{GRR}$	0.0037	0.0035	409.45
			$\hat{\mu}_{exp1}$	0.0037	0.0035	409.47
			$\hat{\mu}_{exp2}$	0.0011	0.0009	1625.35

4.7 Numerical Example

We consider the real population used in Sousa et al (2010) and in Gupta et al. (2012). It is based on the survey on Information and Communication Technologies (ICT) usage in enterprises in 2009 with seat in Portugal (Smilhily and Storm, 2010). This survey intends to promote the development of the national statistical system in the information society and to contribute to a deeper knowledge about the usage of ICT by enterprises. The target population covers all industries with one and more persons employed in the sections of economic activity C (Manufacturing) to N (Administrative and support service activities) and S (Other service activities), from NACE[1] Rev. 2 (Eurostat, 2008). The data are essentially collected using Electronic Data Interchange, applying direct connection between information systems at the respondent and the National Statistics Institute. For some enterprises the paper questionnaire is still used. The questions in the structural business surveys mainly deal with characteristics that can be found in the organisations' annual reports and financial statements, such as employment, turnover and investment.

[1]NACE is derived from the French title "Nomenclature générale des Activités économiques dans les Communautés Européennes" (Statistical classification of economic activities in the European Communities).

In our application the study variable Y is the purchase orders in 2010, collected by the ICT survey in that year. This is typically a confidential variable for enterprises, only known from business surveys. The auxiliary variable X is the turnover of each enterprise. This information can be easily obtained from enterprise records available in the public domain, as administrative information. In 2010 the population survey contained approximately 278000 enterprises and we know the value of X for all these enterprises. The purchase orders information was collected in the ICT survey and we have the values of Y for 5336 enterprises (which answered this question in the ICT survey in 2010). For this study, these 5336 enterprises are considered as our population. The scrambling variable S is taken to be a normal random variable with mean equal to zero and standard deviation equal to 10% of the standard deviation of X, that is $\sigma_S = 0.1\sigma_X$. The reported response is given by $Z = Y + S$ (the purchase order value plus a random quantity). The variables Y and X are strongly correlated so we can take advantage of this correlation by using the ratio and regression estimators.

Population Characteristics:

$N = 5336, \rho_{XY} = 0.9632$
$\mu_X = 22.99, \mu_Y = 30.19, \sigma_X = 172.09, \sigma_Y = 138.65$ (in millions of Euros)
and $\beta_{YX} = 0.7763$

We use the following samples sizes in our simulation study: $n = 100, 500, 1000$ and 2000.

The empirical, theoretical *MSE* and Percent Relative Efficiency (*PRE*) values for all estimators are given in Table 4.3.

Table 4.3: *MSE* and *PRE* for the ratio estimator ($\hat{\mu}_R$), the regression estimator ($\hat{\mu}_{Reg}$), the generalized regression-cum-ratio estimator ($\hat{\mu}_{GRR}$) and the exponential estimator ($\hat{\mu}_{exp1}$) relative to the RRT mean estimator.

Population				***MSE* Estimation**		
N	ρ_{XY}	n	Estimator	Empirical	Theoretical	*PRE*
5336	0.9636	100	$\hat{\mu}_Y$	196.8002	191.5088	100.00
			$\hat{\mu}_R$	11.3683	16.4393	1164.94
			$\hat{\mu}_{Reg}$	16.7963	16.3768	1169.39
			$\hat{\mu}_{GRR}$	11.3909	15.6644	1222.58
			$\hat{\mu}_{exp1}$	12.3339	13.1849	1452.49
		500	$\hat{\mu}_Y$	34.6507	35.3757	100.00
			$\hat{\mu}_R$	2.7259	3.0367	1164.94
			$\hat{\mu}_{Reg}$	3.0170	3.0252	1169.39
			$\hat{\mu}_{GRR}$	2.7509	3.0069	1176.50
			$\hat{\mu}_{exp1}$	2.8631	2.9173	1212.61
		1000	$\hat{\mu}_Y$	15.7543	15.8591	100.00
			$\hat{\mu}_R$	1.3092	1.3614	1164.94
			$\hat{\mu}_{Reg}$	1.3592	1.3562	1169.39
			$\hat{\mu}_{GRR}$	1.3175	1.3526	1172.47
			$\hat{\mu}_{exp1}$	1.3381	1.3345	1188.38
		2000	$\hat{\mu}_Y$	6.2451	6.1008	100.00
			$\hat{\mu}_R$	0.5691	0.5237	1164.94
			$\hat{\mu}_{Reg}$	0.5573	0.5217	1169.39
			$\hat{\mu}_{GRR}$	0.5718	0.5212	1170.55
			$\hat{\mu}_{exp1}$	0.5673	0.5185	1176.62

From Table 4.3 we can say that generalized regression-cum-exponential estimator has the largest *PRE*.

4.8 Conclusions

This paper proposed type estimators using non-sensitive one or two auxiliary variables to improve the efficiency of RRT estimators of mean. The expression for *Bias* and *MSE* are derived. We found that the proposed exponential type estimators are more efficient than the existing estimators in literature. These results are also supported with a simulation study and using a real data.

References

BAHL, S. & TUTEJA, R. K. 1991. Ratio and product type exponential estimators. *Information and Optimization Sciences*, 12(1), 159-163.

EICHHORN, B. H. & HAYRE, L. S. 1983. Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*, 7, 307-316.

EUROSTAT. 2008. NACE Rev. 2 - Statistical classification of economic activities in the European Community. *Official Publications of the European Communities*, 112-285 and 306-311.

GROVER, L.K. 2010. A correction note on improvement in variance estimation using auxiliary information. *Communications in Statistics - Theory and Methods*, 39, 753-764.

GUPTA, S. N., GUPTA, B. C. & SINGH, S. 2002. Estimation of sensitivity level of personal interview survey questions. *Journal of Statistical Planning and Inference*, 100, 239-247.

GUPTA, S. & SHABBIR, J. 2004. Sensitivity estimation for personal interview survey questions. *Statistica*, 64, 643-653.

GUPTA, S., SHABBIR, J. & SEHRA, S. 2010. Mean and sensitivity estimation in optional randomized response models. *Journal of Statistical Planning and Inference*, 140(10), 2870-2874.

GUPTA, S., SHABBIR, J., SOUSA, R. & REAL, P.C. 2012. Estimation of the Mean of a Sensitive Variable in the Presence of Auxiliary Information. *Communications in Statistics - Theory and Methods*, 41(13-14), 2394-2404.

KOYUNCU, N. 2012. Efficient estimators of population mean using auxiliary attributes. *Applied Mathematics and Computation*, 218, 10900-10905.

KOYUNCU, N. & KADILAR, C. 2009. Family of estimators of population mean using two auxiliary variables in stratified random sampling. *Communications in Statistics: Theory and Methods*, 38(14), 2398-2417.

PERRI, P. F. 2008. Modified randomized devices for Simmons' model. *Model Assisted Statistics and Applications*, 3, 233-239.

SAHA, A. 2008. A randomized response technique for quantitative data under unequal probability sampling. *Journal of statistical Theory and Practice*, 2(4), 589-596.

SHABBIR, J. & GUPTA, S. 2007. On improvement in variance estimation using auxiliary information. *Communication in Statistics-Theory and Methods*, 36(12), 2177-2185.

SMILHILY, M. & STORM, H. 2010. ICT usage in enterprises - 2009. *Eurostat Publications*, Issue 1.

SOUSA, R., SHABBIR, J. REAL, P. C. & GUPTA, S. 2010. Ratio estimation of the mean of a sensitive variable in the presence of auxiliary information. *Journal of Statistical Theory and Practice*, 4(3), 495-507.

UPADHYAYA, L.N. & SINGH, H.P. 2006. Almost unbiased ratio and product-type estimators of finite population variance in sample surveys. *Statistics in Transition*, 7(5), 1087-1096.

WU, J-W, TIAN, G-L & TANG, M-L. 2008. Two new models for survey sampling with sensitive characteristic: design and analysis. *Metrika*, 67, 251-263.

Appendix C - R Routines

Listing 4.1: R Code for Simulation Study of Proposed Estimator in Chapter 4

```

proj_exponential <- function(N,sigma,mu)
{

  #Generation of a bivariate normal population
  data_yx <- mvrnorm(N, mu, sigma)

  #Study variable
  Y <- data_yx[,1]
  #Auxiliary variable, correlated with Y
  X <- data_yx[,2]

  #Coefficient of correlation between Y and X
  Ro_YX <- cor(Y,X)

  #Scrambling variable independent of Y and X, with mean=0
  S <- rnorm(N,mean=0,sd=0.1*sd(X))

  #Scrambled response
  Z <- Y+S

  #Coefficient of correlation between Z and X
  Ro_ZX <- Ro_YX/sqrt(1+(var(S)/var(Y)))

  #population
  univ <- data.frame(cbind(Y=Y,S=S,Z=Z,X=X,NRAND=runif(N)))
  univ <- univ[order(univ$NRAND),]

  #Mean of Y
  mz <- mean(univ$Z)
  mx <- mean(univ$X)
  my <- mean(univ$Y)

  mu11 <- sum((univ$Z-mz)*(univ$X-mx))/(N-1)
  mu12 <- sum((univ$Z-mz)*((univ$X-mx)^2))/(N-1)
  mu02 <- sum((univ$X-mx)^2)/(N-1)
  mu03 <- sum((univ$X-mx)^3)/(N-1)

  beta_zx <- Ro_YX*(sd(univ$Y)/sd(univ$X))

  #Samples dimension
  dim_samp <- c(50,100,200,300)

  #Initialize the variables...

  for (i in 1:length(dim_samp))
```

```
  {
    #sample dimension
    n <- dim_samp[i]
    #sample
    samp <- univ[1:n,]
    #Sampling rate
    f <- n/N

    #Ratio
    R <- mean(univ$X)/mean(samp$X)

    #Ordinary meam
    est1 <- mean(samp$Z)
    #Ratio estimator
    est2 <- mean(samp$Z)*(mx/mean(samp$X))
    #Regression estimator
    est3 <- mean(samp$Z)+beta_zx*(mx-mean(samp$X))
    #Regression-cum-ratio estimator
    est4 <- (mean(samp$Z)+beta_zx*(mx-mean(samp$X)))*(mx/mean(samp$X))

    #Coefficient of variation
    c_x <- sd(univ$X)/mx
    c_y <- sd(univ$Y)/my
    c2_x <- c_x^2
    c2_y <- c_y^2
    c2_z <- c2_y+(var(univ$S)/(my^2))
    c_z <- sqrt(c2_z)

    A <- (1+((1-f)/n)*(c2_z+c2_x-2*Ro_ZX*c_z*c_x))
    B <- (((1-f)/n)*(Ro_ZX*c_z*c_x-0.75*c2_x)-2)
    w1 <- (c2_x/2)*(-B-((1-f)/n)*(c2_x-Ro_ZX*c_z*c_x))
        /(A*c2_x-((1-f)/n)*((c2_x-Ro_ZX*c_z*c_x)^2))
    w2 <- (my*c2_x-2*w1*my*(c2_x-Ro_ZX*c_z*c_x))/(2*mx*c2_x)

    #Auxiliar coefficients
    k1 <- (1-((1-f)*c2_x/n))/(1-((1-f)/n)*(c2_x-c2_z*(1-(Ro_ZX^2))))
    k2 <- (my/mx)*(1+k1*((Ro_ZX*c_z/c_x)-2))

    #Generalized Regression-cum-ratio Estimator
    est5 <- (k1*mean(samp$Z)+k2*(mx-mean(samp$X)))*(mx/mean(samp$X))

    #Exponential Type Estimator
    est6 <- (mean(samp$Z)+beta_zx*(mx-mean(samp$X)))
        *exp((mx-mean(samp$X))/(mx+mean(samp$X)))
    #Generalized Exponential Type Estimator
    est7 <- (w1*mean(samp$Z)+w2*(mx-mean(samp$X)))
        *exp((mx-mean(samp$X))/(mx+mean(samp$X)))

    #Mean Square Error of 1st estimator (ordinal mean)
    mse1 <- ((1-f)/n)*(var(univ$Y)+var(univ$S))
```

```

#Bias of ratio estimator - 1st degree approximation
bias2i <- ((1-f)/n)*my*(c2_x-Ro_ZX*c_z*c_x)
#Mean Square Error of ratio estimator - 1st degree approximation
mse2i <- ((1-f)/n)*(my^2)*(c2_z+c2_x-2*Ro_ZX*c_z*c_x)

#Bias of regression estimator - 1st degree approximation
bias3i <- -beta_zx*((1-f)/n)*((mu12/mu11)-(mu03/mu02))
#Mean Square Error of regression estimator - 1st degree approximation
mse3i <- ((1-f)/n)*(my^2)*c2_z*(1-(Ro_ZX^2))

#Bias of regression-cum-ratio estimator - 1st degree approximation
bias4i <- ((1-f)/n)*(my*c2_x-beta_zx*((mu12/mu11)-(mu03/mu02)))
#Mean Square Error of regression-cum-ratio estimator
#1st degree approximation
mse4i <- ((1-f)/n)*(my^2)*(c2_z*(1-(Ro_ZX^2))+c2_x)

#Bias of genetalized regression-cum-ratio estimator
#1st degree approximation
bias5i <- (k1-1)*my+k1*my*((1-f)/n)*(c2_x-Ro_ZX*c_z*c_x)
      +k2*mx*((1-f)/n)*c2_x
#Mean Square Error of generalized regression-cum-ratio estimator
#1st degree approximation
mse5i <- ((k1-1)^2)*(my^2)+(k1^2)*(my^2)*((1-f)/n)
    *(c2_z+3*c2_x-4*Ro_ZX*c_z*c_x)+(k2^2)*(mx^2)*((1-f)/n)
    *c2_x-2*k1*(my^2)*((1-f)/n)*(c2_x-Ro_ZX*c_z*c_x)
    -2*k2*my*mx*((1-f)/n)*c2_x-2*k1*k2*my*mx*((1-f)/n)
    *(Ro_ZX*c_z*c_x-2*c2_x)

#Bias of exponential type estimator - 1st degree approximation
bias6i <- ((1-f)/n)*(beta_zx*((mu03/mu02)-(mu12/mu11))+(3/8)*my*c2_x)
#Mean Square Error of exponential type estimator - 1st degree approximation
mse6i <- ((1-f)/n)*(my^2)*(c2_z*(1-(Ro_ZX^2))+0.25*c2_x)

#Bias of generalized exponential type estimator
#1st degree approximation
bias7i <- (w1-1)*my+((1-f)/n)
      *(0.5*w1*my*(0.75*c2_x-Ro_ZX*c_z*c_x)+0.5*w2*mx*c2_x)
#Mean Square Error of generalized exponential type estimator
#1st degree approximation
mse7i <- (mse3i/(1+(mse3i/(my^2))))-((((1-f)/n)* c2_x*(mse3i
      +((1-f)/n)*(1/16)*c2_x*(my^2)))/(4*(1+mse3i/(my^2))))

#Empirical results
#Simulation of 5000 replicas of estimates
...

#Results
res <- rbind(res,c(N,n,Ro_YX,Ro_ZX,R,
                c_x,c_y,c_z,k1,k2,w1,w2,
```

```
                     mx,my,mz,
                     med_est1,med_est2,med_est3,med_est4,
                     med_est5,med_est6,med_est7,
                     bias2i,bias3i,bias4i,bias5i,
                     bias6i,bias7i,
                     emp_mse1,mse1,emp_mse2,mse2i,
                     emp_mse3,mse3i,emp_mse4,mse4i,
                     emp_mse5,mse5i,emp_mse6,mse6i,
                     emp_mse7,mse7i))
  }
  colnames(res) <- c("N","n","RhoXY","RhoZX","R",
                     "Cx","Cy","Cz","k1","k2","w1","w2",
                     "mX","mY","mZ",
                     "Est1","Est2","Est3","Est4",
                     "Est5","Est6","Est7",
                     "BIAS2I","BIAS3I","BIAS4I","BIAS5I",
                     "BIAS6I","BIAS7I",
                     "EMP_MSE1","MSE1","EMP_MSE2","MSE2I",
                     "EMP_MSE3","MSE3I","EMP_MSE4","MSE4I",
                     "EMP_MSE5","MSE5I","EMP_MSE6","MSE6I",
                     "EMP_MSE7","MSE7I")
  return(res)
}

#Package for generation
require(MASS)
N <- 1000

#Parameters
sigma1 <- matrix(c(9,1.9,1.9,4),2,2)
sigma2 <- matrix(c(10,3,3,2),2,2)
sigma3 <- matrix(c(6,3,3,2),2,2)
mu <- c(2,2)

res <- NULL
for (i in 1:length(N))
{
  res <- rbind(res,proj_exponential(N[i],sigma1,mu))
  res <- rbind(res,proj_exponential(N[i],sigma2,mu))
  res <- rbind(res,proj_exponential(N[i],sigma3,mu))
}
write.table(res,"chapter4_ss_results.txt",sep="\t",dec=",",row.names=FALSE)
```

Listing 4.2: R Code for Numerical Example of Proposed Estimator in Chapter 4

```
proj_exponential_real <- function(Y,X,N)
{
  #Coefficient of correlation between Y and X
  Ro_YX <- cor(Y,X)

  #Scrambling variable independent of Y and X, with mean=0
  S <- rnorm(N,mean=0,sd=sd(X)*0.1)
  #Scrambled response
  Z <- Y+S

  #Coefficient of correlation between Z and X
  Ro_ZX <- Ro_YX/sqrt(1+(var(S)/var(Y)))

  #population
  univ <- data.frame(cbind(Y=Y,S=S,Z=Z,X=X,NRAND=runif(N)))
  univ <- univ[order(univ$NRAND),]

  #Mean of Y
  mz <- mean(univ$Z)
  mx <- mean(univ$X)
  my <- mean(univ$Y)
  ms <- mean(univ$S)

  mu11 <- sum((univ$Z-mz)*(univ$X-mx))/(N-1)
  mu12 <- sum((univ$Z-mz)*((univ$X-mx)^2))/(N-1)
  mu02 <- sum((univ$X-mx)^2)/(N-1)
  mu03 <- sum((univ$X-mx)^3)/(N-1)

  beta_zx <- Ro_YX*(sd(univ$Y)/sd(univ$X))

  #Samples dimension
  dim_samp <- c(100,500,1000,2000)

  #Initialize the variables...

  for (i in 1:length(dim_samp))
  {
    #sample dimension
    n <- dim_samp[i]
    #sample
    samp <- univ[1:n,]
    #Sampling rate
    f <- n/N

    #Ratio
    R <- mean(univ$X)/mean(samp$X)
```

```

    #Ordinary meam
    est1 <- mean(samp$Z)
    #Ratio estimator
    est2 <- mean(samp$Z)*(mx/mean(samp$X))
    #Regression estimator
    est3 <- mean(samp$Z)+beta_zx*(mx-mean(samp$X))
    #Regression-cum-ratio estimator
    est4 <- (mean(samp$Z)+beta_zx*(mx-mean(samp$X)))*(mx/mean(samp$X))


    #Coefficient of variation
    c_x <- sd(univ$X)/mx
    c_y <- sd(univ$Y)/my
    c2_x <- c_x^2
    c2_y <- c_y^2
    c2_z <- c2_y+(var(univ$S)/(my^2))
    c_z <- sqrt(c2_z)

    A <- (1+((1-f)/n)*(c2_z+c2_x-2*Ro_ZX*c_z*c_x))
    B <- (((1-f)/n)*(Ro_ZX*c_z*c_x-0.75*c2_x)-2)
    w1 <- (c2_x/2)*(-B-((1-f)/n)*(c2_x-Ro_ZX*c_z*c_x))
         /(A*c2_x-((1-f)/n)*((c2_x-Ro_ZX*c_z*c_x)^2))
    w2 <- (my*c2_x-2*w1*my*(c2_x-Ro_ZX*c_z*c_x))/(2*mx*c2_x)

    #Auxiliar coefficients
    k1 <- (1-((1-f)*c2_x/n))/(1-((1-f)/n)*(c2_x-c2_z*(1-(Ro_ZX^2))))
    k2 <- (my/mx)*(1+k1*((Ro_ZX*c_z/c_x)-2))

    #Generalized Regression-cum-ratio Estimator
    est5 <- (k1*mean(samp$Z)+k2*(mx-mean(samp$X)))*(mx/mean(samp$X))

    #Exponential Type Estimator
    est6 <- (mean(samp$Z)+beta_zx*(mx-mean(samp$X)))
        *exp((mx-mean(samp$X))/(mx+mean(samp$X)))
    #Generalized Exponential Type Estimator
    est7 <- (w1*mean(samp$Z)+w2*(mx-mean(samp$X)))
        *exp((mx-mean(samp$X))/(mx+mean(samp$X)))

    #Mean Square Error of 1st estimator (ordinal mean)
    mse1 <- ((1-f)/n)*(var(univ$Y)+var(univ$S))

    #Bias of ratio estimator - 1st degree approximation
    bias2i <- ((1-f)/n)*my*(c2_x-Ro_ZX*c_z*c_x)
    #Mean Square Error of ratio estimator - 1st degree approximation
    mse2i <- ((1-f)/n)*(my^2)*(c2_z+c2_x-2*Ro_ZX*c_z*c_x)

    #Bias of regression estimator - 1st degree approximation
    bias3i <- -beta_zx*((1-f)/n)*((mu12/mu11)-(mu03/mu02))
    #Mean Square Error of regression estimator - 1st degree approximation
```

```
    #mse3i <- ((1-f)/n)*var(univ$Y)*((1+(var(univ$S)/var(univ$Y)))-(Ro_YX^2))
    mse3i <- ((1-f)/n)*(my^2)*c2_z*(1-(Ro_ZX^2))

    #Bias of regression-cum-ratio estimator - 1st degree approximation
    bias4i <- ((1-f)/n)*(my*c2_x-beta_zx*((mu12/mu11)-(mu03/mu02)))
    #Mean Square Error of regression-cum-ratio estimator
    #1st degree approximation
    mse4i <- ((1-f)/n)*(my^2)*(c2_z*(1-(Ro_ZX^2))+c2_x)

    #Bias of genetalized regression-cum-ratio estimator
    #1st degree approximation
    bias5i <- (k1-1)*my+k1*my*((1-f)/n)*(c2_x-Ro_ZX*c_z*c_x)
          +k2*mx*((1-f)/n)*c2_x
    #Mean Square Error of generalized regression-cum-ratio estimator
    #1st degree approximation
    mse5i <- ((k1-1)^2)*(my^2)+(k1^2)*(my^2)*((1-f)/n)
         *(c2_z+3*c2_x-4*Ro_ZX*c_z*c_x)+(k2^2)*(mx^2)
         *((1-f)/n)*c2_x-2*k1*(my^2)*((1-f)/n)
         *(c2_x-Ro_ZX*c_z*c_x)-2*k2*my*mx*((1-f)/n)
         *c2_x-2*k1*k2*my*mx*((1-f)/n)*(Ro_ZX*c_z*c_x-2*c2_x)

    #Bias of exponential type estimator - 1st degree approximation
    bias6i <- ((1-f)/n)*(beta_zx*((mu03/mu02)-(mu12/mu11))+(3/8)*my*c2_x)
    #Mean Square Error of exponential type estimator - 1st degree approximation
    mse6i <- ((1-f)/n)*(my^2)*(c2_z*(1-(Ro_ZX^2))+0.25*c2_x)

    #Bias of generalized exponential type estimator
    #1st degree approximation
    bias7i <- (w1-1)*my+((1-f)/n)*(0.5*w1*my*(0.75*c2_x-Ro_ZX*c_z*c_x)
          +0.5*w2*mx*c2_x)
    #Mean Square Error of generalized exponential type estimator
    #1st degree approximation
    mse7i <- (mse5i/(1+(mse5i/(my^2))))
        -((((1-f)/n)*c2_x*(mse5i+((1-f)/n)*(1/16)
        *c2_x*(my^2)))/(4*(1+mse5i/(my^2))))

    #Empirical results
    #Simulation of 5000 replicas of estimates
    ...

    #Results
    res <- rbind(res,c(N,n,Ro_YX,Ro_ZX,R,
                   c_x,c_y,c_z,k1,k2,w1,w2,
                   mx,my,mz,
                   med_est1,med_est2,med_est3,med_est4,
                   med_est5,med_est6,med_est7,
                   bias2i,bias3i,bias4i,bias5i,
                   bias6i,bias7i,
                   emp_mse1,mse1,emp_mse2,mse2i,
                   emp_mse3,mse3i,emp_mse4,mse4i,
```

```
                      emp_mse5,mse5i,emp_mse6,mse6i,
                      emp_mse7,mse7i))
  }
  colnames(res) <- c("N","n","RhoXY","RhoZX","R",
                      "Cx","Cy","Cz","k1","k2","w1","w2",
                      "mX","mY","mZ",
                      "Est1","Est2","Est3","Est4",
                      "Est5","Est6","Est7",
                      "BIAS2I","BIAS3I","BIAS4I","BIAS5I",
                      "BIAS6I","BIAS7I",
                      "EMP_MSE1","MSE1","EMP_MSE2","MSE2I",
                      "EMP_MSE3","MSE3I","EMP_MSE4","MSE4I",
                      "EMP_MSE5","MSE5I","EMP_MSE6","MSE6I",
                      "EMP_MSE7","MSE7I")
  return(res)
}

#Package for generation
require(MASS)

#Import data
data_yx <- read.table("IUTICE10.txt",sep="\t",dec=",",header = T)
#Study variable (purchase, millions of euros)
Y <- data_yx[,3]
#Auxiliary variable, correlated with Y (turnover, millions of euros)
X <- data_yx[,2]

#Data application
N <- dim(data_yx)[1]
res <- proj_exponential_real(Y,X,N)

#Export data
write.table(res,"chapter4_ne_results.txt",sep="\t",dec=",",row.names=FALSE)
```

5

Improved Exponential Type Estimators of the Mean of a Sensitive Variable in the Presence of Non-Sensitive Auxiliary Information

Abstract

Recently Koyuncu et al. (2013) proposed an exponential type estimator to improve the efficiency of mean estimator based on Randomized Response Technique (RRT). In this paper, we propose an improved exponential type estimator which is more efficient than the Koyuncu et al. (2013) estimator, which in turn was shown to be more efficient than the usual mean estimator, ratio estimator, regression estimator, and the Gupta et al. (2012) estimator. Under simple random sampling without replacement (SRSWOR) scheme, *Bias* and Mean Square Error (*MSE*) expressions for the proposed estimator are obtained up to first order of approximation and comparisons are made with the Koyuncu et al. (2013) estimator. A simulation study is used to observe the performances of these two estimators. Theoretical findings are also supported by a numerical example with real data.

Submitted as: GUPTA, S., SHABBIR, J., SOUSA, R. & REAL, P. C. 2013. Improved exponential type estimators of the mean of a sensitive variable in the presence of non-sensitive auxiliary information.

5.1 Introduction

This study proposes an improved exponential type estimator for estimating the population mean of a sensitive variable when information about a non-sensitive auxiliary variable is available. A common problem in conducting a statistical sample survey is that of response bias in the face of sensitive questions. Warner (1965) introduced the Randomized Response Technique (RRT) in order to solve this problem. Our main purpose in this study is to improve the mean estimation of a sensitive variable based on a RRT when some non-sensitive auxiliary information is available.

Many authors such as Kadilar and Cingi (2004), Kadilar et al. (2007), Shabbir and Gupta (2007, 2010) and Nangsue (2009) have presented ratio and regression estimators when both the study variable and the auxiliary variable are directly observable.

In this study we propose an exponential type estimator for the mean of a sensitive variable using known information on a correlated but non-sensitive auxiliary variable. The proposed estimator performs better than the recently introduced estimator by Koyuncu et al. (2013) which was shown to outperform many existing estimators of this type.

5.2 Terminology

Consider a finite population with N units $U = (U_1, U_2, ..., U_N)$ from which a sample of size n is drawn using simple random sampling without replacement (SRSWOR). Let Y be the study variable, a sensitive variable which cannot be observed directly due to respondent bias. Let X be the non-sensitive auxiliary variable which is correlated with Y. Let S be a scrambling variable independent of Y and X. The respondent is asked to report a scrambled response for Y given by $Z = Y + S$ but is asked to provide a true response for X. Let $(\bar{y}, \bar{x})$ be the sample means corresponding to $\left(\bar{Y}, \bar{X}\right)$, the population means of Y and X, respectively. Consider $\bar{Z}$ to be the population mean of the scrambled variable Z.

Let S_x^2 and s_x^2 respectively be the population variance and the sample variance of X. On the other hand, S_{zx}^2 and s_{zx}^2 are the population covariance and the sample covariance between Z and X, respectively.

To obtain the *Bias* and *MSE* expressions, let us define $e_0 = \frac{\bar{z}-\bar{Z}}{\bar{Z}}$, $e_1 = \frac{\bar{x}-\bar{X}}{\bar{X}}$, $e_2 = \frac{s_x^2-S_x^2}{s_x^2}$ and $e_3 = \frac{s_{zx}-S_{zx}}{s_{zx}}$ such that $E(e_i) = 0,\ i = 0, 1, 2, 3$. To first degree of approximations, we have:

$E(e_0^2) = \lambda C_z^2 = v_{20}$, $E(e_1^2) = \lambda C_x^2 = v_{02}$, $E(e_0 e_1) = \lambda C_{zx} = \lambda \rho_{zx} C_z C_x = v_{11}$,
$E(e_1 e_2) = \lambda \dfrac{\mu_{03}}{\bar{X}\mu_{02}}$, $E(e_1 e_3) = \lambda \dfrac{\mu_{12}}{\bar{X}\mu_{11}}$,

where $\lambda = \dfrac{1-f}{n}$, $f = n/N$, $C_{zx} = \rho_{zx} C_z C_x$ and $\mu_{rs} = \dfrac{1}{N-1} \sum_{i=1}^{N} (z_i - \bar{Z})^r (x_i - \bar{X})^s$.

5.3 Difference-cum-exponential Estimator (Koyuncu et al., 2013)

Recently Koyuncu et al. (2013) have suggested a combination of the difference estimator and the exponential estimator with some gain in the efficiency. This estimator is given by

$$\hat{\mu}_{DE} = \left[w_1 \bar{z} + w_2 \left(\bar{X} - \bar{x}\right)\right] exp\left(\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}}\right), \tag{5.1}$$

where w_1 and w_2 are constants.

The *Bias* and *MSE* of $\hat{\mu}_{DE}$, up to first degree of approximation, at optimum values

$$w_{1(opt)} = \frac{1 - \left(\lambda C_x^2/8\right)}{1 + \lambda C_z^2 \left(1 - \rho_{zx}^2\right)} \text{ and } w_{2(opt)} = \frac{\bar{Y}}{\bar{X}} \left\{\frac{1}{2} - w_{1(opt)} \left(1 - \frac{\rho_{zx} C_z}{C_x}\right)\right\}$$

are given by

$$Bias(\hat{\mu}_{DE}) \cong (w_{1(opt)} - 1)\bar{Y} + w_{1(opt)} \bar{Y} \lambda \left\{\frac{3}{8} C_x^2 - \frac{1}{2} \rho_{zx} C_z C_x\right\} + w_{2(opt)} \bar{X} \lambda C x^2, \tag{5.2}$$

and

$$MSE(\hat{\mu}_{DE})_{opt} \cong \bar{Y}^2 \left\{\left(1 - \frac{1}{4}\lambda C_x^2\right) - \frac{\left(1 - \frac{1}{8}\lambda C_x^2\right)^2}{1 + \lambda C_z^2 \left(1 - \rho_{zx}^2\right)}\right\}$$

or

$$MSE(\hat{\mu}_{DE})_{opt} \cong \bar{Y}^2 \left\{\left(1 - \frac{1}{4} v_{02}\right) - \frac{v_{02} \left(8 - v_{02}\right)^2}{64 \left(v_{02} + v_{20} v_{02} - v_{11}^2\right)}\right\}. \tag{5.3}$$

It is shown in Koyuncu et al. (2013) that this estimator is better than all the other similar estimators such as Sousa et al. (2010) and Gupta et al. (2012).

5.4 Proposed estimator

The combined product estimators have shown advantages in efficiency spite of being more biased than the traditional ratio or regression estimators (Koyuncu et al., 2013). Motivated by this fact, we propose a change in the difference-cum-exponential estimator in (5.1) so that when the sample mean $(\bar{x})$ of the auxiliary variable X is close to population mean $(\bar{X})$, the expected value of the proposed estimator is closer to the variable of interest Y. So, our proposed estimator is an improved exponential estimator as an modified version of the difference-cum-exponential estimator in (5.1) and is given by the following expression

$$\hat{\mu}_{IE} = [d_1 \bar{z} + d_2] \, exp \left(\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}} \right), \tag{5.4}$$

where d_1 and d_2 are constants.

Using Taylor's approximation and retaining terms of order up to 2, (5.4) can be rewritten as

$$\hat{\mu}_{IE} - \bar{Z} \cong \left[(d_1 - 1)\bar{Z} + d_1 \bar{Z} e_0 + d_2\right] \left\{1 - \frac{1}{2}e_1 + \frac{3}{8}e_1^2\right\}. \tag{5.5}$$

Recognizing that $\bar{Z} = \bar{Y}$, the optimum *Bias* and *MSE* of $\hat{\mu}_{IE}$, to first degree of approximation, are given by

$$Bias(\hat{\mu}_{IE}) \cong (d_1 - 1)\bar{Y} + d_1 \bar{Y} \left(\frac{3}{8} v_{02} - \frac{1}{2} v_{11} \right) + d_2 \left(1 + \frac{3}{8} v_{02} \right), \tag{5.6}$$

and

$$MSE(\hat{\mu}_{IE}) \cong d_1^2 \bar{Y}^2 A + d_2^2 B - 2d_1 \bar{Y}^2 C - 2d_2 \bar{Y} D + 2d_1 d_2 \bar{Y} E + \bar{Y}^2, \tag{5.7}$$

where $A = 1 + v_{20} + v_{02} - 2v_{11}$, $B = 1 + v_{02}$, $C = 1 + \frac{3}{8}v_{02} - \frac{1}{2}v_{11}$, $D = 1 + \frac{3}{8}v_{02}$, $E = 1 + v_{02} - v_{11}$.

Using (5.7), the optimum values are

$$d_{1(opt)} = \frac{BC - DE}{AB - E^2},$$

and

$$d_{2(opt)} = \frac{\bar{Y}\,(AD - CE)}{AB - E^2}.$$

Considering the *MSE* at optimum values we get

$$MSE(\hat{\mu}_{IE})_{opt} \cong \bar{Y}^2 \left[1 - \frac{BC^2 + AD^2 - 2CDE}{AB - E^2} \right]$$

or

$$MSE(\hat{\mu}_{IE})_{opt} \cong \bar{Y}^2 \left[1 - \frac{v_{20} + \frac{3}{4} v_{20} v_{02} \left(1 - \rho_{zx}^2\right) \left(1 + \frac{3}{16} v_{02}\right) + \frac{1}{64} v_{02} v_{11}^2}{v_{20} + v_{02} v_{20} \left(1 - \rho_{zx}^2\right)} \right], \tag{5.8}$$

where $\rho_{zx} = \frac{v_{11}}{\sqrt{v_{20}}\sqrt{v_{02}}}$.

Comparing the *MSE* of this estimator to the *MSE* of difference-cum-exponential estimator given in (5.3), we note that the proposed estimator will be more efficient if

$$MSE(\hat{\mu}_{IE})_{opt} < MSE(\hat{\mu}_{DE})_{opt}.$$

This will be so if

$$\frac{64v_{20}-48v_{11}^2-8v_{02}v_{11}^2+48v_{20}v_{02}+9v_{20}v_{02}^2}{64\left(v_{20}+v_{02}v_{20}-v_{11}^2\right)}-\frac{v_{02}\left(8-v_{02}\right)^2}{64\left(v_{02}+v_{02}v_{20}-v_{11}^2\right)}-\frac{1}{4}v_{02}>0$$

or if

$$\frac{\left[\begin{array}{c} 64v_{20}^2v_{02}+32v_{20}^2v_{02}^2+8v_{20}v_{02}^3-7v_{20}^2v_{02}^3-v_{20}v_{02}^4+15v_{02}^2v_{20}v_{11}^2-80v_{20}v_{02}v_{11}^2 \\ -16v_{20}v_{02}^2-64v_{20}v_{11}^2-8v_{02}^2v_{11}^2-8v_{02}v_{11}^4+v_{02}^3v_{11}^2+48v_{11}^4+16v_{02}v_{11}^2 \end{array}\right]}{M}>0,$$

where $M=64\left[\left\{v_{20}+v_{20}v_{02}\left(1-\rho_{zx}^2\right)\right\}\left\{v_{02}+v_{20}v_{02}\left(1-\rho_{zx}^2\right)\right\}\right]$,

or if

$$\frac{v_{20}v_{02}\left(1-\rho_{zx}^2\right)\left[8\left(8v_{20}+v_{02}v_{11}^2-6v_{11}^2\right)-v_{02}\left(4-v_{02}\right)^2+v_{20}v_{02}\left(32-7v_{02}\right)\right]}{64\left[\left\{v_{20}+v_{20}v_{02}\left(1-\rho_{zx}^2\right)\right\}\left\{v_{02}+v_{20}v_{02}\left(1-\rho_{zx}^2\right)\right\}\right]}>0. \tag{5.9}$$

The above condition is likely to be true if numerator is positive.

5.5 Simulation Study

In this section, we conduct a simulation study with particular focus on comparing the performance of the proposed combined estimator $\hat{\mu}_P$ to the estimator $\hat{\mu}_{DE}$ suggested by Koyuncu et al. (2013), using the *Bias* and *MSE* results, correct up to first order of approximation.

We consider 2 different bivariate normal distributions for (Y, X). The scrambling variable S is taken to be a normal variable with mean equal to zero and standard deviation equal to 10% of the standard deviation of X. The reported response is given by $Z = Y+S$. The summary statistics about the bivariate normal populations are given below.

Population Statistics:

I	$N=1000, \mu_Y=2, \sigma_Y=\sqrt{10}, \mu_X=2, \sigma_X=\sqrt{2}, \sigma_{XY}=3$ and $\rho_{XY}=0.6708$
II	$N=1000, \mu_Y=2, \sigma_Y=\sqrt{6}, \mu_X=2, \sigma_X=\sqrt{2}, \sigma_{XY}=3$ and $\rho_{XY}=0.8660$

We take samples of size $n = 50, 100, 200$ and 300 from each population to compare the results. We estimate the empirical *Bias* and *MSE* using 5000 samples of various sizes from the study populations. The absolute relative bias (*ARB*) is given by

$$\left|\frac{Bias(\mu_\alpha)}{\bar{Y}}\right|,$$

where $\alpha = DE$ and IE.

The empirical and the theoretical results for the two estimators under study are presented in Table 5.1 and Table 5.2, respectively. From these tables we can observe that the proposed estimator shows reduced *Bias* when compared to other estimator.

Table 5.1: Empirical *ARB* for the difference-cum-exponential estimator ($\hat{\mu}_{DE}$) and for the improved exponential estimator ($\hat{\mu}_{IE}$).

Population			**Empirical *ARB***			
N	ρ_{XY}	Estimator	$n = 50$	$n = 100$	$n = 200$	$n = 300$
1000	0.6867	$\hat{\mu}_{DE}$	0.0267	0.0122	0.0042	0.0012
		$\hat{\mu}_{IE}$	**0.0009**	**0.0007**	**0.0000**	**0.0003**
	0.8713	$\hat{\mu}_{DE}$	0.0064	0.0027	0.0001	0.0009
		$\hat{\mu}_{IE}$	**0.0001**	**0.0001**	**0.0002**	**0.0004**

Table 5.2: Theoretical *ARB* for the difference-cum-exponential estimator ($\hat{\mu}_{DE}$) and for the improved exponential estimator ($\hat{\mu}_{IE}$).

Population			**Theoretical *ARB***			
N	ρ_{XY}	Estimator	$n = 50$	$n = 100$	$n = 200$	$n = 300$
1000	0.6867	$\hat{\mu}_{DE}$	0.0214	0.0103	0.0046	0.0027
		$\hat{\mu}_{IE}$	**0.0013**	**0.0006**	**0.0003**	**0.0002**
	0.8713	$\hat{\mu}_{DE}$	0.0023	0.0011	0.0005	0.0003
		$\hat{\mu}_{IE}$	**0.0006**	**0.0003**	**0.0001**	**0.0001**

As expected, the absolute relative bias generally decreases as the sample size increases, however this effect becomes less pronounced when the correlation between X and Y is higher. Although the proposed estimator is not unbiased, the bias results show a very good performance for this estimator.

Table 5.3 above gives the empirical and theoretical *MSE*'s for the two competing estimators.

The *MSE* values for the proposed estimator are all less than the *MSE* values for the Koyuncu et al. (2013) estimator. The estimators under study get more and more efficient as ρ_{XY} increases. These results were expected from the condition in (5.9).

Table 5.3: Empirical and theoretical *MSE* for the difference-cum-exponential estimator ($\hat{\mu}_{DE}$) and for the improved exponential estimator ($\hat{\mu}_{IE}$).

Population				*MSE* Estimation		
N	ρ_{XY}	n	Estimator	Empirical	Theoretical	*MSE* Condition[1]
1000	0.6867	50	$\hat{\mu}_{DE}$	0.1025	0.1007	0.0253
			$\hat{\mu}_{IE}$	0.0052	0.0050	
		100	$\hat{\mu}_{DE}$	0.0483	0.0484	0.0122
			$\hat{\mu}_{IE}$	0.0024	0.0024	
		200	$\hat{\mu}_{DE}$	0.0217	0.0217	0.0055
			$\hat{\mu}_{IE}$	0.0011	0.0011	
		300	$\hat{\mu}_{DE}$	0.0127	0.0127	0.0032
			$\hat{\mu}_{IE}$	0.0006	0.0006	
	0.8713	50	$\hat{\mu}_{DE}$	0.0285	0.0283	0.0068
			$\hat{\mu}_{IE}$	0.0024	0.0024	
		100	$\hat{\mu}_{DE}$	0.0132	0.0135	0.0032
			$\hat{\mu}_{IE}$	0.0011	0.0011	
		200	$\hat{\mu}_{DE}$	0.0060	0.0060	0.0014
			$\hat{\mu}_{IE}$	0.0005	0.0005	
		300	$\hat{\mu}_{DE}$	0.0035	0.0035	0.0008
			$\hat{\mu}_{IE}$	0.0003	0.0003	

[1] *MSE* comparison based on expression (5.9).

5.6 Numerical Example

In this section, we use real data concerning enterprises for the Monthly Economic Survey (MES) in Portugal. The survey is conducted to provide an accurate picture of business trends of enterprises. It provides short-term indicators on a monthly basis compiled for four sectors: industry, retail trade, construction and service sector. The survey results are broken down by branches according to the NACE[1] Rev. 2 (Eurostat, 2008). In this survey the main questions refer to an assessment of recent trends in production, of the current levels of order books and stocks, as well as expectations about production, selling prices and employment. We consider as population the enterprises collected in the 2010 sample which provided results for the industry sector, taking the monthly salaries as study variable and number of employees as auxiliary variable in each enterprise.

Let Y be the monthly salaries amount in 2010 collected by the MES in that year. This is typically a confidential variable for enterprises, only known from business surveys. The auxiliary variable X is the number of employees available from business data registers. The variables Y and X are strongly correlated so we can take advantage of this correlation by using the estimators under study. The MES provided 26980 monthly salary values in 2010, collected for about 2316 enterprises which answered this survey in that

[1]NACE is derived from the French title "Nomenclature générale des Activités économiques dans les Communautés Européennes" (Statistical classification of economic activities in the European Communities).

same year. We take these 26980 values as our population. For the RRT part, let S be a normal random variable with mean equal to zero and standard deviation equal to 10% of the standard deviation of X. The reported response is given by $Z = Y + S$ (the salary amount plus a random quantity). The summary statistics about the populations are given below.

Population Characteristics:

$N = 26980, \rho_{XY} = 0.8599$
$\mu_X = 113.91, \mu_Y = 167.18$ (in thousands of Euros)
$\sigma_X = 215.8, \sigma_Y = 501.4$ and $\sigma_{XY} = 93043$

We use the following samples sizes in our simulation study: $n = 1000, 2500, 5000$ and 10000.

In Tables 5.4 and 5.5 below we present the empirical and the theoretical *ARB* results, respectively, for the difference-cum-exponential estimator ($\hat{\mu}_{DE}$) and for the proposed estimator ($\hat{\mu}_{IE}$).

Table 5.4: Empirical *ARB* for the difference-cum-exponential estimator ($\hat{\mu}_{DE}$) and for the improved exponential estimator ($\hat{\mu}_{IE}$).

Population			**Empirical *ARB***			
N	ρ_{XY}	Estimator	$n = 1000$	$n = 2500$	$n = 5000$	$n = 10000$
26980	0.8599	$\hat{\mu}_{DE}$	0.0025	0.0022	0.0016	0.0012
		$\hat{\mu}_{IE}$	**0.0003**	**0.0004**	**0.0004**	**0.0003**

Table 5.5: Theoretical *ARB* for the difference-cum-exponential estimator ($\hat{\mu}_{DE}$) and for the improved exponential estimator ($\hat{\mu}_{IE}$).

Population			**Theoretical *ARB***			
N	ρ_{XY}	Estimator	$n = 1000$	$n = 2500$	$n = 5000$	$n = 10000$
26980	0.8599	$\hat{\mu}_{DE}$	0.0008	0.0003	0.0001	0.0001
		$\hat{\mu}_{IE}$	**0.0002**	**0.0001**	**0.0000**	**0.0000**

The ARB results show the good performance for the improved difference-cum-exponential estimator.

The theoretical *MSE* values for both estimators have been obtained using (5.3) and (5.8). These values are given in Table 5.6.

According to the *MSE* results in Table 5.6, the proposed estimator is considerably better than the difference-cum-exponential estimator ($\hat{\mu}_{DE}$). These results are in line with the theoretical findings and the simulation results.

Table 5.6: Empirical and theoretical *MSE* for the difference-cum-exponential estimator ($\hat{\mu}_{DE}$) and for the improved exponential estimator ($\hat{\mu}_{IE}$).

Population				***MSE* Estimation**		
N	ρ_{XY}	n	Estimator	Empirical	Theoretical	*MSE* Condition[2]
26980	0.8599	1000	$\hat{\mu}_{DE}$	62.58	63.34	0.0205
			$\hat{\mu}_{IE}$	6.28	6.3	
		2500	$\hat{\mu}_{DE}$	24.50	23.92	0.0083
			$\hat{\mu}_{IE}$	2.40	2.38	
		5000	$\hat{\mu}_{DE}$	10.88	10.75	0.0041
			$\hat{\mu}_{IE}$	1.09	1.07	
		10000	$\hat{\mu}_{DE}$	4.19	4.15	0.0020
			$\hat{\mu}_{IE}$	0.41	0.41	

[2] *MSE* comparison based on expression (5.9).

5.7 Conclusions

We can conclude from this study that the estimation of the mean of a sensitive variable can be improved by using a correlated non-sensitive auxiliary variable. Our simulation study and the numerical example show that improved difference-cum-exponential estimator can produce further improvement.

In this paper we show that the proposed estimator is more efficient than the difference-cum-exponential estimator recently proposed by Koyuncu et al. (2013), which in turn was better than most of the existing estimators of finite population mean.

References

EUROSTAT. 2008. NACE Rev. 2 - Statistical classification of economic activities in the European Community. *Official Publications of the European Communities*, 112-285 and 306-311.

GUPTA, S., SHABBIR, J., SOUSA, R. & REAL, P.C. 2012. Estimation of the Mean of a Sensitive Variable in the Presence of Auxiliary Information. *Communications in Statistics - Theory and Methods*, 41(13-14), 2394-2404.

KADILAR, C., CANDAN, M. & CINGI, H. 2007. Ratio estimators using robust regression. *Hacettepe Journal of Mathematics and Statistics*, 36(2), 81-188.

KADILAR, C. & CINGI, H. 2004. Ratio estimators in simple random sampling. A*pplied Mathematics and Computation*, 151, 893-902.

KOYUNCU, N., GUPTA, S. & SOUSA, R. 2013. Exponential type estimators of the mean of a sensitive variable in the presence of non-sensitive auxiliary information. *Communications in Statistics - Simulation and Computation*. (*accepted*).

NANGSUE, N. 2009. Adjusted Ratio and Regression Type Estimators for Estimation of Population Mean when some Observations are missing. *World Academy of Science, Engineering and Technology*, 53, 781-784.

SHABBIR, J. & GUPTA, S. 2007. On improvement in variance estimation using auxiliary information. *Communication in Statistics-Theory and Methods*, 36(12), 2177-2185.

SHABBIR, J. & GUPTA, S. 2010. Estimation of the finite population mean in two-phase sampling when auxiliary variables are attribute. *Hacettepe Journal of Mathematics and Statistics*, 39(1), 121-129.

SOUSA, R., SHABBIR, J. REAL, P. C. & GUPTA, S. 2010. Ratio estimation of the mean of a sensitive variable in the presence of auxiliary information. *Journal of Statistical Theory and Practice*, 4(3), 495-507.

WARNER, S. L. 1965. Randomized response: a survey technique for elimination evasive answer bias. *Journal of American Statistical Association*, 60, 63-69.

Appendix D - R Routines

Listing 5.1: R Code for Simulation Study of Proposed Estimator in Chapter 5

```

proj_improved_exp <- function(N,sigma,mu)
{

  #Generation of a bivariate normal population
  data_yx <- mvrnorm(N, mu, sigma)

  #Study variable
  Y <- data_yx[,1]
  #Auxiliary variable, correlated with Y
  X <- data_yx[,2]

  #Coefficient of correlation between Y and X
  Ro_YX <- cor(Y,X)

  #Scrambling variable independent of Y and X, with mean=0
  S <- rnorm(N,mean=0,sd=0.1*sd(X))
  #Scrambled response
  Z <- Y+S

  #Coefficient of correlation between Z and X
  Ro_ZX <- Ro_YX/sqrt(1+(var(S)/var(Y)))

  #population
  univ <- data.frame(cbind(Y=Y,S=S,Z=Z,X=X,NRAND=runif(N)))
  univ <- univ[order(univ$NRAND),]

  #Mean of Y
  mz <- mean(univ$Z)
  mx <- mean(univ$X)
  my <- mean(univ$Y)

  mu11 <- sum((univ$Z-mz)*(univ$X-mx))/(N-1)
  mu12 <- sum((univ$Z-mz)*((univ$X-mx)^2))/(N-1)
  mu02 <- sum((univ$X-mx)^2)/(N-1)
  mu03 <- sum((univ$X-mx)^3)/(N-1)

  beta_zx <- Ro_YX*(sd(univ$Y)/sd(univ$X))

  #Samples dimension
  dim_samp <- c(50,100,200,300)

  #Initialize the variables...

  for (i in 1:length(dim_samp))
  {
```

```
    #sample dimension
    n <- dim_samp[i]
    #sample
    samp <- univ[1:n,]
    #Sampling rate
    f <- n/N

    #Coefficient of variation
    c_x <- sd(univ$X)/mx
    c_y <- sd(univ$Y)/my
    c2_x <- c_x^2
    c2_y <- c_y^2
    c2_z <- c2_y+(var(univ$S)/(my^2))
    c_z <- sqrt(c2_z)

    l <- (1-f)/n

    #Difference-cum-exponential type Estimator
    w1 <- (1-(l*c2_x/8))/(1+l*c2_z*(1-(Ro_ZX^2)))
    w2 <- (my/mx)*(0.5-w1*(1-(Ro_ZX*c_z/c_x)))
    est5 <- (w1*mean(samp$Z)+w2*(mx-mean(samp$X)))
        *exp((mx-mean(samp$X))/(mx+mean(samp$X)))

    #2nd Improved Exponential Estimator
    A <- 1+l*c2_z+l*c2_x-2*l*Ro_ZX*c_z*c_x
    B <- 1+l*c2_x
    C <- 1+(3/8)*l*c2_x-0.5*l*Ro_ZX*c_z*c_x
    D <- 1+(3/8)*l*c2_x
    E <- 1+l*c2_x-l*Ro_ZX*c_z*c_x
    z1 <- (B*C-D*E)/(A*B-(E^2))
    z2 <- my*(A*D-C*E)/(A*B-(E^2))
    est6 <- (z1*mean(samp$Z)+z2)*exp((mx-mean(samp$X))/(mx+mean(samp$X)))

    #Bias of Difference-cum-exponential estimator - 1st degree approximation
    bias5i <- (w1-1)*my+w1*my*l*((3/8)*c2_x-0.5*Ro_ZX*c_z*c_x)
          +w2*mx*l*c2_x
    mse5i <- (my^2)*((1-0.25*l*c2_x)-(((1-(1/8)*l*c2_x)^2)
          /(1+l*c2_z*(1-(Ro_ZX^2)))))

    #Bias of Improved Exponential - 1st degree approximation
    bias6i <- (z1-1)*my+z1*my*((3/8)*l*c2_x
          -0.5*l*Ro_ZX*c_z*c_x)+z2*(1+(3/8)*l*c2_x)
    #Mean Square Error of improved exponential estimator 2
    #1st degree approximation
    mse6i <- (my^2)*(1-((B*(C^2)+A*(D^2)-2*C*D*E)/(A*B-(E^2))))

    #Condition to compare Est6(P) with Est8(DE)
    cond <- (v20*v02*(1-(Ro_ZX^2))*(8*(8*v20+v02*(v11^2)
        -6*(v11^2))-v02*((4-v02)^2)+v20*v02*(32-7*v02)))
        /(64*(v20+v20*v02*(1-(Ro_ZX^2)))*(v02+v20*v02*(1-(Ro_ZX^2))))
```

```

    #Empirical results
    #Simulation of 5000 replicas of estimates
    ...

    #Results
    res <- rbind(res,c(N,n,Ro_YX,Ro_ZX,
                      c_x,c_y,c_z,k1,k2,w1,w2,
                      z1,z2,mx,my,mz,
                      med_est1,med_est2,med_est3,med_est4,
                      med_est5,med_est6,med_est7,
                      bias2i,bias3i,bias4i,bias5i,
                      bias6i,bias6i,
                      emp_mse1,mse1,emp_mse2,mse2i,
                      emp_mse3,mse3i,emp_mse4,mse4i,
                      emp_mse5,mse5i,emp_mse6,mse6i,
                      emp_mse7,mse6i,
                      cond))
  }
  colnames(res) <- c("N","n","RhoXY","RhoZX",
                      "Cx","Cy","Cz","k1","k2","w1","w2",
                      "z1","z2","mX","mY","mZ",
                      "Est1","Est2","Est3","Est4",
                      "Est5","Est6","Est7",
                      "BIAS2I","BIAS3I","BIAS4I","BIAS5I",
                      "BIAS6I","BIAS7I",
                      "EMP_MSE1","MSE1","EMP_MSE2","MSE2I",
                      "EMP_MSE3","MSE3I","EMP_MSE4","MSE4I",
                      "EMP_MSE5","MSE5I","EMP_MSE6","MSE6I",
                      "EMP_MSE7","MSE7I",
                      "COND")
  return(res)
}

#Package for generation
require(MASS)

N <- 1000

#Parameters
sigma1 <- matrix(c(9,1.9,1.9,4),2,2)
sigma2 <- matrix(c(10,3,3,2),2,2)
sigma3 <- matrix(c(6,3,3,2),2,2)

mu <- c(2,2)

res <- NULL
for (i in 1:length(N))
{
  res <- rbind(res,proj_improved_exp(N[i],sigma1,mu))
```

```
  res <- rbind(res,proj_improved_exp(N[i],sigma2,mu))
  res <- rbind(res,proj_improved_exp(N[i],sigma3,mu))
}
write.table(res,"chapter5_ss_results.txt",sep="\t",dec=",",row.names=FALSE)
```

Listing 5.2: R Code for Numerical Example of Proposed Estimator in Chapter 5

```

proj_improved_exp_real <- function(Y,X,N)
{

  #Coefficient of correlation between Y and X
  Ro_YX <- cor(Y,X)

  #Scrambling variable independent of Y and X, with mean=0
  S <- rnorm(N,mean=0,sd=0.1*sd(X))
  #Scrambled response
  Z <- Y+S

  #Coefficient of correlation between Z and X
  Ro_ZX <- Ro_YX/sqrt(1+(var(S)/var(Y)))

  #population
  univ <- data.frame(cbind(Y=Y,S=S,Z=Z,X=X,NRAND=runif(N)))
  univ <- univ[order(univ$NRAND),]

  #Mean of Y
  mz <- mean(univ$Z)
  mx <- mean(univ$X)
  my <- mean(univ$Y)

  mu11 <- sum((univ$Z-mz)*(univ$X-mx))/(N-1)
  mu12 <- sum((univ$Z-mz)*((univ$X-mx)^2))/(N-1)
  mu02 <- sum((univ$X-mx)^2)/(N-1)
  mu03 <- sum((univ$X-mx)^3)/(N-1)

  beta_zx <- Ro_YX*(sd(univ$Y)/sd(univ$X))

  #Samples dimension
  dim_samp <- c(1000,2500,5000,10000)

  #Initialize the variables...

  for (i in 1:length(dim_samp))
  {
    #sample dimension
    n <- dim_samp[i]
    #sample
    samp <- univ[1:n,]
    #Sampling rate
    f <- n/N

    #Coefficient of variation
    c_x <- sd(univ$X)/mx
```

```
    c_y <- sd(univ$Y)/my
    c2_x <- c_x^2
    c2_y <- c_y^2
    c2_z <- c2_y+(var(univ$S)/(my^2))
    c_z <- sqrt(c2_z)

    l <- (1-f)/n

    #Difference-cum-exponential type Estimator
    w1 <- (1-(l*c2_x/8))/(1+l*c2_z*(1-(Ro_ZX^2)))
    w2 <- (my/mx)*(0.5-w1*(1-(Ro_ZX*c_z/c_x)))
    est7 <- (w1*mean(samp$Z)+w2*(mx-mean(samp$X)))
        *exp((mx-mean(samp$X))/(mx+mean(samp$X)))

    #2nd Improved Exponential Estimator
    A <- 1+l*c2_z+l*c2_x-2*l*Ro_ZX*c_z*c_x
    B <- 1+l*c2_x
    C <- 1+(3/8)*l*c2_x-0.5*l*Ro_ZX*c_z*c_x
    D <- 1+(3/8)*l*c2_x
    E <- 1+l*c2_x-l*Ro_ZX*c_z*c_x
    z1 <- (B*C-D*E)/(A*B-(E^2))
    z2 <- my*(A*D-C*E)/(A*B-(E^2))
    est8 <- (z1*mean(samp$Z)+z2)*exp((mx-mean(samp$X))/(mx+mean(samp$X)))

    #Bias of generalized exponential type estimator
    #1st degree approximation
    bias7i <- (w1-1)*my+w1*my*l*((3/8)*c2_x-0.5*Ro_ZX*c_z*c_x)
          +w2*mx*l*c2_x
    mse7i <- (my^2)*((1-0.25*l*c2_x)-(((1-(1/8)*l*c2_x)^2)
          /(1+l*c2_z*(1-(Ro_ZX^2)))))

    #Bias of improved exponential estimator 2 - 1st degree approximation
    bias8i <- (z1-1)*my+z1*my*((3/8)*l*c2_x-0.5*l*Ro_ZX*c_z*c_x)
          +z2*(1+(3/8)*l*c2_x)
    #Mean Square Error of improved exponential estimator 2
    #1st degree approximation
    mse8i <- (my^2)*(1-((B*(C^2)+A*(D^2)-2*C*D*E)/(A*B-(E^2))))

    #Condition to compare Est8(P) with Est8(DE)
    cond <- ((B*(C^2)+A*(D^2)-2*C*D*E)/(A*B-(E^2)))
        -(((1-(1/8)*l*c2_x)^2)/(1+l*c2_z*(1-(Ro_ZX^2))))-0.25*l*c2_x

    #Empirical results
    #Simulation of 5000 replicas of estimates
    ...

    #Results
    res <- rbind(res,c(N,n,Ro_YX,Ro_ZX,
                    c_x,c_y,c_z,k1,k2,w1,w2,
                    z1,z2,mx,my,mz,
```

```
                      med_est7,med_est8,
                      bias7i,bias8i,
                      emp_mse7,mse7i,
                      emp_mse8,mse8i,
                      cond))
  }
  colnames(res) <- c("N","n","RhoXY","RhoZX",
                      "Cx","Cy","Cz","k1","k2","w1","w2",
                      "z1","z2","mX","mY","mZ",
                      "Est7","Est8",
                      "BIAS7I","BIAS8I",
                      "EMP_MSE7","MSE7I",
                      "EMP_MSE8","MSE8I",
                      "COND")
  return(res)
}

#Package for generation
require(MASS)

#Import data
data_yx <- read.table("IVNEI2010.txt",sep="\t",dec=",",header = T)
#Study variable (purchase, millions of euros)
#data_yx <- data_yx[data_yx$MES>=10,]
Y <- data_yx[,5]/1000
#Auxiliary variable, correlated with Y (turnover, millions of euros)
X <- data_yx[,4]

#Data application
N <- dim(data_yx)[1]
res <- proj_improved_exp_real(Y,X,N)

#Export data
write.table(res,"chapter5_ne_results.txt",sep="\t",dec=",",row.names=FALSE)
```

6

Improved Mean Estimation of a Sensitive Variable Using Auxiliary Information in Stratified Sampling

Abstract

Sousa et al. (2010) and Gupta et al. (2012) have recently introduced ratio estimator and regression estimators for the mean of a sensitive variable which perform better than the ordinary mean estimator based on a Randomized Response Technique (RRT). In the present study we extend these estimators to the stratified sampling setting.

The performance of the proposed estimators is compared to the exiting estimators both theoretically and through a simulation study. We also apply the proposed estimators to some real data.

Submitted as: SOUSA, R., GUPTA, S., SHABBIR, J. & REAL, P. C. 2013. Improved Mean Estimation of a Sensitive Variable Using Auxiliary Information in Stratified Sampling. *Journal of Statistics and Management Systems*.

6.1 Introduction

The main goal of this paper is to extend the results of Sousa et al. (2010) and Gupta et al. (2012) to the case of stratified sampling. It is assumed that the study variable is sensitive and the auxiliary variable is non-sensitive.

Many authors have presented ratio and regression estimators when both the study variable Y and the auxiliary variable X are directly observable. These include Kadilar and Cingi (2005), Kadilar et al. (2007), Shabbir and Gupta (2007, 2010) and Nangsue (2009). Gupta and Shabbir (2008) have suggested a general class of ratio estimators when the population parameters of the auxiliary variable are known. These estimators have also been extended by Kadilar and Cingi (2003) to stratified random sampling. In an attempt to improve the estimators, Kadilar and Cingi (2005), Shabbir and Gupta (2005, 2006) and Singh and Vishwakarma (2008) have suggested new ratio estimators in stratified random sampling. Koyuncu and Kadilar (2008, 2009) have proposed a family of combined-type estimators in stratified random sampling based on the family of estimators proposed by Khoshnevisan et al. (2007). Recently Koyuncu and Kadilar (2010) have suggested a family of estimators in stratified random sampling following Kadilar and Cingi (2003).

Some studies on estimation of the mean have been submitted with different sampling schemes, such as Sahoo et al. (2009) and Singh and Kumar (2011) in a two-stage sampling scheme and recently by Singh and Solanki (2012) in a systematic sampling design.

This paper suggests a combined ratio estimator and a combined regression estimator of population mean of a sensitive variable using non-sensitive auxiliary information, using Randomized Response Technique (RRT) methodology (Gupta et al., 2002 and 2010; Warner, 1965) in stratified sampling. The Bias and the Mean Square Error (*MSE*) of the suggested estimators are derived. Both theoretical and empirical findings support the reliability of the present study.

6.2 Terminology

We denote the finite population by $U = \{U_1, U_2, ..., U_N\}$. Consider a stratified random sample s (Cochran, 1977), selected from U with sampling rate $f = \frac{n}{N}$. The study population is divided into L strata with strata sizes N_h, such that $\sum_{h=1}^{L} N_h = N(h = 1, ..., L)$.

Let Y be the sensitive study variable which cannot be observed directly. Let X be a non-sensitive auxiliary variable which is strongly correlated with Y. Let S be a scrambling random variable independent of Y and X. The particular values of S are unknown to the interviewer but its distribution is known. The respondent is asked to report an additively scrambled response for Y given by $Z = Y + S$ and also asked to provide a true response for X.

Consider a stratified random sample of size n be drawn from U such that the sample size in the h^{th} stratum is n_h and $\sum_{h=1}^{L} n_h = n$. Let y_{hi} and x_{hi} respectively be the values of the study variable Y and the auxiliary variable X in the h^{th} stratum, with $i = 1, ..., n_h$.

Let $\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h$, $\bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h$ and $\bar{z}_{st} = \sum_{h=1}^{L} W_h \bar{z}_h$ be the stratified sample means, where $\bar{y}_h = \dfrac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$, $\bar{x}_h = \dfrac{1}{n_h} \sum_{i=1}^{n_h} x_{hi}$ and $\bar{z}_h = \dfrac{1}{n_h} \sum_{i=1}^{n_h} z_{hi}$ are the stratum sample means corresponding to population stratum means $\bar{Y}_h = E(Y_h)$, $\bar{X}_h = E(X_h)$ and $\bar{Z}_h = E(Z_h)$ and $W_h = \dfrac{N_h}{N}$ are the known stratum weights.

To estimate $\bar{Y} = \sum_{i=1}^{N_h} W_h \bar{Y}_h$ we assume that $\bar{X} = \sum_{i=1}^{N_h} W_h \bar{X}_h$ is known. Let $\bar{Z} = \sum_{i=1}^{N_h} W_h \bar{Z}_h$ be the population mean for the scrambled variable Z.

To discuss the properties of different estimators, we define the following error terms. Let $e_{0st} = \dfrac{\bar{z}_{st} - \bar{Z}}{\bar{Z}}$ and $e_{1st} = \dfrac{\bar{x}_{st} - \bar{X}}{\bar{X}}$, $e_{2st} = \dfrac{s^2_{xst} - S^2_{xst}}{S^2_{xst}}$ and $e_{3st} = \dfrac{s^2_{zxst} - S^2_{zxst}}{S^2_{zxst}}$ such that $E(e_{ist}) = 0,\ i = 0, 1, 2, 3$.

6.3 Estimators Review

Below we list some existing mean estimators for simple random sampling.

(i) Ordinary sample mean:

$$\hat{\mu}_y = \bar{z}. \tag{6.1}$$

$$MSE(\hat{\mu}_y) = \frac{1-f}{n}\left(S_y^2 + S_s^2\right), \tag{6.2}$$

where $S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$ and $S_s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(s_i - \bar{S})^2$.

(ii) Sousa et al. (2010) ratio estimator:

$$\hat{\mu}_R = \bar{z}\frac{\bar{X}}{\bar{x}}. \tag{6.3}$$

The *Bias* and *MSE* of $\hat{\mu}_R$ to first degree of approximation are given by

$$Bias(\hat{\mu}_R) \cong \frac{1-f}{n}\bar{Y}\left(C_x^2 - \rho_{zx}C_zC_x\right) \tag{6.4}$$

and

$$MSE(\hat{\mu}_R) \cong \frac{1-f}{n}\bar{Y}^2\left(C_z^2 + C_x^2 - 2\rho_{zx}C_zC_x\right), \tag{6.5}$$

where $C_z^2 = C_y^2 + \dfrac{S_s^2}{\bar{Y}^2}$, $\rho_{zx} = \dfrac{\rho_{yx}}{\sqrt{1 + \dfrac{S_s^2}{S_y^2}}}$ and C_z, C_y and C_x are the coefficients of variation of Z, Y and X, respectively.

(iii) Gupta et al. (2012) regression estimator:

$$\hat{\mu}_{Reg} = \bar{z} + \hat{\beta}_{zx}\left(\bar{X} - \bar{x}\right), \tag{6.6}$$

where $\hat{\beta}_{zx} = \dfrac{S_{zx}}{S_x^2} = \dfrac{S_{yx}}{S_x^2}$ is the sample regression coefficient between Z and X, $S_{yx} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})$ and $S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

The *Bias* and *MSE* of $\hat{\mu}_{Reg}$ to first degree of approximation, are given by

$$Bias(\hat{\mu}_{Reg}) \cong -\beta_{zx}\left(\frac{1-f}{n}\right)\left\{\frac{\mu_{12}}{\mu_{11}} - \frac{\mu_{03}}{\mu_{02}}\right\} \tag{6.7}$$

and

$$MSE(\hat{\mu}_{Reg}) \cong \left(\frac{1-f}{n}\right)\bar{Y}^2 C_z^2\left(1 - \rho_{zx}^2\right), \tag{6.8}$$

where $\mu_{rs} = \frac{1}{N-1}\sum_{i=1}^{N}(z_i - \bar{Z})^r(x_i - \bar{X})^s$.

For a stratified random sample the usual combined sample mean, ignoring the auxiliary information, is given by

$$\hat{\mu}_{Yst} = \bar{z}_{st}, \tag{6.9}$$

which is an unbiased estimator of population mean $\bar{Y}$.

The *MSE* of $\hat{\mu}_{Yst}$ is given by

$$MSE(\hat{\mu}_{Yst}) = \sum_{h=1}^{L} W_h^2 \gamma_h \left\{S_{yh}^2 + S_{sh}^2\right\}, \tag{6.10}$$

where $\gamma_h = \left(\dfrac{1}{n_h} - \dfrac{1}{N_h}\right)$, $S_{yh}^2 = \frac{1}{N_h-1}\sum_{i=1}^{N_h}(y_{hi} - \bar{Y}_h)^2$ and $S_{sh}^2 = \frac{1}{N_h-1}\sum_{i=1}^{N_h}(s_{hi} - \bar{S}_h)^2$.

The remainder of the paper is as follows. In Section 6.4, we introduce a combined ratio estimator and compare it to the ordinary mean estimator and to the ratio estimator (Sousa et al., 2010), considering the *MSE* as an accuracy indicator. In Section 6.5, we propose a combined regression estimator and compare it with other estimators as well as with the regression estimator proposed by Gupta et al. (2012). We present an empirical study in Section 6.6 and a numerical example in Section 6.7 to support the proposed methodology. Section 6.8 provides some concluding remarks.

6.4 Proposed combined ratio estimator

We propose the following combined ratio estimator

$$\hat{\mu}_{Rst} = \bar{z}_{st}\left(\frac{\bar{X}}{\bar{x}_{st}}\right). \tag{6.11}$$

Using Taylor's approximation and retaining terms of order up to 2, (6.11) can be rewritten as

$$\hat{\mu}_{Rst} - \bar{Z} \cong \bar{Z}\left\{e_{0st} - e_{1st} + e_{1st}^2 - e_{0st}e_{1st}\right\}. \tag{6.12}$$

Under the assumption of bivariate normality (see Sukhatme and Sukhatme, 1984), we have:

$E(e_{0st}^2)=\sum_{h=1}^{L} W_h^2\gamma_h C_{zh}^2$, $E(e_{1st}^2)=\sum_{h=1}^{L} W_h^2\gamma_h C_{xh}^2$, $E(e_{0st}e_{1st})=\sum_{h=1}^{L} W_h^2\gamma_h C_{zxh}$, where $C_{zxh} = \rho_{zxh}C_{zh}C_{xh}$, $C_{zh}^2 = C_{yh}^2 + \left(\frac{S_{sh}^2}{\bar{Y}^2}\right)$ and $\rho_{zxh} = \frac{\rho_{yxh}}{\sqrt{1+\left(\frac{S_{sh}^2}{S_{yh}^2}\right)}}$.

Using $\bar{Z} = \bar{Y}$ in (6.12), the *Bias* of $\hat{\mu}_{Rst}$ to first degree of approximation is given by

$$Bias(\hat{\mu}_{Rst}) \cong \bar{Y}\sum_{h=1}^{L} W_h^2\gamma_h\left(C_{xh}^2 - C_{zxh}\right). \tag{6.13}$$

Using (6.12), the *MSE* of $\hat{\mu}_{Rst}$, correct up to first order of approximation, is given by

$$MSE(\hat{\mu}_{Rst}) = E\left\{\hat{\mu}_{Rst} - \bar{Y}\right\}^2 \cong \bar{Y}^2E\left\{e_{0st} - e_{1st} + e_{1st}^2 - e_{0st}e_{1st}\right\}^2.$$

So, if an independent simple random sample is drawn in each stratum, we have

$$MSE(\hat{\mu}_{Rst}) \cong \bar{Y}^2\sum_{h=1}^{L} W_h^2\gamma_h\left\{C_{zh}^2 + C_{xh}^2 - 2C_{zxh}\right\}. \tag{6.14}$$

It can be observed that $MSE\left(\hat{\mu}_{Rst}\right) < MSE\left(\hat{\mu}_{Yst}\right)$ if

$$\sum_{h=1}^{L} W_h^2\gamma_h C_{zxh} - \frac{1}{2}\sum_{h=1}^{L}\gamma_h C_{xh}^2 > 0. \tag{6.15}$$

On the other hand, comparing this estimator with different sampling methods $MSE\left(\hat{\mu}_{Rst}\right) < MSE\left(\hat{\mu}_{R}\right)$ if

$$\sum_{h=1}^{L} W_h^2\gamma_h\left\{C_{zh}^2 + C_{xh}^2 - 2\rho_{zxh}C_{zh}C_{xh}\right\} < \frac{1-f}{n}\left\{C_z^2 + C_x^2 - 2\rho_{zx}C_zC_x\right\}, \tag{6.16}$$

a condition that can be ensured by a suitable stratification.

6.5 Proposed combined regression estimator

Assuming linear relationship between Y and X, we propose the following combined regression estimator for the population mean of Y

$$\hat{\mu}_{Regst} = \bar{z}_{st} + \hat{\beta}_c \left(\bar{X} - \bar{x}_{st}\right), \tag{6.17}$$

where $\hat{\beta}_c = \dfrac{\sum_{h=1}^{L} W_h^2 \gamma_h s_{zxh}}{\sum_{h=1}^{L} W_h^2 \gamma_h s_{xh}^2}$ is the sample regression coefficient between Z and X and $Z = Y + S$ is the scrambled response on Y.

Using Taylor's approximation and retaining terms of order up to 2, (6.17) can be rewritten as

$$\hat{\mu}_{Regst} - \bar{Z} \cong \bar{Z} e_{ost} - \beta_c \bar{X}\left[e_{1st} + e_{1st}e_{3st} - e_{1st}e_{2st}\right], \tag{6.18}$$

where $\beta_c = \dfrac{\sum_{h=1}^{L} W_h^2 \gamma_h s_{zxh}}{\sum_{h=1}^{L} W_h^2 \gamma_h s_{xh}^2}$ is the population regression coefficient between Z on X.

From Mukhopadhyay (1998, p. 123) and considering a random sample selected from each population stratum we can deduce:

$E(e_{1st}e_{2st}) = \dfrac{1}{\bar{X}} \sum_{h=1}^{L} W_h^2 \gamma_h \dfrac{\mu_{03h}}{\mu_{02h}}$ and $E(e_{1st}e_{3st}) = \dfrac{1}{\bar{X}} \sum_{h=1}^{L} W_h^2 \gamma_h \dfrac{\mu_{12h}}{\mu_{11h}}$,
where $\mu_{rsh} = \dfrac{1}{N_h - 1} \sum_{i=1}^{N_h} (z_{hi} - \bar{Z}_h)^r (x_{hi} - \bar{X}_h)^s$.

Recognizing that $\bar{Z} = \bar{Y}$ in Equation (6.18), the *Bias* and *MSE* of $\hat{\mu}_{Regst}$, are given by

$$Bias(\hat{\mu}_{Regst}) \cong -\sum_{h=1}^{L} W_h^2 \gamma_h \beta_c \left\{\frac{\mu_{12h}}{\mu_{11h}} - \frac{\mu_{03h}}{\mu_{02h}}\right\} \tag{6.19}$$

and

$$MSE(\hat{\mu}_{Regst}) \cong \bar{Y}^2 \sum_{h=1}^{L} W_h^2 \gamma_h C_{zh}^2 \left(1 - \rho_c^2\right), \tag{6.20}$$

where $\rho_c = \dfrac{\sum_{h=1}^{L} W_h^2 \gamma_h S_{zxh}}{\sqrt{\sum_{h=1}^{L} W_h^2 \gamma_h C_{zh}^2}\sqrt{\sum_{h=1}^{L} W_h^2 \gamma_h C_{xh}^2}}$.

It can be verified easily that

(i) $MSE(\hat{\mu}_{Regst}) < MSE(\hat{\mu}_{Yst})$ if

$$\sum_{h=1}^{L} W_h^2 \gamma_h C_{zh}^2 \rho_c^2 > 0. \tag{6.21}$$

(ii) $MSE(\hat{\mu}_{Regst}) < MSE(\hat{\mu}_{Rst})$ if

$$\left(\sqrt{\sum_{h=1}^{L} W_h^2 \gamma_h C_{xh}^2} - \frac{\sum_{h=1}^{L} W_h^2 \gamma_h C_{zxh}}{\sqrt{\sum_{h=1}^{L} W_h^2 \gamma_h C_{xh}^2}}\right)^2 > 0. \tag{6.22}$$

These two conditions will always hold true indicating that, up to first order of approximation, the regression estimator performs better than ordinary mean and ratio estimators in stratified random sampling also, as they did in the case of simple random sampling.

On the other hand, we can say that

(iii) $MSE(\hat{\mu}_{Regst}) < MSE(\hat{\mu}_{Reg})$ if

$$\sum_{h=1}^{L} W_h^2 \gamma_h C_{zh}^2 \left(1 - \rho_c^2\right) < \frac{1-f}{n} C_z^2 \left(1 - \rho_{zx}^2\right), \tag{6.23}$$

a condition that can be ensured by a suitable stratification.

6.6 A Simulation Study

In this section, we present a simulation study with particular focus on comparing the performance of the proposed combined estimators $\hat{\mu}_{Rst}$ and $\hat{\mu}_{Regst}$ to the RRT mean estimator $\hat{\mu}_{Yst}$ and to the corresponding estimators in simple random sampling (Sousa et al., 2010; Gupta et al., 2012). For this purpose we rely on *Bias* and *MSE*, correct up to first order of approximation.

We considered three bivariate normal populations with different covariance matrices to represent the distribution of (Y, X). The scrambling variable S is taken to be a normal distribution with mean equal to zero and standard deviation equal to 10% of the standard deviation of X. The reported scrambled response on Y is given by $Z = Y + S$.

All of the simulated populations have theoretical mean of $[Y, X]$ as $\mu = [5, 5]$ and covariance matrices as given below.

Population 1

$N = 1000$

$\Sigma = \begin{bmatrix} 9 & 3.2 \\ 3.2 & 4 \end{bmatrix}, \rho_{XY} = 0.5333.$

Population 2

$N = 1000$

$\Sigma = \begin{bmatrix} 6 & 3.3 \\ 3.3 & 3 \end{bmatrix}, \rho_{XY} = 0.7778.$

Population 3

$N = 1000$

$\Sigma = \begin{bmatrix} 5 & 3 \\ 3 & 2 \end{bmatrix}, \rho_{XY} = 0.9487.$

For each population we considered five sample sizes: $n = 30, 60, 150$ and 300.

The population is divided in two strata according to a certain criteria set for the auxiliary variable. The sample size from each stratum is based on the *Neyman* allocation. We compare the results of stratified random sampling with the corresponding results of simple random sampling.

Table 6.1 below gives the empirical and theoretical *MSE*'s for the proposed combined estimators based on first order approximation. We use the following expression to find the Percent Relative Efficiency (*PRE*) of study estimators as compared to the ordinary sample mean:

$$PRE = \frac{MSE(\hat{\mu}_{Yst})}{MSE(\hat{\mu}_{\alpha})} \times 100,$$

where $\alpha = R_{st}, Reg_{st}$. This measure is calculated using first degree of approximation for *MSE* per unit estimator. We estimate the empirical *MSE* using 5000 samples of size n and considering the average of all the observed values.

Table 6.1: Empirical and Theoretical *MSE*, for the RRT mean estimator, ratio estimator (underlined) and regression estimator (bold); and corresponding *PRE* relative to the RRT mean estimator.

Population					***MSE* Estimation**		
N	N_h	ρ_{XY}	ρ_{XYh}	n	Empirical *MSE*	Theoretical *MSE*	*PRE*
1000	$N_1 = 550$, $N_2 = 450$	0.5395	$\rho_{XY1} = 0.5397$, $\rho_{XY2} = 0.5410$	30	0.3039	0.2897	100.00
					0.2227	0.2104	137.70
					0.2395	**0.2077**	**139.51**
				60	0.1403	0.1404	100.00
					0.1024	0.1019	137.77
					0.1339	**0.0993**	**141.34**
				150	0.0533	0.0508	100.00
					0.0380	0.0369	137.72
					0.0488	**0.0359**	**141.31**
				300	0.0208	0.0209	100.00
					0.0156	0.0152	137.73
					0.0187	**0.0148**	**141.00**
		0.7827	$\rho_{XY1} = 0.7888$, $\rho_{XY2} = 0.7868$	30	0.2028	0.1932	100.00
					0.0803	0.0763	253.17
					0.0791	**0.0792**	**244.04**
				60	0.0937	0.0936	100.00
					0.0371	0.0370	253.29
					0.0430	**0.0367**	**255.33**
				150	0.0355	0.0339	100.00
					0.0137	0.0134	253.21
					0.0150	**0.0131**	**258.38**
				300	0.0139	0.0139	100.00
					0.0057	0.0055	253.21
					0.0061	**0.0054**	**256.64**
		0.9501	$\rho_{XY1} = 0.9522$, $\rho_{XY2} = 0.9478$	30	0.1688	0.1608	100.00
					0.0346	0.0323	497.12
					0.0168	**0.0179**	**899.95**
				60	0.0783	0.0779	100.00
					0.0157	0.0157	496.80
					0.0084	**0.0079**	**979.71**
				150	0.0296	0.0282	100.00
					0.0059	0.0057	496.88
					0.0030	**0.0028**	**1002.03**
				300	0.0115	0.0116	100.00
					0.0024	0.0023	496.98
					0.0012	**0.0012**	**993.07**

According to the results in Table 6.1, all the percent relative efficiencies are greater than 100, indicating that the proposed combined estimators perform better than the ordinary mean estimator. The use of auxiliary information provides a gain for a stratified random sample. Therefore, the proposed estimators increases the accuracy since there is a significant correlation between X and Y.

These results for the stratified estimators agree with the Sousa et al. (2010) and Gupta et al. (2012) findings for a simple random sampling. Clearly the gain with regression estimator is substantial when correlation between the primary and auxiliary variables is high.

6.7 Numerical Example

We now compare the performances of the proposed combined estimators using a real data set. The data come from a sample from the survey on Information and Communication Technologies (ICT) usage in enterprises in 2010 with seat in Portugal (Smilhily and Storm, 2010). This survey intends to promote the development of the national statistical system in the information society and to contribute to a deeper knowledge about the usage of ICT by enterprises. The target population covers all industries with one and more persons employed in the sections of economic activity C (Manufacturing) to N (Administrative and support service activities) and S (Other service activities), from NACE[1] Rev. 2 (Eurostat, 2008).

The ICT survey has an extensive plan of indicators, so the use of auxiliary information on the sampling stage is essential to get a stratified random sample as a proper representation for the population. The enterprises commercialization is directly related to their turnover, so this auxiliary variable is usually used for stratification. In our example we consider three strata: the first one is enterprises with less than 10 million (in euros) of turnover, the second between 10 and less than 30 million of turnover, and the third with 30 million or more of turnover.

In our application the variable of interest Y is the purchase orders in 2010, collected by the ICT survey in that year. This is typically a confidential variable for enterprises, only known from business surveys. On the other hand, the auxiliary variable X is the turnover of each enterprise which is known for all the population and annually available with the statistical institutes as administrative information.

The purchase orders information was collected in the ICT survey and is known for a sample of 1698 small and medium enterprises (at least 10 and no more than 100 employees) in 2010. For this study, these 1698 enterprises are considered as our population. The scrambling variable S is taken to be a normal random variable with mean equal to zero. Given the high magnitude of the auxiliary variable X, we consider a standard deviation of S equal to 1% of the standard deviation of X, that is $\sigma_S = 0.01\sigma_X$. The reported scrambled response is given by $Z = Y + S$ (the purchase order value plus a random quantity). The variables Y and X are strongly correlated so we can take advantage of this correlation by using the combined ratio and regression estimators.

The variables X and Y are expressed in millions of Euros. We test our stratified sample estimators with random sample of sizes $n = 100, 250$ and 500. The sample size of each stratum is allocated proportionally to the dimension of strata population.

[1]NACE is derived from the French title "Nomenclature générale des Activités économiques dans les Communautés Européennes" (Statistical classification of economic activities in the European Communities).

Population Characteristics:

Stratum	N_h	ρ_{XYh}	μ_{Yh}	σ_{Yh}	μ_{Xh}	σ_{Xh}	Population
1	979	0.7802	2.15	2.46	3.12	2.68	$N = 1698, \rho_{XY} = 0.9368,$
2	362	0.7952	16.67	6.86	20.31	6.02	$\beta_{YX} = 0.8284, \mu_Y = 14.44,$
3	357	0.8408	45.88	30.21	56.33	30.18	$\sigma_Y = 22.39, \mu_X = 17.97, \sigma_X = 25.31$

Table 6.2 below presents the results for the empirical *MSE* estimates, the theoretical estimates, correct up to first degree of approximation, and the *PRE* of combined ratio and regression estimators relative to the ordinary sample mean in the stratified sample. For both sampling designs, we estimate the empirical *MSE* using 5000 samples of size n selected from the population.

We also show the Design Effect (*Deff*) comparing the efficiency of study estimators in stratified sample (*Str*) relative to the ordinary sample mean in simple random sample (SRS):

$$Deff = \frac{MSE\,(\hat{\mu}_Y)}{MSE\,(\hat{\mu}_\alpha)} \times 100,$$

where $\alpha = Y_{st}, R_{st}, Reg_{st}$.

Table 6.2: Empirical, theoretical *MSE*, *PRE* for the ratio estimator (underlined) and for the regression estimator (bold) relative to the RRT mean estimator and *PRE* for the simple random sample (*SRS*) relative to the stratified sample (*Str*).

Population				***SRS***		***Str***			
N	N_h	ρ_{XY}	n	Empirical *MSE*	Theoretical *MSE*	Empirical *MSE*	Theoretical *MSE*	*PRE*[1]	*Deff*[2]
1698	$N_1 = 979$, $N_2 = 362$, $N_3 = 357$	0.9368	100	5.6291	5.6291	1.9699	1.9373	100.00	290.56
				0.8027	0.8027	0.5535	0.6577	294.56	855.88
				0.6894	**0.6894**	**0.6290**	**0.5303**	**365.35**	**1061.57**
			250	2.0110	2.0403	0.6994	0.6948	100.00	293.66
				0.2882	0.2909	0.2110	0.2397	289.91	851.35
				0.2493	**0.2499**	**0.2265**	**0.1863**	**373.00**	**1095.35**
			500	0.8502	0.8440	0.2917	0.2903	100.00	290.71
				0.1186	0.1204	0.0855	0.0992	292.66	850.81
				0.1022	**0.1034**	**0.0882**	**0.0785**	**369.91**	**1075.37**

[1] *MSE* comparison of study estimators relative to the ordinary sample mean in *Str*.
[2] *MSE* comparison of study estimators in *Str* relative to the ordinary sample mean in *SRS*.

According to the results in Table 6.2, all of the percent relative efficiencies are greater than 100, so the proposed estimators perform better than the ordinary RRT mean estimator which does not use auxiliary information. Moreover, there is clear reduction of *MSE* if we compare the results based on stratification to those based on simple random sampling. The *Deff* shows an increase in efficiency by using the stratified sample.

Taking into account the large correlation between the variable of interest Y and the auxiliary variable X, the proposed estimators have similar gain regardless of the sample size. However, the gain is more evident in a simple random sample because the stratification already significantly reduces the *MSE* values.

6.8 Conclusions

In the survey research context, using auxiliary information can be essential to improve the accuracy of estimates, mainly when we have to deal with sensitive variables. We can observe from this study that the estimation of the mean of a sensitive variable can be improved by using a non-sensitive auxiliary variable.

The ratio and the regression estimators perform better than the RRT mean estimator in both simple random sampling and stratified sampling also. Although both the ratio and regression estimators perform better than the ordinary RRT mean estimator, the improvement is much larger with the regression estimator.

Regarding the efficiency, the results indicate that the proposed estimators become more and more efficient as the coefficient of correlation increases. When the study and the auxiliary variables are strongly correlated the proposed estimators, particularly the combined regression estimator performs much better, regardless the sample size.
These results agree with the findings of Sousa et al. (2010) and Gupta et al. (2012) in simple random sampling.

All of the study estimators show better performance than the ordinary RRT sample mean. Nevertheless, the gain in accuracy is stronger in the simple random sampling because the stratification already reduces the *MSE* value for the *RRT* mean estimator.

The main conclusion of this study is that the advantage of using the RRT in the presence of auxiliary information still holds in the context of stratified sampling.

References

COCHRAN, W.G. 1997. *Sampling Techniques*, 3^{rd} Ed., New York, Wiley Eastern Ltd.

EUROSTAT. 2008. NACE Rev. 2 - Statistical classification of economic activities in the European Community. *Official Publications of the European Communities*, 112-285 and 306-311.

GUPTA, S. N., GUPTA, B. C. & SINGH, S. 2002. Estimation of sensitivity level of personal interview survey questions. *Journal of Statistical Planning and Inference*, 100, 239-247.

GUPTA, S. & SHABBIR, J. 2008. On improvement in estimating the population mean in simple random sampling. *Journal of Applied Statistics*, 35(5), 559-566.

GUPTA, S., SHABBIR, J. & SEHRA, S. 2010. Mean and sensitivity estimation in optional randomized response models. *Journal of Statistical Planning and Inference*, 140(10), 2870-2874.

GUPTA, S., SHABBIR, J., SOUSA, R. & REAL, P. C. 2012. Estimation of the Mean of a Sensitive Variable in the Presence of Auxiliary Information. *Communications in Statistics - Theory and Methods*, 41(13-14), 2394-2404.

KADILAR, C., CANDAN, M. & CINGI, H. 2007. Ratio estimators using robust regression. *Hacettepe Journal of Mathematics and Statistics*, 36(2), 81-188.

KADILAR, C. & CINGI, H. 2003. Ratio estimator in stratified sampling. *Biometrical Journal*, 45, 218-225.

KADILAR, C. & CINGI, H. 2005. A new estimator in stratified random sampling. *Communication in Statistics-Theory and Methods*, 34, 597-602.

KHOSHNEVISAN, M., SINGH, R., CHAUHAN, P., SAWAN, N. & SMARANDACHE, F. 2007. A general family of estimators for estimating population mean using known value of some population parameter(s). *Far East Journal of Theoretical Statistics*, 22, 181-191.

KOYUNCU, N. & KADILAR, C. 2008. Ratio and product estimators in stratified random sampling. *Journal of Statistical Planning and Inference*, 139(8), 2552-2558.

KOYUNCU, N. & KADILAR, C. 2009. Family of estimators of population mean using two auxiliary variables in stratified random sampling. *Communications in Statistics: Theory and Methods*, 38(14), 2398-2417.

KOYUNCU, N. & KADILAR, C. 2010. On the family of estimators of population mean in stratified random sampling. *Pakistan Journal of Statistics*, 26(2), 427-443.

MUKHOPADHYAY, P. 1998. *Theory and Methods of Survey Sampling*, New Delhi, Prentice-Hall of India.

NANGSUE, N. 2009. Adjusted Ratio and Regression Type Estimators for Estimation of Population Mean when some Observations are missing. *World Academy of Science, Engineering and Technology*, 53, 781-784.

SAHOO, L. N., MAHAPATRA, N. & SENAPATI, S. C. 2009. A Generalized Method of Estimation for Two-stage Sampling Using Two Auxiliary Variables. *Journal of Statistical Theory and Practice*, 3(4), 831-839.

SHABBIR, J. & GUPTA, S. 2005. Improved ratio estimators in stratified sampling. *American Journal of Mathematical and Management Sciences*, 25, 293-311.

SHABBIR, J. & GUPTA, S. 2006. A new estimator of population mean in stratified sampling. *Communication in Statistics Theory and Methods*, 35, 1201-1209.

SHABBIR, J. & GUPTA, S. 2007. On improvement in variance estimation using auxiliary information. *Communication in Statistics-Theory and Methods*, 36(12), 2177-2185.

SHABBIR, J. & GUPTA, S. 2010. Estimation of the finite population mean in two-phase sampling when auxiliary variables are attribute. *Hacettepe Journal of Mathematics and Statistics*, 39(1), 121-129.

SINGH, H. P. & KUMAR, S. 2011. A General Family of Estimators of Finite Population Ratio, Product and Mean Using Two Phase Sampling Scheme in the Presence of Non-Response. *Journal of Statistical Theory and Practice*, 2(4), 677-692.

SINGH, H. P. & SOLANKI, R. S. 2012. An Efficient Class of Estimators for the Population Mean Using Auxiliary Information in Systematic Sampling. *Journal of Statistical Theory and Practice*, 6(2), 274-285.

SINGH, H.P. & VISHWAKARMA, G. K. 2008. A family of estimators of population mean using auxiliary information in stratified sampling. *Communication in Statistics-Theory and Methods*, 37(7), 1038-1050.

SMILHILY, M. & STORM, H. 2010. ICT usage in enterprises - 2009. *Eurostat Publications*, Issue 1.

SOUSA, R., SHABBIR, J., REAL, P. C. & GUPTA, S. 2010. Ratio estimation of the mean of a sensitive variable in the presence of auxiliary information. *Journal of Statistical Theory and Practice*, 4(3), 495-507.

SUKHATME, P.V. & SUKHATME, B.V. 1984. *Sampling theory of surveys with applications*, 3^{rd} Ed., Ames, Iowa, Iowa State University Press.

WARNER, S. L. 1965. Randomized response: a survey technique for elimination evasive answer bias. *Journal of American Statistical Association*, 60, 63-69.

Appendix E - R Routines

Listing 6.1: R Code for Simulation Study of Proposed Estimator in Chapter 6

```
proj3_ratio_st_NeymanAlloc <- function(N,sigma,mu,L)
{

  #Generation of a bivariate normal population
  data_yx <- mvrnorm(10000, mu, sigma)
  data_yx <- data.frame(data_yx)
  colnames(data_yx) <- c("Y","X")

  indices1 <- round(runif(550,0,10000))
  data_yx1 <- data_yx[indices1,]
  data_yx1$ST <- 1
  indices2 <- round(runif(450,0,10000))
  data_yx2 <- data_yx[indices2,]
  data_yx2$ST <- 2

  data_yx <- rbind(data_yx1,data_yx2)

  #Study variable
  Y <- data_yx[,1]
  #Auxiliary variable, correlated with Y
  X <- data_yx[,2]
  #Stratum
  ST <- data_yx[,3]

  #Scrambling variable independent of Y and X, with mean=0
  S <- rnorm(N,mean=0,sd=0.1*sd(X))
  #Scrambled response
  Z <- Y+S

  #Population
  univ <- data.frame(cbind(Y=Y,S=S,Z=Z,X=X,
        ST=ST,NRAND=runif(N)))
  univ <- univ[order(univ$ST,univ$NRAND),]

  #Coefficients of correlation
  Ro_YX <- cor(Y,X)
  Ro_ZX <- Ro_YX/sqrt(1+(var(S)/var(Y)))
  Ro_YXh <- by(cbind(univ$Y,univ$X),
        univ$ST,function(x) {cor(x[,1],x[,2])})
  Ro_ZXh <- Ro_YXh/sqrt(1+(by(univ$S,
        univ$ST,var)/by(univ$Y,univ$ST,var)))

  #Population means
  Mz <- mean(univ$Z)
  Mx <- mean(univ$X)
```

```
My <- mean(univ$Y)

#Information
SY <- sd(univ$Y)
SYh <- by(univ$Y,univ$ST,sd)
SZ <- sd(univ$Z)
SZh <- by(univ$Z,univ$ST,sd)
SX <- sd(univ$X)
SXh <- by(univ$X,univ$ST,sd)

#Samples dimension
dim_samp <- c(30,60,150,300)

#Information for the population
Nh <- by(univ$Z,univ$ST,length)
wh <- Nh/N

res <- NULL
for (i in 1:length(dim_samp))
{

  #sample dimension
  n <- dim_samp[i]

  #sample with Neyman Allocation
  n_total <- 0
  samp <- NULL
  for (l in 1:L)
  {
    n_aux <- round(n*((length(univ$Z[univ$ST==l])
           *sd(univ$Y[univ$ST==l]))/sum(Nh*by(univ$Y,univ$ST,sd))))
    if (l==L) {n_aux<-n-n_total}
    n_total <- n_total+n_aux
    samp <- rbind(samp,univ[univ$ST==l,][1:n_aux,])
  }

  #Sampling rate for each stratum
  nh <- by(samp$Z,samp$ST,length)
  fh <- nh/Nh
  gh <- (1-fh)/nh
  f <- n/N

  #Sampling mean for each stratum
  mzh <- by(samp$Z,samp$ST,mean)
  myh <- by(samp$Y,samp$ST,mean)
  mxh <- by(samp$X,samp$ST,mean)
  msh <- by(samp$S,samp$ST,mean)

  #Population mean for each stratum
  Mzh <- by(univ$Z,univ$ST,mean)
```

```
Myh <- by(univ$Y,univ$ST,mean)
Mxh <- by(univ$X,univ$ST,mean)
Msh <- by(univ$S,univ$ST,mean)

#Sampling mean for each stratum
szh <- by(samp$Z,samp$ST,sd)
sxh <- by(samp$X,samp$ST,sd)

mu11 <- cbind(sum((univ$Z[univ$ST==1]-Mzh[1])
    *(univ$X[univ$ST==1]-Mxh[1]))/(Nh[1]-1),
        sum((univ$Z[univ$ST==2]-Mzh[2])
        *(univ$X[univ$ST==2]-Mxh[2]))/(Nh[2]-1))
mu12 <- cbind(sum((univ$Z[univ$ST==1]-Mzh[1])
    *((univ$X[univ$ST==1]-Mxh[1])^2))/(Nh[1]-1),
        sum((univ$Z[univ$ST==2]-Mzh[2])
        *((univ$X[univ$ST==2]-Mxh[2])^2))/(Nh[2]-1))
mu02 <- cbind(sum((univ$X[univ$ST==1]-Mxh[1])^2)/(Nh[1]-1),
          sum((univ$X[univ$ST==2]-Mxh[2])^2)/(Nh[2]-1))
mu03 <- cbind(sum((univ$X[univ$ST==1]-Mxh[1])^3)/(Nh[1]-1),
          sum((univ$X[univ$ST==2]-Mxh[2])^3)/(Nh[2]-1))

#Ratio
R <- mean(univ$X)/mean(samp$X)

#Ordinary meam
est1 <- sum(wh*mzh)
#Ratio estimator
est2 <- sum(wh*mzh)*(Mx/sum(wh*mxh))
#Regression estimator
betac <- sum((wh^2)*gh*Ro_ZXh*szh*sxh)/sum((wh^2)*gh*Ro_ZXh*(sxh^2))
est3 <- sum(wh*mzh)+betac*(Mx-sum(wh*mxh))

#Coefficient of variation
c_xh <- by(univ$X,univ$ST,sd)/Mxh
c_yh <- by(univ$Y,univ$ST,sd)/Myh
c2_xh <- c_xh^2
c2_yh <- c_yh^2
c_zh <- by(univ$Z,univ$ST,sd)/Mzh
c2_zh <- c_zh^2

#Bias of ratio estimator - 1st degree approximation
bias2i <- My*sum((wh^2)*gh*(c2_xh-Ro_ZXh*c_zh*c_xh))
#Bias of regression estimator - 1st degree approximation
bias3i <- -sum(c((wh^2)*gh*betac)*((mu12/mu11)-(mu03/mu02)))

#Mean Square Error of 1st estimator (ordinal mean)
mse1 <- sum((wh^2)*(gh*(by(univ$Y,univ$ST,var)
      +by(univ$S,univ$ST,var))))
#Mean Square Error of ratio estimator
#1st degree approximation
```

```
    mse2i <- (My^2)*sum((wh^2)*gh*(c2_zh+c2_xh-2*Ro_ZXh*c_zh*c_xh))
    #Mean Square Error of regression estimator
    #1st degree approximation
    rhoc <- sum((wh^2)*gh*Ro_ZXh*szh*sxh)
        /(sqrt(sum((wh^2)*gh*(szh^2)))*sqrt(sum((wh^2)*gh*(sxh^2))))
    mse3i <- (My^2)*sum((wh^2)*gh*c2_zh*(1-(rhoc^2)))

    #Empirical results
    #Simulation of 5000 replicas of estimates
    ...

    #Results
    res <- rbind(res,cbind(Nh,N,nh,n,
                      Ro_YXh,Ro_ZXh,
                      SY,SYh,SX,SXh,
                      Mxh,Mx,Myh,My,Mzh,Mz,
                      med_est1,med_est2,med_est3,
                      bias2i,bias3i,
                      emp_mse1,mse1,
                      emp_mse2,mse2i,
                      emp_mse3,mse3i))
  }
  colnames(res) <- c("Nh","N","nh","n",
                      "RhoXYh","RhoZXh",
                      "SY","SYh","SX","SXh",
                      "MXh","MX","MYh","MY","MZh","MZ",
                      "Est1","Est2","Est3",
                      "BIAS2I","BIAS3I",
                      "EMP_MSE1","MSE1",
                      "EMP_MSE2","MSE2I",
                      "EMP_MSE3","MSE3I")
  return(res)
}

#Package for generation
require(MASS)

#Parameters
#Population dimension
N <- 1000
#Variance-Covariance matrix
sigma1 <- matrix(c(9,3.2,3.2,4),2,2)
sigma2 <- matrix(c(6,3.3,3.3,3),2,2)
sigma3 <- matrix(c(5,3,3,2),2,2)
#Mean vector
mu <- c(5,5)
#Number of strata
L <- 2

res <- NULL
```

```
for (i in 1:length(N))
{
  res <- rbind(res,proj3_ratio_st_NeymanAlloc(N[i],sigma1,mu,L))
  res <- rbind(res,proj3_ratio_st_NeymanAlloc(N[i],sigma2,mu,L))
  res <- rbind(res,proj3_ratio_st_NeymanAlloc(N[i],sigma3,mu,L))
}

write.table(res,"chapter6_ss_results.txt",sep="\t",dec=",",row.names=FALSE)
```

Listing 6.2: R Code for Numerical Example of Proposed Estimator in Chapter 6

```

proj3_ratio_st_real <- function(Y,X,N)
{

  L <- 3
  data_yx <- data.frame(cbind(Y,X))
  colnames(data_yx) <- c("Y","X")

  #Strata
  data_yx$ST <- 0
  data_yx$ST <- ifelse(data_yx$X<10,1,
                  ifelse(data_yx$X>=10 & data_yx$X<30,2,
                      ifelse(data_yx$X>=30,3,0)))

  data_yx <- data_yx[order(data_yx$ST),]
  Y <- data_yx$Y
  X <- data_yx$X
  ST <- data_yx$ST

  S<-NULL
  #Scrambling variable independent of Y and X, with mean=0
  for (s in 1:L)
  {
    S <- c(S,rnorm(sum(ST==s),mean=0,sd=0.01*sd(X[ST==s])))
  }
  #Scrambled response
  Z <- Y+S

  #Population
  univ <- data.frame(cbind(Y=Y,S=S,Z=Z,X=X,ST=ST,NRAND=runif(N)))
  univ <- univ[order(univ$ST,univ$NRAND),]

  #Coefficients of correlation
  Ro_YX <- cor(Y,X)
  Ro_ZX <- Ro_YX/sqrt(1+(var(S)/var(Y)))
  Ro_YXh <- by(cbind(univ$Y,univ$X),univ$ST,
        function(x) {cor(x[,1],x[,2])})
  Ro_ZXh <- Ro_YXh/sqrt(1+(by(univ$S,univ$ST,var)
        /by(univ$Y,univ$ST,var)))

  #Population means
  Mz <- mean(univ$Z)
  Mx <- mean(univ$X)
  My <- mean(univ$Y)

  #Information
  SY <- sd(univ$Y)
```

```
SYh <- by(univ$Y,univ$ST,sd)
SZ <- sd(univ$Z)
SZh <- by(univ$Z,univ$ST,sd)
SX <- sd(univ$X)
SXh <- by(univ$X,univ$ST,sd)

#Samples dimension
dim_samp <- c(100,250,500)

#Information for the population
Nh <- by(univ$Z,univ$ST,length)
wh <- Nh/N
nL <- Nh[L]

res <- NULL
for (i in 1:length(dim_samp))
{
  #sample dimension
  n <- dim_samp[i]

  #sample with Neyman Allocation
  n_total <- 0
  samp <- NULL
  for (l in 1:L)
  {
    n_aux <- round(n*(Nh[l]/N))
    n_total <- n_total+n_aux
    samp <- rbind(samp,univ[univ$ST==l,][1:n_aux,])
  }

  #Sampling rate for each stratum
  nh <- by(samp$Z,samp$ST,length)
  fh <- nh/Nh
  gh <- (1-fh)/nh
  f <- n/N

  #Sampling mean for each stratum
  mzh <- by(samp$Z,samp$ST,mean)
  myh <- by(samp$Y,samp$ST,mean)
  mxh <- by(samp$X,samp$ST,mean)
  msh <- by(samp$S,samp$ST,mean)

  #Population mean for each stratum
  Mzh <- by(univ$Z,univ$ST,mean)
  Myh <- by(univ$Y,univ$ST,mean)
  Mxh <- by(univ$X,univ$ST,mean)
  Msh <- by(univ$S,univ$ST,mean)

  #Sampling mean for each stratum
  szh <- by(samp$Z,samp$ST,sd)
```

```
    sxh <- by(samp$X,samp$ST,sd)

    mu11 <- cbind(sum((univ$Z[univ$ST==1]-Mzh[1])*(univ$X[univ$ST==1]
        -Mxh[1]))/(Nh[1]-1),sum((univ$Z[univ$ST==2]-Mzh[2])
        *(univ$X[univ$ST==2]-Mxh[2]))/(Nh[2]-1),
        sum((univ$Z[univ$ST==3]-Mzh[3])*(univ$X[univ$ST==3]
        -Mxh[3]))/(Nh[3]-1))
    mu12 <- cbind(sum((univ$Z[univ$ST==1]-Mzh[1])*((univ$X[univ$ST==1]
        -Mxh[1])^2))/(Nh[1]-1),sum((univ$Z[univ$ST==2]-Mzh[2])
        *((univ$X[univ$ST==2]-Mxh[2])^2))/(Nh[2]-1),
            sum((univ$Z[univ$ST==3]-Mzh[3])*((univ$X[univ$ST==3]
            -Mxh[3])^2))/(Nh[3]-1))
    mu02 <- cbind(sum((univ$X[univ$ST==1]-Mxh[1])^2)/(Nh[1]-1),
              sum((univ$X[univ$ST==2]-Mxh[2])^2)/(Nh[2]-1),
              sum((univ$X[univ$ST==3]-Mxh[3])^2)/(Nh[3]-1))
    mu03 <- cbind(sum((univ$X[univ$ST==1]-Mxh[1])^3)/(Nh[1]-1),
              sum((univ$X[univ$ST==2]-Mxh[2])^3)/(Nh[2]-1),
              sum((univ$X[univ$ST==3]-Mxh[3])^3)/(Nh[3]-1))

    #Ratio
    R <- mean(univ$X)/mean(samp$X)

    #Ordinary meam
    est1 <- sum(wh*mzh)
    #Ratio estimator
    est2 <- sum(wh*mzh)*(Mx/sum(wh*mxh))
    #Regression estimator
    betac <- sum((wh^2)*gh*Ro_ZXh*szh*sxh)/sum((wh^2)*gh*Ro_ZXh*(sxh^2))
    est3 <- sum(wh*mzh)+betac*(Mx-sum(wh*mxh))

    #Coefficient of variation
    c_xh <- by(univ$X,univ$ST,sd)/Mxh
    c_yh <- by(univ$Y,univ$ST,sd)/Myh
    c2_xh <- c_xh^2
    c2_yh <- c_yh^2
    c_zh <- by(univ$Z,univ$ST,sd)/Mzh
    c2_zh <- c_zh^2

    #Bias of ratio estimator - 1st degree approximation
    bias2i <- My*sum((wh^2)*gh*(c2_xh-Ro_ZXh*c_zh*c_xh))
    #Bias of regression estimator - 1st degree approximation
    bias3i <- -sum(c((wh^2)*gh*betac)*((mu12/mu11)-(mu03/mu02)))

    #Mean Square Error of 1st estimator (ordinal mean)
    mse1 <- sum((wh^2)*(gh*(by(univ$Y,univ$ST,var)+by(univ$S,univ$ST,var))))
    #Mean Square Error of ratio estimator - 1st degree approximation
    mse2i <- (My^2)*sum((wh^2)*gh*(c2_zh+c2_xh-2*Ro_ZXh*c_zh*c_xh))
    #Mean Square Error of regression estimator - 1st degree approximation
    rhoc <- sum((wh^2)*gh*Ro_ZXh*szh*sxh)
        /(sqrt(sum((wh^2)*gh*(szh^2)))*sqrt(sum((wh^2)*gh*(sxh^2))))
```

```
    mse3i <- (My^2)*sum((wh^2)*gh*c2_zh*(1-(rhoc^2)))

    #Empirical results
    #Simulation of 5000 replicas of estimates
    ...

    #Results
    res <- rbind(res,cbind(Nh,N,nh,n,
                    Ro_YXh,Ro_ZXh,
                    SY,SYh,SX,SXh,
                    Mxh,Mx,Myh,My,Mzh,Mz,
                    med_est1,med_est2,med_est3,
                    bias2i,bias3i,
                    emp_mse1,mse1,
                    emp_mse2,mse2i,
                    emp_mse3,mse3i))
  }
  colnames(res) <- c("Nh","N","nh","n",
                    "RhoXYh","RhoZXh",
                    "SY","SYh","SX","SXh",
                    "MXh","MX","MYh","MY","MZh","MZ",
                    "Est1","Est2","Est3",
                    "BIAS2I","BIAS3I",
                    "EMP_MSE1","MSE1",
                    "EMP_MSE2","MSE2I",
                    "EMP_MSE3","MSE3I")
  return(res)
}

#Package for generation
require(MASS)

#Import data
data_yx <- read.table("IUTICE10_BA.txt",sep="\t",dec=",",header = T)
data_yx <- data_yx[data_yx$NPS>=10 & data_yx$NPS<150,]
data_yx <- data_yx[data_yx$turn<=200,]
#Study variable (purchase, millions of euros)
Y <- data_yx$purch
#Auxiliary variable, correlated with Y (turnover, millions of euros)
X <- data_yx$turn
#Data application
N <- dim(data_yx)[1]

res <- proj3_ratio_st_real(Y,X,N)
#Export data
write.table(res,"chapter6_ne_results.txt",sep="\t",dec=",",row.names=FALSE)
```

7

General Discussion

7.1 Summary

Our thesis work is based on the improvement of the mean estimation of sensitive variables (Edwards, 1957; Groves et al., 2004). In the sampling literature (Cochran, 1997; Mukhopadhyay, 1998; Särdnal et al., 1997; Sukhatme and Sukhatme, 1984), researchers have proposed several estimators which use auxiliary information in order to improve their performance. Over the chapters of this thesis we have proposed different estimators which combine the Randomized Response Technique (RRT) method (Eichhorn, 1983; Warner, 1965) with the use of auxiliary information.

In section 7.2 we present a numerical example that aims to make a comparison of the performance of the main proposed estimators. For that purpose we conduct a study with a real dataset and we show the numerical results for the *Bias* and Mean Square Error (*MSE*), as well as graphic evidence which illustrates the performance of each estimator in terms of estimation precision.

7.2 Comparison of the main study estimators

In this section we conduct a study with a real dataset with particular focus on comparing the performance of the main estimators proposed in this thesis, using the *Bias* and the *MSE* results as the criteria.

Consider a real dataset concerning enterprises for the Monthly Economic Survey (MES) in Portugal. The survey is conducted to provide an accurate picture of business trends of enterprises. It provides short-term indicators on a monthly basis compiled for four sectors: industry, retail trade, construction and service sector.

Generally, the enterprises do not want to report the value of their orders. This is typically a confidential variable for enterprises, only known from business surveys. Nevertheless, every year the entity responsible for MES, the Statistics Portugal [1], provides administrative information with the value of the orders for the previous year. Thus, in our numerical example, we consider the value of orders in 2009 as a sensitive variable and the value of orders in 2008 as an auxiliary variable.

Let Y be the annual orders amount in 2009 collected by the MES in that year. The auxiliary variable X is the annual orders amount in 2008, available from business data registers. The variables Y and X are strongly correlated so we can take advantage of this correlation by using an auxiliary variable. We take the 608 business survey respondents common between 2008 and 2009 as our population. For the RRT part, let S be a normal random variable with mean equal to zero and standard deviation equal to 10% of the standard deviation of X. The reported response is given by $Z = Y + S$ (the orders amount in 2008 plus a random quantity). The summary statistics about the populations are given below.

Population Characteristics:

$N = 608, \rho_{XY} = 0.9447$
$\mu_X = 21357.69, \mu_Y = 17828.2$ (in thousands of Euros)
$\sigma_X = 65874.83, \sigma_Y = 57489.53$ and $\sigma_{XY} = 3577597688$

We use the following samples sizes in our simulation study: $n = 50, 100, 200$ and 300.

In this study we compare the results for the ordinary RRT sample mean $(\hat{\mu}_Y)$ to the main estimators proposed in this study: the ratio estimator $(\hat{\mu}_R)$ (Sousa et al., 2010), the regression estimator $(\hat{\mu}_{Reg})$ (Gupta et al., 2012), the generalized regression-cum-ratio estimator $(\hat{\mu}_{GRR})$ (Gupta et al., 2012), the generalized regression-cum-exponential estimator $(\hat{\mu}_{exp1})$ (Koyuncu et al., 2013) and the improved exponential estimator $(\hat{\mu}_{IE})$ (Gupta et al., 2013).

In Table 7.1 below we present the the theoretical *ARB* results for all the estimators in comparison.

Table 7.1: Theoretical *ARB* for the estimators in comparison.

Population			**Theoretical *ARB***			
N	ρ_{XY}	Estimator	$n=50$	$n=100$	$n=200$	$n=300$
		$(\hat{\mu}_R)$	0.0022	0.0010	0.0004	0.0002
		$(\hat{\mu}_{Reg})$	0.0080	0.0036	0.0015	0.0007
608	0.9447	$(\hat{\mu}_{GRR})$	0.0225	0.0104	0.0042	0.0021
		$(\hat{\mu}_{exp1})$	0.0215	0.0093	0.0036	0.0018
		$(\hat{\mu}_{IE})$	0.0042	0.0022	0.0009	0.0005

The *ARB* results show that it is not always the case that estimators with better performance in terms of accuracy are the best performing in terms of *Bias* as well. However, the improved exponential estimator $(\hat{\mu}_{IE})$ manages to combine great precision results with reduction in *Bias* compared to other estimators which use auxiliary information, such as the ratio estimator $(\hat{\mu}_R)$, the regression estimator $(\hat{\mu}_{Reg})$ and the study generalized exponential estimators ($\hat{\mu}_{exp1}$ and $\hat{\mu}_{IE}$).

From our 5000 samples, selected for each sample size and for each estimator, we take the empirical *Bias* and we draw a graph, presented in Figure 7.1, which shows the *Bias* distribution for all the generated estimates.

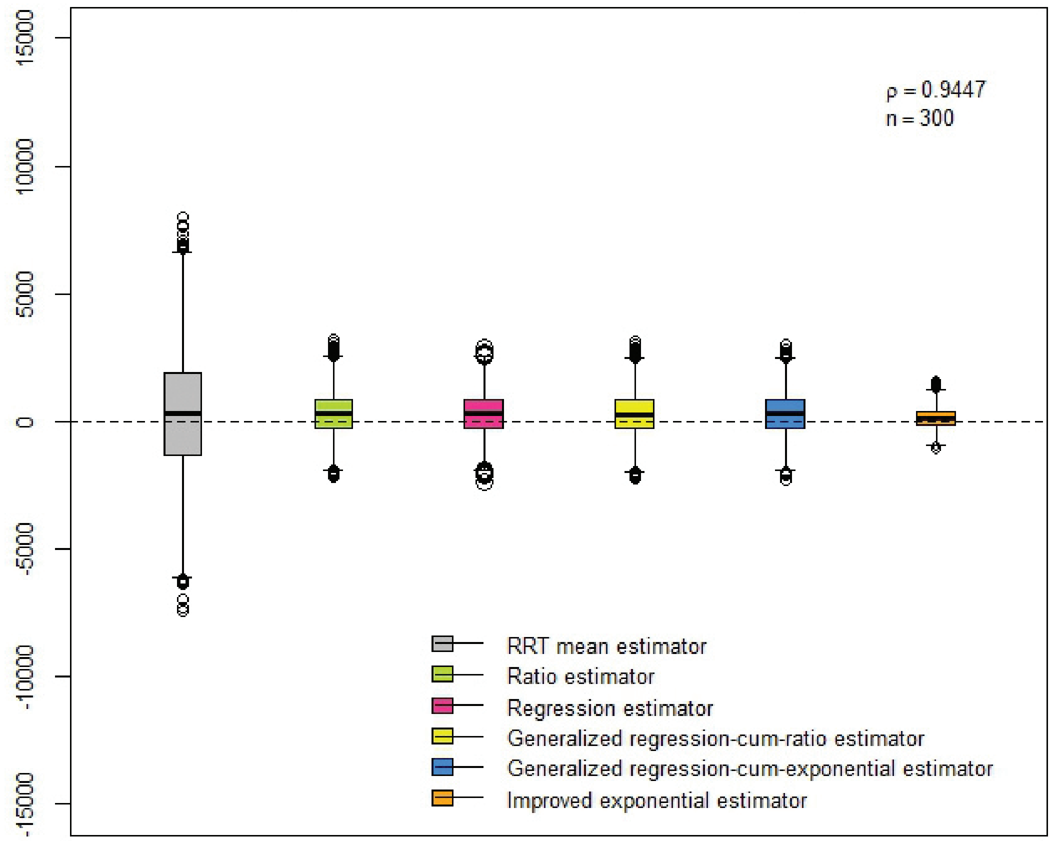

Figure 7.1: Distribution of empirical *Bias*

According to the graph in Figure 7.1 all the estimators have *Bias* about zero but it is the improved exponential estimator which shows less dispersion, with results closer to zero. Despite being unbiased, the RRT mean estimator's greater empirical *Bias* relative to the estimators which use auxiliary information, also based on RRT, is obvious in this graph indicating that just a RRT version of the mean estimator might not be enough.

Table 7.2 below gives empirical and theoretical *MSE*'s based on the first order of approximation for all the estimators considered here. We estimate the empirical *MSE* using 5000 samples of various sizes selected from the study population. We use the following expression to find the Percent Relative Efficiency (*PRE*) of ratio, regression, generalized regression-cum-ratio, generalized regression-cum-exponential and improved exponential estimators as compared to the RRT mean estimator:

$$PRE = \frac{MSE(\hat{\mu}_Y)}{MSE(\hat{\mu}_\alpha)} \times 100,$$

where $\alpha = R, Reg, GRR, exp1, IE.$

Table 7.2: Empirical *MSE*, theoretical *MSE* correct up to 1^{st} order approximation and *PRE* for all the estimators in comparison relative to the RRT mean estimator.

Population				***MSE* Estimation**		
N	ρ_{XY}	n	Estimator	Empirical	Theoretical	*PRE*
608	0.9447	50	$\hat{\mu}_Y$	63985642.27	61487318.16	100.00
			$\hat{\mu}_R$	7098375.83	7357491.69	835.71
			$\hat{\mu}_{Reg}$	7520669.52	7349015.26	836.67
			$\hat{\mu}_{GRR}$	6682678.20	7148757.06	860.11
			$\hat{\mu}_{exp1}$	6308486.16	6721357.39	914.81
			$\hat{\mu}_{IE}$	1633458.04	1334326.31	4608.12
		100	$\hat{\mu}_Y$	27999936.04	27988850.92	100.00
			$\hat{\mu}_R$	3593960.07	3349109.12	835.71
			$\hat{\mu}_{Reg}$	3531015.22	3345250.67	836.67
			$\hat{\mu}_{GRR}$	3431317.25	3307434.75	846.24
			$\hat{\mu}_{exp1}$	3252540.69	3213576.01	870.96
			$\hat{\mu}_{IE}$	794022.08	686993.08	4074.11
		200	$\hat{\mu}_Y$	11444042.86	11239617.30	100.00
			$\hat{\mu}_R$	1475109.02	1344917.84	835.71
			$\hat{\mu}_{Reg}$	1412070.68	1343368.38	836.67
			$\hat{\mu}_{GRR}$	1416483.71	1337528.93	840.33
			$\hat{\mu}_{exp1}$	1370565.25	1322001.04	850.20
			$\hat{\mu}_{IE}$	0333343.39	292135.71	3847.40
		300	$\hat{\mu}_Y$	5728956.33	5656539.42	100.00
			$\hat{\mu}_R$	791784.85	676854.07	835.71
			$\hat{\mu}_{Reg}$	752752.90	676074.28	836.67
			$\hat{\mu}_{GRR}$	765577.73	674615.91	838.48
			$\hat{\mu}_{exp1}$	740274.33	670651.05	843.44
			$\hat{\mu}_{IE}$	175396.47	149770.32	3776.81

According to the *MSE* results in Table 7.2 the component regression shows performance gains. As expected and shown in Chapter 5 the best performance comes from improved exponential estimator because of its large reduction in *MSE*.

From our 5000 samples, selected for each sample size and for each estimator, we take the empirical *MSE* and we draw a graph, presented in Figure 7.2, which shows the precision distribution for all the generated estimates.

According to the graph in Figure 7.2 the use of auxiliary information significantly reduces the magnitude of *MSE*, particularly in the improved exponential estimator which presents results closer to zero.

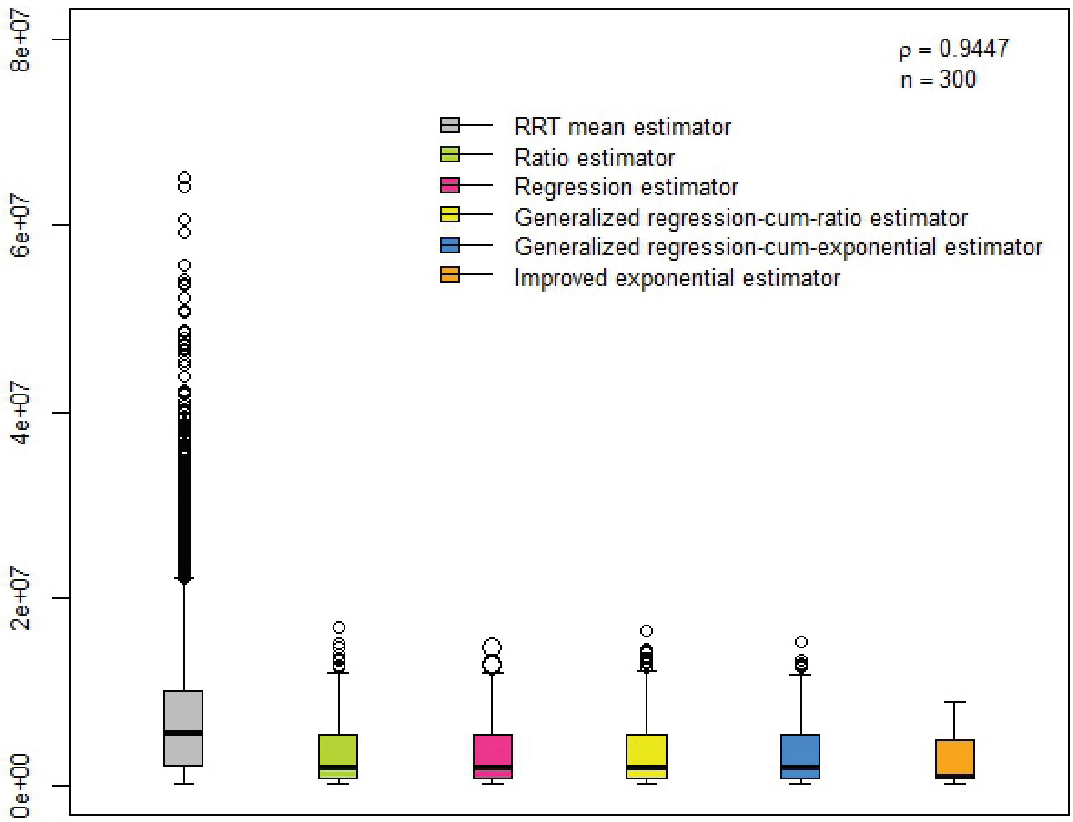

Figure 7.2: Distribution of empirical *MSE*

7.3 Final Remarks

The aim of this project was to develop new methodologies that can potentially improve the mean estimation in the presence of auxiliary information. These new methodologies were proposed in Chapters 2 to 6 and were compared with each other and with the ordinary RRT mean estimator which does not uses the auxiliary information. For that purpose we studied theoretically the proposed estimators, deriving the expressions for the *Bias* and Mean Square Error (*MSE*) correct up to first or second order approximations. Also, R routines [2] were developed for an extensive study by using real data and simulated data for all the estimators under study.

One of the main conclusions was the use of auxiliary information significantly reduces the magnitude of *MSE*, providing a gain for the parameter estimation based on RRT, just as in the context of direct estimation of non-sensitive parameters.

We concluded from this study that the estimation of the mean of a sensitive variable can be improved further by using a correlated non-sensitive auxiliary variable.

When there is a high correlation between the study variable and the auxiliary variable the regression estimator performs better than ratio estimator.

We also found some exponential type estimators more efficient than the ratio and regression type estimators.

We showed that the advantage of using the RRT in the presence of auxiliary information still holds with other sampling designs, such as the stratified sampling (Sousa et al., 2013).

Even though during the thesis project we have tested many new estimators and compared them to existing estimators in literature, we only present those who showed an effective improvement relative to the existing estimators or to the estimators previously proposed for us.

We are aware that in this area there is still much more to explore and part of our future work plans consists of studying other combinations of estimators, as well as the application to different sampling designs and with different techniques which provides confidence to the respondents when they have to answer to sensitive questions. Also, we would like to plan and implement a survey to a group of respondents who struggle with sensitive questions in order to evaluate the performance of proposed estimators with a real application of a RRT.

Even with the natural constraints we face in this kind of research, this study was a great challenge, both in the theoretical context as well as in the practical context of parameter estimation in the presence of auxiliary information.

References

COCHRAN, W.G. 1997. *Sampling Techniques*, 3^{rd} Ed., New York, Wiley Eastern Ltd.

EDWARDS, A. L. 1957. *The social desirability variable in personality assessment and research*, New York: Dryden, Praeger.

EICHHORN, B. H. & HAYRE, L. S. 1983. Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*, 7, 307-316.

GROVES, R. M., FOWLER, F. J., COUPER, M. P., LEPKOWSKI, J. M., SINGER, E. & TOURANGEAU, R. 2004. *Survey Methodology*, New York, Wiley.

GUPTA, S., SHABBIR, J., SOUSA, R. & REAL, P. C. 2012. Estimation of the Mean of a Sensitive Variable in the Presence of Auxiliary Information. *Communications in Statistics - Theory and Methods*, 41(13-14), 2394-2404.

GUPTA, S., SHABBIR, J., SOUSA, R. & REAL, P. C. 2013. Improved exponential type estimators of the mean of a sensitive variable in the presence of non-sensitive auxiliary information. (*submitted*)

KOYUNCU, N., GUPTA, S. & SOUSA, R. 2013. Exponential type estimators of the mean of a sensitive variable in the presence of non-sensitive auxiliary information. *Communications in Statistics - Simulation and Computation*. (*accepted*).

MUKHOPADHYAY, P. 1998. *Theory and Methods of Survey Sampling*, New Delhi, Prentice-Hall of India.

SÄRDNAL, C.-E., SWENSSON, B. & WRETMAN, J. 1997. *Model assisted survey sampling*, Springer series in statistics.

SOUSA, R., GUPTA, S., SHABBIR, J. & REAL, P. C. 2013. Improved Mean Estimation of a Sensitive Variable Using Auxiliary Information in Stratified Sampling. *Journal of Statistics and Management Systems*. (*submitted*).

SOUSA, R., SHABBIR, J., REAL, P. C. & GUPTA, S. 2010. Ratio estimation of the mean of a sensitive variable in the presence of auxiliary information. *Journal of Statistical Theory and Practice*, 4(3), 495-507.

SUKHATME, P.V. & SUKHATME, B.V. 1984. *Sampling theory of surveys with applications*, 3^{rd} Ed., Ames, Iowa, Iowa State University Press.

WARNER, S. L. 1965. Randomized response: a survey technique for elimination evasive answer bias. *Journal of American Statistical Association*, 60, 63-69.

[1] Statistics Portugal: http://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_main&xlang=en.

[2] The R Project for Statistical Computing: www.r-project.org.

Appendix F - R Routines

Listing 7.1: R Code for Numerical Example of Proposed Estimator in Chapter 7

```
comparison_chap7 <- function(Y,X,N)
{
  #Coefficient of correlation between Y and X
  Ro_YX <- cor(Y,X)

  #Scrambling variable independent of Y and X, with mean=0
  S <- rnorm(N,mean=0,sd=0.1*sd(X))
  #Scrambled response
  Z <- Y+S

  #Coefficient of correlation between Z and X
  Ro_ZX <- Ro_YX/sqrt(1+(var(S)/var(Y)))

  #population
  univ <- data.frame(cbind(Y=Y,S=S,Z=Z,X=X,NRAND=runif(N)))
  univ <- univ[order(univ$NRAND),]

  #Mean of Y
  mz <- mean(univ$Z)
  mx <- mean(univ$X)
  my <- mean(univ$Y)

  mu11 <- sum((univ$Z-mz)*(univ$X-mx))/(N-1)
  mu12 <- sum((univ$Z-mz)*((univ$X-mx)^2))/(N-1)
  mu02 <- sum((univ$X-mx)^2)/(N-1)
  mu03 <- sum((univ$X-mx)^3)/(N-1)

  beta_zx <- Ro_YX*(sd(univ$Y)/sd(univ$X))

  #Samples dimension
  dim_samp <- c(50,100,200,300)

  #Initialize the variables...

  for (i in 1:length(dim_samp))
  {
    #sample dimension
    n <- dim_samp[i]
    #sample
    samp <- univ[1:n,]
    #Sampling rate
    f <- n/N

    #Ordinary meam
    est1 <- mean(samp$Z)
```

```
#Ratio estimator
est2 <- mean(samp$Z)*(mx/mean(samp$X))
#Regression estimator
est3 <- mean(samp$Z)+beta_zx*(mx-mean(samp$X))

#Coefficient of variation
c_x <- sd(univ$X)/mx
c_y <- sd(univ$Y)/my
c2_x <- c_x^2
c2_y <- c_y^2
c2_z <- c2_y+(var(univ$S)/(my^2))
c_z <- sqrt(c2_z)

l <- (1-f)/n

#Generalized Regression-cum-ratio Estimator
k1 <- (1-((1-f)*c2_x/n))/(1-((1-f)/n)*(c2_x-c2_z*(1-(Ro_ZX^2))))
k2 <- (my/mx)*(1+k1*((Ro_ZX*c_z/c_x)-2))
est5 <- (k1*mean(samp$Z)+k2*(mx-mean(samp$X)))
    *(mx/mean(samp$X))

#Generalized regression-cum-exponential type Estimator
w1 <- (1-(l*c2_x/8))/(1+l*c2_z*(1-(Ro_ZX^2)))
w2 <- (my/mx)*(0.5-w1*(1-(Ro_ZX*c_z/c_x)))
est7 <- (w1*mean(samp$Z)+w2*(mx-mean(samp$X)))
    *exp((mx-mean(samp$X))/(mx+mean(samp$X)))

#2nd Improved Exponential Estimator
A <- 1+l*c2_z+l*c2_x-2*l*Ro_ZX*c_z*c_x
B <- 1+l*c2_x
C <- 1+(3/8)*l*c2_x-0.5*l*Ro_ZX*c_z*c_x
D <- 1+(3/8)*l*c2_x
E <- 1+l*c2_x-l*Ro_ZX*c_z*c_x
z1 <- (B*C-D*E)/(A*B-(E^2))
z2 <- my*(A*D-C*E)/(A*B-(E^2))
est8 <- (z1*mean(samp$Z)+z2)
    *exp((mx-mean(samp$X))/(mx+mean(samp$X)))

#Mean Square Error of 1st estimator (ordinal mean)
mse1 <- ((1-f)/n)*(var(univ$Y)+var(univ$S))

#Bias of ratio estimator - 1st degree approximation
bias2i <- ((1-f)/n)*my*(c2_x-Ro_ZX*c_z*c_x)
#Mean Square Error of ratio estimator - 1st degree approximation
mse2i <- ((1-f)/n)*(my^2)*(c2_z+c2_x-2*Ro_ZX*c_z*c_x)

#Bias of regression estimator - 1st degree approximation
bias3i <- -beta_zx*((1-f)/n)*((mu12/mu11)-(mu03/mu02))
#Mean Square Error of regression estimator
#1st degree approximation
```

```
mse3i <- ((1-f)/n)*(my^2)*c2_z*(1-(Ro_ZX^2))

#Bias of generalized regression-cum-ratio estimator
#1st degree approximation
bias5i <- (k1-1)*my+k1*my*((1-f)/n)*(c2_x-Ro_ZX*c_z*c_x)
      +k2*mx*((1-f)/n)*c2_x
#Mean Square Error of generalized regression-cum-ratio estimator
#1st degree approximation
mse5i <- ((k1-1)^2)*(my^2)+(k1^2)*(my^2)
     *((1-f)/n)*(c2_z+3*c2_x-4*Ro_ZX*c_z*c_x)
     +(k2^2)*(mx^2)*((1-f)/n)*c2_x-2*k1*(my^2)
     *((1-f)/n)*(c2_x-Ro_ZX*c_z*c_x)
     -2*k2*my*mx*((1-f)/n)*c2_x-2*k1*k2*my*mx
     *((1-f)/n)*(Ro_ZX*c_z*c_x-2*c2_x)

#Bias of generalized exponential type estimator
#1st degree approximation
bias7i <- (w1-1)*my+w1*my*l*((3/8)*c2_x-0.5*Ro_ZX*c_z*c_x)
      +w2*mx*l*c2_x
mse7i <- (my^2)*((1-0.25*l*c2_x)-(((1-(1/8)*l*c2_x)^2)
     /(1+l*c2_z*(1-(Ro_ZX^2)))))

#Bias of improved exponential estimator 2
#1st degree approximation
bias8i <- (z1-1)*my+z1*my*((3/8)*l*c2_x-0.5*l*Ro_ZX*c_z*c_x)
    +z2*(1+(3/8)*l*c2_x)
#Mean Square Error of improved exponential estimator 2
#1st degree approximation
mse8i <- (my^2)*(1-((B*(C^2)+A*(D^2)-2*C*D*E)/(A*B-(E^2))))

#Empirical results
#Simulation of 5000 replicas of estimates
...

#Graphics
#*************************************
emp_bias <- emp-my
emp_arb <- abs(emp_bias)/my
emp_mse <- apply(emp,2,var)+(emp_bias^2)
emp_res <- rbind(emp_res,rbind(cbind(N,n,Ro_YX,1,
  emp_bias[,1],emp_arb[,1],emp_mse[,1]),
  cbind(N,n,Ro_YX,2,emp_bias[,2],emp_arb[,2],emp_mse[,2]),
  cbind(N,n,Ro_YX,3,emp_bias[,3],emp_arb[,3],emp_mse[,3]),
  cbind(N,n,Ro_YX,5,emp_bias[,5],emp_arb[,5],emp_mse[,5]),
  cbind(N,n,Ro_YX,7,emp_bias[,7],emp_arb[,7],emp_mse[,7]),
  cbind(N,n,Ro_YX,8,emp_bias[,8],emp_arb[,8],emp_mse[,8])))
colnames(emp_res) <- c("N","n","RhoXY","EST",
    "EMP_BIAS","EMP_ARB","EMP_MSE")
#*************************************
```

```
    #Results
    res <- rbind(res,c(N,n,Ro_YX,Ro_ZX,
                       c_x,c_y,c_z,k1,k2,w1,w2,
                       z1,z2,mx,my,mz,
                       med_est1,med_est2,med_est3,
                       med_est5,med_est7,med_est8,
                       bias2i,bias3i,bias5i,
                       bias7i,bias8i,
                       emp_mse1,mse1,emp_mse2,mse2i,
                       emp_mse3,mse3i,emp_mse5,mse5i,
                       emp_mse7,mse7i,emp_mse8,mse8i))
  }
  colnames(res) <- c("N","n","RhoXY","RhoZX",
                       "Cx","Cy","Cz","k1","k2","w1","w2",
                       "z1","z2","mX","mY","mZ",
                       "Est1","Est2","Est3",
                       "Est5","Est7","Est8",
                       "BIAS2I","BIAS3I","BIAS5I",
                       "BIAS7I","BIAS8I",
                       "EMP_MSE1","MSE1","EMP_MSE2","MSE2I",
                       "EMP_MSE3","MSE3I","EMP_MSE5","MSE5I",
                       "EMP_MSE7","MSE7I","EMP_MSE8","MSE8I")
  return(list(res=res,emp_res=emp_res))
}

#Package for generation
require(MASS)
#Import data
data_yx <- read.table("ENC0809.txt",sep="\t",dec=",",header = T)
#Study variable (orders in 2009, thousands of euros)
Y <- data_yx[,3]
#Auxiliary variable (orders in 2009, thousands of euros)
X <- data_yx[,2]

#Data application
N <- dim(data_yx)[1]
res <- comparison_chap7(Y,X,N)

res_exp<-res[[1]]
#Export data
write.table(res_exp,"chapter7_ne_results.txt",sep="\t",dec=",",row.names=FALSE)
```