2015

Median Regression for Complex Survey Data

Raphael André Fraser

FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

MEDIAN REGRESSION FOR COMPLEX SURVEY DATA

By

RAPHAEL ANDRÉ FRASER

A Dissertation submitted to the
Department of Statistics
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2015

Raphael André Fraser defended this dissertation on August 10, 2015.
The members of the supervisory committee were:

Debajyoti Sinha
Professor Co-Directing Dissertation

Stuart R. Lipsitz
Professor Co-Directing Dissertation

Elwood Carlson
University Representative

Elizabeth Slate
Committee Member

Fred Huffer
Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with university requirements.

To my parents, Uriel & Edna Fraser.

ACKNOWLEDGMENTS

I have received help from a number of people while writing this dissertation. Foremost among them are my advisors, Professors Debajyoti Sinha and Stuart R. Lipsitz, who provided many insights and constructive comments on initial drafts of my dissertation. Also many thanks to my other committee members. I owe a particular debt of gratitude to Dr. Depdeep Pati for accommodating myriads of questions and facilitating many discussions about Bayesian statistics. Finally, I would like to thank my wife Kayann for her unwavering support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

The following short list of symbols are used throughout the document. The symbols represent quantities that I tried to use consistently.

h	Stratum index; $h = 1, \ldots, L$.
i	PSU index; $i = 1, \ldots, n_h$.
j	Observation index; $j = 1, \ldots, m_{hi}$.
L	Number of strata.
n	Total sample size; total number of PSUs in sample, $\sum_{h=1}^{L} n_h$.
N	Population size; total number of PSUs in population, $\sum_{h=1}^{L} N_h$.
n_h	Sample size in stratum; the number of PSUs sampled from stratum h
N_h	Population size in stratum; the number of PSsU in stratum h
f_h	Sampling rate for stratum h; $f_h = n_h/N_h$.
$\hat{V}_{BRR}(\hat{\beta})$	Estimator of variance $V_{BRR}(\hat{\beta})$ obtained from balanced repeated replication.
$\hat{V}_{RBS}(\hat{\beta})$	Estimator of variance $V_{RBS}(\hat{\beta})$ obtained from Rescaled bootstrap.
t_{hi}^{r}	Bootstrap frequency; the number of times unit h, i is sampled in the r-th replicate.
$\delta_{hij}^{(r)}$	Replicate weight of unit h, i, j in the r-th replicate.
δ_{hij}	Sampling weight of unit h, i, j.
β	Population parameter.
$\hat{\beta}$	Parameter estimate obtained from original survey data.
$\hat{\beta}^{(r)}$	Parameter estimate obtained in the r-th replicate.

ABSTRACT

The ready availability of public-use data from various large national complex surveys has immense potential for the assessment of population characteristics—means, proportions, totals, etcetera. Using a model-based approach, complex surveys can be used to evaluate the effectiveness of treatments and to identify risk factors for important diseases such as cancer. Existing statistical methods based on estimating equations and/or utilizing resampling methods are often not valid with survey data due to design features such as stratification, multistage sampling and unequal selection probabilities. In this paper, we accommodate these design features in the analysis of highly skewed response variables arising from large complex surveys. Specifically, we propose a double-transform-both-sides based estimating equations approach to estimate the median regression parameters of the highly skewed response; the double-transform-both-sides method applies the same transformation twice to both the response and regression function. The usual sandwich variance estimate can be used in our approach, whereas a resampling approach would be needed for a pseudo-likelihood based on minimizing absolute deviations. Furthermore, the double-transform-both-sides estimator is relatively robust to the true underlying distribution, and has much smaller mean square error than the least absolute deviations estimator. The method is motivated by an analysis of laboratory data on urinary iodine concentration from the National Health and Nutrition Examination Survey.

CHAPTER 1

INTRODUCTION

The problem of analyzing skewed data arises in many applied fields, such as medicine, public health, epidemiology and economics. A popular approach to this problem is to transform the response using a transformation function. However, this approach can lead to regression coefficients that are difficult to interpret. Further, there is no guarantee that a suitable transformation even exists. An alternative to the popular approach is to use median regression. Median regression has become increasingly popular since the seminal work of Koenker and Bassett Jr (1978). The median is a simple and meaningful measure of center for skewed distributions. Therefore, median regression is well suited for skewed data. Median (least absolute deviations) regression is appealing because it does not rely on parametric assumptions concerning the error terms as in mean regression.

In conducting large national surveys we often encounter skewed data. For example, in the Medical Expenditure Panel Survey, the variable medical expenditure is skewed. Another example is the Current Population Survey where personal income is skewed. Although median regression for univariate independent and identically distributed data is frequently seen in the literature, median regression is not a common choice for analyzing skewed data for complex sample surveys. Why is this the case? And can median regression be applied to survey data successfully? The answer to these questions is the focal point of this dissertation.

The purpose of this chapter is to introduce the reader to the basic ideas of the material to be encountered in later chapters. The first section introduces complex sample surveys and some important design features. Section 2 discusses how sample weights are calculated for a stratified multistage design. Section 3 highlights two popular ways in complex surveys to estimate variance for a non-differentiable estimator such as the least absolute deviations estimator. Four transformation functions are reviewed along with their properties in the context of transform-both-sides regression. Finally, we review transform-both-sides regression including estimation and inference.

1.1 Complex Sample Surveys

In 1936 when George Gallup correctly predicted Franklin D. Roosevelt as the presidential election winner, from the replies of only 50,000 responses, public opinion surveys entered the age of scientific sampling. This was in direct contradiction to the venerable Literary Digest poll who had correctly predicted the outcome of the last five presidential elections. The Literary Digest polled about 10 million Americans, and received responses from nearly 2.5 million. The polled showed that Alf Landon would likely be the overwhelming winner with 57% of the votes. However, Roosevelt won by a landslide victory with 62% of the votes. The magazine was completely discredited and was soon discontinued. Today Gallup Inc. founded by George Gallup is world renowned for their public opinion polls.

What went wrong? The problem was that the sample used by Literary Digest was not representative of the target population. The magazine had surveyed its own readers, registered automobile owners, and registered telephone users during the period of the Great Depression. These groups had incomes well above the national average of the day which resulted in lists of voters far more likely to support the Republicans than a typical voter of the time. Since the sample was biased, predictions would be inaccurate. Sample surveys have advanced greatly since the days of Gallup, relying on complex probability sampling designs.

Sample surveys differ fundamentally from the rest of statistics in how the data is sampled. Traditional statistics assumes that our data is a simple random sample from an infinite population whereas in sample surveys we sample a fraction of a finite population. In addition, sample surveys may involve dividing the population into strata, sampling clusters with different probabilities of being selected, and multiple stages of sampling. These design features are known as stratification, unequal probability sampling and multistage sampling, respectively. Consequently, weighted analyses are necessary to obtain unbiased or nearly unbiased estimates of population parameters. Variance estimation for estimators depend on the sampling design of the sample survey and requires approximate methods; typically, Taylor series linearization or resampling methods.

Ignoring the fact that we are analyzing sample survey data can lead to incorrect inference. The analysis would result in biased estimates of the population parameters and incorrect variance of

the parameter estimates. Using the sampling weight (defined in section 1.2) we can obtain the appropriate estimates of population parameters. However, the estimated variance would still be incorrect because the variance estimation method does not take into account stratification and/or clustering of the survey design. Consequently, both sampling weights and variance estimation are of paramount concern in sample surveys. What follows is a brief discussion of sampling weights and variance estimation methods.

1.2 Sampling Weights

Sampling weights are used primarily to compensate for unequal probability of selection resulting in unbiased estimates of population parameters. Other secondary uses includes compensating for non-response and non-coverage of the population, etcetera. In sample surveys the sampling weights are defined as the reciprocal of the probability of selection. The probability of selection in turn depends on the sample design used to select a unit. For example, if we sample 2,000 people from Florida with a total population of 20 million under a simple random sample design then any person in Florida has a 1 in 10,000 chance of being sampled. That is, 2,000/20,000,000. Therefore the sample weight for each person in the state is $10,000 = (1/10,000)^{-1}$. We can think of the sample weights as representing 10,000 individuals in the Florida population. The sum of the weights for all 2,000 persons provides an unbiased estimate of the total number of units in the target population (i.e. 2,000$\times$10,000 = 20,000,000). This is known as the Horvitz-Thompson estimator. For the remainder of this chapter, let us suppose $h = 1, ..., L$ is the stratum index, $i = 1, ..., n_h$ is the cluster index and $j = 1, ..., m_{hi}$ is the observation index.

For a stratified multistage designs, the sample weights must reflect the probabilities of selection at each stage in each stratum. For instance, in the case of a stratified two-stage design in which the i-th cluster in stratum h is selected with probability π_{hi} at the first stage, and the j-th observation is selected within a sampled cluster in stratum h with probability $\pi_{j|hi}$ at the second stage, then the overall probability of selection π_{hij} of each observation in the sample is given by $\pi_{hij} = \pi_{hi} \times \pi_{j|hi}$. Hence the sampling weight is $\delta_{hij} = 1/\pi_{hij}$. Suppose an equal probability sample of n_h clusters is selected from a total of N_h clusters at the first stage in stratum h and observations are then

selected from each sampled cluster of size M_{hi}. Therefore, the overall probability of selection of an observation is

$$\pi_{hij} = \pi_{hi} \times \pi_{j|hi} = \frac{n_h}{N_h} \times \frac{m_{hi}}{M_{hi}}. \tag{1.1}$$

It is important to note that for many complex surveys in practice, the sampling probabilities π_{hij} are known only up to a proportionality constant, and π_{hi} and $\pi_{j|hi}$ are usually not available to us for most large national complex surveys such as the National Health and Nutrition Examination Survey (NHANES). Although the survey weights can be used to find a point estimate of any population quantity, the weights are not sufficient information to calculate variance of statistics. Variance estimation depend on the stratification and clustering in the survey design. In most cases, only the stratification and information from the first stage of clustering are used to calculate the variance of estimates.

1.3 Variance Estimation in Complex Surveys

For many sample survey designs that are used in practice, the variance estimation process may involve stratification, several stages of cluster sampling, and other procedures. As a result, the variance of the estimate may not be linear or even a known function of the population parameters. In order to estimate the variances of estimates obtained from complex surveys, one of two general classes of methods has been developed.

The two most widely used methods of variance estimation are Taylor series linearization (Wolter, 2007; Shah, 1998) and resampling techniques (Wolter, 2007; Rust and Rao, 1996). For estimators that are smooth functions of the sample data—means, proportions, totals, etc.—both methods give comparable variance estimates and neither is clearly preferred. For estimators that are non-smooth functions of the sample data—for example, medians—a particular resampling procedure, balanced repeated replication, seems preferred over Taylor series linearization and jackknife, another resampling method (Korn and Graubard, 2011). In the discussion that follows both the Taylor series and the resampling methods are explained briefly. We assume a stratified multistage design and use the same notation as in section 1.2.

1.3.1 Taylor Series Linearization

In the Taylor series approach to variance estimation, the non-linear estimator is expanded as a infinite Taylor series centered at the expected value of the numerator and the expected value of the denominator. The non-linear estimator is approximated using a first order Taylor series by retaining only the leading terms in the infinite series, resulting in a linear function of sample data. That is, the non-linear estimator has been linearized. Hence, the estimated variance of the linearized function can be obtained directly. For a stratified multistage design, the varinace of the linearized function is estimated within each stratum separately and then the stratum specific estimated variances are summed to obtain the variance of the estimator.

A drawback of the Taylor series method is that a unique approximate variance estimation formula needs to be derived for every different non-linear estimator and for each possible survey design. In addition, the method cannot be applied to non-differentiable estimators; the least absolute deviations estimator, for example, does not fit into this framework. For non-differentiable estimators, resampling methods are recommended.

1.3.2 Resampling Methods

Resampling methods for variance estimation of sample surveys estimators are computer-intensive but offer more options than the Taylor series linearization method in terms of the number of different estimators for which estimated variances can be computed. All the resampling methods in this section calculate variance estimates for a sample in which the primary sampling units (PSUs) are sampled without replacement. If the primary sampling units are sampled without replacement, as is often the case in practice, these methods may still be used but are expected to overestimate the variance.

Balanced Repeated Replication. Balanced repeated replication is a specific resampling technique that can be used for very general designs, namely, stratified multistage sampling. However, it was developed for the specific situation of exactly two primary sampling units selected (sampled) per stratum, generally sampled with unequal probability with or without replacement.

With balanced repeated replication (BRR), each replicate contains exactly half of the sample PSUs, one PSU from each stratum. Each replicate is called a half-sample. The total number of

possible different replicates is 2^L, where L is the number of strata. However, when L is large, the number of possible different replicates increases exponentially; requiring more computing time and is simply not practical. But what if a smaller and 'balanced set' of replicates can yield the same variance estimate that would be obtained from all possible replicates? This is known as BRR. R balanced replicates are formed, using a Hadamard matrix (Wolter, 2007), so that each sample PSU appears in the same number of replicates and each pair of sample PSUs from two different strata appears in the same number of replicates. The minimum number R of replicates required is the smallest integer that is greater than or equal to L but divisible by 4. For example, 16 strata, each with two sampled PSUs, would require 20 balanced replicates. Observations in sample PSUs that are omitted in each replicate have a value of zero for the replicate weight variable, and observations in sample PSUs that are included in each replicate have a value that is twice their sampling weight in the full sample.

Bootstrap. Theoretical results of the bootstrap for a simple random sample without replacement were developed by Efron (1979, 1982). Shao and Tu (2012) summarized theoretical results for the bootstrap in complex sample surveys. In this section we describe the rescaling bootstrap of Rao and Wu (1988) for a stratified multistage design.

For the rescaling bootstrap n_h primary sampling units are sampled from stratum h. We select a simple random sample of $(n_h - 1)$ primary sampling units with replacement from the n_h primary sampling units in the observed sample from stratum h. This is done independently for each stratum. The number of times the j-th primary sampling unit of stratum h is selected, $t_{hj}^{(r)}$, for each replicate r is recorded, resulting in the replicate weight for observation i in primary sampling unit j of stratum h as

$$\delta_{hij}^{(r)} = \delta_{hij} \left(\frac{n_h}{n_h - 1} \right) t_{hj}^{(r)},$$

Note that $t_{hj}^{(r)}$ is sometimes referred to as the bootstrap frequency. The procedure is repeated R times resulting in the replicate weight vectors $\delta^{(1)}, \ldots, \delta^{(R)}$. These new weight vectors are then used to estimate $\hat{\theta}_{(1)}^*, \ldots, \hat{\theta}_{(R)}^*$, estimators of θ. Thus, the variance of the estimator is

$$\hat{\mathrm{V}}_{RBS}(\hat{\theta}) = \frac{1}{R-1}\sum_{r=1}^{R}\left(\hat{\theta}_r^* - \hat{\theta}\right)^2, \tag{1.2}$$

where $\hat{\theta}$ is the estimator using the original weights. Theoretical results regarding inconsistencies of bootstrap-based and other resampling based estimator $\hat{\mathrm{V}}_{RBS}(\hat{\theta})$ for complex surveys is well-known (Shao, 1996; Shao et al., 2003; Shao and Tu, 2012). For complex surveys, the sampling probability of clusters π_{hi} and conditional probabilities $\pi_{j|hi}$ within selected cluster (as in section 1.2) are usually not available. This makes implementation of weighted bootstrap

1.4 Transformation Families

Many important results in statistical analysis assume independent and identically distributed normal random errors with mean zero and constant variance. In situations where these assumptions are seriously violated several options are available. However, in many cases, an appropriate transformation can reduce or remove nonnormality and heteroscedasticity from the data. Since the publication of the Box-Cox transformation function there has been many proposed transformation functions. In this section we review four of these transformations and highlight some deficiencies in the context of transform-both-sides regression.

Let $g_\lambda(y)$ be any transformation function with transformation parameter λ. To apply transform-both-sides regression successfully to any data the following properties of the transformation function must be met:

1. $g_\lambda(y)$ is convex in y for $\lambda > 1$ and concave in y for $\lambda < 1$,
2. $g_\lambda(y)$ is a continuous function of y,
3. $g'_\lambda(y)$ and $g''_\lambda(y)$ are continuous functions of y,
4. $g_\lambda(y)$ is a monotone function in y

The first point is essential for any good transformation function. If $g_\lambda(y)$ changes from convex to concave as y changes sign then it will be difficult to predict the effect of the transformation on skewed data. The second point is necessary since in transform-both-sides regression the regression

function is also transformed, meaning the transformation is a function of the regression parameters. If the transformation function is not continuous with respect to the regression parameters then the corresponding estimating equation will be discontinuous and non-differentiable with respect to the regression parameters. The third point is needed to compute valid variance and is closely related to the second point since $g'_\lambda(y)$ is included in the estimating equation expression while $g''_\lambda(y)$ will be a part of the expression that represents the Hessian matrix. Lastly, the transformation function needs to monotone so that the inverse of the transformation is unique. Therefore, we can back-transform to obtain the untransformed response variable which makes prediction and inference of the response possible. Additionally, the inverse transformation is useful for estimating quantiles.

The first transformation we consider is the venerable Box and Cox (1964) transformation function, $g_\lambda : \mathbb{R}^+ \to (-1/\lambda, \infty)$,

$$g_\lambda(y) = \frac{y^\lambda - 1}{\lambda}, \quad \frac{dg}{dy} = y^{\lambda-1}, \quad \frac{d^2g}{dy^2} = (\lambda - 1)y^{\lambda-2}. \tag{1.3}$$

where $\lambda \neq 0$. The first and second derivatives are continuous in y but the function only maps $\mathbb{R}^+$ to $(-1/\lambda, \infty)$ instead of the whole real number line, $\mathbb{R}$. Consequently, distributions that cover the entire real line such as the normal density can only be approximated.

Another transformation is the function proposed by Manly (1976) to handle skewed data, $g_\lambda : \mathbb{R} \to \mathbb{R}$,

$$g_\lambda(y) = \frac{e^{y\lambda} - 1}{\lambda}, \quad \frac{dg}{dy} = e^{y\lambda}, \quad \frac{d^2g}{dy^2} = \lambda e^{y\lambda}. \tag{1.4}$$

where $\lambda \neq 0$. This transformation function is very useful for handling unimodal skewed data. However, one of the limitations of this transformation function is its inability to transform bimodal or U-shaped distributions.

Bickel and Doksum (1981) suggested a modification to the Box-Cox function such that the range of the transformed values spans the entire real line, $g_\lambda : \mathbb{R} \to \mathbb{R}$. For $\lambda > 0$,

$$g_\lambda(y) = \frac{sgn(y) \mid y \mid^\lambda - 1}{\lambda}, \quad \frac{dg}{dy} = \mid y \mid^{\lambda-1}, \quad \frac{d^2g}{dy^2} = sgn(y)(\lambda - 1) \mid y \mid^{\lambda-2}. \tag{1.5}$$

However, this transformation is not without problems. It does not work well for extremely skewed distributions or when the variable y contains both negative and positive values. This is because $g_\lambda(y)$ changes from convex to concave as y changes sign. Yeo and Johnson (2000) gave an example of when y is a mixture distribution of the standard normal and gamma densities that resulted in a bimodal distribution after transformation instead of a distribution close to normality. This leads us to the next transformation.

The Yeo and Johnson (2000) transformation is the most recent transformation function on our list and possesses all of properties necessary for transform-both-sides regression. The function $g_\lambda : \mathbb{R} \to \mathbb{R}$ can be written as

$$g_\lambda(y) = \begin{cases} \{(1+y)^\lambda - 1\}/\lambda & \text{if } (y \geq 0, \lambda \neq 0), \\ -\{(1-y)^{2-\lambda} - 1\}/(2-\lambda) & \text{if } (y < 0, \lambda \neq 2), \end{cases} \tag{1.6}$$

$$\frac{dg}{dy} = \begin{cases} (1+y)^{\lambda-1} & \text{if } (y \geq 0, \lambda \neq 0), \\ (1-y)^{1-\lambda} & \text{if } (y < 0, \lambda \neq 2), \end{cases} \tag{1.7}$$

$$\frac{d^2g}{dy^2} = \begin{cases} (\lambda-1)(1+y)^{\lambda-2} & \text{if } (y \geq 0, \lambda \neq 0), \\ (1-\lambda)(1-y)^{-\lambda} & \text{if } (y < 0, \lambda \neq 2), \end{cases} \tag{1.8}$$

Figure 1.1 highlights the difference between the four transformations presented. Unlike the Bickel-Doksum transformation, the Yeo and Johnson (2000) transformation does not change from convex to concave as y changes sign. Additionally, the first and second derivatives are continuous in y. In the next section, we discuss estimation and inference of the transform-both-sides regression.

1.5 Transform-both-sides Regression

This section and the next is loosely based on chapter 4 of Carroll and Ruppert (1988). Skewed data occur frequently in many disciplines; biology, economics, medicine to name a few. There are two well established ways of modeling skewed data. One approach assumes the error distribution follows a parametric class of skewed densities. For example, the skew-normal density. The other approach assume that the response can be transformed to a symmetric distribution. These are known as transformation models of which the transform-both-sides regression is a member.

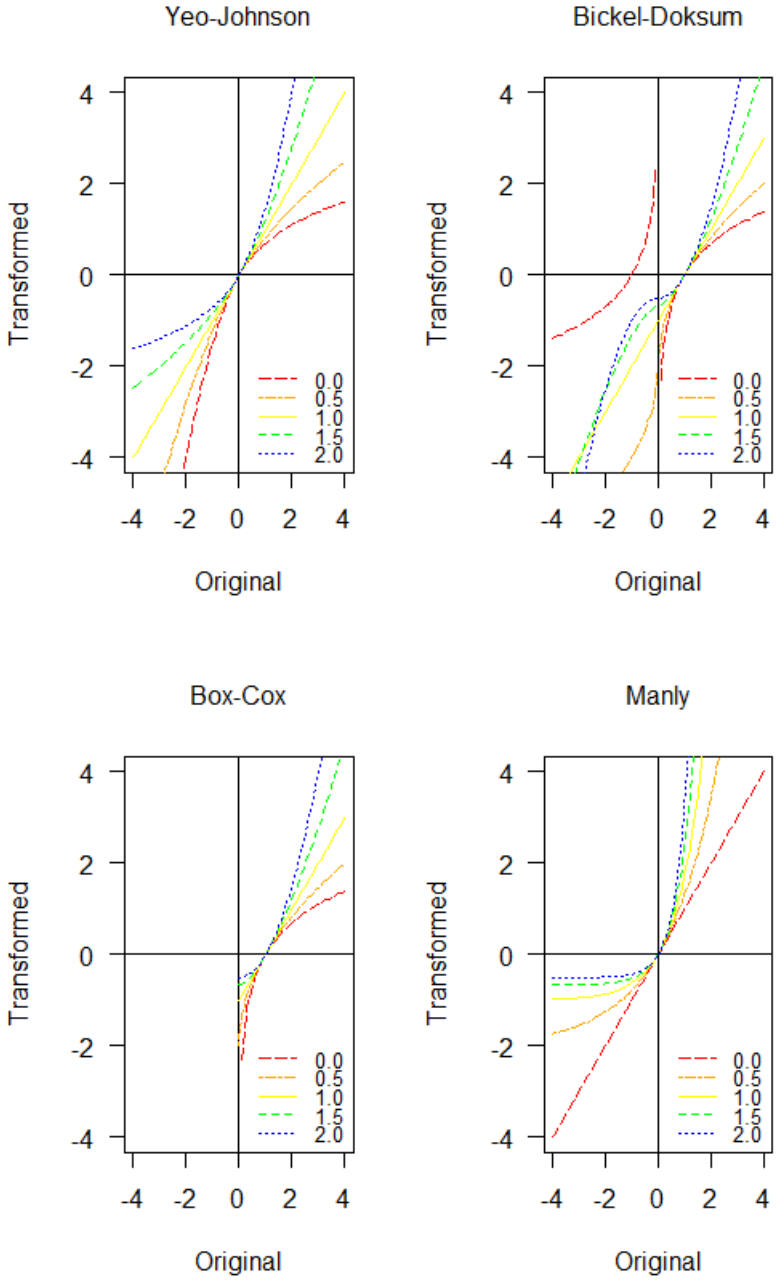


Figure 1.1: A comparison of the Box-Cox, Manly, Yeo-Johnson and Bickel-Doksum transformations with $\lambda = 0.0, 0.5, 1.0, 1.5, 2.0$.

Consider the following regression models

$$y_i = x_i^{\mathsf{T}}\beta + \epsilon_i \quad (i = 1, \ldots, n), \tag{1.9}$$

$$g_\lambda(y_i) = g_\lambda(x_i^{\mathsf{T}}\beta) + \epsilon_i \quad (i = 1, \ldots, n), \tag{1.10}$$

where ϵ_i are independent and identically distributed normal random variables with mean zero and variance σ_ϵ^2, y_i is the response, $x_i^{\mathsf{T}}\beta$ is the model, x_i is a p-vector of covariates, β is a p-vector of unknown parameters and $g_\lambda(\cdot)$ is a monotonic transformation function.

Models (1.9) and (1.10) are known as ordinary least squares regression and transform-both-sides regression, respectively. Model (1.10) is fundamentally different from model (1.9) because it can reduce or remove both skewness and/or heteroscedasticity. For simplicity of our discussion we will assume that the transformation function is the Yeo and Johnson (2000) transformation function such that

$$g_\lambda(y) = \frac{(1+y)^\lambda - 1}{\lambda} \quad (y \in \mathbb{R}^+, \lambda \neq 0).$$

Note that the conditional expectation and variance of (1.9) and (1.10) are given by

$$\mathrm{E}[y_i \mid x_i] = x_i^{\mathsf{T}}\beta, \quad \mathrm{V}[y_i \mid x_i] = \sigma_\epsilon^2 \quad \text{and} \quad \mathrm{E}[g_\lambda(y_i) \mid x_i] = x_i^{\mathsf{T}}\beta, \quad \mathrm{V}[g_\lambda(y_i) \mid x_i] = \sigma_\epsilon^2$$

Hence regression model (1.9) implies that the responses y_i are independent normal random variables with mean $x_i^{\mathsf{T}}\beta$ and variance σ_ϵ^2. Similarly, $g_\lambda(y_i)$ are independent normal random variables with mean $x_i^{\mathsf{T}}\beta$ and variance σ_ϵ^2. Consequently, we can write down the density of both $g_\lambda(y_i)$ and y_i as follows

$$f(g_\lambda(y_i)) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left(-\frac{1}{2\sigma^2}\{g_\lambda(y_i) - g_\lambda(f(x_i,\beta))\}^2\right) \tag{1.11}$$

and

$$f(y_i) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left(-\frac{1}{2\sigma^2}\{g_\lambda(y_i) - g_\lambda(f(x_i,\beta))\}^2\right) \times J(\lambda \mid y_i) \tag{1.12}$$

where $J(\lambda \mid y) = dg_\lambda(y)/dy$ is the Jacobian transformation from y to $g_\lambda(y)$.

In model (1.10) the response and the model are transformed simultaneously with the same transformation with the objective of removing severe heteroscedasticity and/or nonnormality. There are several reasons why model (1.10) should be preferred over model (1.9) when the data exhibits skewness or heteroscedasticity. First, estimation of β based on (1.10) is more efficient than other methods including ordinary least squares. Without the transformation in (1.10) the variance of the ordinary least squares estimator of β will be inflated. Second, it may be necessary to estimate the entire conditional distribution of y given x. Therefore, confidence intervals for the mean and quantiles of y, and prediction intervals for y can be estimated. Third, even when $\hat{\beta}$ is unbiased, confidence intervals and prediction intervals can result in gross error if the residual variation is nonnormal or has a nonconstant variance.

Although transformations can remove heteroscedasticity and nonnormality there is no guarantee that a single transformation can do both. It may be necessary to simultaneously transform to remove skewness and downweight outliers to stabilize the variance.

1.5.1 Estimation and Inference

The loglikelihood of (1.10) can be written as

$$\ell(\beta, \lambda, \sigma) \propto -n \log(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \Big(g_\lambda(y_i) - g_\lambda(x_i^\mathsf{T} \beta) \Big)^2 + (\lambda - 1) \log(y_i) \tag{1.13}$$

The parameters in (1.13) can be estimated using an optimization procedure such as Newton-Raphson algorithm to maximize the loglikelihood function.

Alternatively, we can find the maximum likelihood estimate $\hat{\sigma}^2(\beta, \lambda)$ then maximize the profile loglikelihood function $\ell(\beta, \lambda, \hat{\sigma}^2(\beta, \lambda))$. That is, holding β and λ fixed, we initially maximize $\ell(\beta, \lambda, \sigma)$ over σ by

$$\hat{\sigma}^2(\beta, \lambda) = \frac{1}{n} \sum_{i=1}^{n} (\omega_i - \mu_i)^2 \tag{1.14}$$

where $\omega_i = g_\lambda(y_i)$ and $\mu_i = g_\lambda(x_i^\mathsf{T} \beta)$. The maximum likelihood estimate of β and λ is obtained by maximizing the profile loglikelihood function

$$\ell(\beta, \lambda, \hat{\sigma}(\beta, \lambda)) = \sum_{i=1}^{n} \log J(\lambda \mid y_i) - n \log(\hat{\sigma}(\beta, \lambda)) - \frac{n}{2} \tag{1.15}$$

$$= -n\left(\log(\hat{\sigma}(\beta, \lambda)) - \log\left\{ \prod_{i=1}^{n} J(\lambda \mid y_i)^{1/n} \right\} \right) - \frac{n}{2} \tag{1.16}$$

$$= -\frac{n}{2} \log\left(\frac{\hat{\sigma}(\beta, \lambda)}{\prod_{i=1}^{n} J(\lambda \mid y_i)^{1/n}} \right)^2 - \frac{n}{2}. \tag{1.17}$$

Hence the profile loglikelihood function (1.17) is maximized by minimizing

$$(\hat{\beta}, \hat{\lambda}) = \underset{\beta,\lambda}{\operatorname{argmin}} \sum_{i=1}^{n} \left(\frac{\omega_i - \mu_i}{\eta^{\lambda}} \right)^2 \tag{1.18}$$

where $\eta = (\prod_{i=1}^{n} y_i)^{1/n}$. The maximum likelihood estimate of σ^2 can be recovered using

$$\frac{n}{n - (p+1)} \hat{\sigma}^2(\hat{\beta}, \hat{\lambda}). \tag{1.19}$$

Since $p + 1$ parameters were estimated we correct for the degrees of freedom.

1.5.2 Variance Estimation

Several options are available to estimate the variance of the parameters for the transform-both-sides model. Although, we only discuss the Fisher information, variance can also be estimated via M-estimation and the bootstrap method. Each method gives a large-sample covariance matrix based on the asymptotic normality of the parameter estimates. Confidence intervals for each parameter can then be constructed.

Let $\theta = (\theta_1, \theta_2, \theta_3)^{\mathsf{T}} = (\beta, \lambda, \sigma)^{\mathsf{T}}$ be a vector containing the parameters and $\ell_i(\theta)$ the loglikelihood for the ith observation. Under certain regularity conditions, the maximum likelihood estimator $\hat{\theta}$ is a strongly consistent estimator of θ. The gradient of the loglikelihood function is

$$\nabla \ell_i(\theta) = \left[\frac{\partial}{\partial \theta_j} \ell_i(\theta) \right] \quad (j = 1, 2, 3), \tag{1.20}$$

and the Hessian matrix of the loglikelihood function is

$$\nabla^2 \ell_i(\theta) = \left[\frac{\partial^2}{\partial \theta_j \partial \theta_k} \ell_i(\theta) \right] \quad (j = 1, 2, 3; k = 1, 2, 3). \tag{1.21}$$

The estimator $\hat{\theta}$ is asymptotically normal with mean zero and covariance matrix $\mathrm{V}(\theta) = B(\theta)M(\theta)B(\theta)^{\mathsf{T}}$, where

$$B(\theta) = \mathrm{E}_\theta \left\{ \sum_{i=1}^{n} \nabla^2 \ell_i(\theta) \right\}^{-1}, \quad M(\theta) = \mathrm{E}_\theta \left\{ \sum_{i=1}^{n} (\nabla \ell_i(\theta))(\nabla \ell_i(\theta))^{\mathsf{T}} \right\} \tag{1.22}$$

Therefore

$$\mathrm{V}(\theta) = -\mathrm{E}_\theta \left\{ \sum_{i=1}^{n} \nabla^2 \ell_i(\theta) \right\}^{-1} = \mathrm{I}(\theta)^{-1}. \tag{1.23}$$

The observed Fisher information matrix is

$$\hat{\mathrm{I}}(\hat{\theta}) = -\sum_{i=1}^{n} \nabla^2 \ell_i(\hat{\theta}) \tag{1.24}$$

If the distribution of the errors is not normal then large-sample theory for maximum likelihood is not applicable. However, if the distribution of the errors is close to normality then we expect the fisher information matrix to be only slightly biased.

CHAPTER 2

MEDIAN REGRESSION FOR COMPLEX SAMPLE SURVEYS

2.1 Introduction

Complex sample surveys are increasingly used to produce population-based estimates required in planning health and social services. Complex survey data have also been harnessed by researchers to address important scientific questions, e.g., identifying risk factors for disease. In our motivating example, we use complex survey data to explore the factors that are associated with iodine intake in the US population. Identifying factors associated with iodine intake is scientifically important because iodine deficiency can lead to increased risks of many cancers, including thyroid, breast, endometrial, and ovarian cancer (Feldt-Rasmussen, 2001; Stadel, 1976). During the physical examinations of the 2007-2008 cycle of the National Health and Nutrition Examination Survey (NHANES), spot urine specimens were collected from participants and their urinary iodine (UI) concentration measured. In this motivating example, the response (UI) is extremely right skewed. Therefore ordinary linear regression models for the mean would not be appropriate. A more appealing approach when the response is skewed is to focus on the median regression function. However, in the literature there are very few examples (Geraci, 2013; Chen et al., 2010) of median regression for complex survey data. This is perhaps due to challenges in obtaining consistent variance estimators of the regression estimates for the median functional from complex survey data.

One popular approach for obtaining the estimated median regression parameters is to minimize the sum of absolute deviations (often called LAD or least-absolute-deviation estimator) via a linear programming algorithm (Bassett and Koenker, 1982) while incorporating the sampling weights of the complex survey. However, there still remains the issue of valid variance estimation. Most popular solution for estimating the variance of any estimating equation based estimator is to use the

sandwich estimator (Huber, 1967; White, 1980). However, because the least absolute deviations estimating equation is a discontinuous function of the regression parameters, the sandwich estimate of the variance will not be consistent in this case (Binder, 1983). For the same reason, Taylor series linearisation estimators and jackknife estimators of variance are not consistent for least absolute deviation method. Moreover, use of any resampling method is highly computationally intensive and impractical for large complex surveys. Wang and Opsomer (2011) proposed consistent variance estimators for non-differentiable survey estimators. There is a possibility that this method can be extended to marginal inference on regression parameters, however, it is beyond the scope of this paper. Other major limitations of resampling methods such as the bootstrap and balanced repeated replication (BRR) is that they tend to overestimate variance and variance estimators are usually not consistent (Shao, 1996; Shao et al., 2003; Lohr, 2009). In practice the primary sampling units are sampled without replacement to avoid selecting the same primary sampling unit more than once. However, it is common practice to treat the primary sampling units as if they were sampled with replacement in order to simplify variance estimation calculations. As a result of this approximation the variance may be overestimated. More importantly, it is generally unclear how to extend resampling methods to complex surveys with highly variable sampling weights (Presnell and Booth, 1994).

To estimate the median regression parameters, we propose a double-transform-both-sides (DTBS) regression model where the response and the regression function are transformed simultaneously to ensure an easily interpretable median functional. The DTBS approach applies the same Box-Cox type transformation twice to both the outcome and the linear predictor. After the double transformation, the outcome is assumed to be approximately normal. The median regression parameters are consistently estimated using a pseudo-likelihood based on the normal distribution, which incorporates the weights, but naively assumes observations within a cluster are independent. The usual sandwich estimator can be used to consistently estimate the variance of the parameter estimates, and thus this approach does not involve resampling methods to estimate the variance of the parameter estimates. Previous transform-both-sides approaches (Carroll and Ruppert, 1988; Fitzmaurice et al., 2007) use a single transformation on both sides; in simulations presented in Section 4, we have found that the DTBS is much more robust than a single transform-both-sides model. In par-

ticular, the approach is quite robust to the assumption about the true underlying distribution, and also gives estimators with bias similar to that of least absolute deviations regression for the simple random sample case, but has much smaller mean squared error than the least absolute deviations estimators.

The article is organized as follows. In Section 2, the DTBS regression model is presented along with the transformation function. We also show that the regression parameters of this DTBS approach can be interpreted as median regression parameters. In Section 3, for the proposed method, we derive expressions for the estimating equations and the sandwich variance estimator. In Section 4, we report the results of a simulation study and examine the robustness of the proposed method. Finally, in Section 5, we analyze data pertaining to iodine deficiency in the US population and illustrate some of the consequences of using ordinary least squares regression or least absolute deviations regression with complex survey data. We conclude with a discussion of an alternative approach along with future work on this topic.

2.2 Median Regression Model

For simplicity, we give notation for a weighted, cluster sampling design. Consider a continuous response y_{ij}, for $i = 1, 2, \ldots, n$ clusters and $j = 1, 2, \ldots, m_i$ individuals within the i^{th} cluster. The double transform-both-sides model is given by

$$g_{\lambda_2}(g_{\lambda_1}(y_{ij})) = g_{\lambda_2}(g_{\lambda_1}(x_{ij}^{\mathsf{T}}\beta)) + \epsilon_{ij}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m_i, \tag{2.1}$$

where x_{ij} is a column vector of covariates, β is a $p \times 1$ vector of unknown regression parameters, and $g_{\lambda_2}(\cdot)$ and $g_{\lambda_1}(\cdot)$ are Box-Cox type transformations (discussed later) with unknown transformation parameters λ_1 and λ_2. We assume the transformed outcome $g_{\lambda_2}(g_{\lambda_1}(Y_{ij}))$ is approximately normal, i.e., that ϵ_{ij} is approximately normal with mean 0 and variance σ^2. To obtain consistent estimates of β, we naively assume independence of subjects within a cluster (Binder, 1983; Liang and Zeger, 1986), and as such, do not specify the intra-class correlation of subjects within the same cluster.

Transform-both-sides regression is equivalent to median regression provided the resulting transformed response is symmetric (Fitzmaurice et al., 2007). Taylor (1985) showed that the Box-Cox

transformation is generally the most suitable method for transforming to symmetry. The Box-Cox transformation has been used in linear regression to transform the response variable only with the goal of achieving linearity and homoscedasticity. Alternatively, both the response and the regression function can be transformed (Carroll and Ruppert, 1984). The properties of this median estimator and its robustness to varying degrees of asymmetry in the response variable was studied by (Fitzmaurice et al., 2007). For moderately skewed data such as might arise from the Weibull and gamma distributions, the Box-Cox transformation gave little bias in estimating the regression parameters of the median, even though there is no exact transformation to normality for these distributions. For extremely skewed distributions, such as the Pareto distribution, Fitzmaurice et al. (2007) noted that when the Box-Cox transformation yields an asymmetrical distribution, applying a monotone transformation such as the logarithm function before implementing the Box-Cox transformation can substantially reduce bias. Wang and Ruppert (1995) suggested a non-parametric approach to estimating the transformation function. However, for large complex survey data, this non-parametric approach is difficult to implement.

Let $g_\lambda(y)$ be a family of transformations of the outcome y indexed by the transformation parameter λ, where we assume y is positive. To implement median regression via DTBS we need (1) a monotone transformation, (2) a transformation that can handle negative and positive y, and (3) the first and second derivatives must be a smooth function with respect to y. The first criterion is generally required so that a model for $g_\lambda(y)$ can generate a model for y by finding the inverse of the transformation, $g_\lambda^{-1}(y)$. Otherwise, $g_\lambda^{-1}(y)$ would not be unique. The second criterion becomes important when $x_{ij}^{\mathsf{T}}\beta^{(k)} < 0$, for the k-th iteration, as a result of using an iterative optimization procedure such as Newton-Raphson. Consequently, the regression function vector may temporarily yield negative predicted values of y. Another reason is that the first transformation may yield negative values. Finally, the third criterion allows us to estimate the variance using the sandwich estimator. The basic idea behind transform-both-sides (TBS) regression is to simultaneously transform the response and regression function with the same transformation in order to remove severe heteroscedasticity and/or nonnormality. The goal is to induce symmetric errors with constant variance as well as preserving the relationship between the response and regression function.

Carroll and Ruppert (1988) used the Box-Cox transformation in their transform-both-sides model (Carroll and Ruppert, 1984) and suggested using the Box-Cox transformation with a shift parameter to handle negative y's. Therefore a logical choice when implementing the DTBS model is to use the Box-Cox transformation with shift parameter. The standard practice in using the two parameter Box-Cox transformation is to add a small positive constant to the minimum value of y such that the shift parameter is positive. However this approach has a serious drawback as model parameter estimates are sensitive to the choice of the small arbitrary constant. Cheng and Iles (1987) offered a solution to simultaneously estimate both parameters when transforming the response only but the method cannot be extended to include transformation of the regression function.

Bickel and Doksum (1981) proposed a modification to the Box-Cox transformation to include negative y's but this too is problematic. Carroll and Ruppert (1988) pointed out that the Bickel-Doksum transformation changes from convex to concave as y changes from negative to positive. Therefore it would be difficult to predict its effect on skewed data unless y is either positive or negative. Yeo and Johnson (2000) gave an example that included positive and negative y where the transformation fails to adequately transform the data to normality. Further, it is well known that the Bickel-Doksum transformation is better suited for near symmetric distributions.

A more recent transformation that satisfies all three criteria and can accommodate negative y is the Yeo-Johnson transformation (Yeo and Johnson, 2000); however, it does not appear to work well in practice for the DTBS model. We have examined various combinations of transformations and found that a Box-Cox transformation followed by Yeo-Johnson transformation worked reasonably well. Moreover, we were able to obtain even better results with a modified Bickel-Doksum transformation. This modification allows us to satisfy the condition of a smooth score function. What follows is the development of the modified Bickel-Doksum transformation.

Bickel and Doksum (1981) extended the definition of the power family of transformations to include all real numbers y,

$$g_\lambda(y) = (\text{sgn}(y)|y|^\lambda - 1)/\lambda \quad (\lambda > 0, y \in \mathbb{R}),$$

where $\mathbb{R}$ is the set of real numbers and λ is an unknown transformation parameter to be estimated. The signum function is defined as $\text{sgn}(y) = 1$ if $y > 0$, $\text{sgn}(y) = -1$ if $y < 0$ and zero otherwise.

The transformation $g_\lambda(y)$ is monotone with nonnegative derivative, $g'_\lambda(y) = \mathrm{d}g_\lambda/\mathrm{d}y = |y|^{\lambda-1} \geqslant 0$ for all y. Note that any real number y can be rewritten as the product of the sign function and absolute value function $\mathrm{sgn}(y)|y|$. Therefore, an alternate expression of the Bickel-Doksum transformation is

$$g_\lambda(y) = (y|y|^{\lambda-1} - 1)/\lambda \ .$$

Finally, using the following $|y| \approx (y^2 + \tau)^{1/2}$ to approximate the absolute value function we have the modified Bickel-Doksum transformation

$$g_\lambda(y) = (y(y^2 + \tau)^{(\lambda-1)/2} - 1)/\lambda,$$

where τ is a small positive arbitrary constant. Next, we elucidate why the transform-both-sides model is equivalent to median regression.

We begin with the definition of the median of a random variable. If Y is a continuous random variable then the median of Y is a fixed constant $M \in \mathbb{R}$ such that $P(Y > M) = 1/2$. Since $g_\lambda(\cdot)$ is monotone it follows that

$$P(Y > M) = P(g_\lambda(Y) > g_\lambda(M)) = 1/2.$$

Therefore, as M is the median of Y likewise $g_\lambda(M)$ is the median of $g_\lambda(Y)$. Further, recall that for any symmetric random variable its mean and median are equal. Therefore, if the transformation $g_\lambda(Y)$ yields a symmetric distribution, then $E(g_\lambda(Y)) = g_\lambda(M)$; as a consequence, modeling the mean leads to modeling the median.

Consider the regression setting with a single Box-Cox transformation,

$$g_\lambda(y_{ij}) = g_\lambda(x_{ij}^\mathsf{T}\beta) + \epsilon_{ij}, \quad \text{for } i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, m_i, \tag{2.2}$$

where the error distribution $\epsilon_{ij} \overset{iid}{\sim} f_\epsilon(\cdot)$ is a symmetric density centered at zero with constant variance, y_{ij} is the response variable, x_{ij} is a column vector of covariates, λ is an unknown transformation parameter and β is a $p \times 1$ vector of unknown regression parameters. It follows then that the conditional median of y_{ij} is $x_{ij}^\mathsf{T}\beta$ since

$$
\begin{aligned}
P(y_{ij} > x_{ij}^T\beta) &= P(g_\lambda(y_{ij}) > g_\lambda(x_{ij}^T\beta)) \\
&= P(g_\lambda(y_{ij}) - g_\lambda(x_{ij}^T\beta) > 0) \\
&= P_\epsilon(\epsilon_{ij} > 0) \\
&= 1/2.
\end{aligned}
$$

Consequently, the regression model (2.2) implies that the response y_{ij} comes from a probability distribution whose median is $x_{ij}^T\beta$. Hence we are able to model the median via the monotone transformation $g_\lambda(\cdot)$ applied to both sides of (2.2). Even though we have only considered a single transformation on both sides, the above discussion can be extended to include a double transformation on both sides. The primary motivation for a double transformation is to enhance the symmetry of the errors in (2.2). In the next section, expressions for the estimating equations and sandwich variance estimator are derived.

2.3 Estimating Equations and Variance Estimation

We obtain expressions for the weighted estimating equations and the sandwich variance estimator based on the pseudo-log-likelihood. Naively assuming independence of observations within a cluster, and that $g_{\lambda_2}(g_{\lambda_1}(y_{ij}))$ is normal, the log of the pdf of $g_{\lambda_2}(g_{\lambda_1}(y_{ij}))$ is given by

$$
\log f(y_{ij}|x_{ij},\beta,\sigma^2,\lambda) \propto -\frac{1}{2}\sum_{i=1}^{n} m_i \ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(\omega_{ij}-\mu_{ij})^2 + \ln\left(\prod_{i=1}^{n}\prod_{j=1}^{m_i} J(y_{ij},\lambda)\right), \quad (2.3)
$$

where $f(y_{ij}|x_{ij},\beta,\sigma^2,\lambda)$ is the conditional density of y_{ij} given x_{ij}, $\omega_{ij} = g_{\lambda_2}(g_{\lambda_1}(y_{ij}))$, $\mu_{ij} = g_{\lambda_2}(g_{\lambda_1}(x_{ij}^T\beta))$, $\lambda = (\lambda_1,\lambda_2)$ and $J(y_{ij},\lambda)$ is the Jacobian of the transformation of y_{ij} to $g_{\lambda_2}(g_{\lambda_1}(y_{ij}))$. That is, $f(y_{ij}|x_{ij},\beta,\sigma^2,\lambda) = (2\pi\sigma^2)^{-1/2}\exp\{-(\omega_{ij}-\mu_{ij})^2/(2\sigma^2)\}J(y_{ij},\lambda)$. With rescaled sampling weights, δ_{ij}, the pseudo-log-likelihood is given by

$$
\ell_\pi(\beta,\sigma^2,\lambda) = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\delta_{ij}\log f(y_{ij}|x_{ij},\beta,\sigma^2,\lambda), \quad (2.4)
$$

where the rescaled weights sum to one. The model parameters β, σ^2 and λ can be estimated by maximizing the pseudo-log-likelihood of (2.4) using an iterative optimization technique. For each subject j in cluster i, let $\mu_{ij} = g_{\lambda_2}(g_{\lambda_1}(x_{ij}^T\beta))$ and $\eta_{ij} = g_{\lambda_1}(x_{ij}^T\beta)$ such that $\mu_{ij} = g_{\lambda_2}(\eta_{ij})$ and $\tau_{ij} = x_{ij}^T\beta$. Obtaining the MLE $\hat{\beta}$ of the pseudo-log-likelihood function (2.4) is the same as solving the weighted estimating equation

$$S(\beta) = \frac{\partial}{\partial\beta}\ell_\pi(\beta, \sigma^2, \lambda) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} S_{ij}(\beta) = -\frac{1}{\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\delta_{ij}\frac{\partial\mu_{ij}}{\partial\beta}(\omega_{ij} - \mu_{ij}) = \mathbf{0}, \tag{2.5}$$

where

$$\frac{\partial\mu_{ij}}{\partial\beta} = \frac{\partial\mu_{ij}}{\partial\eta_{ij}}\frac{\partial\eta_{ij}}{\partial\tau_{ij}}\frac{\partial\tau_{ij}}{\partial\beta} = g'_{\lambda_2}(\eta_{ij})g'_{\lambda_1}(\tau_{ij})\boldsymbol{x}_{ij},$$

and x_{ij} is a $p \times 1$ vector. The sandwich estimate of variance of the estimator $\hat{\beta}$ is constructed using $\boldsymbol{V_\beta} = \boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B}^{-T}$ where

$$\boldsymbol{B} = \left[-\frac{\partial S(\beta)}{\partial\beta}\right], \quad \boldsymbol{M} = \sum_{i=1}^{n}\left[\sum_{j=1}^{m_i} S_{ij}(\beta)\right]\left[\sum_{j=1}^{m_i} S_{ij}(\beta^T)\right].$$

Note that $\boldsymbol{M}$ is the covariance matrix of the estimating equation and $\boldsymbol{B}$ is the Hessian matrix.

We now derive expressions for $\boldsymbol{M}$ and $\boldsymbol{B}$ under the naive likelihood model. Using the expression derived for $S_{ij}(\beta)$ in (2.5) the matrix $\boldsymbol{M}$ is easily obtained and is given by

$$\boldsymbol{M} = \frac{1}{\sigma^4}\sum_{i=1}^{n}\boldsymbol{X}_i^T\mathsf{Diag}(\boldsymbol{v}_i^2)\boldsymbol{X}_i,$$

with

$$\boldsymbol{v}_i = \boldsymbol{\delta}_i \circ g'_{\lambda_2}(\boldsymbol{\eta}_i) \circ g'_{\lambda_1}(\boldsymbol{\tau}_i) \circ (\boldsymbol{\omega}_i - \boldsymbol{\mu}_i),$$

where operator $\circ$ denotes the Hadamard product and $\boldsymbol{\delta}_i, \boldsymbol{\eta}_i, \boldsymbol{\tau}_i, \boldsymbol{\omega}_i, \boldsymbol{\mu}_i$ are vectors corresponding to scalars $\delta_{ij}, \eta_{ij}, \tau_{ij}, \omega_{ij}, \mu_{ij}$. Next we obtain $\boldsymbol{B}$ by taking the second partial derivative of the pseudo-log-likelihood function (2.3) with respect to β

$$\begin{aligned} S'(\beta) &= \frac{\partial}{\partial\beta\partial\beta^T}\ell_\pi(\beta, \sigma^2, \lambda) \\ &= -\frac{1}{\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\delta_{ij}\left[(\omega_{ij} - \mu_{ij})\frac{\partial^2\mu_{ij}}{\partial\beta\partial\beta^T} - \frac{\partial\mu_{ij}}{\partial\beta}\left(\frac{\partial\mu_{ij}}{\partial\beta}\right)^T\right]. \end{aligned} \tag{2.6}$$

Hence

$$\boldsymbol{B} = -\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X},$$

where

$$\boldsymbol{\alpha} = \boldsymbol{\delta} \circ \{[g'_{\lambda_2}(\boldsymbol{\eta}) \circ g''_{\lambda_1}(\boldsymbol{\tau}) + g''_{\lambda_2}(\boldsymbol{\eta}) \circ g'_{\lambda_1}(\boldsymbol{\tau})^2] \circ (\boldsymbol{\omega} - \boldsymbol{\mu})\} - \boldsymbol{\delta} \circ (g'_{\lambda_2}(\boldsymbol{\eta}) \circ g'_{\lambda_1}(\boldsymbol{\tau}))^2,$$

$\boldsymbol{D} = \mathsf{Diag}(\boldsymbol{\alpha})$ and $\boldsymbol{\delta}, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\omega}, \boldsymbol{\mu}$ are vectors corresponding to vectors $\boldsymbol{\delta}_i, \boldsymbol{\eta}_i, \boldsymbol{\tau}_i, \boldsymbol{\omega}_i, \boldsymbol{\mu}_i$. Therefore the sandwich estimate of variance is given by

$$\hat{V}_{\hat{\beta}} = \boldsymbol{X}^T\hat{D}\boldsymbol{X}\left(\sum_{i=1}^{n}\boldsymbol{X}_i^T\mathsf{Diag}(\hat{\boldsymbol{v}}_i^2)\boldsymbol{X}_i\right)\boldsymbol{X}^T\hat{D}\boldsymbol{X}/\hat{\sigma}^6. \tag{2.7}$$

Next, we evaluate the performance of the DTBS estimator in finite samples.

2.4 Simulation Study

To investigate the performance of our proposed estimator and its robustness to asymmetry in the response variable, we simulated 1000 samples of size 600 and 6000 under a cluster sampling with equal probabilities design. We also considered several covariates including combinations of different correlations, sample sizes, number of clusters and cluster sizes. For each cluster (i.e. 30 or 60 clusters), we simulated 10, 20, 100, 200 multivariate normal observations with exchangeable correlation 0.01, 0.05 and 0.10. The marginal normal variables were then transformed to the log-normal, exponential, Weibull, gamma and Pareto distributions with median, $m_{ij} = 6.5 + 2x_{ij1} + x_{ij2} + 2x_{ij3}$ where $x_{ij1} \sim U[1, 10]$, $x_{ij2} \sim N(0, 1)$ and $x_{ij1} = 1$ with probability 0.5 and $x_{ij1} = -1$ otherwise. The transformation is given by $y_{ij} = F^{-1}(U_{ij})$ and $U_{ij} = \Phi(Z_{ij})$ such that $Z_{ij} = (\alpha_i + \epsilon_{ij})/(1 + \sigma_\alpha^2)^{1/2}$, $\epsilon_{ij} \sim N(0, 1)$, $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\sigma_\alpha^2 = \rho/(1 - \rho)$ where $\rho = sin(\pi\tau/2)$ is the intra-class correlation and τ is Kendall's tau coefficient. We use Kendall's τ because it is invariant to monotone transformations. Thus the within cluster correlation for the latent normal random variables Z_{ij} and $y_{ij} = F^{-1}(U_{ij})$ will be the same.

Additionally, the five different specifications were simulated in the following manner. The log-normal distribution is given by $\log(y_{ij}) \sim N(\mu_{ij}, 1)$ where $\mu_{ij} = \log(m_{ij})$. The exponential density is $f(y_{ij}|\psi_{ij})$ where $\psi_{ij} = \log(2)/m_{ij}$. The Weibull density is $f(y_{ij}|\alpha, \psi_{ij})$ with shape

parameter $\alpha = 0.9$ and scale parameter $\psi_{ij} = m_{ij}(\log 2)^{-1/\alpha}$. The gamma density is given by $f(y_{ij}|k, \theta_{ij})$ with mean $k\theta_{ij}$ having shape parameter $k = 0.25$ and scale θ_{ij}. To find θ_{ij} we solve the equation $F(m_{ij}|k, \theta_{ij}) - 0.5 = 0$ where F is the cumulative distribution function of the gamma density. Finally, the Pareto distribution is given by $f(y_{ij}|\alpha_{ij}, k)$ with scale parameter $k = 1$ and shape parameter $\alpha_{ij} = \log 2/\log(m_{ij})$. For all simulation configurations, we estimated the median regression parameters for our proposed DTBS model, the single TBS model and also the standard median regression as a comparison. By standard median regression we mean least absolute deviations regression.

The simulation results in Tables 2.1–2.5 indicate that the proposed DTBS method yields estimates that are unbiased and are discernibly more efficient (i.e. smaller mean squared error), when compared to the standard median regression. This is true regardless of the correlations, sample sizes, number of clusters and cluster sizes. Even when we considered the extremely skewed, heavy-tailed gamma and Pareto distributions; bias was small for both the DTBS and MR models and were at most -6.808%, 7.691% and 5.677%, 7.029 %, respectively (Tables 2.1–2.2). In contrast, the TBS model yields biased estimates that were large compared with the DTBS and MR models for the gamma and Pareto distributions, indicating that a single transformation may not be sufficient for extremely skewed distributions. For the Weibull, exponential and log-normal distributions (Table 2.3–2.5) in which the TBS appears unbiased, the mean squared error of the DTBS and TBS are similar, suggesting that the additional parameter estimated in DTBS does not increase the mean squared error. In addition, with the exception of the Pareto distribution the DTBS method shows good coverage probabilities for 95% confidence intervals; coverage probabilities for the standard median regression should be interpreted with caution as the variance of the nondifferentiable LAD estimator is estimated using the bootstrap (He and Hu, 2002). Coverage probability for β_2 in the DTBS models of Table 2.1 is poor when $n = 600$ but is good with increased sample size of $n = 6000$.

2.5 Application: Predictors of Urinary Iodine Concentration in NHANES

The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. NHANES uses a stratified, multistage survey to provide a representative sample of the non-institutionalized US population. It consists of an initial in-person interview at the household, followed by a physical examination in a mobile examination center and follow up questionnaires. During the NHANES physical examinations, spot urine specimens were collected from participants, and aliquots of these specimens were generated and stored cold or frozen until shipped. Our analysis is restricted to the 2007-2008 cycle of NHANES laboratory data involving urinary iodine (UI) concentration. Severe iodine deficiency of UI can lead to increased risks of many cancers, including thyroid, breast, endometrial, and ovarian cancer (Feldt-Rasmussen, 2001; Stadel, 1976) The objective of the analysis is to identify potentially important characteristics of individuals that are associated with urinary iodine concentration; in particular, it is of interest to determine whether females are at a higher risk of iodine deficiency than males.

Our complex survey consists of data on 6802 persons. There are a total of 32 primary sampling units and 16 strata, with 2 primary sampling units per stratum. The average cluster size is 213 with the smallest being 51 and the largest 314. The response variable of interest, urinary iodine concentration measured in $\mu g\, L^{-1}$, is extremely right-skewed with median of 165.7, mean of 413.8 and standard deviation of 9460. The minimum and maximum iodine concentrations are 2.1 and 762,010. The individual characteristics of interest were gender, body mass index (BMI), age at screening, race, total grain intake, dairy consumption, dietary supplements, fish and salt intake. We used a dummy coding scheme for all categorical variables. The continuous variables age, BMI and total grain intake were centered and scaled accordingly: Age - 30, (BMI - 25)/5 and (Total grain - 310)/10.

In Table 2.6 we compare the results of four models: TBS, DTBS, standard median regression, and ordinary least squares (OLS) regression after taking the natural logarithm of the response. All approaches take into account the weights for estimation, and all except for standard least absolute

deviations median regression use the sandwich variance estimator taking account of the stratification, clustering, and weighting. Variances for the standard median regression model estimates were produced using balanced repeated replication (BRR); a description of BRR can be found in Lohr (2009), section 9.3.1. The degrees of freedom for the t tests in Table 2.6 is 16 for all models.

Note that the estimated coefficients for TBS and DTBS in Table 2.6 are discernibly different, suggesting that a single transformation of urinary iodine concentration is not adequate. Moreover, the estimated coefficients in the DTBS and standard median regression models are very similar, indicating that the double transformation is adequate. Overall, the DTBS, standard median regression and ordinary least squares models yield similar results in terms of the covariates associated with iodine concentration, with the exception of the covariates age, fish intake and supplements. Results from the DTBS model showed age, fish intake and supplements to be significantly associated with iodine concentration but the standard median regression model did not reveal these associations to be statistically significant. Similarly, while the results from the ordinary least squares model showed age to be associated with iodine it failed to show any statistically discernible association with fish intake and supplements.

There are at least two reasons for the different pattern of results concerning these three covariates. First, note that the coefficients of the DTBS and standard median regression model are quite similar as we would expect but, in general, their standard errors are somewhat different. With few exceptions, the standard errors for DTBS are similar or substantially smaller than those obtained for the standard median regression (using the BRR method). In light of the efficiency gains seen for DTBS in the simulation results reported in Table **??**, this is most likely an indication of the increased efficiency of the DTBS estimator over standard median regression. The standard errors for age, fish intake and supplements in the standard median regression model are discernibly larger than those of the DTBS model (i.e. $0.183/0.104 = 1.8$, $6.612/5.345 = 1.2$ and $7.963/5.024 = 1.6$); in the case of age, almost twice as large. This explains why age, fish intake and supplements are significantly associated with iodine in the DTBS model but show no association with iodine in the standard median regression model.

Second, the residual plot for the ordinary least squares model shows that even after log transforming the response potential outliers remain (Figure 2.1(b)). In addition, the QQ plot for or-

dinary least squares regression strongly indicates a violation of the assumption of normal errors (Figure 2.1(d)). Therefore, results of the ordinary least squares model should be interpreted with caution. In contrast, the assumptions of normal errors and constant variance seem quite reasonable for the DTBS model (Figure 2.1(a),(c)).

In summary, results from the DTBS model indicate that gender, age, race, BMI, supplements, and fish and dairy intake are significantly associated with urinary iodine concentration. We note that the first three of these factors are non-modifiable, while the remainder are modifiable. When taken together, this set of predictors may be useful for identifying individuals who are at higher risk for iodine deficiency, and hence may potentially have increased risks of many cancers (e.g., thyroid, breast, endometrial, and ovarian cancer), and who would benefit from interventions to modify lifestyle risk behaviors.

2.6 Discussion

The aim of this paper was to present a viable alternative to standard median regression for complex sample survey data where standard theoretically errors can be obtained in a computationally convenient way. One key difference between our transform-both-sides method and standard median regression based on LAD is that our method assumes that the error follows a parametric density (at least approximately) whereas the later makes no distributional assumption about the error term in the model. However, unlike other competing method, our method produces estimates of all quantile functions using only one estimating equation (as explained later in this section). We also note that even if the moments of the original response distribution are not defined as in the case of Cauchy density, our method is still valid as long as the resulting estimating equation based on double-transformation is unbiased. Based on our simulation results, the proposed DTBS estimator is relatively robust to large outliers, and it was found to have comparable bias to that of standard least absolute deviations median regression even when our underlying modeling assumptions are not valid. Further, we demonstrated that our method is robust to varying asymmetric densities of the response variable including the densities that can not be reduced to symmetry even after double-transformation. The DTBS approach also appears to have much smaller mean squared error compared to standard least absolute deviations median regression.

Throughout the paper we assumed that the error terms are normally distributed. Other distributions such as a normal/independent distribution can be used (Lange and Sinsheimer, 1993). For example, the t_ν distribution can be expressed as a scale mixture of normals by letting $\epsilon_i \sim N(0, u_i^{-1}\sigma^2)$ with $u_i \sim Ga(\nu/2, \nu/2)$. The ML estimate of β under the t_ν model has estimating equation $\sum_i \sum_j \eta_{ij} \frac{\partial \mu_{ij}}{\partial \beta}(\omega_{ij} - \mu_{ij})/\sigma^2 = 0$, where $\eta_{ij} = (\nu+1)/(\nu+w_{ij})$ is a weight corresponding to each observation and $w_{ij} = (\omega_{ij} - \mu_{ij})^2/\sigma^2$. The advantage of assuming distributions such as the t_ν distribution is that extreme observations are downweighted, with the end result being transformation and weighting applied simultaneously. However, this would require use of the EM algorithm to estimate model parameters and convergence may be relatively slow. Another incentive though for this approach is that transformations, such as the Yeo-Johnson transformation, a more flexible transformation allowing for negative and/or positive responses, that performed poorly under the normal model may now perform somewhat better.

The proposed method is applicable to any multi-stage complex sampling design. One possible reason for very limited use of median regression tools in current sample survey literature is that one particular quantile functional is not considered a comprehensive summary of a finite population. For example, total sum of response can not be estimated from a median response of a finite population. Existing quantile regression tools only focus on estimating a pre-determined quantile. Even though we have focused on the median, it is, however, possible to use our method to estimate all other quantiles. Unlike existing median regression tools, our method presents the τ-th quantile of y given x as

$$Q_\tau(y \mid x) = g_\lambda^{-1}\left\{ g_\lambda(x^\mathsf{T}\beta) + F_\epsilon^{-1}(\tau) \right\} \tag{2.8}$$

where estimates of λ and β are the same as we have obtained before. This quantile, $Q_\tau(y \mid x)$, for any $0 < \tau < 1$ can be estimated as

$$\hat{Q}_\tau(y \mid x) = g_{\hat{\lambda}}^{-1}\left\{ g_{\hat{\lambda}}(x^\mathsf{T}\hat{\beta}) + \hat{\sigma}_\epsilon \Phi^{-1}(\tau) \right\} \tag{2.9}$$

where $\hat{\sigma}_\epsilon \Phi^{-1}(\tau)$ is a parametric estimate of the τ-th quantile of the error distribution. Alternatively, $Q_\tau(y \mid x)$ can be estimated by replacing $\hat{\sigma}_\epsilon \Phi^{-1}(\tau)$ with the empirical distribution function of the residuals such that

$$\hat{Q}_\tau(y \mid x) = g_{\hat{\lambda}}^{-1}\left\{ g_{\hat{\lambda}}(x^\mathsf{T}\hat{\beta}) + \hat{F}_\epsilon^{-1}(\tau) \right\}. \tag{2.10}$$

Hence, using a single data analysis, our method produces a comprehensive description of the whole population. Finally, the method can also be extended to median regression of longitudinal data from complex sample surveys.

Table 2.1: Simulation study of 1000 replicates (of size 600 and 6000) for the Pareto distribution the standard least absolute deviations median regression (MR), double-transform-both-sides (DTBS), and transform-both-sides (TBS) models.

					MR†			DTBS			TBS		
Kendall's τ	Sample Size	No. of Clusters	Cluster Size		β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
0.01	600	30	20	Relative Bias (%)	-0.980	2.003	-2.962	-0.259	-4.498	3.285	10.327	19.722	15.329
				Mean Squared Error	1.232	5.361	7.192	0.788	2.886	3.946	1.024	3.835	4.456
				Coverage Probability	0.930	0.926	0.937	0.930	0.859	0.924	0.941	0.911	0.960
		60	10	Relative Bias (%)	-3.835	3.262	1.103	1.000	-5.603	5.850	12.677	21.181	19.653
				Mean Squared Error	1.217	4.655	6.577	0.785	2.765	4.063	3.787	6.006	7.542
				Coverage Probability	0.909	0.935	0.948	0.922	0.857	0.937	0.942	0.914	0.967
0.05	600	30	20	Relative Bias (%)	2.061	1.201	1.407	-0.027	-3.536	2.909	12.795	21.176	17.640
				Mean Squared Error	1.525	5.814	7.611	0.910	2.789	4.075	1.195	3.885	4.968
				Coverage Probability	0.910	0.927	0.932	0.926	0.890	0.931	0.945	0.917	0.952
		60	10	Relative Bias (%)	-4.122	1.823	2.038	-2.293	-6.808	3.225	10.691	16.757	17.658
				Mean Squared Error	1.319	5.007	6.516	0.722	2.564	3.615	0.993	3.400	4.644
				Coverage Probability	0.901	0.933	0.939	0.920	0.868	0.928	0.943	0.914	0.963
0.10	600	30	20	Relative Bias (%)	4.719	6.438	3.224	-0.307	-2.512	2.462	16.197	23.733	21.287
				Mean Squared Error	1.902	5.927	8.419	1.405	3.074	4.930	1.592	4.224	5.813
				Coverage Probability	0.883	0.929	0.932	0.881	0.861	0.889	0.950	0.937	0.956
		60	10	Relative Bias (%)	-3.309	2.034	3.547	-0.443	-5.461	3.342	11.785	17.895	18.606
				Mean Squared Error	1.466	5.162	7.301	0.886	2.829	3.955	1.161	3.569	4.885
				Coverage Probability	0.888	0.937	0.913	0.892	0.856	0.912	0.950	0.927	0.963
0.01	6000	30	200	Relative Bias (%)	0.134	-4.027	2.055	1.904	1.549	1.573	20.408	28.592	23.546
				Mean Squared Error	0.151	0.600	0.705	0.098	0.017	0.015	9.956	5.990	7.979
				Coverage Probability	0.902	0.914	0.927	0.953	0.995	0.999	0.909	0.922	0.944
		60	100	Relative Bias (%)	-0.266	-3.462	1.038	-0.846	0.511	1.746	18.834	17.300	37.784
				Mean Squared Error	0.125	0.529	0.652	0.053	0.035	0.039	6.604	0.361	59.850
				Coverage Probability	0.916	0.940	0.933	0.979	0.987	0.993	0.910	0.936	0.938
0.05	6000	30	200	Relative Bias (%)	2.696	-2.847	5.142	3.776	0.758	1.147	18.492	23.814	28.834
				Mean Squared Error	0.326	0.684	0.860	0.288	0.020	0.027	2.187	1.304	8.744
				Coverage Probability	0.762	0.905	0.909	0.916	0.995	0.995	0.956	0.943	0.962
		60	100	Relative Bias (%)	0.589	-3.959	2.639	0.246	0.700	1.134	16.784	25.562	21.881
				Mean Squared Error	0.211	0.551	0.759	0.153	0.039	0.054	2.710	5.997	5.110
				Coverage Probability	0.855	0.923	0.918	0.921	0.986	0.988	0.954	0.948	0.957
0.10	6000	30	200	Relative Bias (%)	5.449	1.336	7.691	5.470	-0.907	0.076	24.050	46.312	58.945
				Mean Squared Error	0.566	0.791	1.121	0.507	0.032	0.045	4.574	22.295	254.527
				Coverage Probability	0.654	0.894	0.876	0.901	0.987	0.991	0.957	0.948	0.956
		60	100	Relative Bias (%)	1.540	-3.197	2.730	1.028	-0.541	0.212	16.980	26.435	23.152
				Mean Squared Error	0.314	0.597	0.892	0.271	0.048	0.065	1.873	3.826	3.310
				Coverage Probability	0.756	0.917	0.888	0.905	0.980	0.989	0.957	0.958	0.954

†Note: Variance for MR is estimated using the bootstrap (He and Hu, 2002). Hence coverage probabilities must be interpreted with caution.

Table 2.2: Simulation study of 1000 replicates (of size 600 and 6000) for the gamma distribution comparing the standard least absolute deviations median regression (MR), double-transform-both-sides (DTBS), and transform-both-sides (TBS) models.

					MR†			DTBS			TBS		
Kendall's τ	Sample Size	No. of Clusters	Cluster Size		β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
0.01	600	30	20	Relative Bias (%)	-5.278	-1.598	0.284	-1.327	-5.322	2.094	-21.532	-25.903	-18.875
				Mean Squared Error	1.178	5.425	7.024	0.573	2.695	3.348	0.582	1.906	2.448
				Coverage Probability	0.921	0.932	0.918	0.941	0.915	0.938	0.875	0.894	0.912
		60	10	Relative Bias (%)	0.144	-4.355	-1.281	1.550	-5.221	3.386	-20.136	-24.793	-18.173
				Mean Squared Error	1.081	5.036	6.927	0.542	2.571	3.433	0.538	1.783	2.452
				Coverage Probability	0.943	0.947	0.932	0.954	0.901	0.943	0.849	0.870	0.903
0.05	600	30	20	Relative Bias (%)	-3.557	-1.137	1.215	-1.850	-0.863	1.940	-22.692	-21.840	-19.075
				Mean Squared Error	1.143	5.537	7.360	0.620	2.667	3.722	0.650	1.853	2.703
				Coverage Probability	0.925	0.926	0.922	0.931	0.902	0.941	0.841	0.903	0.910
		60	10	Relative Bias (%)	-0.601	-5.241	7.029	2.239	-3.111	2.101	-18.773	-25.341	-19.350
				Mean Squared Error	1.224	5.608	7.636	0.634	2.636	3.436	0.579	1.830	2.503
				Coverage Probability	0.927	0.930	0.929	0.942	0.915	0.944	0.862	0.883	0.907
0.10	600	30	20	Relative Bias (%)	1.056	-6.237	2.570	4.332	-5.239	5.677	-17.150	-25.511	-15.192
				Mean Squared Error	1.359	6.117	8.335	0.777	2.719	3.902	0.666	1.924	2.810
				Coverage Probability	0.914	0.934	0.924	0.938	0.903	0.936	0.862	0.893	0.910
		60	10	Relative Bias (%)	-3.480	-0.811	-0.953	-1.641	-2.914	0.066	-21.832	-24.717	-20.614
				Mean Squared Error	1.199	5.873	8.271	0.619	2.484	3.739	0.624	1.714	2.751
				Coverage Probability	0.923	0.923	0.922	0.947	0.921	0.944	0.828	0.893	0.881
0.01	6000	30	200	Relative Bias (%)	-0.044	1.299	1.277	2.198	2.376	2.893	-18.745	-18.486	-18.105
				Mean Squared Error	0.136	0.672	0.756	0.074	0.311	0.341	0.193	0.247	0.366
				Coverage Probability	0.936	0.946	0.941	0.945	0.923	0.939	0.590	0.902	0.864
		60	100	Relative Bias (%)	0.358	-2.699	-1.582	3.105	0.247	1.526	-17.953	-20.189	-19.377
				Mean Squared Error	0.131	0.642	0.748	0.072	0.294	0.339	0.178	0.240	0.387
				Coverage Probability	0.944	0.944	0.935	0.929	0.936	0.947	0.593	0.920	0.859
0.05	6000	30	200	Relative Bias (%)	0.609	-1.347	-0.079	2.959	2.240	2.203	-17.794	-18.459	-18.400
				Mean Squared Error	0.227	0.709	0.788	0.153	0.316	0.407	0.245	0.256	0.434
				Coverage Probability	0.857	0.938	0.928	0.941	0.929	0.955	0.728	0.897	0.855
		60	100	Relative Bias (%)	1.217	-0.176	1.434	3.501	3.085	3.009	-17.609	-17.950	-17.886
				Mean Squared Error	0.179	0.665	0.799	0.107	0.309	0.377	0.202	0.246	0.391
				Coverage Probability	0.904	0.944	0.938	0.954	0.934	0.956	0.715	0.918	0.866
0.10	6000	30	200	Relative Bias (%)	1.291	-2.229	1.954	3.463	0.943	3.297	-17.280	-19.241	-17.372
				Mean Squared Error	0.360	0.769	0.948	0.252	0.339	0.512	0.320	0.279	0.499
				Coverage Probability	0.738	0.920	0.901	0.950	0.934	0.944	0.781	0.893	0.836
		60	100	Relative Bias (%)	0.867	0.616	3.784	3.180	2.164	4.630	-17.655	-18.464	-16.506
				Mean Squared Error	0.221	0.702	0.934	0.147	0.301	0.466	0.238	0.247	0.436
				Coverage Probability	0.852	0.924	0.901	0.950	0.953	0.947	0.759	0.932	0.853

†Note: Variance for MR is estimated using the bootstrap (He and Hu, 2002). Hence coverage probabilities must be interpreted with caution.

Table 2.3: Simulation study of 1000 replicates (of size 600 and 6000) for the Weibull distribution comparing the standard least absolute deviations median regression (MR), double-transform-both-sides (DTBS), and transform-both-sides (TBS) models.

					MR†			DTBS			TBS		
Kendall's τ	Sample Size	No. of Clusters	Cluster Size		β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
0.01	600	30	20	Relative Bias (%)	-0.248	2.165	-2.942	-4.449	-1.956	-4.663	-4.597	-1.738	-4.565
				Mean Squared Error	0.180	1.055	1.080	0.103	0.514	0.542	0.103	0.517	0.543
				Coverage Probability	0.942	0.934	0.942	0.925	0.932	0.943	0.925	0.930	0.942
		60	10	Relative Bias (%)	-1.492	2.649	-0.843	-4.703	-1.547	-3.308	-4.648	-1.858	-3.306
				Mean Squared Error	0.183	0.927	1.007	0.103	0.497	0.556	0.101	0.493	0.557
				Coverage Probability	0.944	0.942	0.956	0.934	0.937	0.951	0.932	0.937	0.952
0.05	600	30	20	Relative Bias (%)	0.071	1.712	-1.407	-4.068	-1.652	-4.497	-4.190	-1.834	-4.401
				Mean Squared Error	0.195	1.094	1.071	0.113	0.513	0.553	0.115	0.512	0.561
				Coverage Probability	0.936	0.939	0.938	0.921	0.938	0.943	0.921	0.937	0.945
		60	10	Relative Bias (%)	1.837	3.953	-0.396	-4.689	-1.845	-3.311	-4.749	-1.800	-3.467
				Mean Squared Error	0.193	0.960	1.041	0.109	0.489	0.564	0.109	0.489	0.562
				Coverage Probability	0.931	0.947	0.943	0.931	0.940	0.950	0.925	0.939	0.947
0.10	600	30	20	Relative Bias (%)	0.279	3.599	-2.146	-3.726	-1.325	-4.240	-3.827	-1.863	-4.130
				Mean Squared Error	0.210	1.051	1.087	0.131	0.513	0.573	0.130	0.526	0.589
				Coverage Probability	0.927	0.945	0.939	0.920	0.939	0.948	0.921	0.940	0.951
		60	10	Relative Bias (%)	-1.870	3.026	-0.542	-4.541	-1.963	-3.348	-4.600	-2.015	-3.531
				Mean Squared Error	0.201	0.938	1.085	0.117	0.485	0.572	0.115	0.488	0.576
				Coverage Probability	0.937	0.945	0.928	0.927	0.937	0.945	0.929	0.936	0.943
0.01	6000	30	200	Relative Bias (%)	-0.123	-1.312	0.605	-4.413	-3.548	-1.771	-4.493	-3.387	-1.700
				Mean Squared Error	0.020	0.099	0.114	0.018	0.039	0.036	0.019	0.051	0.047
				Coverage Probability	0.937	0.927	0.933	0.884	0.948	0.957	0.861	0.936	0.951
		60	100	Relative Bias (%)	-0.126	-1.027	0.231	-4.235	-3.015	-2.005	-4.223	-3.746	-2.682
				Mean Squared Error	0.018	0.098	0.109	0.015	0.040	0.043	0.015	0.052	0.057
				Coverage Probability	0.947	0.948	0.949	0.902	0.963	0.960	0.904	0.944	0.946
0.05	6000	30	200	Relative Bias (%)	0.175	-1.346	1.144	-3.992	-3.220	-2.014	-4.066	-2.693	-1.758
				Mean Squared Error	0.033	0.117	0.122	0.039	0.041	0.044	0.039	0.055	0.059
				Coverage Probability	0.860	0.917	0.932	0.866	0.950	0.954	0.865	0.935	0.941
		60	100	Relative Bias (%)	-0.091	-1.291	0.507	-4.181	-3.172	-2.357	-4.265	-3.622	-2.770
				Mean Squared Error	0.026	0.102	0.118	0.026	0.042	0.049	0.026	0.054	0.063
				Coverage Probability	0.903	0.936	0.934	0.869	0.957	0.959	0.870	0.940	0.941
0.10	6000	30	200	Relative Bias (%)	0.365	-0.283	1.387	-3.430	-2.765	-1.810	-3.618	-2.660	-1.742
				Mean Squared Error	0.049	0.123	0.135	0.065	0.044	0.052	0.063	0.059	0.071
				Coverage Probability	0.780	0.919	0.909	0.855	0.947	0.959	0.868	0.927	0.941
		60	100	Relative Bias (%)	-0.128	-1.394	0.040	-4.003	-3.334	-2.538	-4.138	-3.887	-2.779
				Mean Squared Error	0.034	0.107	0.124	0.038	0.045	0.058	0.038	0.056	0.070
				Coverage Probability	0.843	0.927	0.920	0.871	0.957	0.947	0.871	0.939	0.933

†Note: Variance for MR is estimated using the bootstrap (He and Hu, 2002). Hence coverage probabilities must be interpreted with caution.

Table 2.4: Simulation study of 1000 replicates (of size 600 and 6000) for the exponential distribution comparing the standard least absolute deviations median regression (MR), double-transform-both-sides (DTBS), and transform-both-sides (TBS) models.

					MR†			DTBS			TBS		
Kendall's τ	Sample Size	No. of Clusters	Cluster Size		β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
0.01	600	30	20	Relative Bias (%)	-0.348	-3.200	2.487	-4.497	-1.916	-4.539	-4.372	-1.693	-4.260
				Mean Squared Error	0.147	0.847	0.860	0.086	0.422	0.440	0.086	0.428	0.446
				Coverage Probability	0.946	0.931	0.942	0.933	0.937	0.948	0.929	0.939	0.947
		60	10	Relative Bias (%)	0.883	-4.068	0.341	-4.592	-1.782	-3.351	-4.507	-1.718	-3.349
				Mean Squared Error	0.147	0.758	0.820	0.085	0.405	0.451	0.079	0.409	0.456
				Coverage Probability	0.937	0.950	0.955	0.938	0.938	0.950	0.938	0.940	0.953
0.05	600	30	20	Relative Bias (%)	-0.389	-2.821	1.325	-4.191	-1.458	-4.266	-3.912	-1.562	-4.325
				Mean Squared Error	0.158	0.880	0.837	0.094	0.422	0.456	0.092	0.429	0.459
				Coverage Probability	0.935	0.940	0.944	0.933	0.934	0.940	0.933	0.938	0.944
		60	10	Relative Bias (%)	1.229	-5.623	0.269	-4.629	-1.696	-3.398	-4.525	-1.743	-3.354
				Mean Squared Error	0.158	0.779	0.857	0.086	0.403	0.460	0.084	0.404	0.463
				Coverage Probability	0.937	0.948	0.944	0.934	0.937	0.954	0.931	0.938	0.953
0.10	600	30	20	Relative Bias (%)	-0.388	-4.178	2.031	-3.857	-1.285	-4.152	-3.671	-1.369	-4.104
				Mean Squared Error	0.167	0.854	0.861	0.106	0.423	0.472	0.107	0.429	0.475
				Coverage Probability	0.924	0.944	0.941	0.925	0.935	0.939	0.924	0.937	0.939
		60	10	Relative Bias (%)	1.442	-4.390	0.092	-4.602	-1.978	-3.408	-4.420	-2.148	-3.433
				Mean Squared Error	0.164	0.773	0.898	0.092	0.397	0.465	0.091	0.399	0.469
				Coverage Probability	0.933	0.939	0.927	0.925	0.936	0.945	0.929	0.939	0.947
0.01	6000	30	200	Relative Bias (%)	0.108	0.960	-0.544	-4.295	-1.002	-1.552	-4.223	-1.082	-1.829
				Mean Squared Error	0.016	0.089	0.093	0.015	0.034	0.034	0.015	0.039	0.043
				Coverage Probability	0.943	0.924	0.935	0.873	0.935	0.956	0.885	0.933	0.938
		60	100	Relative Bias (%)	0.075	0.731	-0.177	-4.258	-1.906	-2.186	-4.213	-1.886	-2.283
				Mean Squared Error	0.015	0.079	0.088	0.013	0.039	0.045	0.013	0.043	0.047
				Coverage Probability	0.945	0.953	0.945	0.919	0.936	0.941	0.914	0.935	0.943
0.05	6000	30	200	Relative Bias (%)	0.028	1.038	-0.854	-4.270	-0.641	-2.105	-4.224	-1.808	-2.402
				Mean Squared Error	0.027	0.093	0.099	0.033	0.035	0.040	0.033	0.047	0.054
				Coverage Probability	0.864	0.924	0.927	0.839	0.933	0.955	0.846	0.922	0.936
		60	100	Relative Bias (%)	0.150	0.999	-0.278	-4.004	-1.173	-2.437	-4.059	-1.668	-2.477
				Mean Squared Error	0.021	0.084	0.097	0.022	0.040	0.049	0.021	0.046	0.057
				Coverage Probability	0.902	0.938	0.941	0.886	0.946	0.935	0.875	0.935	0.936
0.10	6000	30	200	Relative Bias (%)	0.093	0.330	-0.826	-4.031	-1.298	-2.575	-4.181	-2.278	-3.059
				Mean Squared Error	0.039	0.098	0.108	0.054	0.042	0.051	0.053	0.053	0.064
				Coverage Probability	0.781	0.923	0.915	0.829	0.921	0.941	0.840	0.909	0.929
		60	100	Relative Bias (%)	0.311	1.227	0.276	-3.671	-1.809	-2.252	-3.888	-1.776	-2.631
				Mean Squared Error	0.028	0.088	0.102	0.032	0.042	0.054	0.031	0.048	0.061
				Coverage Probability	0.853	0.930	0.919	0.865	0.931	0.944	0.872	0.936	0.929

†Note: Variance for MR is estimated using the bootstrap (He and Hu, 2002). Hence coverage probabilities must be interpreted with caution.

Table 2.5: Simulation study of 1000 replicates (of size 600 and 6000) for the log-normal distribution comparing the standard least absolute deviations median regression (MR), double-transform-both-sides (DTBS), and transform-both-sides (TBS) models.

					MR†			DTBS			TBS		
Kendall's τ	Sample Size	No. of Clusters	Cluster Size		β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
0.01	600	30	20	Relative Bias (%)	-0.212	1.646	-2.339	0.507	3.378	0.641	-0.780	1.817	-0.708
				Mean Squared Error	0.111	0.652	0.665	0.066	0.342	0.354	0.064	0.342	0.361
				Coverage Probability	0.948	0.930	0.943	0.943	0.930	0.948	0.934	0.930	0.945
		60	10	Relative Bias (%)	-1.174	2.120	-0.730	0.460	2.519	1.731	-0.765	1.221	0.217
				Mean Squared Error	0.112	0.573	0.619	0.063	0.328	0.370	0.063	0.328	0.379
				Coverage Probability	0.938	0.951	0.960	0.943	0.940	0.956	0.943	0.937	0.955
0.05	600	30	20	Relative Bias (%)	0.071	1.419	-1.126	0.813	3.363	0.836	-0.480	1.852	-0.718
				Mean Squared Error	0.120	0.676	0.661	0.075	0.342	0.372	0.073	0.341	0.374
				Coverage Probability	0.940	0.936	0.936	0.949	0.936	0.951	0.936	0.933	0.945
		60	10	Relative Bias (%)	-1.440	3.035	-0.269	0.420	2.418	1.533	-0.826	0.945	0.281
				Mean Squared Error	0.119	0.593	0.641	0.069	0.325	0.382	0.068	0.326	0.384
				Coverage Probability	0.932	0.948	0.943	0.939	0.931	0.946	0.939	0.933	0.944
0.10	600	30	20	Relative Bias (%)	0.221	2.853	-1.608	0.997	3.442	1.218	-0.190	1.597	-0.712
				Mean Squared Error	0.130	0.650	0.667	0.086	0.345	0.388	0.084	0.343	0.389
				Coverage Probability	0.923	0.950	0.943	0.939	0.932	0.950	0.934	0.937	0.946
		60	10	Relative Bias (%)	-1.459	2.378	-0.394	0.525	2.314	1.414	-0.776	0.776	0.098
				Mean Squared Error	0.124	0.579	0.670	0.075	0.323	0.386	0.073	0.324	0.391
				Coverage Probability	0.934	0.948	0.925	0.943	0.936	0.943	0.937	0.935	0.946
0.01	6000	30	200	Relative Bias (%)	-0.096	-1.036	0.470	0.441	-0.430	0.778	-0.584	3.032	4.210
				Mean Squared Error	0.012	0.068	0.070	0.009	0.036	0.039	0.007	0.036	0.047
				Coverage Probability	0.940	0.927	0.935	0.938	0.936	0.948	0.956	0.939	0.940
		60	100	Relative Bias (%)	-0.100	-0.812	0.189	0.524	0.015	0.644	1.988	2.303	3.249
				Mean Squared Error	0.011	0.060	0.066	0.007	0.033	0.038	0.005	0.032	0.043
				Coverage Probability	0.951	0.949	0.945	0.951	0.950	0.954	0.972	0.951	0.940
0.05	6000	30	200	Relative Bias (%)	0.143	-1.062	0.899	0.651	-0.097	1.006	-0.195	3.180	4.201
				Mean Squared Error	0.020	0.072	0.075	0.017	0.037	0.047	0.021	0.037	0.050
				Coverage Probability	0.860	0.923	0.929	0.940	0.943	0.951	0.905	0.940	0.956
		60	100	Relative Bias (%)	-0.069	-1.016	0.402	0.492	-0.133	0.698	-0.597	1.695	2.949
				Mean Squared Error	0.016	0.062	0.072	0.012	0.034	0.043	0.012	0.033	0.047
				Coverage Probability	0898	0.938	0.933	0.943	0.948	0.945	0.932	0.950	0.948
0.10	6000	30	200	Relative Bias (%)	0.302	-0.227	1.092	0.890	0.260	1.240	0.159	2.747	3.719
				Mean Squared Error	0.030	0.076	0.083	0.028	0.040	0.057	0.037	0.039	0.058
				Coverage Probability	0.779	0.919	0.912	0.939	0.937	0.943	0.895	0.937	0.951
		60	100	Relative Bias (%)	-0.092	-1.090	0.045	0.493	-0.243	0.740	-0.464	1.668	2.995
				Mean Squared Error	0.021	0.065	0.076	0.017	0.035	0.049	0.020	0.035	0.051
				Coverage Probability	0.846	0.931	0.919	0.945	0.956	0.937	0.914	0.952	0.957

†Note: Variance for MR is estimated using the bootstrap (He and Hu, 2002). Hence coverage probabilities must be interpreted with caution.

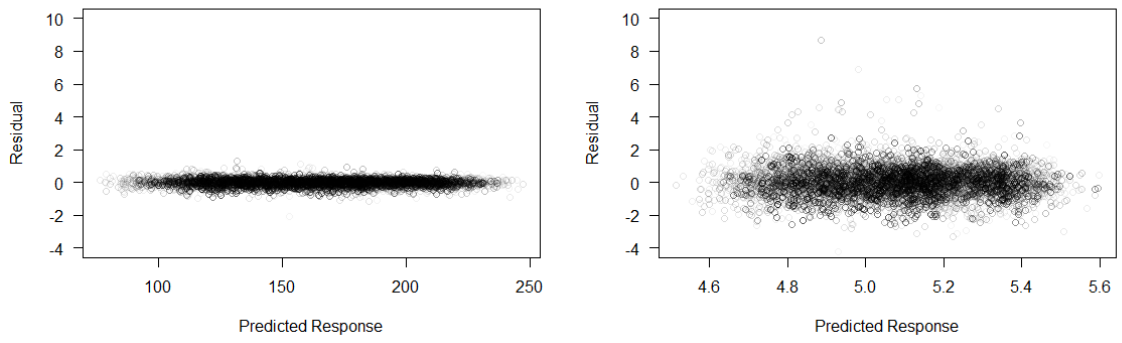
Table 2.6: Point estimates and standard errors for the TBS, DTBS, standard median regression (MR) and ordinary least squares (OLS) regression model applied to the NHANES urinary iodine concentration data.

	TBS[†]			DTBS[‡]			MR			OLS		
Variable	Est.	SE	t	Est.	SE[#]	t	Est.	SE	t	Est.	SE	t
Intercept	150.63	7.469	20.17	147.32	7.459	19.75	147.35	13.041	11.30	5.022	0.075	67.15
Age (years)	-0.133	0.105	-1.27	-0.250	0.104	-2.40	-0.229	0.183	-1.25	-0.002	0.001	-2.28
BMI (kg/m^2)	5.875	1.471	3.99	4.390	1.433	3.06	4.352	1.955	2.23	0.040	0.008	5.06
Total grain (g/day)	-0.215	0.082	-2.62	-0.176	0.085	-2.07	-0.125	0.088	-1.42	-0.001	0.001	-2.11
Gender												
Female	-30.082	4.464	-6.74	-29.375	4.488	-6.55	-29.355	5.655	-5.19	-0.206	0.018	-11.26
Male												
Dairy in-take												
Never/Rare												
Not Often	18.619	5.648	3.30	18.844	5.534	3.41	18.849	8.883	2.12	0.125	0.039	3.18
Often	52.670	5.285	9.97	58.774	5.386	10.91	58.791	10.734	5.48	0.340	0.047	7.24
Fish in-take												
Yes												
No	8.441	5.185	1.63	13.397	5.345	2.51	13.408	6.612	2.03	0.055	0.043	1.27
Race												
White												
Black	-18.046	4.656	-3.88	-18.244	4.606	-3.96	-18.246	10.424	-1.75	-0.106	0.070	-1.50
Hispanic	3.147	4.804	0.66	9.938	4.936	2.01	9.943	7.315	1.36	0.045	0.034	1.31
Other	-7.315	10.31	-0.71	-0.346	10.59	-0.03	-0.346	9.288	-0.04	0.027	0.076	0.36
Salt in-take												
Never/Rarely												
Occasionally	2.032	5.488	0.37	7.762	5.546	1.40	7.771	4.467	1.74	0.020	0.035	0.56
Very Often	7.292	5.705	1.28	7.136	5.638	1.27	7.152	4.644	1.54	0.007	0.037	0.20
Supplements												
Yes												
No	-11.689	5.068	-2.31	-13.966	5.024	-2.78	-13.952	7.963	-1.75	-0.084	0.045	-1.85

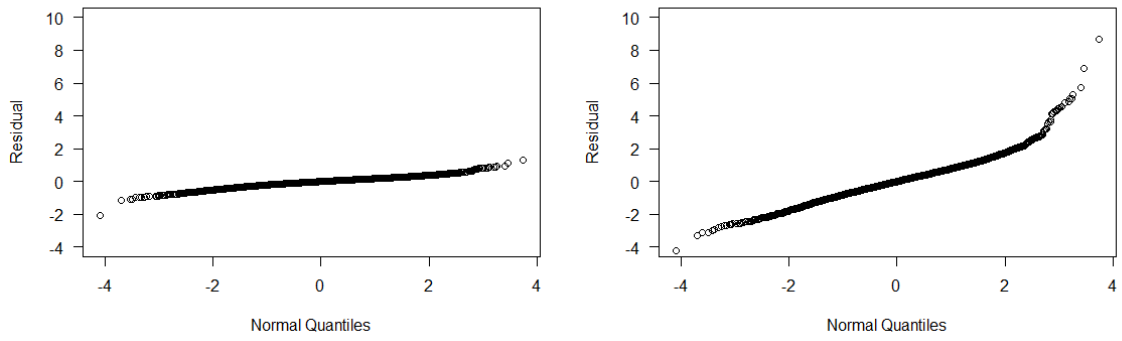
[‡]TBS model yielded estimates of $\lambda = -0.08747$ and $\sigma^2 = 0.3062$.

[†]DTBS model yielded estimates of $\lambda_1 = 0.000468$, $\lambda_2 = 0.6789$ and $\sigma^2 = 0.2711$.

[#] Standard errors (SE) for MR was computed using balanced repeated replication (BRR).



(a) DTBS (b) OLS



(c) DTBS (d) OLS

Figure 2.1: Diagnostic plots. Residual plots (a) and (b) shows predicted response on the untransformed scale and on the predicted log scale, respectively. The intensity of the shading in (a) and (b) are proportional to the sampling weights. Plots (c) and (d) are weighted normal quantile-quantile (QQ) plots.

CHAPTER 3

A NOTE ON ASYMPTOTIC NORMALITY OF L_1 ESTIMATORS FOR COMPLEX SURVEYS

3.1 Introduction

Complex sample surveys are used to collect data on the economic, social, and health status of the individuals in a finite population. These data are generally used to make policy decisions as well as for research purposes. When conducting large national surveys, various design features may be incorporated in the sample survey design. These design features include stratification, clustering and unequal probability sampling. Due to the complex nature of survey data two issues are of paramount concern. The first is the impact of sampling weights on the parameter estimates. The second is the estimation of population variances. Our primary focus is on the latter. Specifically, we address the issue of variance estimation of regression quantiles for complex survey data.

Regression models are widely used in the analysis of complex survey data. In many complex surveys, the outcome of interest may be skewed with no suitable transformation. Under these conditions quantile regression is a viable alternative. However, there are very few examples of quantile regression applied to complex survey data (Geraci, 2013; Chen et al., 2010). One reason for this is that the L_1 estimating equation is a discontinuous function of the regression parameters. Hence the sandwich estimate of variance will not be consistent (Binder, 1983). For the same reason, other well-known estimators such as the Taylor series linearisation estimators and jackknife estimators are not consistent for estimates obtained via the L_1 estimating equation.

Further, resampling methods such as the bootstrap and balanced repeated replication tend to overestimate variance. In practice the primary sampling units are sampled without replacement to avoid selecting the same primary sampling unit more than once. However, it is common practice to treat the primary sampling units as if they were sampled with replacement in order to simplify variance estimation calculations. As a result of this approximation the variance may be overesti-

mated. More importantly, it is generally unclear how to extend resampling methods for complex surveys (Presnell and Booth, 1994).

The objective of this note is to offer an approximation to the estimation of the asymptotic covariance matrix for quantile regression estimators for complex survey data.

3.2 Asymptotic Normality

Consider a stratified multi-stage sampling design in which the primary sampling units are selected with replacement. Let $\mathcal{P}_1, \mathcal{P}_2, \ldots$ denote a sequence of finite populations with H strata in $\mathcal{P}_H$. We assume that the strata sample sizes are fixed in each population and the number of strata is tending to infinity. The linear regression model is

$$y_{khij} = \mathbf{x}_{khij}^{\mathsf{T}}\beta + \epsilon_{khij} \tag{3.1}$$

where $k = 1, 2, \ldots$ is the population index; $h = 1, 2, \ldots, H$ is the stratum index; $i = 1, 2, \ldots, n_h$ is the cluster index within stratum h and $j = 1, 2, \ldots, m_{hi}$ are individuals within cluster i of stratum h. The total number of observations in the sample is $n = \sum_h \sum_i m_{hi}$. The error terms $\{\epsilon_{khij}\}$ are dependent random variables each with median zero. The regression parameter β is a $p \times 1$ vector of unknowns and $\mathbf{x}_{khij}^{\mathsf{T}}$ is a $1 \times p$ vector. Here we let $\mathbf{x}_{khij}^{\mathsf{T}}$ and the sampling weights w_{khij} correspond to the $n \times p$ design matrix $\boldsymbol{X}$ and $n \times n$ diagonal matrix $\boldsymbol{W}$. For simplicity, we omit the population index k in our discussion that follows.

The regression quantiles can be estimated as

$$\hat{\beta}(\tau) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} \rho_\tau(y_{hij} - \mathbf{x}_{hij}^{\mathsf{T}}\beta(\tau)) \tag{3.2}$$

where $\rho_\tau(z) = z(\tau - 1(z < 0))$. The weights w_{hij} are sampling weights and is scaled to sum to one. Note that a minimizer of (3.2) is a solution of the following estimating equation:

$$U(\beta) = \frac{1}{n} \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}(\tau - 1(y_{hij} - \mathbf{x}_{hij}^{\mathsf{T}}\beta(\tau) < 0))\mathbf{x}_{hij} \approx \mathbf{0} \tag{3.3}$$

The regularity conditions necessary to find the limiting distribution of the L_1-estimator are as follows:

(R1) The distribution functions $\{F_{hij}\}$ are absolutely continuous, with continuous densities $f_{hij}(\xi_{hij})$ uniformly bounded away from 0 and ∞ at the points $\xi_{hij}(\tau)$.

(R2) $\lim_{n\to\infty} n^{-1} \sum\sum\sum f_{hij}(\xi_{hij}) \mathbf{x}_{hij} w_{hij} \mathbf{x}_{hij}^{\mathsf{T}} = \boldsymbol{B}(\tau)$ ($p \times p$ positive definite)

(R3) $\max_{hij} |w_{hij}||\mathbf{x}_{hij}^{\mathsf{T}}\theta|/\sqrt{n} \to 0$.

(R4) The higher moments of the convex random functions $G_1^{(1)}(\theta), G_2^{(1)}(\theta), \ldots$ do not increase too rapidly as compared with variance (Lyapunov condition).

(R5) $\max_{h=1,\ldots,H} n_h = O(1)$

(R6) $\max_{h=1,\ldots,H} W_h = O(H^{-1})$ where $W_h = N_h/N$ is the fraction of the population in stratum h

(R7) $n \sum_{h=1}^{H} W_h^2 n_h^{-1} M_h \to \boldsymbol{M}$ ($p \times p$ positive definite)

Theorem 1 *Assume the model* (3.1) *and that the stratum sample sizes are fixed. Under the following regularity conditions R1-R7, as the number of strata tends to infinity the regression quantile* $\hat{\beta}(\tau)$ *converges in distribution to a normal vector with mean zero and covariance matrix given by* $\tau(1-\tau)\boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B}^{-1}$

where

$$\boldsymbol{B} = E[f_{\epsilon_\tau|\boldsymbol{X},\boldsymbol{W}}(0|\boldsymbol{X},\boldsymbol{W})\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}]$$

and

$$\boldsymbol{M} = E\left[\frac{n-1}{n-p}\sum_{h=1}^{H}\frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}(e_{hi\cdot}-\bar{e}_{h\cdot\cdot})^T(e_{hi\cdot}-\bar{e}_{h\cdot\cdot})\right]$$

with

$$
\begin{aligned}
e_{hij} &= w_{hij}\left(\frac{1}{1-\tau}\right)(\tau - 1(y_{hij} - \mathbf{x}_{hij}^{\mathsf{T}}\beta(\tau) < 0))\mathbf{x}_{hij} \\
e_{hi\cdot} &= \sum_{j=1}^{m_{hi}} e_{hij} \\
\bar{e}_{h\cdot\cdot} &= \frac{1}{n_h}\sum_{i=1}^{n_h} e_{hi\cdot}
\end{aligned}
$$

Therefore by incorporating the survey design in the matrix $\boldsymbol{M}$, the covariance matrix of β is given by $V(\beta) = \tau(1-\tau)\boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B}^{-1}$. Here we assume that ϵ_τ is independent of $\boldsymbol{X}$ and $\boldsymbol{W}$. The sparsity function can be written as $s(\tau) = dF_\epsilon^{-1}(\tau)/d\tau$. Numerically, we can approximate this derivative using the following

$$
\begin{aligned}
\frac{dF_\epsilon^{-1}(\tau)}{d\tau} &= \lim_{h\to 0}\frac{F_\epsilon^{-1}(\tau+h) - F_\epsilon^{-1}(\tau-h)}{2h} \\
&\approx \frac{\hat{F}_\epsilon^{-1}(\tau+h) - \hat{F}_\epsilon^{-1}(\tau-h)}{2h} \\
&= \hat{s}(\tau)
\end{aligned}
$$

where h is the bandwidth, $\hat{F}_\epsilon^{-1}(\tau+h) = \epsilon_{(\tau+h)}$ and $\hat{F}_\epsilon^{-1}(\tau-h) = \epsilon_{(\tau-h)}$. Note that the estimator of the sparsity function depends on the bandwidth. This can be computed in a number of ways (Bofingeb, 1975; Hall and Sheather, 1988; Chamberlain, 1994). In this paper we use the very elegant formula of Chamberlain, 1994.

$$
h = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{\tau(1-\tau)}{\sum_h n_h}}
$$

Note that we have replaced the total number of observations in the sample, n with the total number of primary sampling units in the sample, $\sum_h n_h$.

REFERENCES

Gilbert Bassett and Roger Koenker. An empirical quantile function for linear models with iid errors. *Journal of the American Statistical Association*, 77(378):407–415, 1982.

Peter J Bickel and Kjell A Doksum. An analysis of transformations revisited. *Journal of the American Statistical Association*, 76(374):296–311, 1981.

D.A. Binder. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51:279–292, 1983.

Eve Bofingeb. Estimation of a density function using order statistics. *Australian Journal of Statistics*, 17(1):1–7, 1975.

George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252, 1964.

Raymond J Carroll and David Ruppert. Power transformations when fitting theoretical models to data. *Journal of the American Statistical Association*, 79(386):321–328, 1984.

Raymond J Carroll and David Ruppert. *Transformation and weighting in regression*, volume 30. CRC Press, 1988.

Gary Chamberlain. Quantile regression, censoring, and the structure of wages. In *Advances in Econometrics: Sixth World Congress*, volume 2, pages 171–209, 1994.

Qixuan Chen, David H Garabrant, Elizabeth Hedgeman, Roderick JA Little, Michael R Elliott, Brenda Gillespie, Biling Hong, Shih-Yuan Lee, James M Lepkowski, Alfred Franzblau, et al. Estimation of background serum 2, 3, 7, 8-tcdd concentrations by using quantile regression in the umdes and nhanes populations. *Epidemiology*, 21(4):S51–S57, 2010.

RCH Cheng and TC Iles. Corrected maximum likelihood in non-regular problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 95–101, 1987.

Bradley Efron. Bootstrap methods: another look at the jackknife. *The annals of Statistics*, pages 1–26, 1979.

Bradley Efron. *The jackknife, the bootstrap and other resampling plans*, volume 38. SIAM, 1982.

Ulla Feldt-Rasmussen. Iodine and cancer. *Thyroid*, 11(5):483–486, 2001.

Garrett M Fitzmaurice, Stuart R Lipsitz, and Michael Parzen. Approximate median regression via the box-cox transformation. *The American Statistician*, 61(3):233–238, 2007.

Marco Geraci. Estimation of regression quantiles in complex surveys with data missing at random: An application to birthweight determinants. *Statistical methods in medical research, doi: 10.1177/0962280213484401*, 2013.

Peter Hall and Simon J Sheather. On the distribution of a studentized quantile. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 381–391, 1988.

Xuming He and Feifang Hu. Markov chain marginal bootstrap. *Journal of the American Statistical Association*, 97(459):783–795, 2002.

Peter J Huber. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1 (1):221–33, 1967.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

Edward L Korn and Barry I Graubard. *Analysis of health surveys*, volume 323. John Wiley & Sons, 2011.

Kenneth Lange and Janet S Sinsheimer. Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics*, 2(2):175–198, 1993.

Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

Sharon Lohr. *Sampling: design and analysis*. Cengage Learning, 2009.

BFJ Manly. Exponential data transformations. *The Statistician*, pages 37–42, 1976.

Brett Presnell and James G Booth. Resampling methods for sample surveys. *Technical Report 470, Department of Statistics, University of Florida, Gainesville, FL*, 1994.

Jon NK Rao and CFJ Wu. Resampling inference with complex survey data. *Journal of the american statistical association*, 83(401):231–241, 1988.

Keith F Rust and JNK Rao. Variance estimation for complex surveys using replication techniques. *Statistical methods in medical research*, 5(3):283–310, 1996.

Babubhai V Shah. *Encyclopedia of Biostatistics*, volume 3, pages 2276–2279. New York: John Wiley and Sons, 1998.

Jun Shao. Invited discussion paper resampling methods in sample surveys. *Statistics: A Journal of Theoretical and Applied Statistics,*, 27(3-4):203–237, 1996.

Jun Shao and Dongsheng Tu. *The jackknife and bootstrap*. Springer Science & Business Media, 2012.

Jun Shao et al. Impact of the bootstrap on sample surveys. *Statistical Science*, 18(2):191–198, 2003.

BruceV Stadel. Dietary iodine and risk of breast, endometrial, and ovarian cancer. *The Lancet*, 307(7965):890–891, 1976.

Jeremy MG Taylor. Power transformations to symmetry. *Biometrika*, 72(1):145–152, 1985.

Jianqiang C Wang and Jean D Opsomer. On asymptotic normality and variance estimation for nondifferentiable survey estimators. *Biometrika*, 98(1):91–106, 2011.

Naisyin Wang and David Ruppert. Nonparametric estimation of the transformation in the transform-both-sides regression model. *Journal of the American Statistical Association*, 90 (430):522–534, 1995.

Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838, 1980.

Kirk Wolter. *Introduction to variance estimation*. Springer, 2007.

In-Kwon Yeo and Richard A Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, 2000.

BIOGRAPHICAL SKETCH

The author was born in Kingston, Jamaica where he pursued a Bachelor of Science degree in Mathematics at the University of the West Indies–Mona campus. After graduating in 2001 he taught Mathematics at a local high school, Campion College, for 2 years. He then obtained a Masters of Science degree in Biostatistics from the University of the West Indies and had the opportunity to travel overseas to the Medical University of South Carolina in partial fulfilment of the degree. During this period he met Stuart R. Lipsitz who would later become one of his co-advisor for his dissertation. After graduating, he worked as a Biostatistician for 5 years at the Tropical Medicine Research Institute (Kingston, Jamaica) then enrolled in graduate school at Florida State University to pursue a Ph.D. in Statistics.