

ABSTRACT

Title of Dissertation: ANALYSIS OF COMPLEX SURVEY DATA
USING ROBUST MODEL-BASED AND
MODEL-ASSISTED METHODS

Yan Li, Ph.D., 2006

Directed By: Professor, Partha Lahiri
Joint Program in Survey Methodology

Over the past few decades, major advances have taken place in both model-based and model-assisted approaches to inferences in finite population sampling. In the standard model-based approach, the finite population is assumed to be a realization from a superpopulation characterized by a probability distribution, and that the distribution of the sample is identical to that of the finite population. The model-based method could lead to a misleading inference if either assumption is violated. The model-assisted estimators typically are consistent or at least approximately unbiased with respect to the sampling design, and yet more efficient than the customary randomization-based estimators in the sense of achieving smaller variance with respect to the design if the assumed model is appropriate.

Since both approaches rely on the assumed model, there is a need to achieve robustness with respect to the model selection. This is precisely the main theme of this dissertation. This study uses the well-known Box-Cox transformation on the dependent variable to generate certain robust model-based and model-assisted estimators of finite population totals. The robustness is achieved since the appropriate transformation on the dependent variable is determined by the data. Both

Monte Carlo simulation study and real data analyses are conducted to illustrate the robustness properties of the proposed estimation method using two different ways: (i) design-based, and (ii) model-based, wherever appropriate.

A few potential areas of future research within the context of transformations in linear regression models, as well as linear mixed models, for analysis of complex survey data are identified.

ANALYSIS OF COMPLEX SURVEY DATA USING ROBUST MODEL-BASED
AND MODEL-ASSISTED METHODS

By

Yan Li

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2006

Advisory Committee: Arrange the names in alphabetical order
Professor Partha Lahiri, Chair
Professional Lecturer Fritz Scheuren
Professor Katharine Abraham
Professor Paul Smith
Mr. Paul D. Williams
Professor Wolfgang Jank

© Copyright by
Yan Li
2006

Preface

Over the past few decades, major advances have taken place in both model-based and model-assisted approaches to inferences in finite population sampling. In the standard model-based approach, the finite population is assumed to be a realization from a superpopulation characterized by a probability distribution, and that the distribution of the sample is identical to that of the finite population. The model-based method could lead to a misleading inference if either assumption is violated. The model-assisted estimators typically are consistent or at least approximately unbiased with respect to the sampling design, and yet more efficient than the customary randomization-based estimators in the sense of achieving smaller variance with respect to the design if the assumed model well describes the finite population.

Since both approaches rely on the assumed model, there is a need to achieve robustness with respect to the model selection. This is precisely the main theme of this dissertation.

In Chapter 1, the Box-Cox transformation on the dependent variable is used to generate robust model-based estimator of a finite population total. The proposed approach deviates from the usual model-based approach that uses a linear regression model with the normality assumption on the dependent variable or a known transformation, such as the log-transformation or the square root transformation, on the dependent variable. The robustness is achieved because the appropriate transformation on the dependent variable is automatically determined by the data. The proposed research suggests a new way to achieve robustness in addressing various inferential issues in the prediction approach to the finite population theory.

In Chapter 2, an automated generalized regression (AUTOGREG) estimator of a finite population total and its variance estimator under a general unequal probability sampling design are proposed. This estimator maintains the robustness property of the usual GREG estimator in the sense that AUTOGREG is design-consistent even if the underlying model fails. The class of ‘robust models’ is further extended such that an appropriate transformation on the dependent variable is automatically determined by the data using the Box-Cox technique. The AUTOGREG method does not require a linear model on the dependent variable assumed under the GREG theory.

Chapter 3 identifies a few potential future research topics.

This is a “manuscript” style dissertation, in which each chapter is like a paper that is publishable in a statistical journal. However, unlike a standard journal paper, each chapter includes detail explanations and derivations. Since they are intended to be stand-alone papers, each has a brief literature review specific to the topic.

Acknowledgements

First and foremost, I must thank my mom for her support and her love of motherhood. Whenever I need her, she dedicates all her time and helps me out. Her love to me makes me a better student. The past six years was meeting my husband Cheng, falling in love with him, and growing up with him together. Thanks to his love, his encouragement, and his friendship I overcome the initial pain in research and become a better researcher, but also a better person.

I own much gratitude to Partha Lahiri, my master thesis and Ph.D. dissertation advisor, for the past five years. Not only has he given me the necessary and expected assistance on my research, but he makes me interested in the subject and helps me with my career development. I would also thank the director of JPSM: Roger Tourangeau and my committee: Fritz Scheuren, Katharine Abraham, Paul Smith, Paul D. Williams, Wolfgang Jank. They have all provided me whatever support I have asked from them.

Table of Contents

Preface.....	ii
Acknowledgements.....	iv
Table of Contents.....	v
List of Tables.....	vii
List of Figures.....	viii
Chapter 1: A Robust Model-Based Predictor of A Finite Population Total.....	1
1.1 Introduction.....	1
1.2 The Box-Cox transformation.....	5
1.2.1 Different Box-Cox transformation forms.....	5
1.2.2 Estimation of transformation parameter λ , and model parameters β	7
1.2.3 Prediction in the original scale.....	9
1.3 Prediction of the finite population total based on linear and log-linear model.....	11
1.3.1 Prediction under a standard linear model.....	12
1.3.2 Prediction under a loglinear model.....	13
1.4 Robust model-based predictor of the population total.....	15
1.4.1 Estimation of $\theta = (\beta, \lambda, \sigma)'$	17
1.4.2 Estimating the asymptotic variance-covariance matrix of $\hat{\theta}$	19
1.4.3 Prediction of the finite population total.....	21
1.4.4 Estimation of the prediction variance of the population total predictor.....	23
1.4.5 Confidence interval of the population total predictor.....	25
1.5 Simulation study.....	27
1.6 Real data analysis.....	30
1.7 Concluding remarks.....	33
Appendix for Chapter 1.....	42
Chapter 2: Automated Generalized Regression (AUTOGREG) Estimators of A Finite Population Total.....	49
2.1 Introduction.....	49

2.2	Design-based, model-based and GREG estimators of a finite population total	52
2.2.1	Design-based Estimator	53
2.2.2	Model-based estimators	54
2.2.3	GREG estimators	58
2.3	AUTOGREG estimator of the finite population total.....	60
2.3.1	Estimation of model and transformation parameters $\boldsymbol{\varphi} = (\boldsymbol{\beta}, \lambda, \sigma^2)'$ using the PML method	60
2.3.2	AUTOGREG estimator of population total and its variance estimator	66
2.4	A Simulation study	69
2.5	Results.....	71
2.6	Concluding remarks	73
Chapter 3: Summary and Future research.....		83
Appendix: <i>R</i> Programs		86
Bibliography		109

List of Tables

Table 1.1: The prediction mean squared error of six predictors ¹	34
Table 1.2: The prediction relative bias of six predictors ¹	35
Table 1.3: Percentage of loss of prediction mean squared error	36
Table 1.4: Estimates and 95% confidence intervals for β_0, β_1 , and λ	38
Table 1.5: <i>AARD</i> , <i>ALCI</i> , and <i>P</i> for three predictors based on models $\mathbf{M}_1, \mathbf{M}_2$, and \mathbf{M}_3	41
Table 2.1: Relative biases of the different estimators using different sampling designs with varying sample sizes ($\sigma=0.5$)	74
Table 2.2: Root mean square errors of the different estimators using different sampling designs with varying sample sizes ($\sigma=0.5$)	75
Table 2.3: Relative biases of the different estimators using different sampling designs with varying sample sizes ($\sigma=1$)	76
Table 2.4: Root mean squared errors of the different estimators using different sampling designs with varying sampling sizes ($\sigma=1$).....	77
Table 2.5: Relative biases of the different estimators using different sampling designs with varying sample sizes ($\sigma=2$)	78
Table 2.6: Root mean squared errors of the different estimators using different sampling designs with varying sampling sizes ($\sigma=2$).....	79
Table 2.7: Relative biases and root mean square error of $\hat{\lambda}^1$ and $\hat{\lambda}_w^2$ for SSRS sampling with varying sample sizes and standard deviations.	82

List of Figures

Figure 1.1: Histograms for the beef population before and after taking the log-transformation	37
Figure 1.2: Scatter plots for the beef population before and after taking the log-transformation	37
Figure 1.3: Histogram of $\hat{\lambda}$'s estimated using 1000 bootstrap samples	38
Figure 1.4: Scatter plot for the beef population after taking	39
Figure 1.5: Distribution of the absolute value of relative bias and the length of 95% confidence interval for three predictors over the 431 possible samples	40
Figure 2.1: Comparison of root mean square error of \hat{T}_{G-L} , \hat{T}_{G-LOGL} , and \hat{T}_{AG} with varying sampling designs, sample sizes, and standard deviations.....	80
Figure 2.2: Comparison of root mean square error of \hat{T}_{M-BC} vs. \hat{T}_{AG} with varying sampling designs, sample sizes, and standard deviations.	81

Chapter 1: A Robust Model-Based Predictor of A Finite Population Total

1.1 Introduction

The use of a superpopulation to describe a finite population can be traced back at least to Cochran (1939). Brewer (1963) and Royall (1970) considered a prediction approach to estimate the finite population mean, partly motivated by a superpopulation model. For a comprehensive review of the subject, see the books by Bolfarine and Zacks (1992), Valliant, *et. al.* (2000), and Korn and Graubard (1999), and the review paper by Graubard and Korn (2002). We refer to the book by Ghosh and Meeden (1997) for a related Bayesian approach, and Ghosh and Meeden (1985) for an empirical Bayesian approach. Rao (2005) examined the interplay between sample survey theory and practice over the past 60 years or so.

Under the standard superpopulation prediction approach, the finite population is assumed be a realization from a superpopulation generated by a probability model. The superpopulation model is then used to predict the non-sampled units from the knowledge gained through the sample. One nice feature of the prediction approach is that it can lend itself to a conditional inference, i.e. probability statement about the parameter of interest can be made conditional on the data. The main criticism about this approach is that the prediction could be unreliable in case of a model misspecification. Therefore, model robustness is important, a topic studied by researchers from different perspectives. For example, Meeden (1999) proposed a noninformative Bayesian approach for two-stage cluster sampling. Valliant (1985, 1986) extended the model-based estimation to certain

non-linear models. See Hartley and Rao (1968) for a “scale-load” approach and Kott (2005) for a randomization-assisted model-based approach. Ghosh and Lahiri (1986) proposed a robust empirical Bayes estimator of a finite population mean using certain moment assumptions. Arora, Lahiri, and Mukherjee (1997) relaxed the homoscedasticity assumption of Ghosh and Meeden (1986). Ghosh, Lahiri and Tiwari (1989) proposed a nonparametric empirical Bayes method that uses the Dirichlet process prior.

In the mainstream statistics, transformations are often used to achieve normality, linearity, and homoscedasticity (Carroll and Ruppert, 1988). But the literature on transformations in finite population inference is not very rich. There is, however, a growing interest in developing methods that use an appropriate transformation with survey data. Chen and Chen (1996) considered a known transformation on the survey data in order to improve on the precision of the normal approximation. Korn and Graubard (1998) compared different confidence intervals, including intervals based on a logit-transformation, for proportions with small expected number of positive counts. Karlberg (2000) proposed an estimator based on a lognormal-logistic superpopulation model to predict the finite population total of a highly skewed survey variable. Their simulation results indicated that the lognormal-logistic model estimator offers a sensible alternative to other estimators, especially when the sample size is small. Recently, Chambers and Dorfman (2003) discussed the estimation of finite population mean under certain general but known transformation on the continuous data.

Researchers find the transformation technique useful in analyzing survey data. However, the key step is the identification of an appropriate transformation that fits the survey data well. In many applications, the form of transformation is determined

subjectively. However, a priori knowledge or theory may not suggest the transformation to be used. In such situations, it would be convenient to determine the transformation adaptively using the data.

The work of Box and Cox (1964) has led to the development of “data-decide-transformation” methods for constructing models with independently and identically distributed (iid) errors. Their paper and other papers on the subject, including Tukey (1957), John and Draper (1980), and Bickel and Doksum (1981), have inspired a large volume of applied research. Spitzer (1976) estimated the relationship between the demand for money and the liquidity trap with a generalized Box-Cox model. An examination of the incidence of malaria equation by Newman (1977) concluded that the functional specification obtained by using the Box-Cox procedure was superior to earlier specifications. Soybean yield functions have been examined by Miner (1982) and Davison *et al.* (1989) have modeled U.S. soybean export. They concluded that the transformation provides approximately normally distributed error terms. A bibliography of the published research related to the Box-Cox transformation can be found in a review paper by Sakia (1992).

Although there is an extensive literature on parametric estimation of Box-Cox regression models (Egy and Lahiri, 1979; Savin and White, 1978; White, 1972; Zarembka, 1968), the literature on prediction of variables is sparse. These papers mostly focus on the prediction of the conditional mean and/or median of a single future observation (Sakia, 1990; Talyor, 1986; Carrol, 1982; Yang, 1999; Yang 2002; Carrol and Ruppert, 1981). Collins (1991) reviewed and compared different prediction techniques for Box-Cox regression models, including plug-in, mean squared error (MSE)

analysis, predictive likelihood, and stochastic simulation. These techniques take into account non-normality and parameter uncertainty in varying degrees.

Smallwood and Blaylock (1986) considered a predictor for the mean of multiple future observations via a Monte Carlo simulation. However, neither theoretical nor empirical considerations have been given to the properties of their predictor. In this chapter, we use the Box-Cox transformation on the dependent variable to generate robust model-based predictor of a finite population total. Specifically, two issues are addressed in this chapter: prediction variance of the proposed predictor and the associated prediction interval of the finite population total.

Our approach deviates from the usual model-based approach that uses a linear regression model with the normality assumption on the dependent variable or a known transformation, such as the log-transformation or the square root transformation, on the dependent variable. The robustness is achieved because the appropriate transformation on the dependent variable is automatically determined by the data. The proposed research suggests a new way to achieve robustness in addressing various inferential issues in the prediction approach to the finite population theory.

In Section 1.2, Box-Cox transformation is briefly reviewed. In Section 1.3, an overview of the finite population prediction theory that is based on two linear regression models is given. In Section 1.4, we propose the robust model-based approach to the finite population sampling. Our predictor is evaluated using a Monte Carlo simulation study and a real data analysis in Sections 1.5 and 1.6. In Section 1.7 we offer concluding remarks. Unlike other model-based approaches, our approach is adaptive, i.e., the model is determined automatically by the survey data and hence should be appealing to the

practitioners. In addition, as our numerical results suggest, our approach utilizes available auxiliary variables in an efficient way and offers a potential attractive alternative to the relatively more expensive design-based methods that require more samples to achieve the same level of precision.

1.2 The Box-Cox transformation

Many important results in statistical analysis follow the assumptions that the population is normally distributed with a common variance and additive error structure. In situations where the various assumptions are violated, researchers often transform the dependent variable for which the assumptions are more reasonable. Transformation on dependent variable in a linear model is not a new idea. In this section, various transformation forms, the estimation methods for the model and transformation parameters, and the prediction in the original scale are briefly reviewed.

1.2.1 Different Box-Cox transformation forms

Transformations were first introduced for the general linear model for a single response to help validate the assumptions of the model:

$$\mathbf{y}^{(\lambda)} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$
$$\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

Here, \mathbf{y} is an n by 1 vector of observations, \mathbf{X} is an n by p known, constant design matrix with full rank p , and $\boldsymbol{\beta}$ is a p by 1 vector of unknown, constant population parameters. Tukey (1957) discussed a family of power transformations for the response value of a linear model represented by:

$$y_i^{(\lambda)} = \begin{cases} y_i^\lambda & \lambda \neq 0 \\ \log(y_i) & \lambda = 0 \end{cases}; i = 1, \dots, n.$$

Box and Cox (1964) introduced the more general form to take into account the discontinuity at $\lambda = 0$:

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^\lambda - 1}{\lambda} & \lambda \neq 0; y_i > 0; i = 1, \dots, n. \\ \log(y_i) & \lambda = 0 \end{cases} \quad (1.1)$$

The transformation (1.1) holds for only $y_i > 0$; thus, Box and Cox (1964) also proposed a shifted power transformation:

$$y_i^{(\lambda)} = \begin{cases} \frac{(y_i + \lambda_2)^{\lambda_1} - 1}{\lambda_1} & \lambda_1 \neq 0; y_i > -\lambda_2; i = 1, \dots, n. \\ \log(y_i + \lambda_2) & \lambda_1 = 0 \end{cases} \quad (1.2)$$

John and Draper (1980) argued that the Box-Cox family of power transformations does not perform well when the distribution of the data already exhibits symmetry but has long tails. The transformation primarily removes skewness, and data in this case does not need this correction. They proposed the “modulus transformation” to normalize distributions that are originally close to symmetric:

$$y_i^{(\lambda)} = \begin{cases} \text{sign} \left\{ \frac{(|y_i| + 1)^\lambda - 1}{\lambda} \right\} & \lambda \neq 0; i = 1, \dots, n. \\ \log(|y_i| + 1) & \lambda = 0 \end{cases} \quad (1.3)$$

It must be noted that the range of $y_i^{(\lambda)}$ in either (1.1), (1.2), or (1.3) is restricted depending on whether λ (λ_1) is positive or negative. The transformed values consequently do not cover real line, and their distributions have a bounded support. Thus, only approximate normality in errors for the new model with transformed response

can be achieved. Bickel and Doksum (1981) introduced another version of the Box-Cox transformation to account for this limitation:

$$y_i^{(\lambda)} = \frac{|y_i|^\lambda \text{sign}(y_i) - 1}{\lambda}; \lambda > 0. \quad (1.4)$$

This includes distributions of the transformed data with unbounded support, such as the normal distribution.

1.2.2 Estimation of transformation parameter λ , and model parameters β

Estimation of λ using the maximum likelihood (ML) method was discussed by Box and Cox (1964). They derived the estimate of λ by calculating and plotting the log-likelihood values for the fitted model against a large set of λ . The MLE, $\hat{\lambda}$, is that value for which log-likelihood is the maximum. Box and Cox (1964) used a standardization so that the magnitude of the error term does not depend on λ :

$$y_i^{*(\lambda)} = \begin{cases} \frac{y_i^\lambda - 1}{\lambda \tilde{y}^{\lambda-1}} & \lambda \neq 0; y_i > 0; i = 1, \dots, n, \\ \tilde{y} \log(y_i) & \lambda = 0 \end{cases} \quad (1.5)$$

where \tilde{y} is the geometric mean of the response values:

$$\tilde{y} = \left(\prod_{i=1}^n y_i \right)^{1/n}.$$

A Bayesian approach to estimating λ was also proposed and compared to the traditional ML method (Box and Cox, 1964).

Hernandez and Johnson (1980) investigated the large sample behavior of the Box-Cox transformation procedure for attaining normality. They presented a theorem stating that both the MLE and the Bayes estimator of λ proposed by Box and Cox (1964)

converge to the estimate of λ that minimizes the Kullback-Leibler information, a measure of distance between two distributions.

Bickel and Doksum (1981) examined the consistency properties of the Box-Cox MLE of both β and λ , as well as the asymptotic variances of these estimates. Their theoretical and Monte Carlo work indicate that the estimates of β , λ , and θ , a vector of nuisance parameters to be estimated in the analysis, are highly variable and highly correlated, a problem similar to that of multicollinearity.

Some researchers (Box and Cox, 1982; Hinkey and Runger, 1984) argued against the conclusions made by Bickel and Doksum in 1981. They stated that the results by Bickel and Doksum are “scientifically irrelevant” and the interpretation of β and their estimates have no meaning independent of a specific value of λ .

A potential solution to the controversy was proposed by Carroll and Ruppert (1984). They introduced the “transform both sides” (TBS) model, in which the response and the model are transformed simultaneously and identically in order to achieve homoscedasticity and normality. The benefits of TBS is that for estimating β , there is little penalty for estimating λ . However, TBS model is not realistic when applied to situations where the true model cannot be assumed to be known. Thus, selecting a linear model to describe an unknown process is the best option. The Box-Cox transformation model, in which only the dependent variable in a linear model is transformed, is admittedly not perfect, but can be considered a viable method in the model selection process.

Carroll and Ruppert (1985, 1987) considered a weighted, modified maximum likelihood estimation (MMLE) method. Unlike the MLE, this method is relatively insensitive to outliers in both the design and the residual.

Gurka (2004, 2006) estimated the model and transformation parameters for the linear mixed model using the residual maximum likelihood (REML) approach in order to obtain more accurate estimate for θ . The benefits of using REML estimation have been well documented (Patterson and Thompson, 1971; Harville, 1977; Kenward and Roger, 1997; Jiang 1996). The advantage of using REML to estimate model parameters, β , and transformation parameter, λ , was obtained through the more accurate approximation of θ and demonstrated via simulation study and real data analysis.

1.2.3 Prediction in the original scale

It is frequently of interest to predict the conditional mean of a future observation. Talyor (1986) proposed an approximated method to estimate the conditional mean of a future observation under the Box-Cox model. The new method was compared to the smearing method (Duan, 1983), a nonparametric method of estimating the conditional mean when the data follow a linear model after a known transformation. The results showed that smearing estimate and the new estimate are approximately equal except when the transformation parameter is near zero. Sakia (1990) applied the Taylor Series technique to estimate the conditional mean of a future value, along with its variance. He noted that bias may not be a serious problem, but the variances can be inflated. Rather than predicting the mean, some researchers have interest in predicting the median of a single future observation (Carroll, 1982; Carroll and Ruppert, 1981; Yang, 2002). Carroll (1982) considered the situation when the choice of power is restricted to a finite set. He

found that the resulting method can be very different from the unrestricted maximum likelihood method. Carrol and Ruppert (1981) predicted the conditional median and mean of a future observation. Their results indicated that when the transformation must be estimated, the prediction error is not much larger than when the parameter is known. The effect of estimating the transformation parameter is small. Yang (2002) constructed confidence intervals for the median of a future observation at certain values of exogenous variables. He proposed a simple analytical correction on the usual prediction interval, obtained through a simple inverse transformation. The corrected interval provided good small-sample properties.

In the same context of predicting a single future observation, Collins (1991) reviewed and compared prediction techniques for Box-Cox regression models, including plug-in, mean squared error analysis, predictive likelihood, and stochastic simulation. These techniques take account of non-normality and parameter uncertainty in varying degrees. The results from a Monte Carlo simulation indicated that stochastic simulation, as usually carried out, leads to badly biased predictions. A modification of the usual approach was proposed and rendered stochastic simulation predictions largely unbiased.

Prediction through Box-Cox transformation has been applied to different areas. Yang (1999) predicted a future lifetime with different lifetime distributions. The study addressed the effect of: 1) non-normality of the transformed observations, and 2) estimating transformation parameter, on the performance of the prediction interval. The results suggested even if observations can not be transformed to achieve exact normality, the Box-Cox procedure still provides a reasonable approximation to prediction intervals and the procedure is robust against misspecification of the parent distribution. Recently,

Dagne (2003) improved the quality of prediction of small area means by incorporating the power transformation into the mixed-effects model. Hwang (2004) used the Box-Cox power transformation to predict temporally correlated longitudinal data. Results indicated that the prediction ability of the model can be significantly improved by employing power transformation.

Instead of predicting the mean and/or median of one single future observation, Smallwood and Blaylock (1986) examined the small-sample properties and forecasting performance of predictor for the mean of multiple future observations via a Monte Carlo simulation. All models were estimated for 100 samples of size 30 and 60. An additional 10 observations were all generated for each sample for use in evaluating the out-of-sample forecasting performance. It is found that both the sign and the magnitude of the transformation parameters influence the precision of the estimators and the forecasting performance.

1.3 Prediction of the finite population total based on linear and log-linear model

Let $U = \{1, \dots, N\}$ be a finite population of N identifiable units, each of which has a value of a dependent variable y associated with it. The population vector of y 's, i.e. $\mathbf{y} = (y_1, \dots, y_N)'$, is treated as a realization of a random vector $\mathbf{Y} = (Y_1, \dots, Y_N)'$. Let S be the set of all samples of size n , a sample s being a subset of U . Our goal is to predict the finite population total: $T = \sum_{i \in U} y_i$. It is assumed that for the finite population, we have information on $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)'$, where $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ik})'$ is a column vector of k known

auxiliary variables for the unit i . For any sample s of size n , redefine \mathbf{y} and \mathbf{X} so that the first n rows of \mathbf{y} and \mathbf{X} correspond to those in the sample. Write

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_s \\ \mathbf{y}_r \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_s \\ \mathbf{X}_r \end{pmatrix},$$

where

\mathbf{y}_s is a $n \times 1$ column vector of observed dependent variable;

\mathbf{y}_r is a $(N - n) \times 1$ column vector of unobserved dependent variable;

\mathbf{X}_s is a $n \times (k + 1)$ matrix of known auxiliary variables in the sample;

\mathbf{X}_r is a $(N - n) \times (k + 1)$ matrix of known auxiliary variables outside the sample.

In a standard prediction approach, any inference on the finite population characteristic of interest is based solely on the assumed superpopulation model. Under this approach, the sample design is important in the sample selection, but this plays no role at the inference stage.

1.3.1 Prediction under a standard linear model

In the prediction approach, a conceptual infinite superpopulation of y -values is assumed. The observations, y_1, y_2, \dots, y_N , are independent realizations from this superpopulation model and inferences are based on repeated sampling from this model. Consider the following linear model:

$$\mathbf{M}_1 : \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \mathbf{I})$, a N -variate probability distribution with the mean vector $\mathbf{0}$ and variance covariance matrix $\sigma^2 \mathbf{I}$, and \mathbf{I} is the $N \times N$ identity matrix. In this equation,

$\boldsymbol{\beta}$ is the $(k + 1) \times 1$ column vector of regression coefficients. Both σ^2 and $\boldsymbol{\beta}$ are unknown superpopulation parameters.

The finite population total T is predicted as

$$\hat{T} = \sum_{i \in s} y_i + \sum_{i \in r} \hat{y}_i, \text{ and } \hat{y}_i = \mathbf{x}_i' \hat{\boldsymbol{\beta}} \quad (1.6)$$

where $\hat{\boldsymbol{\beta}}$ is the least squares estimator of $\boldsymbol{\beta}$, and r represents the set of unobserved units in the finite population.

The prediction variance of \hat{T} under model \mathbf{M}_1 is given by

$$Var(\hat{T} - T) = Var\left(\sum_{i \in r} y_i\right) + Var\left(\sum_{i \in r} \hat{y}_i\right),$$

where Var denotes the variance with respect to model \mathbf{M}_1 . In the prediction approach, the confidence intervals can be produced using the asymptotic normality of the predictor \hat{T} . Valliant *et. al.* (2000) provided the regularity conditions under which the prediction error, $\hat{T} - T$, is asymptotically normal. Any violation of the regularity conditions could affect the efficiency of the confidence intervals. A concise summary about prediction theory in finite population sampling using linear models can be found in Bolfarine and Zacks (1992), Lohr (1999), Valliant *et. al.* (2000), and Chambers and Skinner (2003).

1.3.2 Prediction under a loglinear model

In many applications, especially in business and agricultural surveys, a linear model may not be appropriate for y , but may be appropriate for a strictly monotonic transformation of y . For the data set given in Royall and Cumberland (1981), Chen and Chen (1996) observed that the finite population distribution was severely skewed and that

the log-transformation helped achieving symmetry. In addition, the scatter plot of $\log(y)$ vs. $\log(x)$ showed a better linear relationship than that of y vs. x . The need and the benefit of taking log-transformation were obvious.

In this subsection, the important case where the log-transformation is used on the dependent variable is briefly reviewed. Consider the following model:

$$\mathbf{M}_2 : \log \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta}$ is $(k+1) \times 1$ a vector of regression coefficients, and $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \mathbf{I})$. A transformation on \mathbf{X} often improves the fit. But, since this does not affect the form of the density function of \mathbf{Y} , the notation \mathbf{X} is retained for simplicity.

The prediction of the population total involves prediction of $\log \mathbf{Y}$ for all the nonsample units in the finite population. By simple back-transformation, the population total is given by

$$\hat{T}_A = \sum_{i \in s} y_i + \sum_{i \in r} \hat{y}_i,$$

where $\hat{y}_i = \sum_{i \in r} e^{x_i \hat{\boldsymbol{\beta}}}$, and $\hat{\boldsymbol{\beta}}$ is the least square estimator of $\boldsymbol{\beta}$ under model \mathbf{M}_2 . Chamber and Dorfman (2003) called \hat{T}_A the naïve back-transformation predictor. Under the assumption of normality of $\boldsymbol{\varepsilon}$, the prediction bias of \hat{T}_A is given by (see appendix for details)

$$E(\hat{T}_A - T) = \sum_{i \in r} e^{x_i \boldsymbol{\beta}} \left(e^{\frac{1}{2} x_i \text{Var}(\hat{\boldsymbol{\beta}}) x_i} - e^{\frac{1}{2} \sigma^2} \right).$$

Note that the bias of \hat{T}_A is not necessarily zero. When $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)'$ is known, the best predictor (BP) under model \mathbf{M}_2 is given by

$$\hat{T}^{BP} = \hat{T}^{BP}(\boldsymbol{\theta}) = \sum_{i \in s} y_i + \sum_{i \in r} \hat{y}_i,$$

where $\hat{y}_i = \hat{y}_i^{BP}(\boldsymbol{\theta}) = e^{\mathbf{x}_i' \boldsymbol{\beta} + \sigma^2/2}$. Note that $\hat{y}_i^{BP}(\boldsymbol{\theta})$ is not a random variable. The BP is an unbiased predictor of T under model \mathbf{M}_2 .

In practice, $\boldsymbol{\theta}$ is unknown and needs to be estimated from the data. The parameter $\boldsymbol{\theta}$ can be estimated by

$$\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)',$$

where $\hat{\sigma}^2 = (n - k - 1)^{-1} (\log \mathbf{Y}_s - \mathbf{X}_s \hat{\boldsymbol{\beta}})' (\log \mathbf{Y}_s - \mathbf{X}_s \hat{\boldsymbol{\beta}})$. An empirical best predictor (EBP) can be obtained as:

$$\hat{T}^{EBP} = \hat{T}^{EBP}(\hat{\boldsymbol{\theta}}) = \sum_{i \in s} y_i + \sum_{i \in r} \hat{y}_i \quad \text{and} \quad \hat{y}_i = \hat{y}_i^{EBP}(\hat{\boldsymbol{\theta}}) = e^{\mathbf{x}_i' \hat{\boldsymbol{\beta}} + \hat{\sigma}^2/2}. \quad (1.7)$$

Note that $\hat{y}_i^{EBP}(\hat{\boldsymbol{\theta}})$ is a random variable. Chambers and Dorfman (2003) considered alternate EBP's with least prediction biases under model \mathbf{M}_2 .

The prediction variance of the above EBP is given by:

$$\text{Var}(\hat{T} - T) = \text{Var}\left(\sum_{i \in r} \hat{y}_i\right) + \text{Var}\left(\sum_{i \in r} y_i\right),$$

where $\hat{T} = \hat{T}^{EBP}$, $\text{Var}\left(\sum_{i \in r} y_i\right) = e^{\sigma^2} (e^{\sigma^2} - 1) \sum_{i \in r} e^{2(\mathbf{x}_i' \boldsymbol{\beta})}$, $\text{Var}\left(\sum_{i \in r} \hat{y}_i\right) = \sum_{i \in r} \sum_{j \in r} \text{Cov}(\hat{y}_i, \hat{y}_j)$ with

$$\text{Cov}(\hat{y}_i, \hat{y}_j) \approx \left(\frac{\partial}{\partial \hat{\boldsymbol{\theta}}} \hat{y}_i \right)'_{\hat{\boldsymbol{\theta}}} \text{Var}(\hat{\boldsymbol{\theta}}) \left(\frac{\partial}{\partial \hat{\boldsymbol{\theta}}} \hat{y}_j \right)_{\hat{\boldsymbol{\theta}}} \quad (\text{see appendix for details}).$$

1.4 Robust model-based predictor of the population total

The transformation-based predictor described in Section 1.3 requires a subjective specification of the transformation to be applied on the dependent variable. This may be

okay in some problems where the transformation to be used is known either from some prior empirical evidence or from some theory. In absence of any prior knowledge about the transformation to be used, an appropriate family of transformations needs to be determined by the data. Thus, our approach is essentially a transformation-based adaptive technique that lets the data decide on the transformation.

Tukey (1957) considered the following family of power transformations:

$$y^{(\lambda)} = \begin{cases} y^\lambda & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases},$$

where $y > 0$. In order to take care of the discontinuity at $\lambda = 0$, Box and Cox (1964) proposed the following family of transformations:

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1) / \lambda & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases}, \quad (1.8)$$

where $y > 0$. The parameter λ determines the nature of transformation. For example, $\lambda = 1, 0, 0.5, -1$ correspond to no transformation, log-transformation, square root transformation, and reciprocal transformation, respectively. The transformation parameter λ is estimated by the data. The Box-Cox analysis may lead to a log-transformation, but may equally lead to some other transformation in the above family – it depends on the actual data observed.

Consider the following superpopulation model for the transformed dependent variable:

$$\mathbf{M}_3 : \mathbf{Y}^{(\lambda)} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$.

Note that the Box-Cox transformation (1.8) requires that the dependent variable must be positive. Also, both the magnitude and the sign of λ affect the range of the dependent variable. When $\lambda > 0$, $y_i^{(\lambda)} = \frac{y_i^\lambda - 1}{\lambda} > \frac{0 - 1}{\lambda} = \frac{-1}{\lambda}$; When $\lambda < 0$, $y_i^{(\lambda)} = \frac{y_i^\lambda - 1}{\lambda} < \frac{0 - 1}{\lambda} = \frac{-1}{\lambda}$ since $y_i > 0$. Thus $y_i^{(\lambda)}$ is bounded from above or below depending on the sign of λ . Hence only approximate normality of ε can be assumed. If the distribution of $y_i^{(\lambda)}$ is truncated normal, parameter estimation may be seriously affected. However, as Zarembka (1974) points out, “if probability of such large negative values [of ε_i] is quite low, the error term may still approximately be normal.” Researchers considered different modifications of the original Box-Cox model in order to allow negative values of the dependent variable. See Box and Cox (1964), Manly (1976), John and Draper (1980), Bickel and Doksum (1981).

1.4.1 Estimation of $\theta = (\beta, \lambda, \sigma)'$

For computational advantages, Box and Cox (1964) suggested the following scaled transformation:

$$y^{*(\lambda)} = \begin{cases} (y^\lambda - 1) / \lambda \tilde{y}^{\lambda-1} & \lambda \neq 0 \\ \tilde{y} \cdot \log(y) & \lambda = 0 \end{cases}, \quad (1.9)$$

where $\tilde{y} = \left(\prod_{i=1}^n y_i \right)^{1/n}$, the geometric sample mean of the sample observations. Consider the following scaled model:

$$\mathbf{M}_4 : \mathbf{Y}^{*(\lambda)} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}^*,$$

where $\boldsymbol{\varepsilon}^*$ is approximately $N(0, \sigma^{*2} \mathbf{I})$. The scaling avoids large numbers and simplifies the log-likelihood function, and thus the model \mathbf{M}_4 has computational advantages over \mathbf{M}_3 in estimating the respective parameters. Note that this scaling is different from the one used by Zarembka (1968, note 8) who suggested dividing $y^{(\lambda)}$ by \tilde{y}^λ . Schlesselman (1971) showed the maximum likelihood estimator of $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^*, \lambda, \sigma^*)'$ is scale invariant so that rescaling the original observations y 's leads to the same log-likelihood function under model \mathbf{M}_4 so long as the regression model contains an intercept term.

Box-Cox (1964) discussed the estimation of $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^*, \lambda, \sigma^*)'$. The density function of y_i based on model \mathbf{M}_4 is:

$$f(y_i, \boldsymbol{\theta}^*) = (2\pi\sigma_e^{*2})^{-1/2} \exp\left\{-\frac{1}{2\sigma_e^{*2}}(y_i^{*(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}^*)^2\right\} \cdot (y_i / \tilde{y})^{\lambda-1}.$$

The corresponding log-likelihood function is:

$$L(\boldsymbol{\theta}^*) = (2\pi\sigma_e^{*2})^{-n/2} \exp\left\{-\frac{1}{2\sigma_e^{*2}} \sum_s (y_i^{*(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}^*)^2\right\}.$$

To estimate $\boldsymbol{\theta}^*$, the method requires the maximization of the approximate log-likelihood function, given by

$$l(\boldsymbol{\theta}^*) = \log L(\boldsymbol{\theta}^*) = -\frac{1}{2} \sum_s \log(2\pi\sigma_e^{*2}) - \frac{1}{2\sigma_e^{*2}} \sum_s (y_i^{*(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}^*)^2.$$

The above log-likelihood function is approximate because the distribution of the error term in model \mathbf{M}_4 is not exactly normal. The maximum likelihood estimate of λ can be obtained by a grid search method. That is, for a large set of values of λ , model \mathbf{M}_4 can be fit. This is a simple task since model \mathbf{M}_4 is a linear model for a given λ . The computations and plotting of the log-likelihood values for the fitted model against the set

of values for λ locate the maximum likelihood estimate, $\hat{\lambda}$, of the transformation parameter λ . The maximum likelihood estimates of $\boldsymbol{\beta}^*$ and σ^{*2} are then given by:

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}_s' \mathbf{X}_s)^{-1} \mathbf{X}_s' \mathbf{Y}_s^{*(\hat{\lambda})} \text{ and}$$

$$\hat{\sigma}^{*2} = \frac{1}{n} (\mathbf{Y}_s^{*(\hat{\lambda})} - \mathbf{X}_s' \hat{\boldsymbol{\beta}}^*)' (\mathbf{Y}_s^{*(\hat{\lambda})} - \mathbf{X}_s' \hat{\boldsymbol{\beta}}^*).$$

Using $\hat{\boldsymbol{\theta}}^* = (\hat{\boldsymbol{\beta}}^*, \hat{\lambda}, \hat{\sigma}^*)'$, the maximum likelihood estimator of $\boldsymbol{\theta}$ under model \mathbf{M}_3 is obtained as: $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\lambda}, \hat{\sigma})' = (\tilde{y}^{\hat{\lambda}-1} \hat{\boldsymbol{\beta}}^*, \hat{\lambda}, \tilde{y}^{\hat{\lambda}-1} \hat{\sigma}^*)'$. In the appendix, it is proved that the maximum likelihood estimates of λ with respect to \mathbf{M}_3 and \mathbf{M}_4 are equivalent. The above algorithm, originally proposed by Box and Cox (1964), supplements the four different algorithms of obtaining the maximum likelihood estimator of $\boldsymbol{\theta}$ under model \mathbf{M}_3 considered in Spitzer (1982a, b). Bickel and Doksum (1981) provided precise conditions under which the maximum likelihood estimator of $\boldsymbol{\theta}$ is consistent.

1.4.2 Estimating the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\theta}}$

Note that for known λ , \mathbf{M}_4 is simply the standard linear regression model and so one can suggest standard variance estimators for $\hat{\boldsymbol{\beta}}^*$ and $\hat{\sigma}^{*2}$. However, these variance estimators would underestimate the true uncertainties of $\hat{\boldsymbol{\beta}}^*$ and $\hat{\sigma}^{*2}$ since such variance estimators treat $\hat{\lambda}$ as the true value. See Bickel and Doksum (1981) and Hinkley and Runger (1984). In the context of estimation of $(\boldsymbol{\beta}, \lambda)'$, Spitzer (1982a) incorporated the additional uncertainty due to estimation of λ by considering the inverse of the observed Fisher's information matrix. However, his model is different from the Box-Cox scaled model. Unlike Spitzer (1982a), this study includes additional uncertainties involving $\hat{\sigma}^{*2}$

for inference. A standard consistent estimator of the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\theta}}^*$ is given by:

$$\text{var}(\hat{\boldsymbol{\theta}}^*) = i(\hat{\boldsymbol{\theta}}^*)^{-1},$$

where $i(\hat{\boldsymbol{\theta}}^*)$ is the observed Fisher information matrix, i.e. the negative of the matrix of second partial derivatives of $l(\boldsymbol{\theta}^*)$ with respect to $\boldsymbol{\theta}^*$, evaluated at the MLE, $\hat{\boldsymbol{\theta}}^*$:

$$\begin{aligned} i(\hat{\boldsymbol{\theta}}^*) &= -\left(\partial^2 l(\boldsymbol{\theta}^*) / \partial(\boldsymbol{\theta}^*)^2\right)_{\hat{\boldsymbol{\theta}}^*} \\ &= \begin{bmatrix} i_{11} & i_{12} & i_{13} \\ i_{12}' & i_{22} & i_{23} \\ i_{13}' & i_{23}' & i_{33} \end{bmatrix}_{\hat{\boldsymbol{\theta}}^*}, \end{aligned}$$

and

$$\begin{aligned} i_{11} &= -\partial^2 l(\boldsymbol{\theta}^*) / \partial \boldsymbol{\beta}^* \partial \boldsymbol{\beta}^{*'} = \frac{1}{\sigma^{*2}} \sum_s \mathbf{x}_i \mathbf{x}_i', \\ i_{12} &= -\partial^2 l(\boldsymbol{\theta}^*) / \partial \boldsymbol{\beta}^* \partial \lambda = -\frac{1}{\sigma^{*2}} \sum_s \left(\frac{\partial}{\partial \lambda} y_i^{*(\lambda)} \right) \mathbf{x}_i, \\ i_{13} &= -\partial^2 l(\boldsymbol{\theta}^*) / \partial \boldsymbol{\beta}^* \partial \sigma^* = \frac{2}{\sigma^{*3}} \sum_s (y_i^{*(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}^*) \mathbf{x}_i, \\ i_{22} &= -\partial^2 l(\boldsymbol{\theta}^*) / \partial \lambda^2 = \frac{1}{\sigma^{*2}} \sum_s \left[(y_i^{*(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}^*) \left(\frac{\partial^2}{\partial \lambda^2} y_i^{*(\lambda)} \right) + \left(\frac{\partial}{\partial \lambda} y_i^{*(\lambda)} \right)^2 \right], \\ i_{23} &= -\partial^2 l(\boldsymbol{\theta}^*) / \partial \lambda \partial \sigma^* = -\frac{2}{\sigma^{*3}} \sum_s (y_i^{*(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}^*) \left(\frac{\partial}{\partial \lambda} y_i^{*(\lambda)} \right), \\ i_{33} &= -\partial^2 l(\boldsymbol{\theta}^*) / \partial (\sigma^*)^2 = -n \frac{1}{\sigma^{*2}} + \frac{3}{\sigma^{*4}} \sum_s (y_i^{*(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}^*)^2, \end{aligned}$$

and

$$\begin{aligned} \frac{\partial y_i^{*(\lambda)}}{\partial \lambda} &= \lambda^{-1} \tilde{y}^{1-\lambda} s_i^{(\lambda)} - [\lambda^{-1} + \log(\tilde{y})] y_i^{*(\lambda)}, \\ \frac{\partial^2 y_i^{*(\lambda)}}{\partial \lambda^2} &= \lambda^{-1} \tilde{y}^{1-\lambda} [t_i^{(\lambda)} - 2\{\lambda^{-1} + \log(\tilde{y})\} s_i^{(\lambda)}] + [\{\lambda^{-1} + \log(\tilde{y})\}^2 + \lambda^{-2}] y_i^{*(\lambda)} \end{aligned}$$

$\mathbf{s}^{(\lambda)}$ and $\mathbf{t}^{(\lambda)}$ are $n \times 1$ column vectors such that the i^{th} element is $s_i^{(\lambda)} = y_i^\lambda \log y_i$ and $t_i^{(\lambda)} = y_i^\lambda (\log y_i)^2$, respectively.

The goal of this study is to obtain an estimate of the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\lambda}, \hat{\sigma})'$. Using the relationship $\hat{\boldsymbol{\beta}} = \tilde{y}^{\hat{\lambda}-1} \hat{\boldsymbol{\beta}}^*$, $\hat{\sigma} = \tilde{y}^{\hat{\lambda}-1} \hat{\sigma}^*$ and the fact that $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \lambda, \sigma)'$ under model \mathbf{M}_3 , we have

$$\text{var}(\hat{\boldsymbol{\theta}}) = \mathbf{J} \text{var}(\hat{\boldsymbol{\theta}}^*) \mathbf{J}',$$

where

$$\mathbf{J} = \begin{bmatrix} \frac{\partial}{\partial \hat{\boldsymbol{\beta}}^*} \hat{\boldsymbol{\beta}} & \frac{\partial}{\partial \hat{\lambda}} \hat{\boldsymbol{\beta}} & \frac{\partial}{\partial \hat{\sigma}^*} \hat{\boldsymbol{\beta}} \\ \frac{\partial}{\partial \hat{\boldsymbol{\beta}}^*} \hat{\lambda} & \frac{\partial}{\partial \hat{\lambda}} \hat{\lambda} & \frac{\partial}{\partial \hat{\sigma}^*} \hat{\lambda} \\ \frac{\partial}{\partial \hat{\boldsymbol{\beta}}^*} \hat{\sigma} & \frac{\partial}{\partial \hat{\lambda}} \hat{\sigma} & \frac{\partial}{\partial \hat{\sigma}^*} \hat{\sigma} \end{bmatrix} = \begin{bmatrix} \tilde{y}^{\hat{\lambda}-1} \mathbf{I} & \log(\tilde{y}) \hat{\boldsymbol{\beta}} & \mathbf{0} \\ \mathbf{0}^T & 1 & 0 \\ \mathbf{0}^T & \hat{\sigma} \log(\tilde{y}) & \tilde{y}^{\hat{\lambda}-1} \end{bmatrix}.$$

Following Spitzer (1982a), one could have applied the Taylor Series to obtain $\text{var}(\hat{\boldsymbol{\theta}})$ from $\text{var}(\hat{\boldsymbol{\theta}}^*)$. However, it should be noted that such an argument is hard to justify since \mathbf{J} is a random matrix. Instead, a direct method to obtain $\text{var}(\hat{\boldsymbol{\theta}})$ is applied (see the appendix for details).

1.4.3 Prediction of the finite population total

The finite population total is predicted by

$$\hat{T} = \sum_{i \in s} y_i + \sum_{i \in r} \hat{y}_i,$$

where \hat{y}_i denotes a predicted value of the unobserved y_i . Define $\hat{T} - T$ the prediction error of the predictor \hat{T} . The bias, variance and mean squared error of the prediction error are defined as follows:

$$\text{Prediction bias: } B(\hat{T} - T) = E(\hat{T} - T),$$

$$\text{Prediction variance: } \text{Var}(\hat{T} - T) = E(\hat{T} - E(T))^2,$$

$$\text{Prediction mean squared error (MSE): } \text{MSE}(\hat{T} - T) = E(\hat{T} - T)^2,$$

where all the expectations above are taken with respect to model \mathbf{M}_3 . Note that $\text{MSE}(\hat{T} - T) = \text{Var}(\hat{T} - T) + B^2(\hat{T} - T)$. For an unbiased predictor, i.e., for a predictor with $B(\hat{T} - T) = 0$, $\text{MSE}(\hat{T} - T) = \text{Var}(\hat{T} - T)$.

The best predictor (BP) of T , i.e., the predictor which minimizes the prediction MSE, is obtained when $\hat{y}_i = \hat{y}_i^{BP}(\boldsymbol{\theta})$, where

$$\hat{y}_i^{BP}(\boldsymbol{\theta}) = E(y_i) = \int_{-\infty}^{\infty} [\lambda(\mathbf{x}_i' \boldsymbol{\beta} + \sigma z) + 1]^{\frac{1}{\lambda}} \phi(z) dz,$$

and $\phi(z)$ is the density of the standard normal deviate. The above integral can be evaluated by numerical integration or the following Monte Carlo approximation:

$$\hat{y}_i^{BP}(\boldsymbol{\theta}) \approx \frac{1}{M} \sum_{j=1}^M [\lambda(\mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_{ij}) + 1]^{\frac{1}{\lambda}}, \quad (1.10)$$

where M denotes the number of independent simulation runs and $\varepsilon_{ij} \sim N(0, \sigma^2)$.

In practice, $\boldsymbol{\theta}$ is unknown. Replacing $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$ in $\hat{y}_i^{BP}(\boldsymbol{\theta})$, an empirical best predictor (EBP) of T is obtained. In EBP, $\hat{y}_i = \hat{y}_i^{EBP}(\hat{\boldsymbol{\theta}})$, where

$$\hat{y}_i^{EBP}(\hat{\boldsymbol{\theta}}) = \int_{-\infty}^{\infty} [\hat{\lambda}(\mathbf{x}_i' \hat{\boldsymbol{\beta}} + \hat{\sigma} z) + 1]^{\frac{1}{\hat{\lambda}}} \phi(z) dz \approx \frac{1}{M} \sum_{j=1}^M [\hat{\lambda}(\mathbf{x}_i' \hat{\boldsymbol{\beta}} + \hat{\varepsilon}_{ij}) + 1]^{\frac{1}{\hat{\lambda}}} \quad (1.11)$$

and $\hat{\varepsilon}_{ij} \sim iid N(0, \hat{\sigma}^2)$.

When σ^2 is small, $\hat{y}_i^{BP}(\boldsymbol{\theta})$ is approximated by the Taylor Series expansion and the approximate best predictor (ABP) of T is obtained when $\hat{y}_i = \hat{y}_i^{ABP}(\boldsymbol{\theta})$ with

$$\hat{y}_i^{ABP}(\boldsymbol{\theta}) = [\lambda(\mathbf{x}_i' \boldsymbol{\beta}) + 1]^{1/\lambda} \approx \hat{y}_i^{BP}(\boldsymbol{\theta}) \quad (1.12)$$

Replacing $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$ in $\hat{y}_i^{ABP}(\boldsymbol{\theta})$, the following approximated empirical best predictor (AEBP) of T when $\hat{y}_i = \hat{y}_i^{AEBP}(\hat{\boldsymbol{\theta}})$ is obtained, where

$$\hat{y}_i^{AEBP}(\hat{\boldsymbol{\theta}}) = [\hat{\lambda}(\mathbf{x}_i' \hat{\boldsymbol{\beta}}) + 1]^{1/\hat{\lambda}}. \quad (1.13)$$

This approach is easy to implement in terms of CPU time.

1.4.4 Estimation of the prediction variance of the population total predictor

First note that for each $i \in r$, any arbitrary predictor \hat{y}_i is a function of y_i , $i \in s$ and hence independent of all y_i , $i \in r$ under model \mathbf{M}_3 . Thus, using the fact that

$$\hat{T} - T = \sum_{i \in r} \hat{y}_i - \sum_{i \in r} y_i,$$

$$Var(\hat{T} - T) = Var\left(\sum_{i \in r} y_i\right) + Var\left(\sum_{i \in r} \hat{y}_i\right). \quad (1.14)$$

If $\hat{y}_i = \hat{y}_i^{BP}(\boldsymbol{\theta})$ or $\hat{y}_i^{ABP}(\boldsymbol{\theta})$, the second term of right side of (1.14) is zero since \hat{y}_i is non-stochastic. Thus, for both BP and ABP prediction variances are identical given by

$$Var(\hat{T} - T) = \sum_{i \in r} Var(y_i),$$

where $\sum_{i \in r} \text{Var}(y_i) = \sum_{i \in r} \left[\int_{-\infty}^{\infty} [\lambda(\mathbf{x}'_i \boldsymbol{\beta} + \sigma z) + 1]^{\frac{2}{\lambda}} \phi(z) dz - E^2(y_i) \right]$.

The BP and ABP differ in terms of their prediction biases, and accordingly the prediction MSE. Evidently, the prediction bias of the BP is zero and thus for the BP the prediction MSE is same as the prediction variance. On the other hand, ABP suffers from prediction bias, but as noted in Subsection 1.4.3 the bias is negligible for small σ^2 .

If $\hat{y}_i = \hat{y}_i^{EBP}(\hat{\boldsymbol{\theta}})$ or $\hat{y}_i^{AEBP}(\hat{\boldsymbol{\theta}})$, the prediction variance of \hat{T} is given by

$$\text{Var}(\hat{T} - T) = \text{Var}\left(\sum_{i \in r} y_i\right) + \text{Var}\left(\sum_{i \in r} \hat{y}_i(\hat{\boldsymbol{\theta}})\right), \quad (1.15)$$

where

$$\text{Var}\left(\sum_{i \in r} \hat{y}_i(\hat{\boldsymbol{\theta}})\right) = \sum_{i \in r} \sum_{j \in r} \text{Cov}(\hat{y}_i(\hat{\boldsymbol{\theta}}), \hat{y}_j(\hat{\boldsymbol{\theta}})).$$

Note the second term of (1.15), $\text{Var}\left(\sum_{i \in r} \hat{y}_i(\hat{\boldsymbol{\theta}})\right)$, captures the variability due to the estimation of $\hat{\boldsymbol{\theta}}$, which approaches to zero as sample size $n \rightarrow \infty$. Using the Taylor Series expansion argument, the following variance estimator is proposed:

$$\text{var}\left(\sum_{i \in r} \hat{y}_i(\hat{\boldsymbol{\theta}})\right) \approx \left(\sum_{i \in r} \frac{\partial}{\partial \hat{\boldsymbol{\theta}}} \hat{y}_i(\hat{\boldsymbol{\theta}}) \right)' \text{var}(\hat{\boldsymbol{\theta}}) \left(\sum_{j \in r} \frac{\partial}{\partial \hat{\boldsymbol{\theta}}} \hat{y}_j(\hat{\boldsymbol{\theta}}) \right).$$

For $\hat{y}_i(\hat{\boldsymbol{\theta}}) = \hat{y}_i^{AEBP}(\hat{\boldsymbol{\theta}})$,

$$\frac{\partial}{\partial \hat{\lambda}} \hat{y}_i^{AEBP}(\hat{\boldsymbol{\theta}}) = \hat{\lambda}^{-1}(\mathbf{x}'_i \hat{\boldsymbol{\beta}}) [\hat{\lambda}(\mathbf{x}'_i \hat{\boldsymbol{\beta}}) + 1]^{1/\hat{\lambda}-1} - \hat{\lambda}^{-2} [\hat{\lambda}(\mathbf{x}'_i \hat{\boldsymbol{\beta}}) + 1]^{1/\hat{\lambda}} \log[\hat{\lambda}(\mathbf{x}'_i \hat{\boldsymbol{\beta}}) + 1],$$

and

$$\frac{\partial}{\partial \hat{\boldsymbol{\beta}}} \hat{y}_i^{AEBP}(\hat{\boldsymbol{\theta}}) = [\hat{\lambda}(\mathbf{x}'_i \hat{\boldsymbol{\beta}}) + 1]^{1/\hat{\lambda}-1} \mathbf{x}_i.$$

For $\hat{y}_i(\hat{\boldsymbol{\theta}}) = \hat{y}_i^{EBP}(\hat{\boldsymbol{\theta}})$,

$$\frac{\partial}{\partial \hat{\lambda}} \hat{y}_i^{EBP}(\hat{\boldsymbol{\theta}}) = \int_{-\infty}^{\infty} \left[\hat{\lambda}^{-2} \hat{w}_i^{1/\hat{\lambda}} (1 - \hat{w}_i^{-1} - \log \hat{w}_i) \right] \phi(z) dz,$$

$$\frac{\partial}{\partial \hat{\boldsymbol{\beta}}} \hat{y}_i^{EBP}(\hat{\boldsymbol{\theta}}) = \int_{-\infty}^{\infty} \hat{w}_i^{1/\hat{\lambda}-1} \mathbf{x}_i \phi(z) dz,$$

$$\frac{\partial}{\partial \hat{\sigma}} \hat{y}_i^{EBP}(\hat{\boldsymbol{\theta}}) = \int_{-\infty}^{\infty} \hat{w}_i^{1/\hat{\lambda}-1} z \phi(z) dz,$$

where $\hat{w}_i = \hat{\lambda}(\mathbf{x}_i' \hat{\boldsymbol{\beta}} + \hat{\sigma} z) + 1$.

1.4.5 Confidence interval of the population total predictor

In this subsection, the construction of the prediction interval of T based on the asymptotic distribution of \hat{T}^{EBP} is illustrated. Consider an asymptotic set-up when N and $n \rightarrow \infty$, $f \rightarrow 0$ – set-up common in finite population sampling (Valliant, *et. al.* 2000).

The following theorem is needed:

Theorem: Assume model \mathbf{M}_3 and the following regularity conditions:

(i) $\hat{y}_i^{BP}(\boldsymbol{\theta})$ is a smooth function of $\boldsymbol{\theta}$ in the sense that it permits bounded continuous first two derivatives with respect to the components of $\boldsymbol{\theta}$.

(ii) $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(0, \text{Var}(\hat{\boldsymbol{\theta}}))$ and $\text{Var}(\hat{\boldsymbol{\theta}}) = O\left(\frac{1}{n}\right)$.

Then, as N and $n \rightarrow \infty$, $f \rightarrow 0$, we have

$$\frac{\hat{T}^{EBP} - T}{\sqrt{V(\hat{T}^{EBP} - T)}} \xrightarrow{d} N(0,1).$$

Proof:

First note that

$$\frac{\hat{T}^{EBP} - T}{\sqrt{V(\hat{T}^{EBP} - T)}} = \frac{\hat{T}^{BP} - T}{\sqrt{V(\hat{T}^{EBP} - T)}} + \frac{\hat{T}^{EBP} - \hat{T}^{BP}}{\sqrt{V(\hat{T}^{EBP} - \hat{T}^{BP})}} \frac{\sqrt{V(\hat{T}^{EBP} - \hat{T}^{BP})}}{\sqrt{V(\hat{T}^{EBP} - T)}},$$

where

$$\hat{T}^{BP} - T = \sum_{i \in r} \hat{y}_i^{BP}(\boldsymbol{\theta}) - \sum_{i \in r} y_i,$$

$$\hat{T}^{EBP} - \hat{T}^{BP} = \sum_{i \in r} \hat{y}_i^{EBP}(\hat{\boldsymbol{\theta}}) - \sum_{i \in r} \hat{y}_i^{BP}(\boldsymbol{\theta}), \text{ and}$$

$\hat{T}^{EBP} - \hat{T}^{BP}$ and $\hat{T}^{BP} - T$ are independent.

Define $V(\hat{T}^{EBP} - T) = V(\boldsymbol{\theta})$ and $V(\boldsymbol{\theta}) = V_1(\boldsymbol{\theta}) + V_2(\boldsymbol{\theta})$,

where

$$V_1(\boldsymbol{\theta}) = V(\hat{T}^{BP} - T) = \sum_{i \in r} \text{Var}(y_i) \text{ and}$$

$$V_2(\boldsymbol{\theta}) = V(\hat{T}^{EBP} - \hat{T}^{BP}) \approx \left(\sum_{i \in r} \frac{\partial}{\partial \hat{\boldsymbol{\theta}}} \hat{y}_i(\hat{\boldsymbol{\theta}}) \right)'_0 \text{Var}(\hat{\boldsymbol{\theta}}) \left(\sum_{j \in r} \frac{\partial}{\partial \hat{\boldsymbol{\theta}}} \hat{y}_j(\hat{\boldsymbol{\theta}}) \right)_0.$$

Under the model assumptions, $V_1(\boldsymbol{\theta}) = O(N - n)$. Under the model assumptions

and the regularity conditions, we have $V_2(\boldsymbol{\theta}) = O\left(\frac{(N - n)^2}{n}\right)$. Thus, we have

$$V(\boldsymbol{\theta}) = O(N - n) + O\left(\frac{(N - n)^2}{n}\right) = O\left(\frac{(N - n)^2}{n}\right).$$

Therefore,

$$\frac{\sqrt{V_2(\boldsymbol{\theta})}}{\sqrt{V(\boldsymbol{\theta})}} = \frac{\sqrt{V(\hat{T}^{EBP} - \hat{T}^{BP})}}{\sqrt{V(\hat{T}^{EBP} - T)}} \xrightarrow{p} 1.$$

Now using Taylor Series expansion we get

$$\hat{T}^{EBP} - \hat{T}^{BP} = \sum_{i \in r} (\hat{y}_i^{EBP} - \hat{y}_i^{BP}) \approx \sum_{i \in r} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})' \frac{\partial \hat{y}_i^{BP}}{\partial \boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})' \sum_{i \in r} \frac{\partial \hat{y}_i^{BP}}{\partial \boldsymbol{\theta}}. \quad \text{Thus, under the}$$

regularity conditions, $\hat{T}^{EBP} - \hat{T}^{BP} \xrightarrow{d} N(0, V_2(\boldsymbol{\theta}))$. Therefore, the second term

$$\frac{\hat{T}^{EBP} - \hat{T}^{BP}}{\sqrt{V(\hat{T}^{EBP} - \hat{T}^{BP})}} \frac{\sqrt{V(\hat{T}^{EBP} - \hat{T}^{BP})}}{\sqrt{V(\hat{T}^{EBP} - T)}} \xrightarrow{d} N(0,1).$$

Turning to the first term, we have

$$E\left(\frac{\hat{T}^{BP} - T}{\sqrt{V(\boldsymbol{\theta})}}\right) = 0, \text{ and } V\left(\frac{\hat{T}^{BP} - T}{\sqrt{V(\boldsymbol{\theta})}}\right) = \frac{V_1(\boldsymbol{\theta})}{V(\boldsymbol{\theta})} = \frac{O(N-n)}{O\left(\frac{(N-n)^2}{n}\right)} = O\left(\frac{n}{N-n}\right) \xrightarrow{p} 0.$$

$$\text{Hence, } \frac{\hat{T}^{BP} - T}{\sqrt{V(\hat{T}^{EBP} - T)}} \xrightarrow{p} 0.$$

The theorem then follows from an application of the Slutsky's theorem.

The above theorem suggests the following $100(1-\alpha)\%$ prediction interval for T based on the EBP: $\hat{T}^{EBP} \pm z_{\alpha/2} \sqrt{\text{var}(\hat{T}^{EBP} - T)}$, where $\text{var}(\hat{T}^{EBP} - T)$ is a consistent estimator of $\text{Var}(\hat{T}^{EBP} - T)$ and $z_{\alpha/2}$ is the upper $100\frac{\alpha}{2}\%$ point of the normal deviate.

1.5 Simulation study

The purpose of this simulation study is to evaluate different predictors of a finite population total for different values of the sample size and the error standard deviation σ of model \mathbf{M}_3 . In this simulation exercise, consider a finite population of size $N=431$ that corresponds to the sample size of the Australian Agricultural and Grazing Industries Survey (AAGIS). This survey data contains information on the number of cattle (y) and

farm area (x) for each of the 431 farms. This simulation is based on repeated generation of the finite population from the following model:

$$y_i^{(\lambda)} = \frac{y_i^\lambda - 1}{\lambda} = \beta_0 + \beta_1 \times \log(x_i) + \varepsilon_i, \quad (1.16)$$

where ε_i are *iid* from $N(0, \sigma^2)$. In order to mimic a true situation, we choose $\lambda=0.1$, $\beta_0=4.20$, and $\beta_1=2.66$ which are the estimates obtained by fitting the real data to model **M₃**. Strictly speaking, truncated normal distribution of y is considered, and all negative values of y generated are discarded. The effect of this is negligible since less than 0.1% of the generated y values are discarded. The same phenomena were found by Talyor (1986).

A random sample of size n is first drawn from a finite population of size $N = 431$. The units and the associated x -values in the sample are not changed in the simulation experiment. Then $R=1,000$ finite populations are generated using the model (1.16). While the units and the associated x -values are unchanged, $R=1,000$ sets of sample y -values, each of size n , is obtained.

We investigate the performance of six predictors of population total $T = \sum_{i=1}^N y_i$:

$$\hat{T} = \sum_{i \in s} y_i + \sum_{i \in r} \hat{y}_i,$$

where \hat{y}_i is an arbitrary predictor among (1.6), (1.7), (1.10)-(1.13). Six different \hat{T} 's are denoted by no-transformation predictor (NTP), log-transformation predictor (LTP), BP, EBP, ABP, and AEBP, respectively.

The six predictors are evaluated in terms of the relative bias and MSE criteria, approximated by the Monte Carlo simulation method as follows:

$$E_M \left(\frac{\hat{T} - T}{T} \right) \approx \frac{1}{R} \sum_{r=1}^R \frac{\hat{T}_r - T_r}{T_r}, \text{ and}$$

$$E_M (\hat{T} - T)^2 \approx \frac{1}{R} \sum_{r=1}^R (\hat{T}_r - T_r)^2,$$

respectively. The BP is the best predictor and thus the remaining five are compared using the percent relative loss (PRL) in terms of MSE given by:

$$PRL_{est} = \frac{MSE_{est} - MSE_{BP}}{MSE_{BP}} \times 100\%,$$

where subscript *est* denotes any of the five predictors.

Table 1.1 displays the MSE of various predictors for different sample sizes (n) and model standard deviation (σ). We report *PRL* in Table 1.3. As we expected, the BP performs the best for all n and σ . As σ increases, the performance of ABP compared to the BP worsens. When $\sigma = 2$, ABP performs much worse than the BP even for a large sample size – this is consistent with our theory. It is interesting to note that although in general EBP performs better than AEBP, the difference in the performance is not that prominent as that between the BP and ABP. The NTP and LTP perform worse than the EBP (and AEBP when σ is small). In our experiment $\lambda = .1$ (very close 0 which corresponds to the log-transformation) and yet LTP performs worse than NTP. In practice the BP or ABP can not be produced since the true values of model parameters are unknown. The EBP emerges as the best among AEBP, NTP and LTP. AEBP is a sensible alternative predictor to EBP when σ is small, which can simplify the computation substantially.

Table 1.2 displays the relative bias of various predictors for different sample sizes (n) and model standard deviation (σ). AEBP and/or ABP tend to provide

underestimation of the finite population total, and they perform much worse than EBP and/or BP when σ is large. The difference in the performance between AEBP and ABP is not as prominent as that between the EBP and BP. It implies that a better estimator of σ should be considered in the future research.

1.6 Real data analysis

In this section, the actual survey data from the AAGIS is treated as an artificial finite population of $N=431$ farms. For each farm, the information on the number of beef cattle (dependent variable, y) and the farm area (auxiliary variable, x) is available. In Figure 1.1, the histogram of y and $\log(y)$ is plotted. It is clear that the distribution of y is highly skewed and the log-transformation is useful in achieving nearly normal distribution. In Figure 1.2, the scatter plots of y (or $\log y$) vs. x (or $\log x$) is displayed. The log-transformation is exhibiting a better linear fit. The adjusted R^2 for the log-transformed data is .74 compared to .45 for the original data.

The benefit of taking log-transformation is obvious, but the question is whether the superpopulation can be described by a better model. To this end, the Box-Cox model \mathbf{M}_3 with the log-transformation on the auxiliary variable x is considered. Note that no transformation and the log-transformation belong to the class of Box-Cox transformations when $\lambda = 1$ and 0 respectively. Table 1.4 reports the estimates and the corresponding 95% confidence intervals (*CI*) for model parameters β 's and transformation parameter λ . Note that $\hat{\lambda}$ is 0.1, and the 95% asymptotic *CI* for λ is (0.05, 0.16) which does not cover both $\lambda = 1$ and 0. Figure 1.3 plots the histogram of $\hat{\lambda}$'s, estimated using 1,000 bootstrap samples, each bootstrap sample (of size 431) being selected by a simple random sampling

with replacement from the finite population. The histogram is nearly symmetric and the 95% bootstrap *CI* for λ is (0.04, 0.17). Again, $\lambda = 1$ and 0 are both excluded. This implies that both the untransformed and the log-transformed data may not adequately describe the finite population.

Figure 1.4 displays the scatter plot of $y^{(\hat{\lambda})} = \frac{y^{\hat{\lambda}} - 1}{\hat{\lambda}}$ vs. $\log(x)$ and it is similar to that for $\log(y)$ vs. $\log(x)$ in Figure 1.3, and the adjusted R^2 is just a little larger (0.75 vs. 0.74). Thus, the Box-Cox transformation does not appear to perform better than the log-transformation in describing the finite population. But, our purpose is the prediction of the finite population total based on a sample from this finite population, or equivalently, the prediction of the unobserved part of the finite population. Next we study the predictive power of different models by the well-known cross-validation method in which we drop one unit at a time and using the remaining units we predict the unit deleted.

We compare three predictors (NTP, LTP, and EBP) of the total number of beef cattle in $N = 431$ farms based on the three different models: no transformation model \mathbf{M}_1 , log-transformation model \mathbf{M}_2 , and the Box-Cox transformation model \mathbf{M}_3 . For a fair comparison, a log-transformation is taken on x for both transformation models: \mathbf{M}_2 and \mathbf{M}_3 . The cross-validation sample can be viewed as a simple random sample of size $n=430$ from the population. We estimate the prediction variance $Var(\hat{T} - T)$, construct 95% asymptotic *CI* for T , and calculate the length of confidence interval (I) for each of the 431 possible cross-validation samples from the finite population.

We use the 431 possible cross-validation samples to plot the absolute value of relative bias of \hat{T} , defined as $\left| \frac{\hat{T} - T}{T} \right|$, and the length of 95% confidence interval, defined as $2 \times 1.96 \times \sqrt{\text{Var}(\hat{T} - T)}$, for each of the three predictors in Figure 1.5. It can be observed that the absolute relative errors for predictors based on the log-transformation and Box-Cox transformation model (LTP and EBP) are closer to zero compared to the predictor based on no transformation model (NTP) for most samples; whereas for the rest of samples the LTP tends to give extreme large absolute relative errors. Figure 1.5 also displays the distribution of the length of 95% confidence interval over 431 samples. We observe that the length produced by EBP is shortest for most samples. For the remaining samples (about 70 samples), NTP provides the shortest length. Among the remaining samples, however, the proportion of the times that the true value of T included in the 95% confidence interval is only 78% for NTP but 94% for EBP.

In order to have an overview of the performance of the three predictors, the following evaluation statistics for each predictor are calculated:

- Average Absolute Relative Deviation (*AARD*) = $\frac{1}{431} \sum_{i=1}^{431} \left| \frac{\hat{T}_i - T}{T} \right|$,
- Average length of 95% confidence interval (*ALCI*) = $\frac{1}{431} \sum_{i=1}^{431} I_i$, and
- Proportion of times the true value included in the 95% confidence interval

$$(P) = \frac{1}{431} \sum_{i=1}^{431} I\{T \in CI_i\},$$

where the subscript i denotes the i^{th} sample selected from the beef population, $I\{\}$ is an indicator function which is equal to one if the true value T is included in CI , and zero otherwise.

Table 1.5 reports the *AARD*, *ALCI*, and P for three predictors based on different models. It can be observed that based on the both *AARD* and *ALCI* criteria, the log-transformation improves on the no transformation model, but the improvement is not substantial. The predictor based on the Box-Cox model achieves the smallest *AARD* (.0012). Also, for this method the *ALCI* is the shortest (332,542), about three-fifth of the *ALCI* based on the no-transformation model (563,656). The proportion of times that the true value is included for three predictors are all 95%, same as the nominal 95% confidence interval.

1.7 Concluding remarks

It is interesting to note that survey researchers have not used the well-known Box-Cox method of adaptive transformation. In this chapter, we have found such a method useful in achieving robustness in finite population inference. One of the challenges here is how to handle the fact that the observations may not be selected with equal probability and also may be clustered. We sidestep much of this complexity by employing a prediction approach in this chapter, but we still have to face this issue when we look at the robustness of our approach.

Table 1.1: The prediction mean squared error of six predictors¹.

σ	n	NTP ² .	LTP ² .	AEBP ² .	ABP ² .	EBP ² .	BP ² .
0.1	50	0.50	5.10	0.08	0.01	0.08	0.01
	100	3.67	13.35	0.05	0.01	0.05	0.01
	150	4.17	4.14	0.03	0.01	0.03	0.01
0.5	50	2.82	8.65	2.38	0.40	2.29	0.25
	100	4.83	16.60	1.54	0.35	1.48	0.22
	150	4.96	5.54	0.71	0.30	0.64	0.20
1	50	11.36	22.59	11.29	3.50	9.84	1.10
	100	8.38	29.75	6.85	2.90	5.91	0.89
	150	7.49	10.38	3.96	2.50	2.63	0.89
2	50	57.03	129.22	72.84	50.29	61.62	5.91
	100	29.67	112.18	54.96	43.17	27.58	5.30
	150	22.72	53.49	36.95	30.70	15.01	4.30

¹. Prediction mean squared errors are scaled down by multiplying 10^{-10} .

². NTP and LTP are predictors based on non-transformation model and log-transformation model. BP, EBP, ABP, and AEBP are based on Box-Cox model, denoting best predictor, empirical best predictor, approximate best predictor and approximate empirical best predictor.

Table 1.2: The prediction relative bias of six predictors¹.

σ	n	NTP ² .	LTP ² .	AEBP ² .	ABP ² .	EBP ² .	BP ² .
0.1	50	46.57	61.35	-0.33	-0.47	0.06	0.00
	100	-39.18	101.57	-0.16	-0.35	0.24	0.26
	150	-3.23	79.99	-0.14	-0.12	0.26	0.16
0.5	50	44.22	63.97	-10.16	-10.60	-0.27	-0.78
	100	-37.18	109.32	-7.15	-9.16	2.66	0.42
	150	-2.79	84.79	-8.66	-10.00	1.15	-0.77
1	50	46.15	81.60	-34.23	-37.25	5.20	1.81
	100	-35.42	126.90	-32.10	-38.50	7.11	0.01
	150	-2.91	99.72	-36.76	-37.72	1.80	0.42
2	50	46.72	155.46	-132.44	-141.66	19.24	1.72
	100	-35.84	202.75	-133.72	-138.30	13.02	4.10
	150	7.60	173.15	-131.04	-136.26	16.65	10.25

¹. Prediction relative biases of six predictors are scale up by multiplying 10^3 .

². NTP and LTP are predictors based on non-transformation model and log-transformation model. BP, EBP, ABP, and AEBP are based on Box-Cox model, denoting best predictor, empirical best predictor, approximate best predictor and approximate empirical best predictor.

Table 1.3: Percentage of loss of prediction mean squared error

relative to the best predictor

σ	n	AEBP ¹	ABP ¹	EBP ¹
0.1	50	6.82	0.02	6.77
	100	3.53	0.01	3.51
	150	2.21	0.01	2.19
0.5	50	8.65	0.63	8.29
	100	5.88	0.56	5.61
	150	2.56	0.51	2.18
1	50	9.22	2.16	7.90
	100	6.67	2.24	5.61
	150	3.43	1.80	1.95
2	50	11.31	7.50	9.42
	100	9.38	7.15	4.21
	150	7.58	6.13	2.49

¹EBP, ABP, and AEBP are based on Box-Cox model, denoting empirical best predictor, approximate best predictor and approximate empirical best predictor.

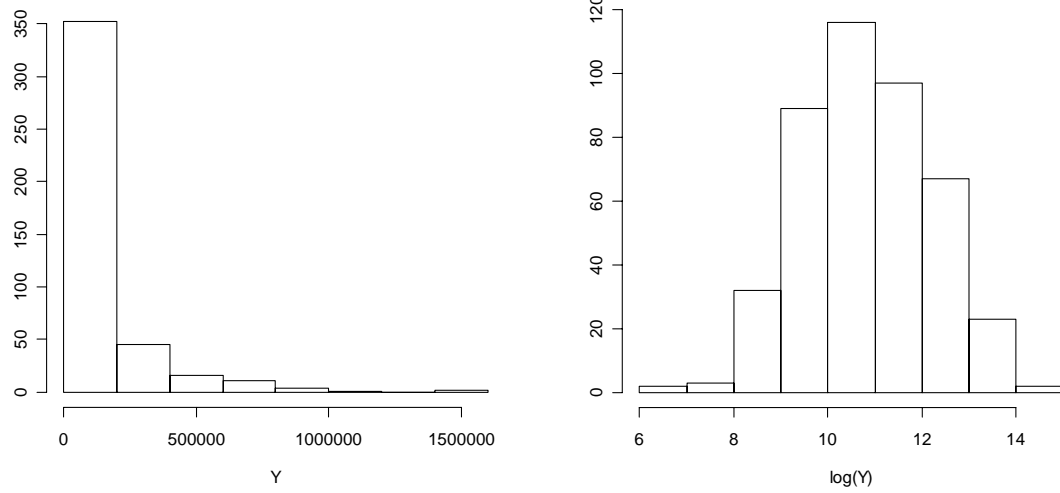


Figure 1.1: Histograms for the beef population before and after taking the log-transformation

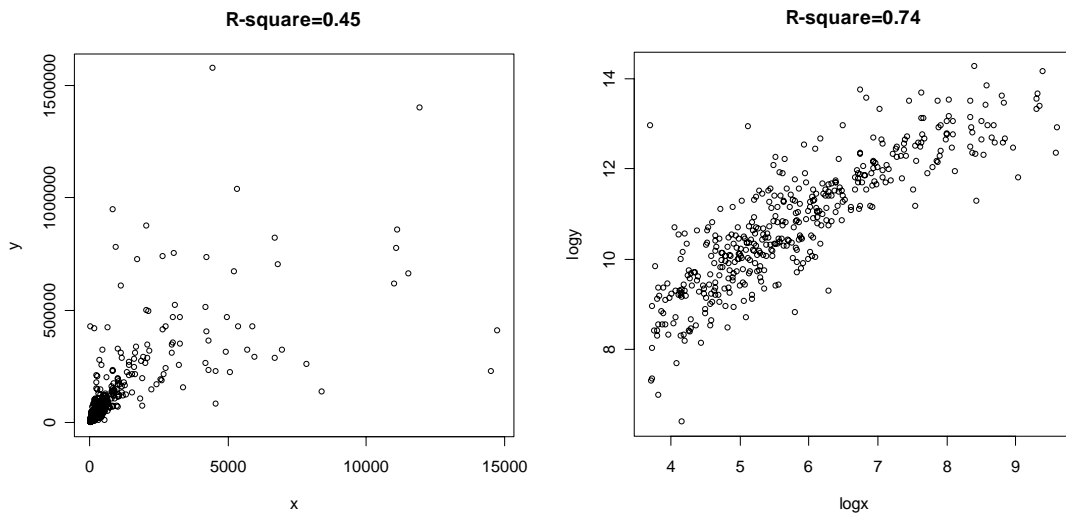


Figure 1.2: Scatter plots for the beef population before and after taking the log-transformation

Table 1.4: Estimates and 95% confidence intervals for β_0, β_1 , and λ .

	Estimate	Confidence Interval	
		Lower limit	Upper limit
β_0	4.20	3.34	5.06
β_1	2.67	2.53	2.81
λ	0.10	0.05	0.16

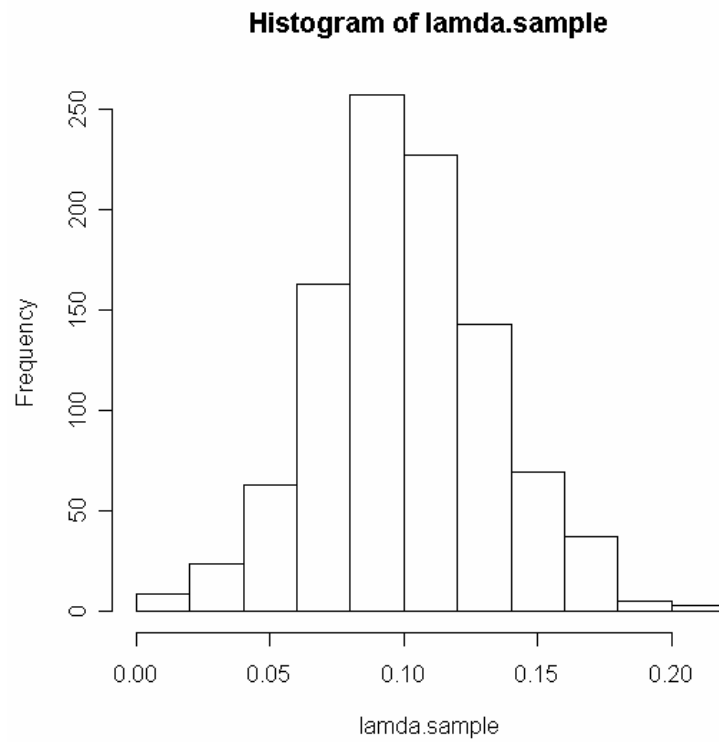


Figure 1.3: Histogram of $\hat{\lambda}$'s estimated using 1000 bootstrap samples

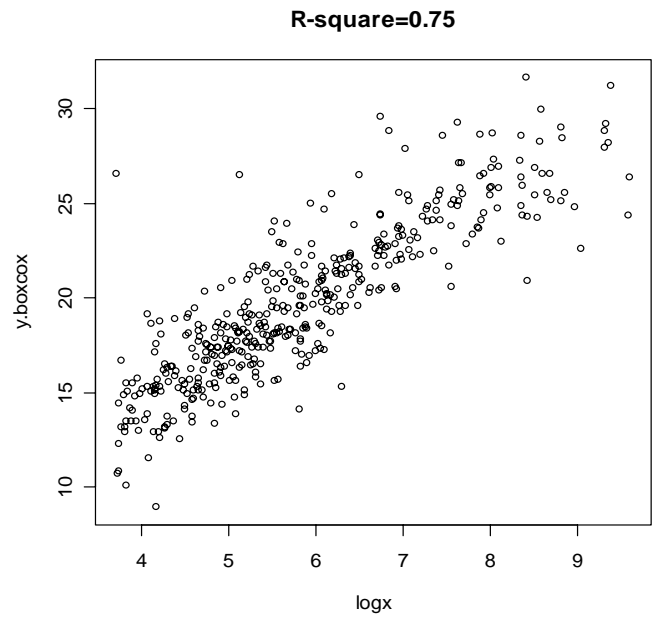


Figure 1.4: Scatter plot for the beef population after taking the Box-Cox transformation on y

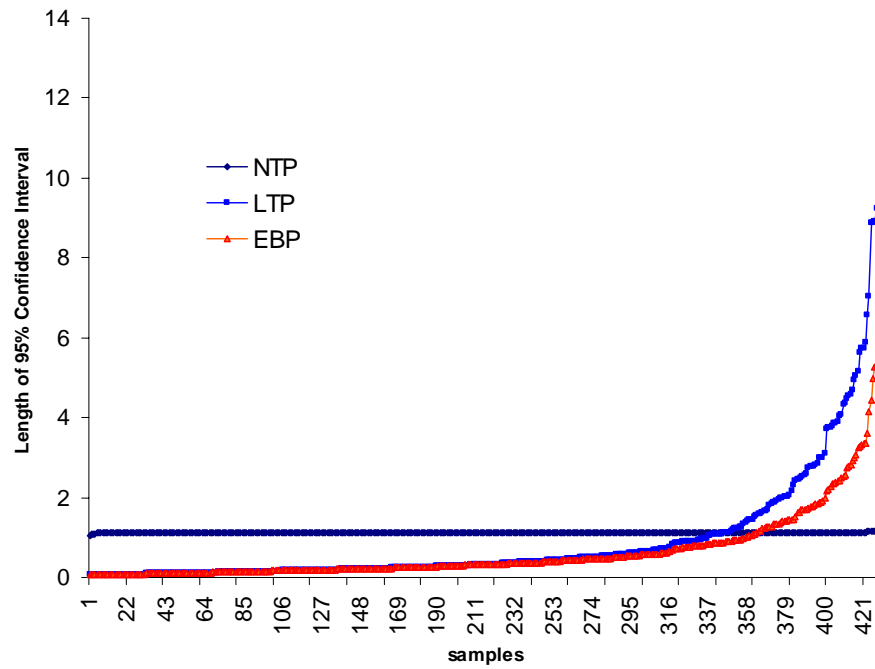
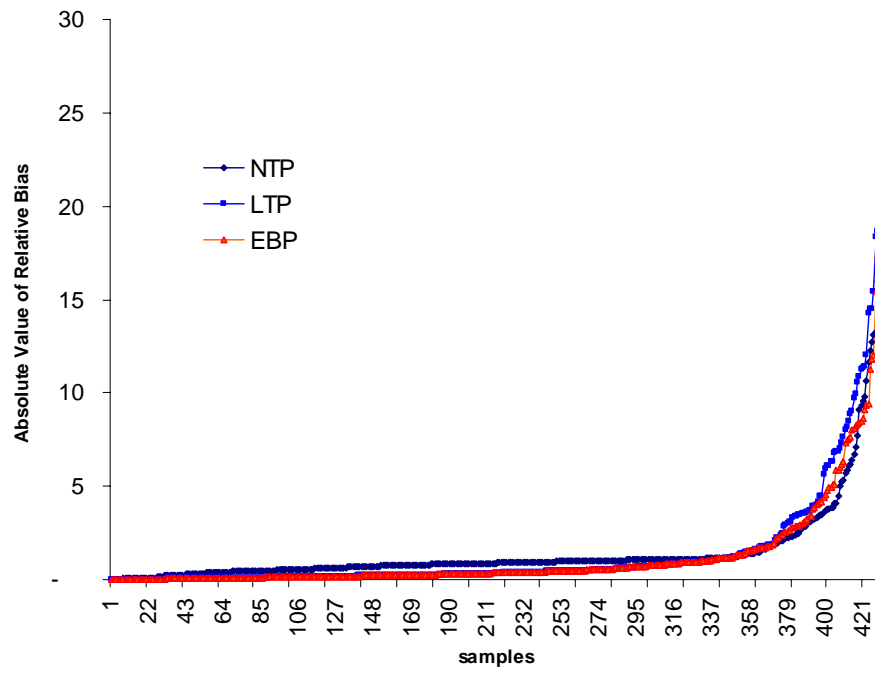


Figure 1.5: Distribution of the absolute value of relative bias and the length of 95% confidence interval for three predictors over the 431 possible samples

Table 1.5: *AARD*, *ALCI*, and *P* for three predictors based on models \mathbf{M}_1 , \mathbf{M}_2 , and \mathbf{M}_3

	<i>AARD</i>	<i>ALCI</i>	<i>P</i>
NTP ¹ (\mathbf{M}_1)	0.0015	563,656	0.95
LTP ¹ (\mathbf{M}_2)	0.0014	468,563	0.95
EBP ¹ (\mathbf{M}_3)	0.0012	332,542	0.95

¹. NTP, LTP, and EBP are predictors based on no-transformation model, log-transformation model, and Box-Cox model, respectively.

Appendix for Chapter 1

Prediction variance of \hat{T} under model M_1

$$\hat{T} = \sum_{i \in s} y_i + \sum_{i \in r} \hat{y}_i \text{ and } \hat{y}_i = \mathbf{x}_i' \hat{\boldsymbol{\beta}},$$

$$\begin{aligned} \text{Var}(\hat{T} - T) &= \text{Var}\left(\sum_{i \in r} (\mathbf{x}_i' \hat{\boldsymbol{\beta}} - y_i)\right) \\ &= \text{Var}\left(\sum_{i \in r} \mathbf{x}_i' \hat{\boldsymbol{\beta}}\right) + \text{Var}\left(\sum_{i \in r} y_i\right) \\ &= \sigma^2 \left(\sum_{i \in r} \mathbf{x}_i'\right) (\mathbf{X}_s' \mathbf{X}_s)^{-1} \left(\sum_{i \in r} \mathbf{x}_i\right) + (N - n)\sigma^2 \end{aligned}$$

Prediction bias of \hat{T} when $\hat{y}_i = e^{\mathbf{x}_i' \hat{\boldsymbol{\beta}}}$ under model M_2

$$\hat{T} = \sum_{i \in s} y_i + \sum_{i \in r} \hat{y}_i \text{ and } \hat{y}_i = e^{\mathbf{x}_i' \hat{\boldsymbol{\beta}}}$$

$$E(\hat{T} - T) = E\left(\sum_{i \in r} (e^{\mathbf{x}_i' \hat{\boldsymbol{\beta}}} - y_i)\right)$$

$$E(y_i) = e^{\mathbf{x}_i' \boldsymbol{\beta} + \frac{1}{2}\sigma^2}$$

$$\mathbf{x}_i' \hat{\boldsymbol{\beta}} \sim N(\mathbf{x}_i' \boldsymbol{\beta}, \mathbf{x}_i' \text{Var}(\hat{\boldsymbol{\beta}}) \mathbf{x}_i)$$

By moment generating function,

$$E(e^{\mathbf{x}_i' \hat{\boldsymbol{\beta}}}) = e^{\mathbf{x}_i' \boldsymbol{\beta} + \frac{1}{2} \mathbf{x}_i' \text{Var}(\hat{\boldsymbol{\beta}}) \mathbf{x}_i}.$$

Therefore,

$$E(\hat{T} - T) = \sum_{i \in r} (e^{\mathbf{x}_i' \boldsymbol{\beta} + \frac{1}{2} \mathbf{x}_i' \text{Var}(\hat{\boldsymbol{\beta}}) \mathbf{x}_i} - e^{\mathbf{x}_i' \boldsymbol{\beta} + \frac{1}{2} \sigma^2}),$$

which is not necessary to be zero.

Prediction variance of \hat{T} when $\hat{y}_i = e^{\mathbf{x}_i' \hat{\boldsymbol{\beta}} + \frac{1}{2} \hat{\sigma}^2}$ under model \mathbf{M}_2

$$\hat{T} = \sum_{i \in S} y_i + \sum_{i \in r} \hat{y}_i \text{ and } \hat{y}_i = e^{\mathbf{x}_i' \hat{\boldsymbol{\beta}} + \frac{1}{2} \hat{\sigma}^2},$$

where

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}_s' \mathbf{X}_s)^{-1} \mathbf{X}_s' (\mathbf{log} \mathbf{Y}_s) \text{ and}$$

$$\hat{\sigma}^2 = \frac{1}{n - (k + 1)} (\mathbf{log} \mathbf{Y}_s - \mathbf{X}_s' \hat{\boldsymbol{\beta}})' (\mathbf{log} \mathbf{Y}_s - \mathbf{X}_s' \hat{\boldsymbol{\beta}}).$$

Assume the normality for the errors.

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}_s' \mathbf{X}_s)^{-1},$$

$$\text{Var}(\hat{\sigma}^2) = \left(\frac{\sigma^2}{n - k - 1} \right)^2 2(n - k - 1) = \frac{2\sigma^4}{n - k - 1},$$

$$\text{Cov}(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = 0.$$

The moment generating function for $\log y_i$ is:

$$E(e^{t \log y_i}) = e^{\mathbf{x}_i' \boldsymbol{\beta} + (1/2) t^2 \sigma^2}.$$

Let t equal to one, we have

$$E(y_i) = e^{\mathbf{x}_i' \boldsymbol{\beta} + (1/2) \sigma^2}.$$

Let t equal to two, we have

$$E(y_i^2) = e^{2(\mathbf{x}_i' \boldsymbol{\beta} + \sigma^2)}.$$

Therefore,

$$\begin{aligned} \text{Var}(y_i) &= e^{2(\mathbf{x}_i'\boldsymbol{\beta} + \sigma^2)} - e^{2\mathbf{x}_i'\boldsymbol{\beta} + \sigma^2} \\ &= e^{\sigma^2} (e^{\sigma^2} - 1) e^{2\mathbf{x}_i'\boldsymbol{\beta}} \end{aligned}$$

$$\begin{aligned} \text{Var}(\hat{T} - T) &= \text{Var}\left(\sum_{i \in r} (e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}} + \frac{1}{2}\hat{\sigma}^2} - y_i)\right) \\ &= \text{Var}\left(\sum_{i \in r} e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}} + \frac{1}{2}\hat{\sigma}^2}\right) + \text{Var}\left(\sum_{i \in r} y_i\right) \end{aligned}$$

$$\text{Var}\left(\sum_{i \in r} y_i\right) = \sum_{i \in r} \text{Var}(y_i) = (e^{2\sigma^2} - e^{\sigma^2}) \sum_{i \in r} e^{2(\mathbf{x}_i'\boldsymbol{\beta})}$$

$$\text{Var}\left(\sum_{i \in r} \hat{y}_i\right) = \sum_{i \in r} \sum_{j \in r} \text{Cov}(\hat{y}_i, \hat{y}_j),$$

By Taylor Series,

$$\text{Cov}(\hat{y}_i, \hat{y}_j) \approx \left(\frac{\partial}{\partial \hat{\boldsymbol{\theta}}}\hat{y}_i\right)'_{\hat{\boldsymbol{\theta}}_0} \text{Var}(\hat{\boldsymbol{\theta}}) \left(\frac{\partial}{\partial \hat{\boldsymbol{\theta}}}\hat{y}_j\right)_{\hat{\boldsymbol{\theta}}_0} \text{ and } \hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\sigma}^2),$$

where

$$\frac{\partial}{\partial \hat{\boldsymbol{\beta}}}\hat{y}_i = e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}} + \frac{1}{2}\hat{\sigma}^2} \mathbf{x}_i,$$

$$\frac{\partial}{\partial \hat{\sigma}^2}\hat{y}_i = \frac{1}{2} e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}} + \frac{1}{2}\hat{\sigma}^2},$$

$$\text{Var}(\hat{\boldsymbol{\theta}}) = \begin{pmatrix} \sigma^2 (\mathbf{X}_s' \mathbf{X}_s)^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{2\sigma^4}{n-k-1} \end{pmatrix}.$$

A direct method to obtain $\text{var}(\hat{\boldsymbol{\theta}})$

$$\text{var}(\hat{\boldsymbol{\theta}}^*) = \mathbf{i}(\boldsymbol{\theta}^*)^{-1}, \text{ where } \mathbf{i}(\boldsymbol{\theta}^*) = \left(-\frac{\partial^2 l^*}{(\partial \boldsymbol{\theta}^*)(\partial \boldsymbol{\theta}^*)'} \right) \Big|_{\hat{\boldsymbol{\theta}}^*}.$$

We know the relationship $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^*, \lambda, \boldsymbol{\sigma}^*)' = (\tilde{\mathbf{y}}^{1-\lambda} \boldsymbol{\beta}, \lambda, \tilde{\mathbf{y}}^{1-\lambda} \boldsymbol{\sigma})'$.

Let $\boldsymbol{\theta} = f(\boldsymbol{\theta}^*) = (\tilde{\mathbf{y}}^{\lambda-1} \boldsymbol{\beta}^*, \lambda, \tilde{\mathbf{y}}^{\lambda-1} \boldsymbol{\sigma}^*)'$.

Therefore,

$$\begin{aligned} \frac{\partial^2 l^*}{(\partial f(\boldsymbol{\theta}^*))(\partial f(\boldsymbol{\theta}^*))'} &= \frac{\partial}{\partial f(\boldsymbol{\theta}^*)} \left(\frac{\partial l^*}{\partial \boldsymbol{\theta}^*} \frac{\partial \boldsymbol{\theta}^*}{\partial f(\boldsymbol{\theta}^*)} \right) \\ &= \left(\frac{\partial}{\partial f(\boldsymbol{\theta}^*)} \left(\frac{\partial l^*}{\partial \boldsymbol{\theta}^*} \right) \right) \frac{\partial \boldsymbol{\theta}^*}{\partial f(\boldsymbol{\theta}^*)} + \frac{\partial l^*}{\partial \boldsymbol{\theta}^*} \left(\frac{\partial}{\partial f(\boldsymbol{\theta}^*)} \left(\frac{\partial \boldsymbol{\theta}^*}{\partial f(\boldsymbol{\theta}^*)} \right) \right) \\ &= \left(\frac{\partial \boldsymbol{\theta}^*}{\partial f(\boldsymbol{\theta}^*)} \right)' \left(\frac{\partial^2 l^*}{(\partial \boldsymbol{\theta}^*)(\partial \boldsymbol{\theta}^*)'} \right) \left(\frac{\partial \boldsymbol{\theta}^*}{\partial f(\boldsymbol{\theta}^*)} \right) + \left(\frac{\partial l^*}{\partial \boldsymbol{\theta}^*} \right) \left(\frac{\partial^2 \boldsymbol{\theta}^*}{(\partial f(\boldsymbol{\theta}^*))(\partial f(\boldsymbol{\theta}^*))'} \right) \end{aligned}$$

$$\mathbf{i}(\boldsymbol{\theta}) = \mathbf{i}(f(\boldsymbol{\theta}^*))$$

$$\begin{aligned} &= \left(-\frac{\partial^2 l^*}{(\partial f(\boldsymbol{\theta}^*))(\partial f(\boldsymbol{\theta}^*))'} \right) \Big|_{\hat{\boldsymbol{\theta}}^*} \\ &= \left(\frac{\partial \boldsymbol{\theta}^*}{\partial f(\boldsymbol{\theta}^*)} \right)' \left(-\frac{\partial^2 l^*}{(\partial \boldsymbol{\theta}^*)(\partial \boldsymbol{\theta}^*)'} \right) \left(\frac{\partial \boldsymbol{\theta}^*}{\partial f(\boldsymbol{\theta}^*)} \right) \Big|_{\hat{\boldsymbol{\theta}}^*} + \left(\frac{\partial l^*}{\partial \boldsymbol{\theta}^*} \right) \left(-\frac{\partial^2 \boldsymbol{\theta}^*}{(\partial f(\boldsymbol{\theta}^*))(\partial f(\boldsymbol{\theta}^*))'} \right) \Big|_{\hat{\boldsymbol{\theta}}^*} \\ &= \left(\frac{\partial \boldsymbol{\theta}^*}{\partial f(\boldsymbol{\theta}^*)} \right)' \left(-\frac{\partial^2 l^*}{(\partial \boldsymbol{\theta}^*)(\partial \boldsymbol{\theta}^*)'} \right) \left(\frac{\partial \boldsymbol{\theta}^*}{\partial f(\boldsymbol{\theta}^*)} \right) \Big|_{\hat{\boldsymbol{\theta}}^*} \text{ due to } \left(\frac{\partial l^*}{\partial \boldsymbol{\theta}^*} \right) \Big|_{\hat{\boldsymbol{\theta}}^*} = 0, \end{aligned}$$

We have:

$$\begin{aligned}
\text{var}(\hat{\boldsymbol{\theta}}) &= \mathbf{i}(f(\boldsymbol{\theta}^*))^{-1} = \left(\left(\frac{\partial \boldsymbol{\theta}^*}{\partial f(\boldsymbol{\theta}^*)} \right)' \left(-\frac{\partial^2 l^*}{(\partial \boldsymbol{\theta}^*)(\partial \boldsymbol{\theta}^*)'} \right) \left(\frac{\partial \boldsymbol{\theta}^*}{\partial f(\boldsymbol{\theta}^*)} \right) \right)^{-1} \Bigg|_{\boldsymbol{\theta}^*} \\
&= \left(\frac{\partial f(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^*} \right)' \left(-\frac{\partial^2 l^*}{(\partial \boldsymbol{\theta}^*)(\partial \boldsymbol{\theta}^*)'} \right)^{-1} \left(\frac{\partial f(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^*} \right) \Bigg|_{\boldsymbol{\theta}^*}, \\
&= \left(\frac{\partial f(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^*} \right)' \text{var}(\hat{\boldsymbol{\theta}}^*) \left(\frac{\partial f(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^*} \right) \Bigg|_{\boldsymbol{\theta}^*}
\end{aligned}$$

$$\text{where } \left(\frac{\partial f(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^*} \right) = \begin{pmatrix} \tilde{y}^{\lambda-1} & 0 & \cdots & 0 & \beta_0 \log(\tilde{y}) & 0 \\ 0 & \tilde{y}^{\lambda-1} & \cdots & 0 & \beta_1 \log(\tilde{y}) & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \tilde{y}^{\lambda-1} & \beta_k \log(\tilde{y}) & 0 \\ 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & \cdots & 0 & \sigma \log(\tilde{y}) & \tilde{y}^{\lambda-1} \end{pmatrix}.$$

Equivalence of the MLE of λ under models \mathbf{M}_3 and \mathbf{M}_4

The log-likelihood function with respect to model \mathbf{M}_3 is given by

$$l(\boldsymbol{\theta}) = \log L = -\frac{1}{2} \sum_s \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_s (y_i^{(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta})^2 + (\lambda - 1) \sum_s \log y_i .$$

The log-likelihood function with respect to model \mathbf{M}_4 is given by

$$l(\boldsymbol{\theta}^*) = \log L^* = -\frac{1}{2} \sum_s \log(2\pi\sigma^{*2}) - \frac{1}{2\sigma^{*2}} \sum_s (y_i^{*(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}^*)^2 .$$

We want to show that the MLE estimates of λ with respect to $l(\boldsymbol{\theta})$ and $l(\boldsymbol{\theta}^*)$ are equivalent, i.e.,

$$\begin{aligned} \frac{\partial l(\boldsymbol{\theta})}{\partial \lambda} &= -\frac{1}{\sigma^2} \sum_s (y_i^{(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}) \frac{\partial y_i^{(\lambda)}}{\partial \lambda} + \sum_s \log y_i = 0 \\ \Leftrightarrow \frac{\partial l(\boldsymbol{\theta}^*)}{\partial \lambda} &= -\frac{1}{\sigma^{*2}} \sum_s (y_i^{*(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}^*) \frac{\partial y_i^{*(\lambda)}}{\partial \lambda} = 0. \end{aligned}$$

We know the relationship of $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^*, \lambda, \sigma^*)' = (\tilde{\mathbf{y}}^{1-\lambda} \boldsymbol{\beta}, \lambda, \tilde{\mathbf{y}}^{1-\lambda} \sigma)'$

Therefore,

$$\begin{aligned} \frac{\partial l(\boldsymbol{\theta}^*)}{\partial \lambda} &= 0 \\ \Leftrightarrow \frac{1}{\tilde{\mathbf{y}}^{2(1-\lambda)} \sigma^2} \sum_s (\tilde{\mathbf{y}}^{(1-\lambda)} y_i^{(\lambda)} - \mathbf{x}_i' \tilde{\mathbf{y}}^{(1-\lambda)} \boldsymbol{\beta}) \frac{\partial (\tilde{\mathbf{y}}^{(1-\lambda)} y_i^{(\lambda)})}{\partial \lambda} &= 0 \\ \Leftrightarrow \frac{1}{\tilde{\mathbf{y}}^{(1-\lambda)} \sigma^2} \sum_s (y_i^{(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}) \left(-\tilde{\mathbf{y}}^{(1-\lambda)} \ln \tilde{\mathbf{y}} \cdot y_i^{(\lambda)} + \tilde{\mathbf{y}}^{(1-\lambda)} \frac{\partial (y_i^{(\lambda)})}{\partial \lambda} \right) &= 0 \\ \Leftrightarrow \frac{-\ln \tilde{\mathbf{y}}}{\sigma^2} \sum_s (y_i^{(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}) (y_i^{(\lambda)}) + \frac{1}{\sigma^2} \sum_s (y_i^{(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}) \left(\frac{\partial (y_i^{(\lambda)})}{\partial \lambda} \right) &= 0 \end{aligned}$$

$$\begin{aligned}
&\Leftrightarrow \frac{-1}{n\sigma^2} \sum_s \log y_i \left(\sum_s (y_i^{(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta})(y_i^{(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}) + \sum_s (y_i^{(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta})(\mathbf{x}_i' \boldsymbol{\beta}) \right) + \\
&\quad \frac{1}{\sigma^2} \sum_s (y_i^{(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}) \left(\frac{\partial (y_i^{(\lambda)})}{\partial \lambda} \right) = 0 \\
&\Leftrightarrow -\sum_s \log y_i - \frac{1}{n\sigma^2} \sum_s \log y_i \left(\sum_s (y_i^{(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta})(\mathbf{x}_i' \boldsymbol{\beta}) \right) + \\
&\quad \frac{1}{\sigma^2} \sum_s (y_i^{(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}) \left(\frac{\partial (y_i^{(\lambda)})}{\partial \lambda} \right) = 0
\end{aligned}$$

Accordingly,

$$\frac{\partial l(\boldsymbol{\theta}^*)}{\partial \lambda} = 0 \Leftrightarrow \frac{\partial l(\boldsymbol{\theta})}{\partial \lambda} = 0 \text{ if and only if}$$

$$\sigma^2 = \frac{1}{n} \sum_s (y_i^{(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta})^2, \text{ and}$$

$$\frac{1}{n\sigma^2} \sum_s \log y_i \left(\sum_s (y_i^{(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta})(\mathbf{x}_i' \boldsymbol{\beta}) \right) = 0, \text{ i.e.,}$$

$$\sum_s (y_i^{(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}) \mathbf{x}_i' = 0.$$

This completes the proof.

Chapter 2: Automated Generalized Regression (AUTOGREG)

Estimators of A Finite Population Total

2.1 Introduction

Generalized regression (GREG) estimators of finite population totals and means are derived using regression models. Models are used to construct estimators, but randomization must be used to select the sample, and statistical properties are evaluated with respect to the probability sampling distribution. The GREG is essentially a model-assisted estimator which has the desirable design-consistency property. Concise discussion of the GREG estimator can be found in Valliant, *et. al.* (2000, Ch2) and Rao (2003, Ch2). For design-consistent estimators which use a general mixed model, see Jiang and Lahiri (2006).

In constructing a GREG estimator, weighting of the sample observations is necessary to obtain design-consistent estimators of model parameters. The ordinary least square method that ignores population structure such as clustering and stratification can provide misleading results when the sampling rates depend upon the outcome variable (Korn and Graubard, 1995; Holt, *et. al.* 1980; Pfeffermann and Holmes, 1985; Nathan and Holt, 1980). Different approaches for incorporating the weights in the inference process were studied (Pfefferman, 1993). The pseudo-maximum likelihood (PML) is a method that accounts for the sampling weights in estimating parameters of a regression model. The method uses sampling weights to estimate the finite population likelihood equation. The basic idea of PML had its origin in Kish and Frankel (1974). Binder (1983) and Godambe and Thompson (1986) made major contributions in this general

research area. The PML method has been used in a variety of models such as the logistic model (Chambless and Boyle 1985; Scott and Wild 1989), loglinear model (Rao and Thomas 1989), GLM (Nordberg 1989) and the proportional hazards model (Binder 1992; Chambless and Boyle 1985; Kasprzyk, *et. al.* 1989). Binder (1983) developed a general method for estimating the randomization variance covariance matrix of the PML estimator.

The GREG estimator is approximately unbiased or design-consistent for the target quantity irrespective of whether the assumptions of the model are true or false. Examples may be found in Särndal, *et. al.* (1992), Estevao, *et. al.* (1995), Fuller, *et. al.* (1994), and Jayasuriya and Valliant (1996). On the other hand, the appropriateness of the model is crucial to achieve a small variance. If the assumed model can well describe the finite population, the GREG estimator can bring about a large variance reduction, as compared to the Narain-Horvitz-Thompson estimator (Särndal, *et. al.* 1992). Therefore, there is a need to achieve robustness with respect to model selection for GREG estimators.

In the mainstream statistics, transformations on the dependent variable in the assumed model are often used to achieve normality, linearity, and homoscedasticity (Carroll and Ruppert, 1988), but the literature on transformations in finite population inference is not very rich. There is, however, a growing interest in developing methods that use an appropriate transformation with survey data. Chen and Chen (1996) considered transformed survey data in order to improve on the precision of the normal approximation. Korn and Graubard (1998) compared different confidence intervals, including intervals based on a logit-transformation, for proportions with small expected number of positive counts. Karlberg (2000) proposed an estimator based on a lognormal-

logistic superpopulation model to predict the finite population total of a highly skewed survey variable. The simulation results indicated that the lognormal-logistic model estimator offers a sensible alternative to other estimators, especially when the sample size is small. Chambers and Dorfman (2003) discussed the estimation of a finite population mean under certain general but known transformation on the continuous data.

Researchers find the transformation technique useful in analyzing survey data. However, the key step is the identification of an appropriate transformation that fits the survey data well. In many applications, the form of transformation is determined subjectively. Now, prior knowledge or theory may not suggest the transformation to be used. In such situations, it would be convenient to determine the transformation adaptively using the data.

The work of Box and Cox (1964) has led to the development of “data-decide-transformation” methods for constructing models with independently and identically distributed errors. Techniques for estimation of Box-Cox model parameters, β , and transformation parameter, λ , have been developed extensively in mainstream statistics. However, in survey sampling context, there is a lack of studies on the estimation methods for β and λ . One potential estimation method is to estimate β and λ in the Box-Cox transformation model using PML technique.

In this chapter, we propose an automated generalized regression (AUTOGREG) estimator of a finite population total and its variance estimator under a general unequal probability sampling design. The proposed estimator maintains the robustness property of GREG estimator in the sense that AUTOGREG is design-consistent even if the underlying model fails. The class of ‘robust models’ is further extended such that an

appropriate transformation on the dependent variable is automatically determined by the data using the Box-Cox technique. The AUTOGREG does not require a linear model on the dependent variable assumed under the GREG theory.

The bias and variance of the new estimator with respect to the design are investigated analytically. We compare the AUTOGREG to the design-based estimator, model-based estimators, and the usual GREG estimators via Monte Carlo simulations. Section 2.2 briefly reviews the design-based, model-based and GREG estimators of a finite population total. The new estimator is presented in Section 2.3, along with its variance. A simulation study is conducted in Section 2.4, and the results are showed in Section 2.5. Finally, we give some concluding remarks in Section 2.6.

2.2 Design-based, model-based and GREG estimators of a finite population total

Suppose that the quantity of interest is the finite population total

$$T = \sum_{i \in U} y_i ,$$

where i indexes population units, N is the population size and y_i is values of the variable of interest associated with each unit. Write $\mathbf{y} = (y_1, \dots, y_N)'$. To estimate T , a sample s of size n is drawn from the finite population $U = \{1, \dots, N\}$ using a probability sampling scheme. A sampling design $p(s)$ is the probability of selecting the sample s . Thus, $p(\cdot)$ defines a discrete probability distribution on S , the set of all samples, and hence satisfies the two basic conditions; (i) $p(s) \geq 0$ for all $s \in S$ and (ii) $\sum_{s \in S} p(s) = 1$. Commonly used sampling designs include simple random sampling, probability proportional to size

sampling, stratified simple random sampling, and stratified multistage sampling. The first-order inclusion probabilities, given by $\pi_i = P(i \in s)$ ($i = 1, \dots, N$), are assumed to be all non-zeroes. We assume that we have information on $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)'$, where $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ik})'$ is a column vector of k known auxiliary variables for the unit i . For any sample s of size n , we redefine \mathbf{y} and \mathbf{X} so that the first n rows of \mathbf{y} and \mathbf{X} correspond to those in the sample. Write

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_s \\ \mathbf{y}_r \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_s \\ \mathbf{X}_r \end{pmatrix},$$

where

\mathbf{y}_s is a $n \times 1$ column vector of observed dependent variable;

\mathbf{y}_r is a $(N - n) \times 1$ column vector of unobserved dependent variable;

\mathbf{X}_s is a $n \times (k + 1)$ matrix of known auxiliary variables in the sample;

\mathbf{X}_r is a $(N - n) \times (k + 1)$ matrix of known auxiliary variables outside the sample.

In this section, we shall briefly review the existing design-based, model-based, and model-assisted estimators of the finite population total T . We use E_d and V_d to denote expected value and variance with respect to the design $p(s)$.

2.2.1 Design-based Estimator

Horvitz and Thmopson (1952) proposed the following well-celebrated estimator:

$$\hat{T}_D = \sum_{i \in s} \frac{y_i}{\pi_i}. \quad (2.1)$$

The estimator \hat{T}_D is design unbiased. Rao (2005) pointed out that this estimator was independently proposed by Narain (1951) and hence should be called Narain-Horvitz-Thompson estimator. The variance of the Narain-Horvitz-Thompson is given by

$$V_d(\hat{T}_D) = \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{y_i}{\pi_i} \frac{y_j}{\pi_j},$$

where $\pi_{ij} = P(i \in s \text{ and } j \in s)$ ($i, j = 1, \dots, N$) is the second-order inclusion probability, i.e. the probability that both units i and j are included in the sample. If the inclusion probability π_i can be chosen to be proportional to the y_i the variance of this estimator will be reduced.

2.2.2 Model-based estimators

We shall consider

$$\hat{T}_M = \sum_{i \in s} y_i + \sum_{i \in r} \hat{y}_i, \quad (2.2)$$

where r represents the set of unobserved units in the finite population and \hat{y}_i is the predictor for the i^{th} unobserved unit. Three different models are motivated in this subsection. The most commonly used model is the standard linear regression model

$$\mathbf{M}_1 : \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \mathbf{I})$, a N -variate probability distribution with the mean vector $\mathbf{0}$ and variance covariance matrix $\sigma^2 \mathbf{I}$, and \mathbf{I} is the $N \times N$ identity matrix. In this equation, $\boldsymbol{\beta}$ is the $(k+1) \times 1$ column vector of regression coefficients. Both σ^2 and $\boldsymbol{\beta}$ are unknown superpopulation parameters. The predictor for the i^{th} unobserved unit is given by

$$\hat{y}_i = \mathbf{x}_i' \hat{\boldsymbol{\beta}}, \quad (2.3)$$

where $\hat{\beta}_o$ is the ordinary least square (OLS) estimator of β under \mathbf{M}_1 and $\hat{\beta}_o = (\mathbf{X}_s' \mathbf{X}_s)^{-1} \mathbf{X}_s' \mathbf{Y}_s$.

In some applications, especially in business and agricultural surveys, a linear model may not be appropriate for y , but may be appropriate for a strictly monotonic transformation of y . For the data set given in Royall and Cumberland (1981), Chen and Chen (1996) observed that the finite population distribution was severely skewed and that the log-transformation helped achieving symmetry. The need and the benefit of taking the log-transformation were obvious. Therefore, we consider the log-linear regression model where the log-transformation is used on the dependent variable

$$\mathbf{M}_2 : \log \mathbf{Y} = \mathbf{X} \beta + \varepsilon ,$$

where $\varepsilon \sim (\mathbf{0}, \sigma^2 \mathbf{I})$. The predictor for the i^{th} unobserved unit is given by

$$\hat{y}_i = e^{x_i \hat{\beta}_l} , \tag{2.4}$$

where $\hat{\beta}_l$ are OLS estimator under the model \mathbf{M}_2 and $\hat{\beta}_l = (\mathbf{X}_s' \mathbf{X}_s)^{-1} \mathbf{X}_s' \log \mathbf{Y}_s$.

The model \mathbf{M}_2 requires a subjective specification of the transformation to be applied on the dependent variable. This may be okay in some problems where we know the transformation to be used either from prior empirical evidence or from theory. In absence of any prior knowledge about the transformation to be used, we can consider an appropriate family of transformations to be determined by the data.

Tukey (1957) considered the following family of power transformations:

$$y^{(\lambda)} = \begin{cases} y^\lambda & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases} ,$$

where $y > 0$. In order to take care of the discontinuity at $\lambda = 0$, Box and Cox (1964) proposed the following family of transformations:

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1) / \lambda & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases},$$

where $y > 0$. The parameter λ determines the nature of transformation. For example, $\lambda = 1, 0, 0.5, -1$ correspond to no transformation, log-transformation, square root transformation, and reciprocal transformation, respectively. The transformation parameter λ is estimated by the data. The Box-Cox analysis may lead to a log-transformation, but may equally lead to some other transformation in the above family – it depends on the actual data observed.

We consider the following superpopulation model for the transformed dependent variable:

$$\mathbf{M}_3: \mathbf{Y}^{(\lambda)} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$. The predictor for the i^{th} unobserved unit is

$$\hat{y}_i = (\hat{\lambda} \mathbf{x}_i' \hat{\boldsymbol{\beta}} + 1)^{1/\hat{\lambda}}, \quad (2.5)$$

where $\hat{\lambda}$ and $\hat{\boldsymbol{\beta}}$ are OLS estimators under \mathbf{M}_3 .

Denote model-based estimators \hat{T}_M with different predictor \hat{y}_i 's, defined by (2.3) – (2.5), as \hat{T}_{M-L} , \hat{T}_{M-LOGL} , and \hat{T}_{M-BC} , respectively.

Define $\boldsymbol{\theta}$ as the unknown parameter of the finite population and $\hat{\boldsymbol{\theta}}$ the OLS estimator of the superpopulation parameter with respect to the underlying model. $\hat{\boldsymbol{\theta}}$ is model-unbiased, and further design-consistent under equal probability of selection method (EPSEM) sampling design. For example, under the model \mathbf{M}_1 , $\boldsymbol{\theta} = \mathbf{B}_0$, where \mathbf{B}_0

is the finite population parameter and $\mathbf{B}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$; while $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\beta}}_o$. We know $\hat{\boldsymbol{\beta}}_o$ is a model unbiased estimator of $\boldsymbol{\beta}$, the superpopulation parameter defined in the model \mathbf{M}_1 . Furthermore, for EPSEM sampling, $\hat{\boldsymbol{\beta}}_o$ is also a design-consistent estimator of \mathbf{B}_0 . When the model holds, $\mathbf{B}_0 = \boldsymbol{\beta} + O_p(N^{-1/2})$, and for N large enough, the distinction between $\boldsymbol{\beta}$ and \mathbf{B}_0 can be ignored (Holt, *et. al.* 1980). Evidently, $E_d(\hat{\boldsymbol{\beta}}_o) \approx \mathbf{B}_0$. Therefore, $\hat{\boldsymbol{\beta}}_o$ should be a reasonable estimator of $\boldsymbol{\beta}$. More discussion about the properties of LS procedure can be found in Fuller (1973), Brewer and Mellor (1973), and Harley and Silken (1975).

Define $f_i(\hat{\boldsymbol{\theta}})$ as an arbitrary predictor for the i^{th} unobserved unit among (2.3) – (2.5). By Taylor Series approximation,

$$\begin{aligned} E_d(\hat{T}_M) &= E_d \left\{ \sum_{i \in s} y_i + \sum_{i \in r} f_i(\hat{\boldsymbol{\theta}}) \right\} \\ &\approx E_d \left\{ \sum_{i \in s} y_i + \sum_{i \in r} f_i(\boldsymbol{\theta}) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \sum_{i \in r} \partial f_i(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \right\} \end{aligned}$$

Assume $\sum_{i \in r} \partial f_i(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ is bounded and $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \xrightarrow{p} 0$. Thus, by ignoring the last term, the

bias of \hat{T}_M is given by

$$\begin{aligned} E_d(\hat{T}_M - T) &\approx \sum_{i \in U} (\pi_i - 1) y_i + \sum_{i \in U} (1 - \pi_i) f_i(\boldsymbol{\theta}) \\ &= \sum_{i \in U} (1 - \pi_i) \{ f_i(\boldsymbol{\theta}) - y_i \} \end{aligned}$$

For EPSEM sampling, we have $\pi_i = \pi$ for each unit i . Under the model \mathbf{M}_1 ,

$f_i(\boldsymbol{\theta}) = \mathbf{x}_i' \mathbf{B}_0$ and the bias of \hat{T}_{M-L} is given by

$$E_d(\hat{T}_{M-L} - T) \approx (1 - \pi) \sum_{i \in U} (\mathbf{x}_i' \mathbf{B}_0 - y_i) = 0,$$

as

$$\sum_{i \in U} (\mathbf{x}_i' \mathbf{B}_0 - y_i) = 0.$$

Therefore, \hat{T}_{M-L} is approximately unbiased.

Under the model \mathbf{M}_2 , $f_i(\boldsymbol{\theta}) = e^{\mathbf{x}_i' \mathbf{B}_l}$, where $\mathbf{B}_l = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \log \mathbf{Y}$, and thus the bias of \hat{T}_{M-LOGL} is given by

$$E_d(\hat{T}_{M-LOGL} - T) \approx (1 - \pi) \sum_{i \in U} (e^{\mathbf{x}_i' \mathbf{B}_l} - y_i),$$

which is not necessary to be zero. Therefore, \hat{T}_{M-LOGL} is biased.

Under the model \mathbf{M}_3 , $f_i(\boldsymbol{\theta}) = (\lambda \mathbf{x}_i' \mathbf{B}_{bc} + 1)^{1/\lambda}$, where $\mathbf{B}_{bc} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{Y}^{(\lambda)}$, and thus the bias of \hat{T}_{M-BC} is

$$E_d(\hat{T}_{M-BC} - T) \approx (1 - \pi) \sum_{i \in U} \{(\lambda \mathbf{x}_i' \mathbf{B}_{bc} + 1)^{1/\lambda} - y_i\},$$

which again is not necessary to be zero. Therefore, \hat{T}_{M-BC} is also biased. It is interesting to note that the biases of both \hat{T}_{M-LOGL} and \hat{T}_{M-BC} do not tend to zero even for large sample with EPSEM sampling design.

2.2.3 GREG estimators

The GREG estimator is defined as

$$\hat{T}_G = \sum_{i \in U} \hat{y}_{i,w} + \sum_{i \in s} \frac{y_i - \hat{y}_{i,w}}{\pi_i}, \quad (2.6)$$

where $\hat{y}_{i,w}$ are predictors based on the models \mathbf{M}_1 and \mathbf{M}_2 . Regardless of how well the underlying model describes the population, GREG estimators of the finite population total are design-consistent (Särndal, *et. al.*1992).

Unlike \hat{y}_i in model-based estimators, $\hat{y}_{i,w}$ incorporates the sampling weights.

Under the model \mathbf{M}_1 ,

$$\hat{y}_{i,w} = \mathbf{x}_i' \hat{\boldsymbol{\beta}}_{wo}, \quad (2.7)$$

and under the model \mathbf{M}_2 ,

$$\hat{y}_{i,w} = e^{\mathbf{x}_i' \hat{\boldsymbol{\beta}}_{wl}}, \quad (2.8)$$

where $\hat{\boldsymbol{\beta}}_{wo}$ and $\hat{\boldsymbol{\beta}}_{wl}$ are weighted least square estimators with respect to the models \mathbf{M}_1 and \mathbf{M}_2 , respectively. Denote GREG estimators under the models \mathbf{M}_1 and \mathbf{M}_2 as \hat{T}_{G-L} and \hat{T}_{G-LOGL} , respectively.

For EPSEM sampling, note that

$$\begin{aligned} \hat{T}_{G-L} &= \sum_{i \in U} \hat{y}_{i,w} + \pi^{-1} \sum_{i \in s} (y_i - \hat{y}_{i,w}) \\ &= \sum_{i \in s} y_i + \pi^{-1} (1 - \pi) \sum_{i \in s} (y_i - \hat{y}_{i,w}) + \sum_{i \in r} \hat{y}_{i,w} \\ &= \sum_{i \in s} y_i + \sum_{i \in r} \hat{y}_{i,w} \\ &= \hat{T}_{M-L} \end{aligned}$$

as

$$\hat{y}_{i,w} = \hat{y}_i \text{ and } \sum_{i \in s} (y_i - \hat{y}_{i,w}) = \sum_{i \in s} (y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_{wo}) = 0.$$

Thus, \hat{T}_{G-L} and \hat{T}_{M-L} are equal.

2.3 AUTOGREG estimator of the finite population total

Here the main idea is to adjust the GREG estimator so as to achieve model robustness. In this section, we propose a new estimator, called AUTOGREG, which is more robust than the GREG. In addition to the design-consistency property of the GREG, our AUTOGREG uses a robust model automatically chosen by the Box-Cox method. Therefore, AUTOGREG estimator has a nice double robustness property. We define AUTOGREG estimator of the finite population total as

$$\hat{T}_{AG} = \sum_{i \in U} \hat{y}_{i,w} + \sum_{i \in s} (y_i - \hat{y}_{i,w}) / \pi_i, \quad (2.9)$$

where $\hat{y}_{i,w}$ is a predictor of y_i based on the model \mathbf{M}_3 . The AUTOGREG estimator is different from the GREG because the data dictates the transformation that is needed on the dependent variable before a linear regression model is used. \hat{T}_{AG} is design-consistent for the finite population total T under the randomization approach, and it uses an appropriate robust model to borrow strength from the relevant covariates to achieve a small variance.

2.3.1 Estimation of model and transformation parameters $\boldsymbol{\varphi} = (\boldsymbol{\beta}, \lambda, \sigma^2)'$

using the PML method

In order to ease the estimation of λ using existing computational procedures, one must replace $\mathbf{Y}^{(\lambda)}$ in the model \mathbf{M}_3 by a scaled transformation $\mathbf{Y}^{*(\lambda)}$. For the i^{th} unit,

$$y_i^{*(\lambda)} = \begin{cases} (y_i^\lambda - 1) / \lambda \tilde{y}^{\lambda-1} & \lambda \neq 0 \\ \tilde{y} \log(y_i) & \lambda = 0 \end{cases}$$

where \tilde{y} is the geometric mean of y 's. The following calculation will be based on the new scaled model:

$$\mathbf{M}_4 : \mathbf{Y}^{*(\lambda)} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}^*,$$

where $\boldsymbol{\varepsilon}^* \sim N(\mathbf{0}, \sigma_e^{*2}\mathbf{I})$. Let $\boldsymbol{\varphi}^* = (\boldsymbol{\beta}^*, \lambda, \sigma_e^{*2})'$.

The maximum likelihood estimator (MLE) for the $\boldsymbol{\varphi}^*$ maximizes the log-likelihood

$$l(\boldsymbol{\varphi}^*) = \sum_i \log f(y_i; \boldsymbol{\varphi}^*, \tilde{y}),$$

where

$$f(y_i; \boldsymbol{\varphi}^*, \tilde{y}) = (2\pi\sigma_e^{*2})^{-1/2} \exp\left\{-\frac{1}{2\sigma_e^{*2}}(y_i^{*(\lambda)} - \mathbf{x}_i'\boldsymbol{\beta}^*)^2\right\} \cdot (y_i / \tilde{y})^{\lambda-1}.$$

Suppose that we wish to allow for a complex design, but retain $\boldsymbol{\varphi}^*$ as the vector of unknown parameter of the finite population. Skinner, *et. al.* (1989) redefines $\boldsymbol{\varphi}^*$ as that value of $\tilde{\boldsymbol{\varphi}}^*$ which maximizes $l(\tilde{\boldsymbol{\varphi}}^*) = \sum_{i \in U} \log f(y_i, \tilde{\boldsymbol{\varphi}}^*)$, where the sum is over all units in the finite population. Thus, among all possible models $f(y_i, \tilde{\boldsymbol{\varphi}}^*)$, the one which “best fits” the finite population is chosen. If we choose the $f(y_i, \tilde{\boldsymbol{\varphi}}^*)$ family poorly, this best fit will still be poor, but our inference treats it as the target we are trying to hit with our sample data. Thus it is important to select appropriate choices for $f(y_i, \tilde{\boldsymbol{\varphi}}^*)$.

For the finite population, $\boldsymbol{\varphi}^*$ satisfies

$$\dot{l}_U(\boldsymbol{\varphi}^*) = \sum_{i \in U} [\partial \log f(y_i; \boldsymbol{\varphi}^*, \tilde{y}) / \partial \boldsymbol{\varphi}^*] = 0,$$

where $\tilde{y} = \prod_{i=1}^N y_i^{1/N}$. For a given $\boldsymbol{\varphi}^*$, let $\dot{l}_U(\boldsymbol{\varphi}^*)$, summation of the first derivative of the log-likelihood with respect to $\boldsymbol{\varphi}^*$, be a finite population parameter. We take a sample, and, by approximating $\log f(y_i; \boldsymbol{\varphi}^*, \tilde{y})$ for each unit i in the sample by $\log f(y_i; \boldsymbol{\varphi}^*, \tilde{y}_w)$, we estimate the population total, $\dot{l}_U(\boldsymbol{\varphi}^*)$, by $\dot{l}_s(\hat{\boldsymbol{\varphi}}_{PML}^*)$:

$$\dot{l}_s(\hat{\boldsymbol{\varphi}}_{PML}^*) = \sum_{i \in s} w_i \left[\partial \log f(y_i; \boldsymbol{\varphi}^*, \tilde{y}_w) / \partial \boldsymbol{\varphi}^* \right]_{\boldsymbol{\varphi}^* = \hat{\boldsymbol{\varphi}}_{PML}^*}, \quad (2.10)$$

where $\tilde{y}_w = \prod_{i \in s} y_i^{w_i / \sum w_i}$, the weighted geometric mean of y 's in the sample, and $\hat{\boldsymbol{\varphi}}_{PML}^*$ is the pseudo maximum likelihood estimator of $\boldsymbol{\varphi}^*$, satisfying $\dot{l}_s(\hat{\boldsymbol{\varphi}}_{PML}^*) = 0$. The PML estimator, $\hat{\boldsymbol{\varphi}}_{PML}^* = (\hat{\boldsymbol{\beta}}_w^*, \hat{\lambda}_w, \hat{\sigma}_{e,w}^{*2})'$, can be obtained by grid search method. That is, calculating and plotting the weighted log likelihood values,

$$\log L(\boldsymbol{\varphi}^*) = \sum_{i \in s} w_i \log f(y_i; \boldsymbol{\varphi}^*, \tilde{y}_w) \quad (2.11)$$

against the set of values for λ will locate the PML estimate, $\hat{\lambda}_w$, of the transformation parameter. When we evaluate the log-likelihood function at each fixed value of λ in the sampling context, $\boldsymbol{\beta}^*$ and σ_e^{2*} are estimated by incorporating the sampling weights as:

$$\hat{\boldsymbol{\beta}}_w^* = \left(\sum_{i \in s} w_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i \in s} w_i \mathbf{x}_i y_i^{*(\lambda)} \right), \text{ and}$$

$$\hat{\sigma}_{e,w}^{*2} = \sum_{i \in s} w_i (y_i^{*(\lambda)} - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_w^*)^2 / \sum_{i \in s} w_i.$$

Since the model \mathbf{M}_3 is of the interest, converting $\hat{\boldsymbol{\varphi}}_{PML}^*$ that maximize (2.11) back to

$\hat{\boldsymbol{\varphi}}_{PML}$ in the model \mathbf{M}_3 is necessary and $\hat{\boldsymbol{\beta}}_w = \tilde{y}_w^{\hat{\lambda}-1} \hat{\boldsymbol{\beta}}_w^*$, $\hat{\sigma}_{e,w}^2 = \tilde{y}_w^{2(\hat{\lambda}-1)} \hat{\sigma}_{e,w}^{*2}$.

In obtaining the PML estimators for $\boldsymbol{\varphi}$, the development of its variance covariance matrix would be ideal as well. A robust estimator which a) recognizes the covariance structure caused by the complex sample design; and b) is robust to misspecified density d_i (i.e., the variance is right even if d_i is not a good description of the data) is the linearization estimator, proposed by Royall (1986). This estimator can be extended naturally to estimate $Var(\hat{\boldsymbol{\varphi}}_{\text{PML}}^*)$

$$\text{var}_L(\hat{\boldsymbol{\varphi}}_{\text{PML}}^*) = I(\hat{\boldsymbol{\varphi}}_{\text{PML}}^*)^{-1} \text{var}_L[\hat{T}(\hat{\boldsymbol{\varphi}}_{\text{PML}}^*)]I(\hat{\boldsymbol{\varphi}}_{\text{PML}}^*)^{-1},$$

where

$$\hat{T}(\hat{\boldsymbol{\varphi}}_{\text{PML}}^*) = \left\{ \begin{array}{l} \partial \log L(\boldsymbol{\varphi}^*) / \partial \boldsymbol{\beta}^* \\ \partial \log L(\boldsymbol{\varphi}^*) / \partial \lambda \\ \partial \log L(\boldsymbol{\varphi}^*) / \partial \sigma_e^{*2} \end{array} \right\}_{\boldsymbol{\varphi}^* = \hat{\boldsymbol{\varphi}}_{\text{PML}}^*},$$

$$\hat{T}(\hat{\boldsymbol{\varphi}}_{\text{PML}}^*) = \left\{ \begin{array}{l} \sigma_e^{*(-2)} \sum_{i \in s} w_i (y_i^{*(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}^*) \mathbf{x}_i \\ -\sigma_e^{*(-2)} \sum_{i \in s} w_i (y_i^{*(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}^*) (\partial y_i^{*(\lambda)} / \partial \lambda) \\ -1/2\sigma_e^{*(-2)} \sum_{i \in s} w_i + 1/2\sigma_e^{*(-4)} \sum_{i \in s} w_i (y_i^{*(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}^*)^2 \end{array} \right\}_{\boldsymbol{\varphi}^* = \hat{\boldsymbol{\varphi}}_{\text{PML}}^*},$$

$$I(\hat{\boldsymbol{\varphi}}_{\text{PML}}^*) = -\left(\partial \hat{T}(\boldsymbol{\varphi}^*) / \partial \boldsymbol{\varphi}^* \right)_{\boldsymbol{\varphi}^* = \hat{\boldsymbol{\varphi}}_{\text{PML}}^*} = \left[\begin{array}{ccc} I_{11} & I_{12} & I_{13} \\ I_{12}' & I_{22} & I_{23} \\ I_{13}' & I_{23}' & I_{33} \end{array} \right]_{\boldsymbol{\varphi}^* = \hat{\boldsymbol{\varphi}}_{\text{PML}}^*},$$

where

$$\left\{ \begin{array}{l} I_{11} = \sigma_e^{*(-2)} \sum_{i \in s} w_i \mathbf{x}_i \mathbf{x}_i' \\ I_{12} = -\sigma_e^{*(-2)} \sum_{i \in s} w_i (\partial y_i^{*(\lambda)} / \partial \lambda) \mathbf{x}_i \\ I_{13} = \sigma_e^{*(-4)} \sum_{i \in s} w_i (y_i^{*(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}^*) \mathbf{x}_i \\ I_{22} = \sigma_e^{*(-2)} \sum_{i \in s} w_i \left[(y_i^{*(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}^*) (\partial^2 y_i^{*(\lambda)} / \partial \lambda^2) + (\partial y_i^{*(\lambda)} / \partial \lambda)^2 \right] \\ I_{23} = -\sigma_e^{*(-4)} \sum_{i \in s} w_i (y_i^{*(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}^*) (\partial y_i^{*(\lambda)} / \partial \lambda) \\ I_{33} = -(2\sigma_e^{*4})^{-1} \sum_{i \in s} w_i + (\sigma_e^{*6})^{-1} \sum_{i \in s} w_i (y_i^{*(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}^*)^2 \end{array} \right.$$

and

$$\begin{aligned} \partial y_i^{*(\lambda)} / \partial \lambda &= \lambda^{-1} \tilde{y}^{1-\lambda} s_i^{(\lambda)} - [\lambda^{-1} + \log(\tilde{y})] y_i^{*(\lambda)} \\ \partial^2 y_i^{*(\lambda)} / \partial \lambda^2 &= \lambda^{-1} \tilde{y}^{1-\lambda} [t_i^{(\lambda)} - 2\{\lambda^{-1} + \log(\tilde{y})\} s_i^{(\lambda)}] + [\{\lambda^{-1} + \log(\tilde{y})\}^2 + \lambda^{-2}] y_i^{*(\lambda)} \end{aligned}$$

$s^{(\lambda)}$ and $t^{(\lambda)}$ are $n \times 1$ vectors such that the i^{th} element is $s_i^{(\lambda)} = y_i^\lambda \log(y_i)$ and $t_i^{(\lambda)} = y_i^\lambda (\log(y_i))^2$, respectively.

Note that $\text{var}_L(\hat{\boldsymbol{\Phi}}_{\text{PML}}^*)$ is a sandwich estimator. The middle matrix, $\text{var}_L[\hat{T}(\hat{\boldsymbol{\Phi}}_{\text{PML}}^*)]$, can be estimated by linearization, balanced repeated replication (BRR) method, or jackknife method. For example, using jackknife method, we have:

$$\text{var}_L[\hat{T}(\hat{\boldsymbol{\Phi}}_{\text{PML}}^*)] = \sum_{h=1}^H l_h (l_h - 1)^{-1} \sum_{d=1}^{l_h} (z_{hd} - \bar{z}_h)(z_{hd} - \bar{z}_h)',$$

where

$$z_{hd} = \left\{ \begin{array}{l} \sigma_e^{*(-2)} \sum_{i \in shd} w_i (y_i^{*(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}^*) \mathbf{x}_i \\ -\sigma_e^{*(-2)} \sum_{i \in shd} w_i (y_i^{*(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}^*) (\partial y_i^{*(\lambda)} / \partial \lambda) \\ -1/2\sigma_e^{*(-2)} \sum_{i \in shd} w_i + 1/2\sigma_e^{*(-4)} \sum_{i \in shd} w_i (y_i^{*(\lambda)} - \mathbf{x}_i' \boldsymbol{\beta}^*)^2 \end{array} \right\}_{\boldsymbol{\varphi}^* = \hat{\boldsymbol{\varphi}}_{\text{PML}}}$$

The notation $\sum_{i \in shd}$ represents the sum over sample units in PSU $d(=1, \dots, l_h)$ in stratum h , and l_h denotes the number of the sampled PSUs in the stratum h . For example, in the stratified simple random sampling,

$$\text{var}_L[\hat{T}(\hat{\boldsymbol{\varphi}}_{\text{PML}}^*)] = \sum_{h=1}^H n_h (n_h - 1)^{-1} \sum_{k=1}^{n_h} (z_{hk} - \bar{z}_h)(z_{hk} - \bar{z}_h)', \text{ and}$$

$$z_{hk} = \left\{ \begin{array}{l} \sigma_e^{*(-2)} w_{hk} (y_{hk}^{*(\lambda)} - \mathbf{x}_{hk}' \boldsymbol{\beta}^*) \mathbf{x}_{hk} \\ -\sigma_e^{*(-2)} w_{hk} (y_{hk}^{*(\lambda)} - \mathbf{x}_{hk}' \boldsymbol{\beta}^*) (\partial y_{hk}^{*(\lambda)} / \partial \lambda) \\ -1/2\sigma_e^{*(-2)} w_{hk} + 1/2\sigma_e^{*(-4)} w_{hk} (y_{hk}^{*(\lambda)} - \mathbf{x}_{hk}' \boldsymbol{\beta}^*)^2 \end{array} \right\}_{\boldsymbol{\varphi}^* = \hat{\boldsymbol{\varphi}}_{\text{PML}}},$$

where n_h denotes the number of the sampled units in the stratum h . The design-consistency of $\text{var}_L(\hat{\boldsymbol{\varphi}}_{\text{PML}}^*)$ does not depend on the assumption that d_i is the true probability density function (Royall, 1986).

In order to obtain $\text{var}(\hat{\boldsymbol{\varphi}}_{\text{PML}}^*)$ from $\text{var}(\hat{\boldsymbol{\varphi}}_{\text{PML}}^*)$ we can apply the fact that

$$\hat{\boldsymbol{\varphi}}_{\text{PML}} = (\hat{\boldsymbol{\beta}}_w, \hat{\lambda}_w, \hat{\sigma}_{e,w}^2)' = (\tilde{y}_w^{\lambda-1} \hat{\boldsymbol{\beta}}_w^*, \hat{\lambda}_w, \tilde{y}_w^{2(\lambda-1)} \hat{\sigma}_{e,w}^{*2})',$$

and:

$$\text{var}(\hat{\boldsymbol{\varphi}}_{\text{PML}}) = \mathbf{J} \text{var}(\hat{\boldsymbol{\varphi}}_{\text{PML}}^*) \mathbf{J}'$$

where

$$\mathbf{J} = \begin{pmatrix} \tilde{y}_w^{\hat{\lambda}_w-1} & 0 & \cdots & 0 & \hat{\beta}_{0,w} \log(\tilde{y}_w) & 0 \\ 0 & \tilde{y}_w^{\hat{\lambda}_w-1} & \cdots & 0 & \hat{\beta}_{1,w} \log(\tilde{y}_w) & 0 \\ 0 & 0 & \ddots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \cdots & \tilde{y}_w^{\hat{\lambda}_w-1} & \hat{\beta}_{(p-1),w} \log(\tilde{y}_w) & 0 \\ \vdots & \vdots & \cdots & 0 & 1 & 0 \\ 0 & 0 & \cdots & 0 & 2\hat{\sigma}_{e,w}^2 \log(\tilde{y}_w) & \tilde{y}_w^{2(\hat{\lambda}_w-1)} \end{pmatrix}.$$

2.3.2 AUTOGREG estimator of population total and its variance estimator

In Subsection 2.3.1, we obtain the PML estimator of $\boldsymbol{\varphi}$, along with its variance estimator. It is, therefore, possible for us to estimate the finite population total T by the AUTOGREG estimator defined by (2.9), and the predicted value $\hat{y}_{i,w}$ is obtained by:

$$\hat{y}_{i,w} = g(\hat{\boldsymbol{\beta}}_w, \hat{\lambda}_w) = (\hat{\lambda}_w \mathbf{x}_i' \hat{\boldsymbol{\beta}}_w + 1)^{1/\hat{\lambda}_w},$$

where $\hat{\lambda}_w$ and $\hat{\boldsymbol{\beta}}_w$ are the PML estimators under the model \mathbf{M}_3 . Note that this predictor is produced by simple back-transformation from the Box-Cox transformation.

It is well-known that the GREG estimator has the nice property of design-consistency (Särndal, *et. al.* 1992). For AUTOGREG estimator, this property is maintained.

Write

$$\begin{aligned} \hat{T}_{AG} &= \sum_U \hat{y}_{i,w} + \sum_s (y_i - \hat{y}_{i,w}) / \pi_i, \\ &= \sum_{i \in s} y_i / \pi_i + \left(\sum_{i \in U} \hat{y}_{i,w} - \sum_{i \in s} \hat{y}_{i,w} / \pi_i \right), \end{aligned}$$

By Taylor Series expansion, we have

$$\hat{y}_{i,w} \approx g_i(\mathbf{B}, \Lambda) + \left(\partial g_i(\hat{\boldsymbol{\beta}}_w, \hat{\lambda}_w) / \partial \hat{\boldsymbol{\beta}}_w \right)' (\hat{\boldsymbol{\beta}}_w - \mathbf{B}) + \left(\partial g_i(\hat{\boldsymbol{\beta}}_w, \hat{\lambda}_w) / \partial \hat{\lambda}_w \right) (\hat{\lambda}_w - \Lambda),$$

where $\partial g_i(\hat{\boldsymbol{\beta}}_w, \hat{\lambda}_w) / \partial \hat{\boldsymbol{\beta}}_w$ is a $(k+1) \times 1$ column vector, \mathbf{B} and Λ are finite population parameters. $\hat{\boldsymbol{\beta}}_w$ and $\hat{\lambda}_w$ are design-consistent estimators of \mathbf{B} and Λ , i.e.,

$$\hat{\boldsymbol{\beta}}_w \rightarrow \mathbf{B} \text{ in probability and } \hat{\lambda}_w \rightarrow \Lambda \text{ in probability}$$

under certain regularity conditions using the argument similar to Binder (1983).

Therefore,

$$\begin{aligned} \hat{T}_{AG} &= \sum_{i \in s} y_i / \pi_i + \left(\sum_{i \in U} \hat{y}_{i,w} - \sum_{i \in s} \hat{y}_{i,w} / \pi_i \right) \\ &\approx \sum_{i \in s} y_i / \pi_i + \sum_{i \in U} \left\{ g_i(\mathbf{B}, \Lambda) + \left(\partial g_i(\hat{\boldsymbol{\beta}}_w, \hat{\lambda}_w) / \partial \hat{\boldsymbol{\beta}}_w \right)_{\hat{\boldsymbol{\beta}}_w = \mathbf{B}, \hat{\lambda}_w = \Lambda} (\hat{\boldsymbol{\beta}}_w - \mathbf{B}) + \right. \\ &\quad \left. \left(\partial g_i(\hat{\boldsymbol{\beta}}_w, \hat{\lambda}_w) / \partial \hat{\lambda}_w \right)_{\hat{\boldsymbol{\beta}}_w = \mathbf{B}, \hat{\lambda}_w = \Lambda} (\hat{\lambda}_w - \Lambda) \right\} - \sum_{i \in s} (\pi_i)^{-1} \left\{ g_i(\mathbf{B}, \Lambda) + \right. \\ &\quad \left. \left(\partial g_i(\hat{\boldsymbol{\beta}}_w, \hat{\lambda}_w) / \partial \hat{\boldsymbol{\beta}}_w \right)_{\hat{\boldsymbol{\beta}}_w = \mathbf{B}, \hat{\lambda}_w = \Lambda} (\hat{\boldsymbol{\beta}}_w - \mathbf{B}) + \left(\partial g_i(\hat{\boldsymbol{\beta}}_w, \hat{\lambda}_w) / \partial \hat{\lambda}_w \right)_{\hat{\boldsymbol{\beta}}_w = \mathbf{B}, \hat{\lambda}_w = \Lambda} (\hat{\lambda}_w - \Lambda) \right\} \\ &= \sum_{i \in s} y_i / \pi_i + \left\{ \sum_{i \in U} g_i(\mathbf{B}, \Lambda) - \sum_{i \in s} (\pi_i)^{-1} g_i(\mathbf{B}, \Lambda) \right\} + \\ &\quad \left\{ \sum_{i \in U} \partial g_i(\hat{\boldsymbol{\beta}}_w, \hat{\lambda}_w) / \partial \hat{\boldsymbol{\beta}}_w - \sum_{i \in s} (\pi_i)^{-1} \partial g_i(\hat{\boldsymbol{\beta}}_w, \hat{\lambda}_w) / \partial \hat{\boldsymbol{\beta}}_w \right\}_{\hat{\boldsymbol{\beta}}_w = \mathbf{B}, \hat{\lambda}_w = \Lambda} (\hat{\boldsymbol{\beta}}_w - \mathbf{B}) + \\ &\quad \left\{ \sum_{i \in U} \partial g_i(\hat{\boldsymbol{\beta}}_w, \hat{\lambda}_w) / \partial \hat{\lambda}_w - \sum_{i \in s} (\pi_i)^{-1} \partial g_i(\hat{\boldsymbol{\beta}}_w, \hat{\lambda}_w) / \partial \hat{\lambda}_w \right\}_{\hat{\boldsymbol{\beta}}_w = \mathbf{B}, \hat{\lambda}_w = \Lambda} (\hat{\lambda}_w - \Lambda) \end{aligned}$$

Isaki and Fuller (1982) give sufficient conditions for the Horvitz-Thompson estimator to be design consistent. Under certain sufficient conditions:

$$N^{-1} \sum_{i \in s} (\pi_i)^{-1} g_i(\mathbf{B}, \Lambda) = N^{-1} \sum_{i \in U} g_i(\mathbf{B}, \Lambda) + O_p(1/\sqrt{n});$$

$$N^{-1} \sum_{i \in s} (\pi_i)^{-1} \left(\partial g_i / \partial \hat{\boldsymbol{\beta}}_w \right)_{\hat{\boldsymbol{\beta}}_w = \mathbf{B}, \hat{\lambda}_w = \Lambda} = N^{-1} \sum_{i \in U} \left(\partial g_i / \partial \hat{\boldsymbol{\beta}}_w \right)_{\hat{\boldsymbol{\beta}}_w = \mathbf{B}, \hat{\lambda}_w = \Lambda} + O_p(1/\sqrt{n});$$

$$N^{-1} \sum_{i \in s} (\pi_i)^{-1} \left(\partial g_i / \partial \hat{\lambda}_w \right)_{\hat{\boldsymbol{\beta}}_w = \mathbf{B}, \hat{\lambda}_w = \Lambda} = N^{-1} \sum_{i \in U} \left(\partial g_i / \partial \hat{\lambda}_w \right)_{\hat{\boldsymbol{\beta}}_w = \mathbf{B}, \hat{\lambda}_w = \Lambda} + O_p(1/\sqrt{n}).$$

Eventually,

$$N^{-1}\hat{T}_{AG} = N^{-1}\sum_{i \in S} y_i / \pi_i + O_p(1/\sqrt{n}).$$

The design-expectation of $N^{-1}(\hat{T}_{AG} - T)$ is

$$E_d \left\{ N^{-1}(\hat{T}_{AG} - T) \right\} = O(1/\sqrt{n});$$

Since the design variance for the Horvitz-Thompson estimator has the order of $1/n$,

$$V_d \left\{ N^{-1} \left(\sum_{i \in S} y_i / \pi_i - T \right) \right\} = O(1/n) \text{ under some regularity conditions (Isaki and Fuller,}$$

1982), the variance of the AUTOGREG estimator

$$V_d \left\{ N^{-1}(\hat{T}_{AG} - T) \right\} = V_d \left\{ N^{-1} \left(\sum_{i \in S} y_i / \pi_i - T \right) + O(1/\sqrt{n}) \right\}$$

has the order of $O(1/n)$.

Therefore, the estimator $N^{-1}(\hat{T}_{AG} - T)$ is design-consistent, and

$$N^{-1}(\hat{T}_{AG} - T) = O_p(1/\sqrt{n}).$$

The design expectation of \hat{T}_{AG} is approximately T .

The design variance of \hat{T}_{AG} is

$$\begin{aligned} V_d(\hat{T}_{AG}) &\approx V_d \left\{ \sum_{i \in S} y_i / \pi_i + \left[\sum_{i \in U} g_i(\mathbf{B}, \Lambda) - \sum_{i \in S} g_i(\mathbf{B}, \Lambda) / \pi_i \right] \right\} \\ &= V_d \left\{ \sum_{i \in S} (y_i - g_i(\mathbf{B}, \Lambda)) / \pi_i \right\} \\ &= \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{(y_i - g_i(\mathbf{B}, \Lambda))(y_j - g_j(\mathbf{B}, \Lambda))}{\pi_i \pi_j}. \end{aligned}$$

A variance estimator is given by

$$\hat{V}_d(\hat{T}_{AG}) = \sum_{i \in S} \sum_{j \in S} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}} \frac{(y_i - g_i(\hat{\mathbf{B}}_w, \hat{\lambda}_w))(y_j - g_j(\hat{\mathbf{B}}_w, \hat{\lambda}_w))}{\pi_i \pi_j}.$$

2.4 A Simulation study

The purpose of this simulation study is to evaluate different estimators of a finite population total for varying values of the sample size n and the standard deviation σ . In this simulation exercise, a finite population from the Australian Agricultural and Grazing Industries Survey (AAGIS) is generated. This survey data contains information on the number of cattle (y) and farm area (x) for each of the 431 farms.

We consider a finite population of size $N=4,000$ that is generated from the following model:

$$y_i^{(\lambda)} = (y_i^\lambda - 1) / \lambda = \beta_0 + \beta_1 \times \log(x_i) + \varepsilon_i,$$

where ε_i 's are independent with $N(0, \sigma^2)$, and x_i is generated from an exponential distribution with mean μ_x and standard error σ_x . In order to mimic a true situation, we choose $\lambda=0.1$, $\beta_0=4.20$, and $\beta_1=2.66$ which are the estimates obtained by fitting the real survey data to the model \mathbf{M}_3 , and $\mu_x=1,040$, $\sigma_x=1,000$ to ensure $y_i > 0$ for each unit i . Strictly speaking, truncated normal distribution of y is considered, and all negative values of y generated are discarded. The effect of this is negligible since less than 0.1% of the generated y values are discarded. The same phenomena were found by Taylor (1986).

Simulation is based on repeated sampling from the generated finite population. Two sample designs are investigated: simple random sampling (SRS) and stratified SRS (SSRS). When a sample is selected by SSRS, unequal selection probabilities among different strata are applied. We define two strata using the boundary value: median of y values in the finite population. For stratum h of size N_h , a simple random sample of size

n_h is selected. Define p_1 and p_2 selection probabilities for stratum 1 and stratum 2, respectively. We specify $p_1 = 2 \times p_2$. For fixed sample size n , $n_1 = N_1 \times p_1$, and $n_2 = N_2 \times p_2$.

We are interested in estimating the finite population total $T = \sum_{i \in U} y_i$. In this simulation, we study the performance of AUTOGREG estimator (\hat{T}_{AG}), defined by (2.9), along with the design-based estimator (\hat{T}_D), defined by (2.1), model-based estimators (\hat{T}_{M-L} , \hat{T}_{M-LOGL} , and \hat{T}_{M-BC}), defined by (2.2)—(2.5), and GREG estimators (\hat{T}_{G-L} and \hat{T}_{G-LOGL}), defined by (2.6)—(2.8), under different models.

One thousand samples are selected from the simulated finite population for each of the sample size $n \in (30, 50, 80, 100, 130, 150)$. Seven estimators (\hat{T}_D , three \hat{T}_M 's, two \hat{T}_G 's, and \hat{T}_{AG}) are produced using each selected sample. Estimator for the finite population transformation parameter Λ is also produced using each sample. For the purpose of comparison, two methods are used to estimate Λ . Let $\hat{\lambda}$ and $\hat{\lambda}_w$ be the OLS/ML and PML estimators of Λ , respectively. Over all the 1,000 samples, we compute the empirical percentage relative biases (*RelBias*) and root mean square errors (*rmse*) to evaluate these estimators. *RelBias* is defined as the average over the samples of $(\hat{\omega} - \omega) / \omega$, where $\hat{\omega}$ represents an arbitrary estimators of the finite population parameter ω , and *rmse* is the square root of the average over the samples of $(\hat{\omega} - \omega)^2$:

$$RelBias = B^{-1} \sum_{b=1}^B (\hat{\omega}_b - \omega) / \omega, \text{ and } rmse = \sqrt{B^{-1} \sum_{b=1}^B (\hat{\omega}_b - \omega)^2},$$

where B is the number of the replications in the Monte Carlo simulation.

2.5 Results

In Table 2.1 we present the *RelBias* of seven estimators using different sampling designs with varying sample sizes when $\sigma=0.5$. All seven estimators give *RelBias* close to zero (maximum of the absolute values of *RelBias* in Table 2.1 is 0.03). Among them, AUTOGREG estimator \hat{T}_{AG} has the smallest *RelBias* over different sampling sizes and sampling designs. For SRS sampling, as we expected, model-based estimator \hat{T}_{M-L} and GREG estimator \hat{T}_{G-L} under the standard linear model are identical; whereas for SSRS sampling, the *RelBias* of \hat{T}_{G-L} is closer to zero than that of \hat{T}_{M-L} . This is because GREG estimators take account of sampling weights, and for unequal selection probability sampling design, GREG estimators is approximately unbiased when sample size approaches to infinity; while this is not true for the model-based estimator \hat{T}_{M-L} . This nice property of GREG estimators can also be observed when we compare \hat{T}_{M-LOGL} vs. \hat{T}_{G-LOGL} and \hat{T}_{M-BC} vs. \hat{T}_{AG} for both sampling designs. Thus, compared to the model-based estimators, GREG and AUTOGREG estimators are protected from possible model failure.

Table 2.2 reports the *rmse* of seven estimators with varying sample sizes when $\sigma=0.5$. We note that the estimator \hat{T}_D has the largest *rmse* over different sample sizes and sampling designs. Again, \hat{T}_{M-L} and \hat{T}_{G-L} are same for SRS. Two estimators \hat{T}_{M-BC} and \hat{T}_{AG} based on the Box-Cox model perform equally well, and \hat{T}_{AG} is slightly better than \hat{T}_{M-BC} as sample size increases. Compared to the GREG estimators (\hat{T}_{G-L} and \hat{T}_{G-LOGL}), AUTOGREG estimator \hat{T}_{AG} has smaller *RelBias* and *rmse*. This result implies

that AUTOGREG estimator is protected by the Box-Cox technique and achieves a smaller *RelBias* and *rmse*. Thus Tables 2.1 and 2.2 nicely demonstrate double robustness property of the AUTOGREG estimator (\hat{T}_{AG}) discussed in Section 2.3.

Tables 2.3 – 2.6 show the *RelBias* and *rmse* under the same conditions when $\sigma = 1$ and $\sigma = 2$. The same patterns can be observed as those in Tables 2.1 and 2.2. One point worth to notice is that the superiority of \hat{T}_{AG} over \hat{T}_{M-BC} is more obvious because absolute *RelBias* and *rmse* of \hat{T}_{AG} are becoming smaller than those of \hat{T}_{M-BC} as σ increases (see, for example, column 5 vs. column 8 in Tables 2.5 and 2.6).

In order to have a close look at the performance of different estimators with different sample sizes and standard deviations, we plot Figures 2.1 – 2.2. Figure 2.1 presents the *rmse* for GREG estimators (\hat{T}_{G-L} and \hat{T}_{G-LOGL}) and AUTOGREG estimator \hat{T}_{AG} using different sampling designs. We note that \hat{T}_{AG} consistently has the smallest *rmse* when $\sigma = 0.5$ and 1; when σ is large, \hat{T}_{G-LOGL} and \hat{T}_{AG} perform equally well as sample size increases. Thus, a robust model chosen by the Box-Cox method brings about the *rmse* reduction, especially when σ is small. The comparison between *rmse*'s of \hat{T}_{M-BC} vs. \hat{T}_{AG} is also investigated in Figure 2.2. When σ is small, \hat{T}_{M-BC} and \hat{T}_{AG} are similar and both predictors tend to zero. For large σ , however, the *rmse* of \hat{T}_{AG} has the tendency to zero; whereas the *rmse* of \hat{T}_{M-BC} has not. This result implies that regardless of correctness of the model, design-consistency of \hat{T}_{AG} is maintained, but this might not be true for \hat{T}_{M-BC} .

Table 2.7 presents the *RelBias* and *rmse* of $\hat{\lambda}$ and $\hat{\lambda}_w$ for SSRS sampling with varying sample sizes and standard deviations. When σ is small, that is, when the simulated data are well fitted to the assumed model, $\hat{\lambda}_w$ gives *RelBias* closer to zero, but $\hat{\lambda}_w$ and $\hat{\lambda}$ perform equally well in terms of the *rmse*. When σ is large, however, $\hat{\lambda}_w$ consistently gives smaller absolute values of *RelBias* and *rmse*, as compared to $\hat{\lambda}$.

2.6 Concluding remarks

In this chapter, we consider a new adjustment to the generalized regression estimator. The proposed new estimator possesses the nice property of double robustness: 1) Design-consistency even under the failure of the underlying model; and 2) Variance reduction when the appropriate model is automatically chosen by the data using the Box-Cox technique. This property is evaluated analytically and via Monte Carlo simulation study. Extension of our method to incorporate clustering effect needs further investigation.

Table 2.1: Relative biases of the different estimators using different sampling designs
with varying sample sizes ($\sigma = 0.5$)

	\hat{T}_D ^{1.}	\hat{T}_{M-L} ^{2.}	\hat{T}_{M-LOGL} ^{2.}	\hat{T}_{M-BC} ^{2.}	\hat{T}_{G-L} ^{3.}	\hat{T}_{G-LOGL} ^{3.}	\hat{T}_{AG} ^{4.}
Simple random sampling ($\times 0.001$)							
n=30	4.37	-9.82	8.12	-5.50	-9.82	5.46	3.62
n=50	-1.51	-8.75	5.49	-9.29	-8.75	2.00	0.14
n=80	1.43	-3.67	4.86	-9.15	-3.67	1.24	0.65
n=100	-0.64	-5.19	5.03	-9.53	-5.19	1.30	0.32
n=130	-2.32	-3.98	4.92	-9.26	-3.98	1.23	0.38
n=150	-2.60	-1.24	5.02	-9.26	-1.24	1.22	0.53
Stratified simple random sampling ($\times 0.001$)							
n=30	6.01	-34.71	8.86	-13.80	-5.82	3.84	1.92
n=50	18.02	-29.71	7.72	-14.70	-1.73	2.49	1.63
n=80	9.93	-30.25	8.78	-14.80	-1.01	3.04	0.98
n=100	-7.07	-33.71	8.43	-16.05	-2.92	2.16	0.55
n=130	-5.40	-33.24	6.99	-17.46	-2.65	0.60	-0.30
n=150	1.75	-32.02	7.34	-17.06	-1.85	1.06	0.42

^{1.} \hat{T}_D is design-based estimator.

^{2.} \hat{T}_{M-L} , \hat{T}_{M-LOGL} , and \hat{T}_{M-BC} are model-based estimators based on standard linear model, log-linear model, and Box-Cox model, respectively.

^{3.} \hat{T}_{G-L} and \hat{T}_{G-LOGL} are GREG estimators based on standard linear model and log-linear model.

^{4.} \hat{T}_{AG} is AUTOGREG estimator based on Box-Cox model.

Table 2.2: Root mean square errors of the different estimators using different sampling designs with varying sample sizes ($\sigma=0.5$)

	\hat{T}_D ^{1.}	\hat{T}_{M-L} ^{2.}	\hat{T}_{M-LOGL} ^{2.}	\hat{T}_{M-BC} ^{2.}	\hat{T}_{G-L} ^{3.}	\hat{T}_{G-LOGL} ^{3.}	\hat{T}_{AG} ^{4.}
Simple random sampling ($\times 10^7$)							
n=30	7.17	3.54	2.02	1.92	3.54	2.02	1.94
n=50	5.69	2.71	1.47	1.39	2.71	1.44	1.33
n=80	4.26	2.09	1.16	1.16	2.09	1.18	1.09
n=100	4.13	1.89	1.07	1.09	1.89	1.08	0.99
n=130	3.48	1.57	0.91	0.94	1.57	0.91	0.82
n=150	3.20	1.58	0.88	0.91	1.58	0.88	0.79
Stratified simple random sampling ($\times 10^7$)							
n=30	5.63	4.04	2.21	2.19	3.67	2.29	2.11
n=50	4.47	3.20	1.70	1.75	2.84	1.80	1.63
n=80	3.51	2.78	1.46	1.45	2.31	1.43	1.28
n=100	3.04	2.62	1.29	1.38	1.98	1.28	1.14
n=130	2.81	2.51	1.13	1.35	1.83	1.14	1.03
n=150	2.49	2.28	1.03	1.25	1.59	1.01	0.90

^{1.} \hat{T}_D is design-based estimator.

^{2.} \hat{T}_{M-L} , \hat{T}_{M-LOGL} , and \hat{T}_{M-BC} are model-based estimators based on standard linear model, log-linear model, and Box-Cox model, respectively.

^{3.} \hat{T}_{G-L} and \hat{T}_{G-LOGL} are GREG estimators based on standard linear model and log-linear model.

^{4.} \hat{T}_{AG} is AUTOGREG estimator based on Box-Cox model.

Table 2.3: Relative biases of the different estimators using different sampling designs with varying sample sizes ($\sigma = 1$)

	\hat{T}_D ^{1.}	\hat{T}_{M-L} ^{2.}	\hat{T}_{M-LOGL} ^{2.}	\hat{T}_{M-BC} ^{2.}	\hat{T}_{G-L} ^{3.}	\hat{T}_{G-LOGL} ^{3.}	\hat{T}_{AG} ^{4.}
Simple random sampling ($\times 0.001$)							
n=30	5.01	-10.87	-16.87	-30.46	-10.87	6.10	3.20
n=50	4.62	-6.19	-20.42	-33.34	-6.19	3.11	1.66
n=80	-2.57	-5.54	-21.66	-34.38	-5.54	2.60	1.24
n=100	-2.11	-3.64	-22.14	-34.94	-3.64	1.24	0.51
n=130	-2.18	-2.89	-23.35	-36.42	-2.89	-0.79	-1.34
n=150	-1.52	-1.91	-21.85	-34.48	-1.91	1.94	1.06
Stratified simple random sampling ($\times 0.001$)							
n=30	-4.37	-54.83	-42.30	-59.14	-10.01	3.22	-0.41
n=50	17.90	-45.69	-38.68	-58.15	-1.83	6.01	3.92
n=80	8.82	-48.34	-42.27	-61.39	-3.53	2.66	1.20
n=100	-7.95	-49.37	-43.37	-62.62	-2.23	2.17	0.95
n=130	-7.40	-50.84	-45.94	-65.44	-3.98	-0.84	-2.22
n=150	-1.66	-49.86	-45.40	-64.93	-3.11	-0.06	-1.16

^{1.} \hat{T}_D is design-based estimator.

^{2.} \hat{T}_{M-L} , \hat{T}_{M-LOGL} , and \hat{T}_{M-BC} are model-based estimators based on standard linear model, log-linear model, and Box-Cox model, respectively.

^{3.} \hat{T}_{G-L} and \hat{T}_{G-LOGL} are GREG estimators based on standard linear model and log-linear model.

^{4.} \hat{T}_{AG} is AUTOGREG estimator based on Box-Cox model.

Table 2.4: Root mean squared errors of the different estimators using different sampling designs with varying sampling sizes ($\sigma=1$)

	\hat{T}_D ^{1.}	\hat{T}_{M-L} ^{2.}	\hat{T}_{M-LOGL} ^{2.}	\hat{T}_{M-BC} ^{2.}	\hat{T}_{G-L} ^{3.}	\hat{T}_{G-LOGL} ^{3.}	\hat{T}_{AG} ^{4.}
Simple random sampling ($\times 10^7$)							
n=30	7.86	4.54	3.71	3.82	4.54	3.71	3.64
n=50	6.20	3.52	3.02	3.17	3.52	2.86	2.77
n=80	4.82	2.74	2.39	2.70	2.74	2.17	2.13
n=100	4.36	2.59	2.28	2.63	2.59	2.04	1.99
n=130	3.81	2.18	2.08	2.48	2.18	1.73	1.66
n=150	3.57	1.99	1.94	2.33	1.99	1.58	1.54
Stratified simple random sampling ($\times 10^7$)							
n=30	6.38	5.54	4.66	5.09	4.90	4.40	4.24
n=50	5.28	4.47	3.77	4.27	3.88	3.42	3.26
n=80	4.10	3.97	3.30	3.98	3.12	2.68	2.58
n=100	3.55	3.74	3.15	3.88	2.77	2.36	2.24
n=130	3.27	3.61	3.09	3.89	2.50	2.06	1.98
n=150	2.85	3.38	3.02	3.82	2.20	1.96	1.84

^{1.} \hat{T}_D is design-based estimator.

^{2.} \hat{T}_{M-L} , \hat{T}_{M-LOGL} , and \hat{T}_{M-BC} are model-based estimators based on standard linear model, log-linear model, and Box-Cox model, respectively.

^{3.} \hat{T}_{G-L} and \hat{T}_{G-LOGL} are GREG estimators based on standard linear model and log-linear model.

^{4.} \hat{T}_{AG} is AUTOGREG estimator based on Box-Cox model.

Table 2.5: Relative biases of the different estimators using different sampling designs
with varying sample sizes ($\sigma=2$)

	\hat{T}_D ^{1.}	\hat{T}_{M-L} ^{2.}	\hat{T}_{M-LOGL} ^{2.}	\hat{T}_{M-BC} ^{2.}	\hat{T}_{G-L} ^{3.}	\hat{T}_{G-LOGL} ^{3.}	\hat{T}_{AG} ^{4.}
Simple random sampling ($\times 0.001$)							
n=30	-8.73	-18.11	-132.37	-132.90	-18.11	-1.20	-1.58
n=50	-0.58	-6.05	-132.55	-133.49	-6.05	4.98	3.11
n=80	5.17	-0.39	-133.29	-133.15	-0.39	5.47	4.62
n=100	1.09	-4.86	-140.05	-139.33	-4.86	-2.86	-2.78
n=130	-5.42	-7.56	-140.46	-140.16	-7.56	-3.21	-4.22
n=150	1.80	-4.07	-140.21	-139.08	-4.07	-1.64	-1.70
Stratified simple random sampling ($\times 0.001$)							
n=30	1.48	-101.64	-218.86	-222.97	-4.77	11.35	7.74
n=50	14.58	-97.35	-220.20	-224.49	-0.50	11.09	8.42
n=80	13.18	-96.22	-224.14	-228.11	1.64	7.45	5.99
n=100	-5.95	-103.31	-229.89	-233.84	-3.35	2.69	0.87
n=130	-5.67	-103.94	-230.11	-233.39	-3.20	2.50	1.04
n=150	-2.05	-102.47	-229.67	-233.21	-2.68	0.82	-0.28

^{1.} \hat{T}_D is design-based estimator.

^{2.} \hat{T}_{M-L} , \hat{T}_{M-LOGL} , and \hat{T}_{M-BC} are model-based estimators based on standard linear model, log-linear model, and Box-Cox model, respectively.

^{3.} \hat{T}_{G-L} and \hat{T}_{G-LOGL} are GREG estimators based on standard linear model and log-linear model.

^{4.} \hat{T}_{AG} is AUTOGREG estimator based on Box-Cox model.

Table 2.6: Root mean squared errors of the different estimators using different sampling designs with varying sampling sizes ($\sigma=2$)

	\hat{T}_D ^{1.}	\hat{T}_{M-L} ^{2.}	\hat{T}_{M-LOGL} ^{2.}	\hat{T}_{M-BC} ^{2.}	\hat{T}_{G-L} ^{3.}	\hat{T}_{G-LOGL} ^{3.}	\hat{T}_{AG} ^{4.}
Simple random sampling ($\times 10^7$)							
n=30	10.93	8.33	10.79	10.63	8.33	8.07	8.11
n=50	8.64	6.53	9.52	9.45	6.53	6.05	6.05
n=80	6.85	5.31	9.02	8.95	5.31	4.95	4.98
n=100	5.98	4.35	9.07	8.99	4.35	4.07	4.08
n=130	5.13	3.85	8.96	8.89	3.85	3.58	3.56
n=150	4.98	3.68	8.89	8.80	3.68	3.48	3.48
Stratified simple random sampling ($\times 10^7$)							
n=30	10.21	10.49	14.78	14.91	9.41	9.43	9.32
n=50	7.89	8.65	14.08	14.25	7.02	7.00	6.86
n=80	6.46	7.72	13.91	14.08	5.63	5.29	5.25
n=100	5.57	7.73	14.14	14.34	5.11	4.93	4.85
n=130	4.89	7.38	14.04	14.21	4.37	4.11	4.09
n=150	4.54	7.18	13.99	14.18	4.11	3.88	3.85

^{1.} \hat{T}_D is design-based estimator.

^{2.} \hat{T}_{M-L} , \hat{T}_{M-LOGL} , and \hat{T}_{M-BC} are model-based estimators based on standard linear model, log-linear model, and Box-Cox model, respectively.

^{3.} \hat{T}_{G-L} and \hat{T}_{G-LOGL} are GREG estimators based on standard linear model and log-linear model.

^{4.} \hat{T}_{AG} is AUTOGREG estimator based on Box-Cox model.

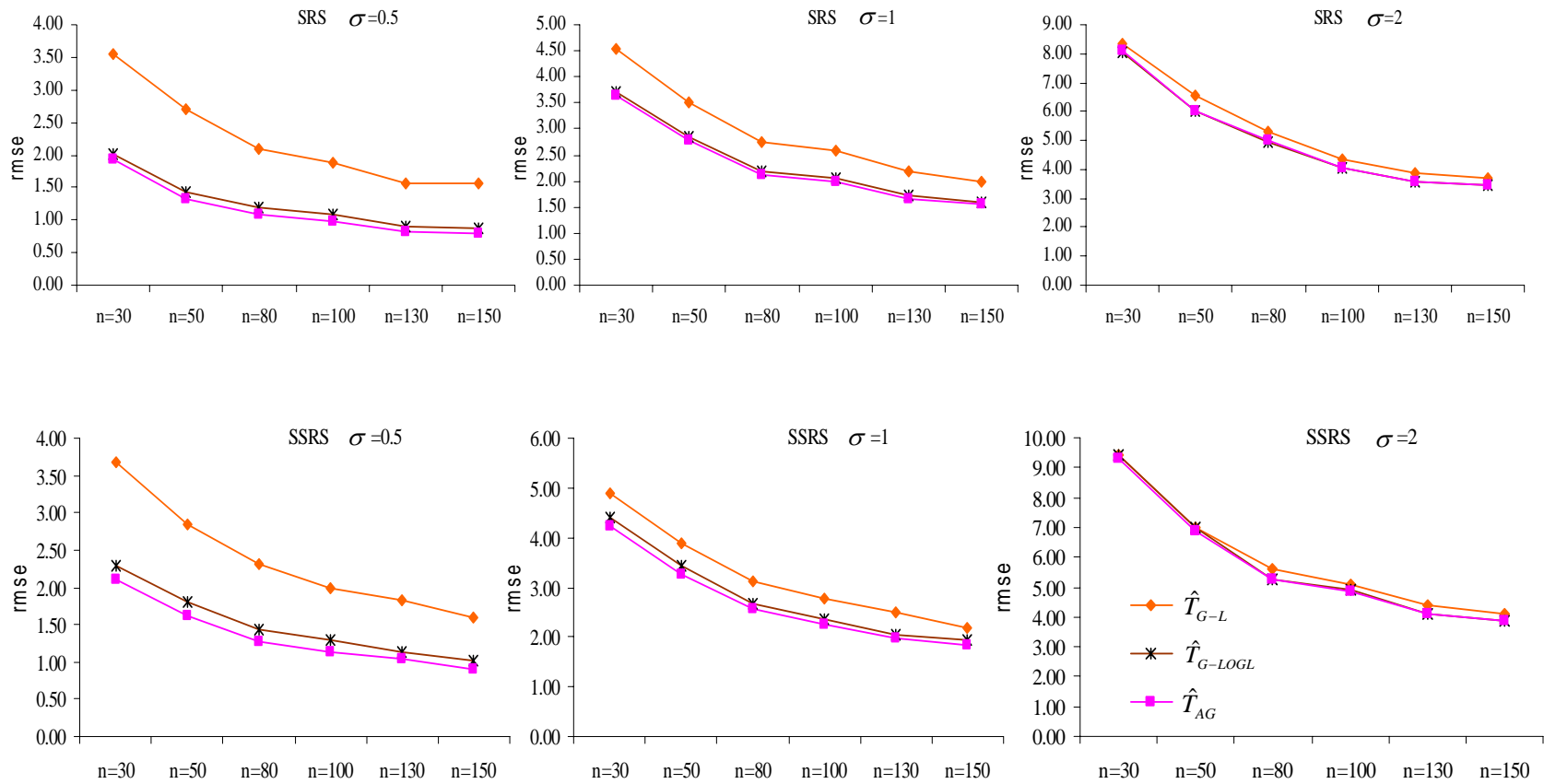


Figure 2.1: Comparison of root mean square error of \hat{T}_{G-L} , \hat{T}_{G-LOGL} , and \hat{T}_{AG} with varying sampling designs, sample sizes, and standard deviations.

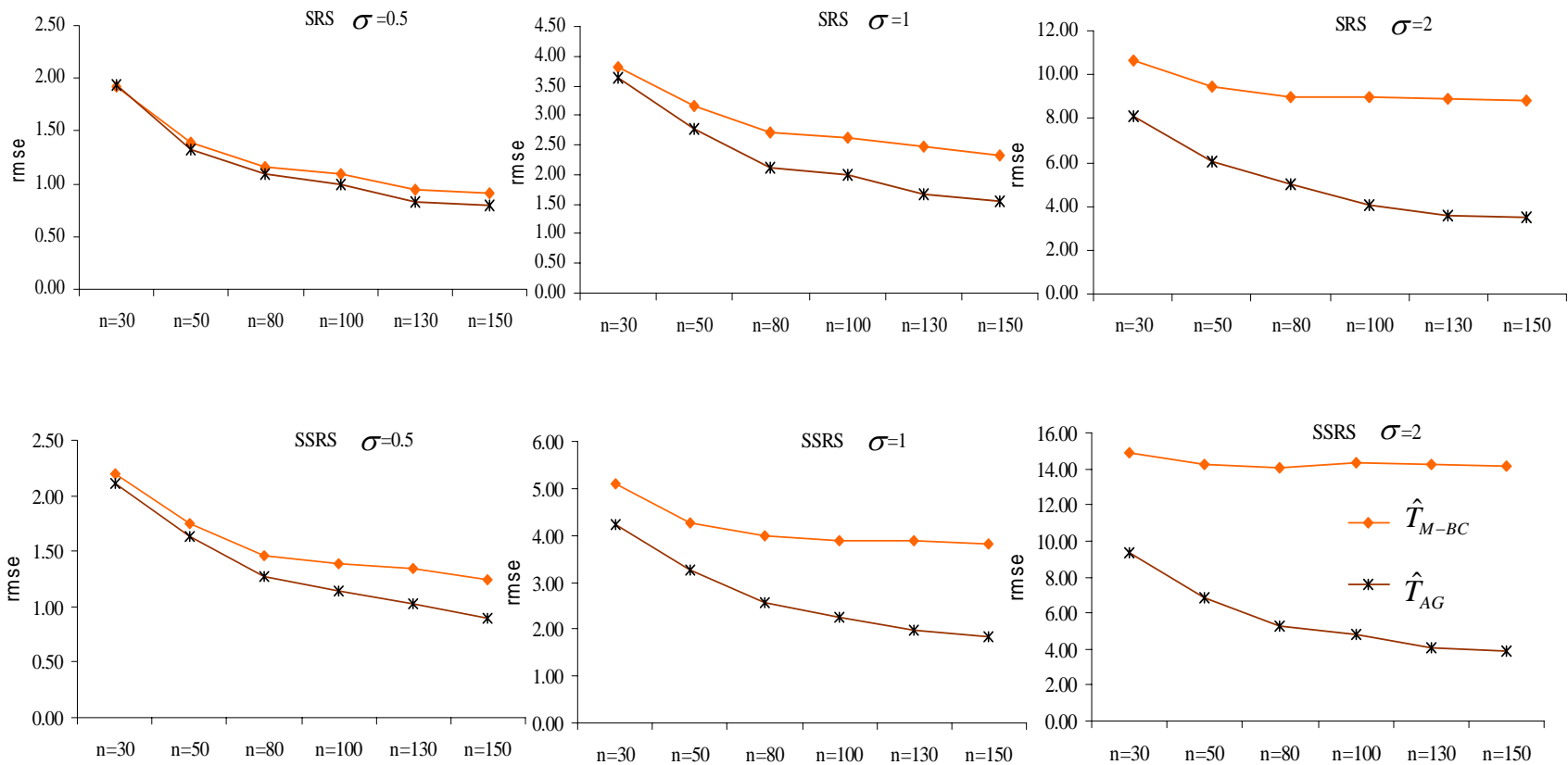


Figure 2.2: Comparison of root mean square error of \hat{T}_{M-BC} vs. \hat{T}_{AG} with varying sampling designs, sample sizes, and standard deviations.

Table 2.7: Relative biases and root mean square error of $\hat{\lambda}^1$ and $\hat{\lambda}_w^2$ for SSRS sampling with varying sample sizes and standard deviations.

	$\sigma=2$		$\sigma=1$		$\sigma=0.5$	
	$\hat{\lambda}$	$\hat{\lambda}_w$	$\hat{\lambda}$	$\hat{\lambda}_w$	$\hat{\lambda}$	$\hat{\lambda}_w$
Relative biases						
n=30	-0.58	0.13	-0.28	0.10	-0.16	-0.01
n=50	-0.50	0.13	-0.20	0.13	-0.15	-0.02
n=80	-0.42	0.14	-0.19	0.11	-0.13	-0.02
n=100	-0.43	0.07	-0.20	0.08	-0.13	-0.02
n=130	-0.45	0.06	-0.19	0.07	-0.11	-0.01
n=150	-0.39	0.10	-0.16	0.09	-0.10	-0.01
Root mean square error						
n=30	0.14	0.12	0.11	0.11	0.07	0.07
n=50	0.10	0.09	0.08	0.08	0.05	0.05
n=80	0.08	0.07	0.06	0.06	0.04	0.04
n=100	0.07	0.06	0.06	0.05	0.04	0.04
n=130	0.07	0.05	0.05	0.05	0.03	0.03
n=150	0.06	0.05	0.05	0.04	0.03	0.03

¹. Estimator $\hat{\lambda}$ is obtained using ordinary least square method/maximum likelihood method;

². Estimator $\hat{\lambda}_w$ is obtained by pseudo-maximum likelihood method.

Chapter 3: Summary and Future research

In this dissertation, both robust model-based and model-assisted estimators of a finite population total are proposed. The robustness property of the two proposed estimators is evaluated through the simulation studies under design-based or model-based framework, wherever appropriate. The two estimators possess all the properties rendered by the respective framework.

Design-based inference is based on the principle of randomization. It has a number of strengths that make it popular with practitioners. It automatically takes into account features of the survey design; it provides reliable inferences in large samples; and it requires few assumptions. However, design-based inference is essentially asymptotic, requiring large sample; it sacrifices efficiency; and there are situations in which the rigorous probability sampling is difficult or costly.

Under model-based inferences, a superpopulation model is assumed, and the finite population is a realization from the superpopulation model. Likelihood-based approaches render the model-based predictor the property of large-sample efficiency, and hence match or outperform design-based inferences if the model is correctly specified. The challenge with the model-based inference is that how exactly to specify the model? In practice all models are simplifications, and there are varying degrees for each model to be misspecified. Seriously misspecified model will lead to inferences that are worse (much worse) than design-based inferences.

Based on the strengths and limits of both design-based and model-based inferences, model-assisted design-based inference was proposed. Models are used to motivate the estimator so that information from auxiliary variables can be utilized to

increase the efficiency. Principle of randomization, however, is applied to select the sample and make the inferences. Therefore, even if the underlying model fails, design-consistency property of the estimator can still be obtained. It seems that model-assisted approach provides the potential solution to the controversy between design-based and model-based inferences.

The practitioner may wonder which of the two proposed estimators in this dissertation (model-based and model-assisted estimators) should be used in practice. The decision should be made based on the practical situations. If the sample size is small and the model has been widely accepted, model-based estimator can achieve high efficiency and might be more preferable; otherwise, model-assisted estimator can provide more conservative inferences, and the estimator is protected from the possible model failure.

The Use of the Box-Cox technique to achieve the robustness of the model-based and model-assisted estimators is the main theme of this dissertation. The Box-Cox transformation has received considerable attention over the last five decades. Past research includes estimation of the transformation parameter, hypothesis tests on the transformation parameter, variance heteroscedasticity and autocorrelation of the error structure, effect of outliers and influential cases, and prediction in the original scale. However, despite its popularity, the Box-Cox transformation has not been used in the context of the finite population sampling.

Research areas that have been thoroughly studied in main stream statistics are still widely open for future research in the finite population sampling context. For example, small-sample properties of models containing heteroscedastic errors or serially correlated

errors could be examined; Hypothesis tests and other inferences on transformation parameters could be studied by incorporating clustering effects and sampling weights, etc.

Different prediction techniques for the Box-Cox regression models were reviewed by Collins (1991), including plug-in, mean squared error analysis, predictive likelihood, and stochastic simulation. The four techniques take account of non-normality and parameter uncertainty in varying degrees. Although Collins has investigated different prediction techniques for a single future observation, to be useful for practitioners these techniques might be extended to generate prediction intervals for the finite population total. The prediction technique proposed in Chapter 1 is one such extension, based on mean squared error analysis. The other three prediction techniques (plug-in, predictive likelihood, and stochastic simulation) can be certainly extended too.

Appendix: R Programs

```
#-----BOX.COX.LAMDA-----
#DISCRIPTION: ML ESTIMATORS OF PARAMETERS OF BOX-COX MODEL USING GRID
#           SEARCH METHOD.
#INPUT:     "DATASET"— INCLUDING DEPENDENT VARIABLE Y, INDEPENDENT
#           VARIABLES X,
#           "DIM" – DIMENSIONS OF X
#OUTPUT:    ML ESTIMATES OF PARAMETERS OF BOX-COX MODEL
#-----
boxcox.lamda <- function(dataset, dim)
{
  N=nrow(dataset)
  geomean=1
  for (i in (1:N)) {geomean=geomean*dataset$y[i]^(1/N)}

  length=0.05
  lambda=seq(-2,2,length)
  times=2/length*2+1
  loglikhod=lambda

  for (i in (1:times))
  {
    if(lambda[i]!=0) Z=(dataset$y^lambda[i]-1)/(lambda[i]*geomean^(lambda[i]-1))
    else Z=log(dataset$y)*geomean

    X=cbind(1,dataset$x)
    Bhat= solve(t(X)%*%X) %*% t(X)%*%Z
    RSS=t(Z-X)%*%Bhat)%*%(Z-X)%*%Bhat
    sigma2=RSS/N
    loglikhod[i]=(-N/2)*(log(2*pi*sigma2)+1)
  }
  boxcox.data=cbind(lambda, loglikhod)

  lamda = boxcox.data[,"lambda"][which.max(boxcox.data[,"loglikhod"])]
  if(lamda!=0) Z=(dataset$y^lamda-1)/lamda
  else Z=log(dataset$y)
```

```

X=cbind(1,dataset$x)
B= solve(t(X)%*%X) %*% t(X)%*%Z
RSS=t(Z-X%*%B)%*%(Z-X%*%B)
sigma=sqrt(RSS/(N-dim))
sigma2=RSS/(N-dim)
list(lamda=lamda, B=B,sigma=sigma,sigma2=sigma2)
}

```

```

#-----BOXCOX.PML-----
#DISCRIPTION: PML ESITMATORS OF PARAMETERS (B, Λ ) OF BOX-COX MODEL USING
#              GRID SEARCH METHOD.;
#INPUT:        "DATASET"— INCLUDING DEPENDEND VARIABLE Y, INDEPEDNENT
#              VARIABLES X, AND SAMPLING WEIGHT W;
#              "DIM" – DIMENSTIONS OF X;
#OUTPUT:       PML ESTIMATES OF PARAMETERS OF BOX-COX MODEL.
#-----

```

```

boxcox.pml <- function(dataset, dim)
{
  y.sam=dataset$y
  w.sam=dataset$w
  x.sam=dataset$x
  N=sum(w.sam)
  n=nrow(dataset)

  wt.geomean=1
  for(i in 1:n) wt.geomean=wt.geomean*(y.sam[i]^(w.sam[i]/N))

  length=0.05
  lambda=seq(-2,2,length)
  times=2/length*2+1
  loglikhod=lambda

  for (i in (1:times))
  {
    if(lambda[i]!=0) Z=(y.sam^lambda[i]-1)/(lambda[i]*wt.geomean^(lambda[i]-1))
    if(lambda[i]==0) Z=log(y.sam)*wt.geomean
  }
}

```

```

X=cbind(1,x.sam)
W=diag(w.sam)
Bstar= solve(t(X)%*%W%*%X) %*% t(X)%*%W%*%Z
RSS=t(Z-X%*%Bstar)%*%W%*%(Z-X%*%Bstar)
sigma2=RSS/N
loglikhod[i]=(-N/2)*(log(2*pi*sigma2)+1)
}

boxcox.data=cbind(lambda, loglikhod)
lamda = boxcox.data[,"lambda"][which.max(boxcox.data[,"loglikhod"])]
if(lamda!=0) Z=(y.sam^lamda-1)/lamda
if(lamda==0) Z=log(y.sam)
X=cbind(1,x.sam)
B= solve(t(X)%*%W%*%X) %*% t(X)%*%W%*%Z
RSS=t(Z-X%*%B)%*%W%*%(Z-X%*%B)
sigma=sqrt(RSS/(N-dim))
sigma2=(RSS/(N-dim))

list(lamda=lamda, B=B,sigma=sigma, sigma2=sigma2)
}

```

#-----CHAPTER 1: SIMULATION STUDY-----

#*DISCRIPTION*: CALCULATE THE RELATIVE ERROR AND MSE of $\hat{T} - T$ FOR SIX PREDICTORS
WITH VARYING SAMPLE SIZES AND STANDARD ERRORS, SEE METHOD
AND RESULTS IN THE CHAPTER 2, SECTION OF SIMULATION STUDY
#*OUTPUT*: MSE of $\hat{T} - T$ FOR SIX PREDICTORS WITH VARYING
SAMPLE SIZES AND STANDARD ERRORS

#-----

```

REPTIME=1000
MCTIMES=1000
beefpop=read.table("I:/dissertation/BOXCOX PROJECT/data/beefpop.txt")
N=431

simubeef=beefpop
beefpop0=beefpop
names(simubeef)=c("truex","y")
simubeef$x=log(simubeef$truex)#CHANGE NAME TO CALL BOXCOX.LAMDA SUBROUTINE
lamda0=boxcox.lamda(simubeef,2)$lamda

```

```

B0=boxcox.lamda(simubeef,2)$B

Sample=c(1:N)
dim=2
nn=1
arr.mse=matrix(0,12,6)
arr.rb=matrix(0,12,6)

#COLUMNS STORE SIX PREDICTORS; AND ROWS STORE 6 COMBINATIONS
#DEFINED BY SAMPLE SIZE AND STANDARD ERROR (3*2);
for (n in c(50,100,150))
{
#-----FIRST STEP, SELECT SAMPLE OF SIZE n-----
sample=sort(sample(Sample,n))
X.nonsam=log(beefpop0[-sample,]$V1)
X.sam=log(beefpop0[sample,]$V1)

#-----CALC MSE for BP, ABP
T.r.ABP=ifelse(lamda0==0, sum(exp(X.nonsam*B0[2]+B0[1])),
sum(((X.nonsam*B0[2]+B0[1])*lamda0+1)^(1/lamda0)))
for (sigma0 in c(0.1, 0.5, 1, 2))
{
T.r.BP=0
count.pos=0
while (count.pos < MCTIMES)
{
err=sigma0*rnorm(N-n)
pos.BP=((X.nonsam*B0[2]+B0[1]+err)*lamda0+1)
if (any(pos.BP>0))
{
tmp.BP=ifelse(lamda==0, sum(exp(X.nonsam*B0[2]+B0[1]+err)),
sum(pos.BP*(1/lamda0)))
T.r.BP=T.r.BP+1/MCTIMES*tmp.BP
count.pos=count.pos+1
}
}
}
#-----

```

```

MSE.BP=0
MSE.ABP=0
MSE.NTP=0
MSE.LTP=0
MSE.AEBP=0
MSE.EBP=0

RB.BP=0
RB.ABP=0
RB.NTP=0
RB.LTP=0
RB.AEBP=0
RB.EBP=0

count=rep(0,REPTIME)
T.r.i.EBP=matrix(0,REPTIME,MCTIMES)
T.i.r=rep(0,REPTIME)
B.i=matrix(0,REPTIME,dim)
lamda.i=rep(0,REPTIME)

for (r in (1:REPTIME))
  {
#-----SECOND STEP, Generate y by the M1 and T=sum(y)-----
    simubeef$y.lamda=(rep(-1,N)-1)/lamda0
    while (any((simubeef$y.lamda*lamda0+1)<0))
      {
        simubeef$y.lamda=B0[1]+B0[2]*simubeef$x+sigma0*rnorm(N)
        simubeef$y=(simubeef$y.lamda*lamda0+1)**(1/lamda0)
      }

    beefpop=simubeef
#-----THIRD STEP: Using the sample fixed in Step 1 calculate the estimators-----
    temp.sample=beefpop[sample,]
    T.r=sum(beefpop[-sample,]$y)

#-----CALC MSE for NTP and LTP
    lm.org=lm(y~x,data=temp.sample)

```

```
T.r.NTP=sum(cbind(1,X.nonsam)%*%lm.org$coefficients)
```

```
temp.sample$logy=log(temp.sample$y)
```

```
lm.log=lm(logy~x,data=temp.sample)
```

```
sigma2.log=sum(lm.log$residuals^2)/(n-2)
```

```
T.r.LTP=sum(exp(cbind(1,X.nonsam)%*%lm.log$coefficients+1/2*sigma2.log))
```

```
#-----CALC MSE for AEBP, EBP
```

```
lamda=boxcox.lamda(temp.sample,2)$lamda
```

```
B=boxcox.lamda(temp.sample,2)$B
```

```
sigma=boxcox.lamda(temp.sample,2)$sigma
```

```
B.i[r,]=B
```

```
lamda.i[r]=lamda
```

```
T.r.AEBP=ifelse(lamda==0, sum(exp(X.nonsam*B[2]+B[1])),  
sum(((X.nonsam*B[2]+B[1])*lamda+1)^(1/lamda)))
```

```
T.r.EBP=0
```

```
count.pos=0
```

```
while (count.pos < MCTIMES)
```

```
{
```

```
err=sigma*rnorm(N-n)
```

```
pos.EBP=((X.nonsam*B[2]+B[1]+err)*lamda+1)
```

```
if (!any(pos.EBP<=0))
```

```
{
```

```
tmp.EBP=ifelse(lamda==0,
```

```
sum(exp(X.nonsam*B[2]+B[1]+err)),
```

```
sum(pos.EBP**(1/lamda)))
```

```
T.r.EBP=T.r.EBP+1/MCTIMES*tmp.EBP
```

```
count.pos=count.pos+1
```

```
T.r.i.EBP[r,count.pos]=tmp.EBP
```

```
}
```

```
count[r]=count[r]+1
```

```
}
```

```
T.i.r[r]=T.r
```

```

MSE.BP=MSE.BP+(T.r.BP-T.r)**2/REPTIME
MSE.ABP=MSE.ABP+(T.r.ABP-T.r)**2/REPTIME
MSE.NTP=MSE.NTP+(T.r.NTP-T.r)**2/REPTIME
MSE.LTP=MSE.LTP+(T.r.LTP-T.r)**2/REPTIME
MSE.AEBP=MSE.AEBP+(T.r.AEBP-T.r)**2/REPTIME
MSE.EBP=MSE.EBP+(T.r.EBP-T.r)**2/REPTIME
RB.BP=RB.BP+(T.r.BP-T.r)/T.r/REPTIME
RB.ABP=RB.ABP+(T.r.ABP-T.r)/T.r/REPTIME
RB.NTP=RB.NTP+(T.r.NTP-T.r)/T.r/REPTIME
RB.LTP=RB.LTP+(T.r.LTP-T.r)/T.r/REPTIME
RB.AEBP=RB.AEBP+(T.r.AEBP-T.r)/T.r/REPTIME
RB.EBP=RB.EBP+(T.r.EBP-T.r)/T.r/REPTIME
}
arr.rb [nn,]=c(MSE.NTP,MSE.LTP,MSE.AEBP,MSE.ABP,MSE.EBP,MSE.BP)
arr.rb[nn,]=c(RB.NTP,RB.LTP,RB.AEBP,RB.ABP,RB.EBP,RB.BP)
nn=nn+1
}#END SIGMA
}#END n
colnames(arr.mse)=c("NTP","LTP","AEBP","ABP","EBP","BP")
rownames(arr.mse)=c("n=50","n=50","n=50","n=50","n=100","n=100","n=100","n=100","n=150",
,"n=150","n=150","n=150")
arr.mse
colnames(arr.rb)=c("NTP","LTP","AEBP","ABP","EBP","BP")
rownames(arr.rb)=c("n=50","n=50","n=50","n=50","n=100","n=100","n=100","n=100","n=150","n=150","
n=150","n=150")
arr.rb

#-----CHAPTER 1: REAL DATA ANALYSIS-----
#DISCRPTION: CALCULATE THE RELATIVE ERROR, CONFIDENCE INTERVAL, AND THE
#          ACTUAL PROBABLITY CI COVERS THE TRUE PARAMETERS FOR NTP, LTP,
#          AND EBP;
#          SEE METHOD AND RESULTS IN THE CHAPTER 2, SECTION OF REAL DATA
#          ANALYSIS.
#-----
beefpop=read.table("I:/dissertation/BOXCOX PROJECT/data/beefpop.txt")
names(beefpop)=c("x","y")
beefpop$logy=log(beefpop$y)

```

```
beefpop$logx=log(beefpop$x)
```

```
T=sum(beefpop[, "y"])
```

```
N=431
```

```
n=430
```

```
yhat.boxcox=rep(0,N)
```

```
yhat.org=rep(0,N)
```

```
yhat.log=rep(0,N)
```

```
yhat.log2=rep(0,N)
```

```
That=rep(0,N)
```

```
That.org=rep(0,N)
```

```
That.log=rep(0,N)
```

```
That.log2=rep(0,N)
```

```
RelErr=rep(0,N)
```

```
RelErr.org=rep(0,N)
```

```
RelErr.log=rep(0,N)
```

```
RelErr.log2=rep(0,N)
```

```
CI.lower=rep(0,N)
```

```
CI.upper=rep(0,N)
```

```
inout90=rep(0,N)
```

```
inout95=rep(0,N)
```

```
CI.lower.org=rep(0,N)
```

```
CI.upper.org=rep(0,N)
```

```
inout90.org=rep(0,N)
```

```
inout95.org=rep(0,N)
```

```
CI.lower.log=rep(0,N)
```

```
CI.upper.log=rep(0,N)
```

```
inout90.log=rep(0,N)
```

```
inout95.log=rep(0,N)
```

```
lamda.sample=rep(0,N)
```

```
MCTIMES=1000
```

```
dim=2
```



```

for (unsam in c(1:N))
{
beef.sample=beefpop[-unsam,]
beef.nonsample=beefpop[unsam,]

temp.sample=beef.sample
temp.sample$x=temp.sample$logx # IN ORDER TO CALL SUBFUNCTION
BOXCOX.LAMDA

lamda=boxcox.lamda(temp.sample,2)$lamda
lamda.sample[unsam]=lamda
B=boxcox.lamda(temp.sample,2)$B
sigma=boxcox.lamda(temp.sample,2)$sigma

count.pos=0
y.pred=rep(0,N-n)
y2.pred=rep(0,N-n)
dy.beta=matrix(0,N-n,dim)
dy.sigma=rep(0,N-n)
dy.lamda=rep(0,N-n)

while (count.pos < MCTIMES)
{
z.norm=rnorm(N-n)
err=sigma*z.norm
pos=((beef.nonsample[,"logx"]*B[2]+B[1]+err)*lamda+1)
if (any(pos>0))
{
tmp1=ifelse(lamda==0, (exp(beef.nonsample[,"logx"]*B[2]+B[1]+err)),
(pos**(1/lamda))
tmp2=ifelse(lamda==0, (exp(2*(beef.nonsample[,"logx"]*B[2]+B[1]+err))),
(pos**(2/lamda))
tmp.dy.beta=pos^(1/lamda-1)*%c(1,beef.nonsample[,"logx"])
tmp.dy.sigma=pos^(1/lamda-1)*z.norm
#tmp.dy.lamda=(pos-1)/lamda/lamda*pos^(1/lamda-1) -
1/lamda^2*(pos^(1/lamda))*log(pos)

```

```

tmp.dy.lamda=1/(lamda^2)*(pos^(1/lamda))*(1-pos^(-1) - log(pos))

y.pred=y.pred+1/MCTIMES*tmp1
y2.pred=y2.pred+1/MCTIMES*tmp2
dy.beta=dy.beta+1/MCTIMES*tmp.dy.beta
dy.sigma=dy.sigma+1/MCTIMES*tmp.dy.sigma
dy.lamda=dy.lamda+1/MCTIMES*tmp.dy.lamda

count.pos=count.pos+1
}
}

#-----BOX-COX METHOD: EBP-----
yhat.boxcox[unsam]=y.pred
That.r=sum(y.pred)
var.y=sum(y2.pred-y.pred^2)

T.r=sum(beef.nonsample[, "y"])
X=cbind(1,beef.sample[, "x"])
logX=cbind(1,log(beef.sample[, "x"]))
Y=beef.sample[, "y"]
logY=beef.sample[, "logy"]
T.s=sum(Y)

That[unsam]=That.r+T.s
RelErr[unsam]=(That[unsam]-T)/T

#-----BOX-COX METHOD: VARIABILITY-----
geomean=1
for (i in (1:n)) { geomean=geomean*Y[i]^(1/n) }
if(lamda!=0) Z=(Y^lamda-1)/lamda/geomean^(lamda-1)
if(lamda==0) Z=log(Y)*geomean
Bstar=solve(t(logX)%*%logX) %*% t(logX)%*%Z
sigma.star=as.numeric(sqrt(t(Z-logX)%*%Bstar)%*%(Z-logX)%*%Bstar)/(n-dim))

#-----CALCULATE THE ASYMPOTOTIC VARIANCE OF LAMDA AND BETA-----
zi=rep(0,n)

```

```

si=rep(0,n)
ti=rep(0,n)
ri=rep(0,n)
d.zi=rep(0,n)
d2.zi=rep(0,n)
I.obs=matrix(0,dim+2,dim+2)
Jacob=matrix(0,dim+2,dim+2)
varcov=matrix(0,dim+2,dim+2)
varcov.scale=matrix(0,dim+2,dim+2)

for (i in (1:n))
{
xi=logX[i,]
yi=Y[i]
zi[i]=(yi^lamda-1)/lamda/(geomean^(lamda-1))
ri[i]=zi[i]-xi%%Bstar
si[i]=yi^lamda*log(yi)
ti[i]=yi^lamda*log(yi)*log(yi)
d.zi[i]=lamda^(-1)*(geomean^(1-lamda))*si[i]-zi[i]*(lamda^(-1)+log(geomean))
d2.zi[i]= lamda^(-1)*(geomean^(1-lamda))*(ti[i]-2*si[i]*(log(geomean)+lamda^(-
1)))+zi[i]*(lamda^(-2)+(lamda^(-1)+log(geomean))^2)
}

I.obs[1:dim, 1:dim]= sigma.star^(-2)*(t(logX)%%logX)
I.obs[1:dim, 1+dim]= -sigma.star^(-2)*(t(logX)%%d.zi)
I.obs[1:dim, 2+dim]= 2*sigma.star^(-3)*(t(logX)%%ri)
I.obs[1+dim, 1:dim]=t(I.obs[1:dim, 1+dim])
I.obs[1+dim, 1+dim]= sigma.star^(-2)*(d.zi%%d.zi + ri%%d2.zi)
I.obs[1+dim, 2+dim]= -2*sigma.star^(-3)*(ri%%d.zi)
I.obs[2+dim, 1:dim]= t(I.obs[1:dim, 2+dim])
I.obs[2+dim, 1+dim]= t(I.obs[1+dim, 2+dim])
I.obs[2+dim, 2+dim]= -sigma.star^(-2)*n+3*sigma.star^(-4)*(ri%%ri)

varcov.scale=solve(I.obs)

Jacob[1:dim, 1:dim]=diag(geomean^(lamda-1),dim,dim)
Jacob[1:dim, 1+dim]=log(geomean)*B
Jacob[1:dim, 2+dim]=0

```

```
Jacb[1+dim, 1:dim]=0
```

```
Jacb[1+dim, 1+dim]=1
```

```
Jacb[1+dim, 2+dim]=0
```

```
Jacb[2+dim, 1:dim]=0
```

```
Jacb[2+dim, 1+dim]=log(geomean)*(sigma)
```

```
Jacb[2+dim, 2+dim]=geomean^(lamda-1)
```

```
varcov=Jacb %*% varcov.scale %*% t(Jacb)
```

```
#-----CALCULATE THE CONFIDENCE INTERVAL FOR T.HAT USING BOX-COX METHOD----
```

```
dy=matrix(0,N-n,dim+2) #1 is saved for lamda and sigma
```

```
dy[,1:dim]=dy.beta
```

```
dy[,dim+1]=dy.lamda
```

```
dy[,dim+2]=dy.sigma
```

```
var.y.pred=sum(dy%*% varcov%*%t(dy))
```

```
var.That.T=var.y.pred+var.y
```

```
CI.lower[unsam]=That[unsam]-1.648*sqrt(var.That.T)
```

```
CI.upper[unsam]=That[unsam]+1.648*sqrt(var.That.T)
```

```
inout90[unsam]=ifelse(T<CI.upper[unsam] & T>CI.lower[unsam], 1, 0)
```

```
CI.lower[unsam]=That[unsam]-1.96*sqrt(var.That.T)
```

```
CI.upper[unsam]=That[unsam]+1.96*sqrt(var.That.T)
```

```
inout95[unsam]=ifelse(T<CI.upper[unsam] & T>CI.lower[unsam], 1, 0)
```

```
#----- NO-TRANSFORMATION METHOD-----
```

```
lm.org=lm(y~x,data=beef.sample)
```

```
y.pred.org=c(1,beef.nonsample["x"])%*%lm.org$coefficients
```

```
That.r.org=sum(y.pred.org)
```

```
yhat.org[unsam]=y.pred.org
```

```
That.org[unsam]=That.r.org+T.s
```

```
RelErr.org[unsam]=(That.org[unsam]-T)/T
```

```
B= solve(t(X)%*%X) %*% t(X)%*% Y
```

```
RSS=t(Y-X%*%B)%*%(Y-X%*%B)
sigma2=RSS/(n-dim)
```

```
# -----IN OUR CASE, THERE ARE ONLY ONE UNSAMPLED UNIT AND r=1-----
```

```
var.That.T.org=sigma2*(c(1,beef.nonsample["x"])%*%solve(t(X)%*%X)%*%c(1,beef.nonsamp
le["x"])+1)
```

```
CI.lower.org[unsam]=That.org[unsam]-1.648*sqrt(var.That.T.org)
CI.upper.org[unsam]=That.org[unsam]+1.648*sqrt(var.That.T.org)
inout90.org[unsam]=ifelse(T<CI.upper.org[unsam] & T>CI.lower.org[unsam], 1, 0)
CI.lower.org[unsam]=That.org[unsam]-1.96*sqrt(var.That.T.org)
CI.upper.org[unsam]=That.org[unsam]+1.96*sqrt(var.That.T.org)
inout95.org[unsam]=ifelse(T<CI.upper.org[unsam] & T>CI.lower.org[unsam], 1, 0)
```

```
# -----LOG-TRANSFORMATION METHOD1-----
```

```
B= solve(t(logX)%*%logX)%*% t(logX)%*%logY
RSS=t(logY-logX%*%B)%*%(logY-logX%*%B)
sigma2=RSS/(n-dim)
X.i=c(1,beef.nonsample["logx"])
```

```
lm.log=lm(logy~logx,data=beef.sample)
y.pred.log=exp(X.i%*%lm.log$coefficients+1/2*sigma2)
That.r.log=sum(y.pred.log)
That.log[unsam]=That.r.log+T.s
RelErr.log[unsam]=(That.log[unsam]-T)/T
```

```
yhat.log[unsam]=y.pred.log
```

```
var.That.T.log=exp(2*X.i%*%B+sigma2)*(sigma2*X.i%*%solve(t(logX)%*%logX)%*%X.i+0.
5*sigma2^2/(n-dim)+exp(sigma2)-1)
CI.lower.log[unsam]=That.log[unsam]-1.648*sqrt(var.That.T.log)
CI.upper.log[unsam]=That.log[unsam]+1.648*sqrt(var.That.T.log)
inout90.log[unsam]=ifelse(T<CI.upper.log[unsam] & T>CI.lower.log[unsam], 1, 0)
CI.lower.log[unsam]=That.log[unsam]-1.96*sqrt(var.That.T.log)
CI.upper.log[unsam]=That.log[unsam]+1.96*sqrt(var.That.T.log)
inout95.log[unsam]=ifelse(T<CI.upper.log[unsam] & T>CI.lower.log[unsam], 1, 0)
```

```

#----- LOG-TRANSFORMATION METHOD2: Chambers and Dorfman-----
a.ii=X.i%%solve(t(logX)%%logX)%%X.i
k.i=1+sigma2*a.ii/2+sigma2^2/4/n
y.pred.log2=exp(X.i%%lm.log$coefficients+1/2*sigma2)/k.i

That.r.log2=sum(y.pred.log2)
That.log2[unsam]=That.r.log2+T.s
RelErr.log2[unsam]=(That.log2[unsam]-T)/T

yhat.log2[unsam]=y.pred.log2
}

org=cbind(ave.That=mean(That.org),ave.AARD=mean(abs(RelErr.org)),ave.ASRD=mean((RelErr.org)^2)
, ste.CI90=sqrt(var(CI.upper.org-CI.lower.org)), ave.CI90=mean(CI.upper.org-
CI.lower.org),ave.inout90=mean(inout90.org),ave.inout95=mean(inout95.org))
log=cbind(ave.That=mean(That.log),ave.AARD=mean(abs(RelErr.log)),ave.ASRD=mean((RelErr.log)^2),
ste.CI90=sqrt(var(CI.upper.log-CI.lower.log)), ave.CI90=mean(CI.upper.log-
CI.lower.log),ave.inout90=mean(inout90.log),ave.inout95=mean(inout95.log))
log2=cbind(ave.That=mean(That.log),ave.AARD=mean(abs(RelErr.log)),ave.ASRD=mean((RelErr.log)^2
))
boxcox=cbind(ave.That=mean(That),ave.AARD=mean(abs(RelErr)),ave.ASRD=mean((RelErr)^2),ste.CI9
0=sqrt(var(CI.upper-CI.lower)), ave.CI90=mean(CI.upper-
CI.lower),ave.inout90=mean(inout90),ave.inout95=mean(inout95))
comp=rbind(org,log,boxcox)
rownames(comp)=c("org","log","boxcox")
comp

#-----CHAPTER 2: SIMULATION STUDY-----
#DISCRIPTION: CALCULATE THE RELATIVE BIAS AND ROOT MEAN SQUARE ERROR FOR
# DESIGN-BASED, 3 MODEL-BASED, 2 GREG, AND AUTOGREG PREDICTORS OF
# POPULATION TOTAL;
# SEE METHOD AND RESULTS IN THE CHAPTER 3, SECTION OF SIMULATION
# STUDY.
# TWO SAMPLING DESIGNS ARE CONSIDERED: SRS AND SSRS
#-----

#-----SUBFUNCTION OF ESTIMATES, CALLED BY THE MAIN PROGRAM-----
estimates <- function(pop,samp)

```

```

{
T.r=sum(pop[-samp,"y"])
T.s=sum(pop[samp,"y"])
beef.sample=pop[samp,]
beef.nonsample=pop[-samp,]

#-----DESIGN-BASED-----
T.D=sum(beef.sample["y"]*beef.sample["w"])

#-----MODEL-BASED UNDER LINEAR REGRESSION MODEL
lm.r=lm(beef.sample$y~beef.sample$x)
y.pred=cbind(1,beef.nonsample["x"])*%*%lm.r$coefficients
That.r=sum(y.pred)
T.ML=That.r+T.s

#----- MODEL-BASED UNDER LOG-LINEAR REGRESSION MODEL
lm.log=lm(log(beef.sample$y)~beef.sample$x)
y.pred=exp(cbind(1,beef.nonsample["x"])*%*%lm.log$coefficients)
That.r=sum(y.pred)
T.MLOG=That.r+T.s

#-----MODEL-BASED UNDER BOX-COX MODEL
lamda=boxcox.lamda(beef.sample,2)$lamda
B=boxcox.lamda(beef.sample,2)$B
sigma=boxcox.lamda(beef.sample,2)$sigma
if (lamda==0)y.pred=exp(beef.nonsample["x"]*B[2]+B[1])
if (lamda!=0)y.pred=((beef.nonsample["x"]*B[2]+B[1])*lamda+1)^(1/lamda)

pos.M=(pop["x"]*B[2]+B[1])*lamda+1
T.MBC=-999
if(!any(pos.M<0))
{
That.r=sum(y.pred)
T.MBC=That.r+T.s
}

#-----GREG UNDER STANDARD LINEAR MODEL

```

```

lm.r=lm(beef.sample$y~beef.sample$x, weights=beef.sample$w)
y.s.pred=cbind(1,beef.sample[, "x"])%%lm.r$coefficients
y.r.pred=cbind(1,beef.nonsample[, "x"])%%lm.r$coefficients
T.GL=sum(y.r.pred)+sum(y.s.pred)+sum((beef.sample[, "y"]-y.s.pred)*beef.sample[, "w"])

#-----GREG UNDER LOG-LINEAR MODEL
lm.log=lm(log(beef.sample$y)~beef.sample$x, weights=beef.sample$w)
y.s.pred=exp(cbind(1,beef.sample[, "x"])%%lm.log$coefficients)
y.r.pred=exp(cbind(1,beef.nonsample[, "x"])%%lm.log$coefficients)
T.GLOG=sum(y.r.pred)+sum(y.s.pred)+sum((beef.sample[, "y"]-y.s.pred)*beef.sample[, "w"])

#-----AUTOGREG UNDER BOX-COX MODEL
lamda.w=boxcox.pml(beef.sample,2)$lamda
B.w=boxcox.pml(beef.sample,2)$B
sigma.w=boxcox.pml(beef.sample,2)$sigma
if (lamda.w==0)
  {
    y.r.pred=exp(beef.nonsample[, "x"]*B.w[2]+B.w[1])
    y.s.pred=exp(beef.sample[, "x"]*B.w[2]+B.w[1])
  }
if (lamda.w!=0)
  {
    y.r.pred=((beef.nonsample[, "x"]*B.w[2]+B.w[1])*lamda.w+1)^(1/lamda.w)
    y.s.pred=((beef.sample[, "x"]*B.w[2]+B.w[1])*lamda.w+1)^(1/lamda.w)
  }

pos=(pop[, "x"]*B.w[2]+B.w[1])*lamda.w+1
T.GBC=-999
if(!any(pos<0))
  {
    T.GBC=sum(y.r.pred)+sum(y.s.pred)+sum((beef.sample[, "y"]-
    y.s.pred)*beef.sample[, "w"])
    #T.AG=T.D+sum(y.r.pred)+sum(y.s.pred)-sum(y.s.pred*beef.sample[, "w"])
  }

list(sam=samp, T.D=T.D, T.ML=T.ML, T.MLOG=T.MLOG, T.MBC=T.MBC, T.GL=T.GL,
T.GLOG=T.GLOG, T.GBC=T.GBC, lamda=lamda, lamda.w=lamda.w, B=B, B.w=B.w,
sigma=sigma, sigma.w=sigma.w)

```



```

    }

#-----MAIN PROGRAM-----
beefpop=read.table("I:/dissertation/BOXCOX PROJECT/data/beefpop.txt")
names(beefpop)=c("exp", "y")
beefpop$x=log(beefpop$exp)
dim=2;H=2;

tmp=boxcox.lamda(beefpop,2)
lamda0=tmp$lamda
B0=tmp$B
N=4000

x.u=mean(beefpop$exp)
x.sigma=sqrt(var(beefpop$exp))
x.simu=sort(40+1000*rexp(N))
x=log(x.simu)

sigma0=2

y=(lamda0*(B0[1]+B0[2]*x+sigma0*rnorm(N))+1)**(1/lamda0)
T=sum(y)

beefpop=data.frame(x, y)
pop1=beefpop[order(beefpop$y),][1:floor(N/2),]
pop2=beefpop[order(beefpop$y),][(floor(N/2)+1):N,]

REP=1000
result.RB.SRS=matrix(0,6,7) # FIRST DIMENSION: DIFFERENT SAMPLE SIZES AND
                             # SECOND DIMENSION: SEVEN DIFFERENT ESTIMATORS
result.RB.SSRS=matrix(0,6,7)
result.rmse.SRS=matrix(0,6,8)
result.rmse.SSRS=matrix(0,6,8)

result.RB.phi.SRS=matrix(0,6,8) # FIRST DIMENSION: DIFFERENT SAMPLE SIZES;
                                  # SECOND DIMENSION: WEIGHTED AND UNWEIGHTED
                                  # ESTIMATORS FOR 4 MODEL PARAMETERS.

```

```

result.rmse.phi.SRS=matrix(0,6,8)
result.RB.phi.SSRS=matrix(0,6,8)
result.rmse.phi.SSRS=matrix(0,6,8)

NN=c(nrow(pop1),nrow(pop2))
ssize=1

for (n in c(30, 50, 80, 100, 130, 150))
  {
#-----SRS-----
    beefpop$w=N/n

    T.D.SRS=numeric(REP)
    T.ML.SRS=numeric(REP)
    T.MLOG.SRS=numeric(REP)
    T.MBC.SRS=numeric(REP)
    T.GL.SRS=numeric(REP)
    T.GLOG.SRS=numeric(REP)
    T.GBC.SRS=numeric(REP)
    lamda.SRS=numeric(REP)
    lamda.w.SRS=numeric(REP)
    B.SRS=matrix(0,REP,2)
    B.w.SRS=matrix(0,REP,2)
    sigma.SRS=numeric(REP)
    sigma.w.SRS=numeric(REP)
    interm=matrix(0,n,REP)

    count=0
    for (r in c(1:REP))
      {
        flag=0
        while (flag==0){
          sam=sample(1:N,n)
          temp=estimates(beefpop,sam)
          T.D.SRS[r]=temp$T.D
          T.ML.SRS[r]=temp$T.ML
          T.MLOG.SRS[r]=temp$T.MLOG
        }
      }
  }

```

```

T.MBC.SRS[r]=temp$T.MBC
T.GL.SRS[r]=temp$T.GL
T.GLOG.SRS[r]=temp$T.GLOG
T.GBC.SRS[r]=temp$T.GBC
lamda.SRS[r]=temp$lamda
lamda.w.SRS[r]=temp$lamda.w
B.SRS[r,]=temp$B
B.w.SRS[r,]=temp$B.w
sigma.SRS[r]=temp$sigma
sigma.w.SRS[r]=temp$sigma.w
interm[,r]=temp$sam
if (temp$T.MBC!=-999&temp$T.GBC!=-999)      flag=1
}
count=count+1
}
RelB.lamda.SRS=(mean(lamda.SRS)-lamda0)/lamda0
rmse.lamda.SRS=sqrt(mean((lamda.SRS-lamda0)^2))
RelB.B1.SRS=(mean(B.SRS[,1])-B0[1])/B0[1]
rmse.B1.SRS=sqrt(mean((B.SRS[,1]-B0[1])^2))
RelB.B2.SRS=(mean(B.SRS[,2])-B0[2])/B0[2]
rmse.B2.SRS=sqrt(mean((B.SRS[,2]-B0[2])^2))
RelB.sigma.SRS=(mean(sigma.SRS)-sigma0)/sigma0
rmse.sigma.SRS=sqrt(mean((sigma.SRS-sigma0)^2))

RelB.lamda.w.SRS=(mean(lamda.w.SRS)-lamda0)/lamda0
rmse.lamda.w.SRS=sqrt(mean((lamda.w.SRS-lamda0)^2))
RelB.B1.w.SRS=(mean(B.w.SRS[,1])-B0[1])/B0[1]
rmse.B1.w.SRS=sqrt(mean((B.w.SRS[,1]-B0[1])^2))
RelB.B2.w.SRS=(mean(B.w.SRS[,2])-B0[2])/B0[2]
rmse.B2.w.SRS=sqrt(mean((B.w.SRS[,2]-B0[2])^2))
RelB.sigma.w.SRS=(mean(sigma.w.SRS)-sigma0)/sigma0
rmse.sigma.w.SRS=sqrt(mean((sigma.w.SRS-sigma0)^2))

result.RB.phi.SRS[ssize,]=cbind(RelB.lamda.SRS, RelB.lamda.w.SRS, RelB.B1.SRS,
RelB.B1.w.SRS, RelB.B2.SRS, RelB.B2.w.SRS, RelB.sigma.SRS, RelB.sigma.w.SRS)
result.rmse.phi.SRS[ssize,]=cbind(rmse.lamda.SRS, rmse.lamda.w.SRS, rmse.B1.SRS,
rmse.B1.w.SRS, rmse.B2.SRS, rmse.B2.w.SRS, rmse.sigma.SRS, rmse.sigma.w.SRS)

```

```
RelB.D.SRS=(mean(T.D.SRS)-T)/T
rmse.D.SRS=sqrt(mean((T.D.SRS-T)^2))
```

```
RelB.ML.SRS=(mean(T.ML.SRS)-T)/T
rmse.ML.SRS=sqrt(mean((T.ML.SRS-T)^2))
RelB.MLOG.SRS=(mean(T.MLOG.SRS)-T)/T
rmse.MLOG.SRS=sqrt(mean((T.MLOG.SRS-T)^2))
RelB.MBC.SRS=(mean(T.MBC.SRS)-T)/T
rmse.MBC.SRS=sqrt(mean((T.MBC.SRS-T)^2))
```

```
RelB.GL.SRS=(mean(T.GL.SRS)-T)/T
rmse.GL.SRS=sqrt(mean((T.GL.SRS-T)^2))
RelB.GLOG.SRS=(mean(T.GLOG.SRS)-T)/T
rmse.GLOG.SRS=sqrt(mean((T.GLOG.SRS-T)^2))
RelB.GBC.SRS=(mean(T.GBC.SRS)-T)/T
rmse.GBC.SRS=sqrt(mean((T.GBC.SRS-T)^2))
```

```
result.RB.SRS[ssize,]=cbind(RelB.D.SRS, RelB.ML.SRS, RelB.MLOG.SRS, RelB.MBC.SRS,
RelB.GL.SRS, RelB.GLOG.SRS, RelB.GBC.SRS)
result.rmse.SRS[ssize,]=cbind(rmse.D.SRS, rmse.ML.SRS,rmse.MLOG.SRS,rmse.MBC.SRS,
rmse.GL.SRS, rmse.GLOG.SRS, rmse.GBC.SRS, count)
```

```
#-----SSRS-----
```

```
beefpop=rbind(pop1,pop2)

prob=c(n*2/3/NN[1], n*1/3/NN[2])
pop1$w=1/prob[1]
pop2$w=1/prob[2]
nn=c(round(NN[1]*prob[1]), round(NN[2]*prob[2]))
beefpop.O=rbind(pop1,pop2)

T.D.SSRS=numeric(REP)
T.ML.SSRS=numeric(REP)
T.MLOG.SSRS=numeric(REP)
T.MBC.SSRS=numeric(REP)
T.GL.SSRS=numeric(REP)
```

```

T.GLOG.SSRS=numeric(REP)
T.GBC.SSRS=numeric(REP)
lamda.SSRS=numeric(REP)
lamda.w.SSRS=numeric(REP)
B.SSRS=matrix(0,REP,2)
B.w.SSRS=matrix(0,REP,2)
sigma.SSRS=numeric(REP)
sigma.w.SSRS=numeric(REP)
interm=matrix(0,n,REP)

count=0
for (r in c(1:REP))
  {
  flag=0
  while (flag==0)
    {
    sam1=sample(1:NN[1],nn[1],replace=F)
    sam2=sample((NN[1]+1):N,nn[2],replace=F)
    sam.O=c(sam1,sam2)
    temp=estimates(beefpop.O, sam.O)
    T.D.SSRS[r]=temp$T.D
    T.ML.SSRS[r]=temp$T.ML
    T.MLOG.SSRS[r]=temp$T.MLOG
    T.MBC.SSRS[r]=temp$T.MBC
    T.GL.SSRS[r]=temp$T.GL
    T.GLOG.SSRS[r]=temp$T.GLOG
    T.GBC.SSRS[r]=temp$T.GBC
    lamda.SSRS[r]=temp$lamda
    lamda.w.SSRS[r]=temp$lamda.w
    B.SSRS[r,]=temp$B
    B.w.SSRS[r,]=temp$B.w
    sigma.SSRS[r]=temp$sigma
    sigma.w.SSRS[r]=temp$sigma.w

    interm[,r]=temp$sam
    if (temp$T.MBC!=-999&temp$T.GBC!=-999)      flag=1
    }
  }

```

```

        count=count+1
    }
    RelB.lamda.SSRS=(mean(lamda.SSRS)-lamda0)/lamda0
    rmse.lamda.SSRS=sqrt(mean((lamda.SSRS-lamda0)^2))
    RelB.B1.SSRS=(mean(B.SSRS[,1])-B0[1])/B0[1]
    rmse.B1.SSRS=sqrt(mean((B.SSRS[,1]-B0[1])^2))
    RelB.B2.SSRS=(mean(B.SSRS[,2])-B0[2])/B0[2]
    rmse.B2.SSRS=sqrt(mean((B.SSRS[,2]-B0[2])^2))
    RelB.sigma.SSRS=(mean(sigma.SSRS)-sigma0)/sigma0
    rmse.sigma.SSRS=sqrt(mean((sigma.SSRS-sigma0)^2))

    RelB.lamda.w.SSRS=(mean(lamda.w.SSRS)-lamda0)/lamda0
    rmse.lamda.w.SSRS=sqrt(mean((lamda.w.SSRS-lamda0)^2))
    RelB.B1.w.SSRS=(mean(B.w.SSRS[,1])-B0[1])/B0[1]
    rmse.B1.w.SSRS=sqrt(mean((B.w.SSRS[,1]-B0[1])^2))
    RelB.B2.w.SSRS=(mean(B.w.SSRS[,2])-B0[2])/B0[2]
    rmse.B2.w.SSRS=sqrt(mean((B.w.SSRS[,2]-B0[2])^2))
    RelB.sigma.w.SSRS=(mean(sigma.w.SSRS)-sigma0)/sigma0
    rmse.sigma.w.SSRS=sqrt(mean((sigma.w.SSRS-sigma0)^2))

    result.RB.phi.SSRS[ssize,]=cbind(RelB.lamda.SSRS, RelB.lamda.w.SSRS, RelB.B1.SSRS,
    RelB.B1.w.SSRS, RelB.B2.SSRS, RelB.B2.w.SSRS, RelB.sigma.SSRS, RelB.sigma.w.SSRS)
    result.rmse.phi.SSRS[ssize,]=cbind(rmse.lamda.SSRS, rmse.lamda.w.SSRS, rmse.B1.SSRS,
    rmse.B1.w.SSRS, rmse.B2.SSRS, rmse.B2.w.SSRS, rmse.sigma.SSRS, rmse.sigma.w.SSRS)

    RelB.D.SSRS=(mean(T.D.SSRS)-T)/T
    rmse.D.SSRS=sqrt(mean((T.D.SSRS-T)^2))

    RelB.ML.SSRS=(mean(T.ML.SSRS)-T)/T
    rmse.ML.SSRS=sqrt(mean((T.ML.SSRS-T)^2))
    RelB.MLOG.SSRS=(mean(T.MLOG.SSRS)-T)/T
    rmse.MLOG.SSRS=sqrt(mean((T.MLOG.SSRS-T)^2))
    RelB.MBC.SSRS=(mean(T.MBC.SSRS)-T)/T
    rmse.MBC.SSRS=sqrt(mean((T.MBC.SSRS-T)^2))

    RelB.GL.SSRS=(mean(T.GL.SSRS)-T)/T
    rmse.GL.SSRS=sqrt(mean((T.GL.SSRS-T)^2))

```

```

RelB.GLOG.SSRS=(mean(T.GLOG.SSRS)-T)/T
rmse.GLOG.SSRS=sqrt(mean((T.GLOG.SSRS-T)^2))
RelB.GBC.SSRS=(mean(T.GBC.SSRS)-T)/T
rmse.GBC.SSRS=sqrt(mean((T.GBC.SSRS-T)^2))

result.RB.SSRS[ssize,]=cbind(RelB.D.SSRS, RelB.ML.SSRS, RelB.MLOG.SSRS,
RelB.MBC.SSRS, RelB.GL.SSRS, RelB.GLOG.SSRS, RelB.GBC.SSRS)
result.rmse.SSRS[ssize,]=cbind(rmse.D.SSRS,
rmse.ML.SSRS,rmse.MLOG.SSRS,rmse.MBC.SSRS, rmse.GL.SSRS, rmse.GLOG.SSRS,
rmse.GBC.SSRS, count)

ssize=ssize+1
}#----END n=30, 50, 80,100, 130, 150----

rownames(result.RB.SRS)=c("n=30","n=50","n=80","n=100","n=130","n=150")
rownames(result.RB.SSRS)=c("n=30","n=50","n=80","n=100","n=130","n=150")
colnames(result.RB.SRS)=c("DESIGN","MBASED_L","MBASED_LLOG","MBASED_BC","GREG_L",
"GREG_LLOG","GREG_BC")
result.RB=rbind(result.RB.SRS,result.RB.SSRS)
rownames(result.rmse.SRS)=c("n=30","n=50","n=80","n=100","n=130","n=150")
rownames(result.rmse.SSRS)=c("n=30","n=50","n=80","n=100","n=130","n=150")
colnames(result.rmse.SRS)=c("DESIGN","MBASED_L","MBASED_LLOG","MBASED_BC","GREG_L",
"GREG_LLOG","GREG_BC","COUNT")
result.rmse=rbind(result.rmse.SRS,result.rmse.SSRS)

rownames(result.RB.phi.SRS)=c("n=30","n=50","n=80","n=100","n=130","n=150")
rownames(result.RB.phi.SSRS)=c("n=30","n=50","n=80","n=100","n=130","n=150")
colnames(result.RB.phi.SRS)=c("lamda","lamda.w","B1","B1.w","B2","B2.w","sigma","sigma.w")
result.RB.phi=rbind(result.RB.phi.SRS,result.RB.phi.SSRS)
rownames(result.rmse.phi.SRS)=c("n=30","n=50","n=80","n=100","n=130","n=150")
rownames(result.rmse.phi.SSRS)=c("n=30","n=50","n=80","n=100","n=130","n=150")
colnames(result.rmse.phi.SRS)=c("lamda","lamda.w","B1","B1.w","B2","B2.w","sigma","sigma.w")
result.rmse.phi=rbind(result.rmse.phi.SRS,result.rmse.phi.SSRS)

result.RB; result.rmse; result.RB.phi; result.rmse.phi
B0; lamda0; sigma0

```

Bibliography

- Arora, V., Lahiri, P. and Mukherjee, K. (1997), Empirical Bayes estimation of finite population means from complex surveys, *Journal of the American Statistical Association*, **92**, 1555-62.
- Bickel, P. J. and Doksum, K. A. (1981), An Analysis of Transformations Revisited, *Journal of the American Statistical Association*, **76**, 296-311.
- Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex survey. *International Statistical Review*, **51**, 279-92.
- Binder, D. A. (1992), Fitting Cox's proportional hazards models from survey data, *Biometrika*, **79**, 139-47.
- Brewer, K. R. W., and Mellor, R. W. (1975), The Effect of Sample Structure on Analytical Surveys. Int. Assoc. of Survey Statisticians, 1st Meeting, Vienna, August 1973.
- Bolfarine, H. and Zacks, S. (1992), *Prediction Theory for Finite Populations*, Springer: Verlag.
- Box, G. E. and Cox, D. R. (1964), An Analysis of Transformations, *Journal of the Royal Statistical Society, Series B*, **26**, 211-52.
- Brewer, K.R.W. (1963), Ratio Estimation and Finite Populations: Some Results Deducible from the Assumption of an Underlying Stochastic Process, *Australian Journal of Statistics*, **5**, 93-105.
- Brewer, K.R.W., and Mellor, R.W. (1975), The Effect of Sample Structure on Analytical Surveys, International Association of Survey Statisticians, 1st Meeting, Vienna, August 1973.
- Campbell, C. (1977), Properties of ordinary and weighted least squares estimators for two stage samples, *Proceedings of the social statistics section, American Statistical Association*, 800-05.
- Carroll, R. J. and Ruppert, D. (1981), *Robust Estimation in Heteroscedastic Linear Models*, NTIS: Springfield.
- Carroll, R. J. and Ruppert, D. (1988), *Transformations and Weighting in Regression*, London: Chapman and Hall.
- Chambers, R. L. and Dorfman, A. H. (2003), Transformed Variables in Survey Sampling, Technical Report.

- Chambers, R. L. and Skinner, C. J. (2003), *Analysis of Survey Data*, Wiley: Chichester.
- Chambless, L. E. and Boyle, K. E. (1985), Maximum likelihood methods for complex sample data: logistic regression and discrete proportional hazards models, *Communications in Statistics – Theory and Methods*, **14**, 1377-92.
- Chen, G. and Chen, J. (1996), A Transformation Method for Finite Population Sampling Calibrated with Empirical Likelihood, *Survey Methodology*, **22**, 139-46.
- Cochran, W. G. (1939), The Use of Analysis of Variance in Enumeration by Sampling, *Journal of the American Statistical Association*, **34**, 492-510.
- Collins, M. and Butcher, B. (1982), Interviewer and clustering effects in an attitude survey, *Journal of the Market Research Society*, **25**, 39–58.
- Dagne, G. A. (2003), The use of transformations in small area estimation, *Journal of Applied Statistics*, **30**, 411-23.
- Davison, C. W., Arnade, C. A. and Hallahan, C. B. (1989), Box-Cox Estimation of U.S. Soyabean Exports, *Journal of Agricultural Economics Research*, **41**, 8-15.
- Estevao, V., Hidioglou, M., and Särndal C.E. (1995). Methodological principles for a generalized estimation system at statistics Canada. *Journal of Official Statistics*, **11**, 181-204.
- Feather, J. (1973), A Study of Interviewer Variance, WHO International Collaborative Study of Medical Care Utilization, Saskatchewan Study Area Reports, Series II, Monograph No. 3.
- Fellegi, I.P. (1964), Response variance and its estimation, *Journal of the American Statistical Association*, **59**, 1016–41.
- Fuller, W. A. (1973), Regression Analysis for Sample Surveys, Int. Assoc. of Survey Statisticians, 1st Meeting, Vienna, August 1973.
- Fuller, W.A., and Battese, G.E. (1973), Transformations for estimation of linear models with nested error structures, *Journal of the American Statistical Association*, **68**, 626-32.
- Fuller, W., Loughin, M., and Baker, H. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 nationwide food consumption survey. *Survey Methodology* **20**, 75-85.
- Ghosh, M. and Lahiri, P. (1987), Robust empirical Bayes estimation of means from stratified samples, *Journal of the American Statistical Association*, **82**, 1153-62.

- Ghosh, M., Lahiri, P. and Tiwari, R.C. (1989), Nonparametric Bayes and empirical Bayes estimation of the distribution function and the mean, *Communications in Statistics: Theory and Methods*, **18**, 121-46.
- Ghosh, M., and Meeden, G. (1986), Empirical Bayes estimation in finite population sampling, *Journal of the American Statistical Association*, **81**, 1058-1062.
- Ghosh, M. and Meeden, G. (1997), *Bayesian Methods for Finite Population Sampling*, Chapman and Hall, London.
- Godambe, V.P., and Thompson, M.E. (1986). Parameters of super populations and survey population: their relationship and estimation. *International Statistical Review* **54**, 37-59.
- Graubard, B. and Korn, E. (2002), Inference for Superpopulation Parameters using Sample Surveys, *Statistical Science*, **17**, 73-96.
- Gray, P.G. (1956), Examples of interviewer variability taken from two sample surveys, *Applied Statistics*, **V**, 73–85.
- Graybill, F.A. (1983), *Matrices with applications in statistics*, 2nd edn. Belmont, CA: Wadsworth.
- Groves, R.M. (1989), *Survey Errors and Survey Costs*, New York: Wiley.
- Gurka, M. (2004), *The Box-Cox Transformation in the General Linear Mixed Model for Longitudinal Data*. Ph.D. Dissertation.
- Gurka, M. (2004). Expanding the Box-Cox transformation to the linear mixed model, *Journal of the Royal Statistical Society, Series A*, **169**, 273-88.
- Hansen, M.H., Hurwitz, W.N., and Bershad, M.A. (1961), Measurement Errors in Censuses and Surveys, *Bulletin of the International Statistical Institute*, **38**, 359-74.
- Hartley, H.O. and Rao, J.N.K. (1968), A new estimation theory for sample surveys, *Biometrika*, **55**, 547-557.
- Harley, H. O. and Silken, R. L. (1975), A “Super-population Viewpoint” for Finite Population Sampling, *Biometrics*, **31**, 411-22.
- Henderson, C.R. (1953), Estimation of variance and covariance components, *Biometrics*, **9**, 226-52.
- Hinkley, D. V. and Runger, G. (1984), The Analysis of Transformed Data (with discussion). *Journal of the American Statistical Association*, **79**, 302-20.

- Holt, D., and Scott, A.J. (1981), Regression analysis using survey data, *The Statistics*, **30**, 169-78.
- Holt, D., Smith, T. M. F., and Winter, P. D. (1980), Regression Analysis of Data from Complex Surveys, *Journal of the Royal Statistical Society, Series A*, **143**, 474-87.
- Horvitz, D.G. and Thompson, D.J. (1952), A Generalization of Sampling Without Replacement From A Finite Population, *Journal of American Statistical Association*, **47**, 663-685.
- Isaki, C. T. and Fuller, W. A. (1982), Survey Design Under the Regression Superpopulation Model, *Journal of the American Statistical Association*, **77**, 89-96.
- Jayasuriya, B., and Valliant, R. (1996). An application of restricted regression estimation to post-stratification in a household survey. *Survey Methodology* **22**, 127-37.
- Jiang, J. (1996), REML estimation: asymptotic behavior and related topics, *Annals of Statistics*, **24**, 255-86.
- Jiang, J., and Lahiri, P. (2006), Estimation of Finite Population Domain Means - A Model-Assisted Empirical Best Prediction Approach, *Journal of the American Statistical Association*, **101**, 301-311.
- John, J. A. and Draper, N. R. (1980), An Alternative Family of Transformations, *Applied Statistics*, **29**, 190-97.
- Karlberg, F. (2000), Survey Estimation for Highly Skewed Population in the Presence of Zeroes, *Journal of Official Statistics*, **16**, 229-41.
- Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M. P. eds. (1989), *Panel Surveys*, New York: John Wiley.
- Kish, L. (1962), Studies of interviewer variance for attitudinal variables, *Journal of the American Statistical Association*, **57**, 92-115.
- Kish, L., and Frankel, M.R. (1974). Inference from complex samples (with discussion). *Journal of the Royal Statistical Society, Series B* **36**, 1-37.
- Korn, E. L. and Graubard, B. I. (1995), Examples of differing weighted and unweighted estimates from a sample survey, *The American Statistician*, **49**, 291-95.
- Korn, E. and Graubard, B. (1998), Confidence Intervals for Proportions with Small Expected Number of Positive Counts Estimated from Survey Data, *Survey Methodology*, **24**, 193-201.

- Korn, E. and Graubard, B. (1999), *Analysis of Health Surveys*, New York: John Wiley & Sons, Inc.
- Kott, P.S. (2005), Randomized-assisted model-based survey sampling, *Journal of Statistical Planning and Inference*, **129**, 263-277.
- Lohr, S. L. (1999), *Sampling: Design and Analysis*, Pacific Grove, California: Duxbury.
- Meeden, G. (1999), A noninformative Bayesian approach for two-stage cluster sampling, *Sankhya*, Series B, **61**, 133-144.
- Miner, A. G. (1982), The Contribution of Weather and Technology to U.S. Soybean Yield, *Unpublished Dissertation*, University of Minnesota.
- Narain, R.D. (1951), On Sampling Without Replacement With Varying Probabilities, *Journal of Indian Social Agricultural Statistician*, **3**, 169-174.
- Nathan, G. and Holt, D. (1980), The effect of survey design on regression analysis, *Journal of the Royal Statistical Society*, B **42**, 377-86.
- Newman, P. (1977), Malaria and Mortality, *Journal of the American Statistical Association*, **72**, 257-63.
- Nordberg, L. (1989), Generalized linear modeling of sample survey data, *Journal of Official Statistics*, **5**, 223-39
- Pfefferman, D. (1993), The role of sampling weights when modeling survey data, *International Statistical Review*, **61**, 317-37.
- Pfeffermann, D. and Holmes, D. J. (1985), Robustness considerations in the choice of a method of inference fro regression analysis of survey data, *Journal of the Royal Statistical Society*, A **148**, 268-78.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York: Wiley.
- Rao, J.N.K. (2005), Interplay between sample survey theory and practice: an appraisal, *Survey Methodology*, **31**, 117-138.
- Rao, J.N.K., Sutradhar, B.C., and Yue, K. (1993), Generalized least squares F test in regression analysis with two-stage cluster samples, *Journal of the American Statistical Association*, **88**, 1388-91.
- Rao, J.N.K. and Thomas, D.R. (1989), Chi-Squared Tests For Contingency Tables, Skinner, C.J., Holt, D., and Smith, T.M.F. (eds.). *Analysis of Data From Complex Surveys*. Chichester: Wiley, 89-114.

- Royall, R. M. (1970), On Finite Population Sampling Under Certain Linear Regression Models, *Biometrika*, **57**, 377-87.
- Royall, R. M. (1976), The Linear Least Squares Prediction Approach to Two-Stage Sampling, *Journal of the American Statistical Association*, **71**, 657-64.
- Royall, R. M. (1986), Model Robust Confidence Intervals Using Maximum Likelihood Estimators, *International Statistical Review*, **54**, 221-26.
- Royall, R. M. and Cumberland, W. G. (1981), The Finite Population Linear Regression Estimator and Estimators of its Variance – An Empirical Study, *Journal of the American Statistical Association*, **76**, 924-30.
- Sakia, R. M. (1990), Retransformation Bias: a Look at the Box-Cox Transformation to Linear Balanced Mixed Balanced Model, *Metrika*, **37**, 345-351.
- Sakia, R. M. (1992), The Box-Cox Transformation Technique: a review, *The Statistician*, **41**, 169-78.
- Särndal, C.E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Schlesselman, J. (1971), Power Families: A Note on the Box and Cox Transformation, *Journal of the Royal Statistical Society, Series B* **33**, 307-11.
- Schnell, R. and Kreuter, F. (2005), Separating interviewer and sampling point effects, to appear in *Journal of Official Statistics*.
- Scott, A.J., and Holt, D. (1982), The effect of two-stage sampling on ordinary least squares methods, *Journal of the American statistical Association*, **77**, 848-54.
- Scott, A.J. and Wild, C.J. (1989), Hypothesis Testing in Case-Control Studies, *Biometrika*, **76**, 806-808.
- Spitzer, J. J. (1976), The Demand for Money, the Liquidity Trap and Functional Forms, *The International Economics Review*, **17**, 220-27.
- Spitzer, J. J. (1982a), A Primer on Box-Cox Estimation, *Review of the Economics and Statistics*, **64**, 307-13.
- Spitzer, J. J. (1982b), A Fast and Efficient Algorithm for the Estimation of Parameters in Models with the Box and Cox Transformation, *Journal of the American Statistical Association*, **77**, 760-66.
- Thomas, D.R., and Rao, J.N.K. (1987), Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling, *Journal of the American Statistical Association*, **82**, 630-36.

- Tukey, J. W. (1957), The Comparative Anatomy of Transformation, *Annals of Mathematical Statistics*, **28**, 601-32.
- Taylor, J. M. G. (1986), The Retransformed Mean After a Fitted Power Transformation, *Journal the American Statistical Association*, **81**, 114-118.
- Valliant, R. (1985), Nonlinear Prediction Theory and the Estimation of Proportions in a Finite Population, *Journal of the American Statistical Association*, **80**, 631-41.
- Valliant, R. (1986), Mean Squared Error Estimation Finite Populations under Nonlinear Models, *Communications in Statistics A*, **15**, 1975-93.
- Valliant, R., Dorfman, A. H. and Royall, R. M. (2000), *Finite Population Sampling and Inference: a Prediction Approach*, New York: Wiley.
- Wu, C.F.J., Holt, D., and Holmes, D.J. (1988), The effect of two-stage sampling on the F statistics, *Journal of the American Statistical Association*, **83**, 150-59.
- Zarembka, P. (1968), Functional Form in the Demand for Money, *Journal of the American Statistical Association*, **63**, 502-11.
- Zarembka, P. (1974), Transformation of Variables in Econometrics, *Frontiers in Econometrics*, ed. Paul Zarembka, New York: Academic Press.