Published in final edited form as:

Biometrics. 2016 December ; 72(4): 1336–1347. doi:10.1111/biom.12517.

Approximate Median Regression for Complex Survey Data with Skewed Response

Raphael André Fraser^{1,*}, Stuart R. Lipsitz³, Debajyoti Sinha², Garrett M. Fitzmaurice³, and Yi Pan⁴

¹Division of Biostatistics, Medical College of Wisconsin, Milwaukee, Wisconsin, U.S.A

²Department of Statistics, Florida State University, Tallahassee, Florida, U.S.A

³Harvard Medical School, Boston, Massachusetts, U.S.A

⁴Department of Biostatistics, Rollins School of Public Health, Emory University, Atlanta, Georgia, U.S.A

Summary

The ready availability of public-use data from various large national complex surveys has immense potential for the assessment of population characteristics using regression models. Complex surveys can be used 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 complex survey design features. That is, stratification, multistage sampling and weighting. 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 (DTBS) based estimating equations approach to estimate the median regression parameters of the highly skewed response; the DTBS approach applies the same Box-Cox type transformation twice to both the outcome 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 (MAD). Furthermore, the approach is relatively robust to the true underlying distribution, and has much smaller mean square error than a MAD approach. The method is motivated by an analysis of laboratory data on urinary iodine (UI) concentration from the National Health and Nutrition Examination Survey.

Keywords

Complex survey; Median regression; Quantile regression; Sandwich estimator; Transform-both-sides

* rfraser@mcw.edu.

Supplementary Materials

Web Appendix A, Web Table 1 and Web Table 2 referenced in Sections 4 and 5 are available with this paper at the Biometrics website on Wiley Online Library. Additionally, the SAS code for implementing the new method is also available at the Biometrics website.

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. The 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 and hence non-differentiable, 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 the least absolute deviation method. Moreover, use of any resampling method is computationally intensive and may be 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, this is a topic 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 the 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 regression function (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 sampling 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 particular, 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 estimators but with much smaller mean squared error.

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. Median Regression Model

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

$$g_{\lambda_2}(g_{\lambda_1}(y_{ij})) = g_{\lambda_2}(g_{\lambda_1}(x_{ij}^T \beta)) + \varepsilon_{ij}, \quad i=1, 2, \dots, n, \quad j=1, 2, \dots, m_i, \quad (1)$$

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 zero 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 $x_{ij}^T \beta^{(k)} < 0$, 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 all positive or all 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,

$g'_{\lambda}(y) = dg_{\lambda}/dy = |y|^{\lambda-1} \geq 0$ for all y . Note that any real number y can be rewritten as the product of the sign function and absolute value function $\text{sgn}(y)|y|$. 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,

$$g_\lambda(y_{ij}) = g_\lambda(x_{ij}^T \beta) + \varepsilon_{ij}, \quad \text{for } i=1, 2, \dots, n, \quad j=1, 2, \dots, m_i, \quad (2)$$

where the error distribution $\varepsilon_{ij} \stackrel{iid}{\sim} f_\varepsilon(\cdot)$ 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 $x_{ij}^T \beta$ since

$$P(y_{ij} > x_{ij}^T \beta) = P(g_\lambda(y_{ij}) > g_\lambda(x_{ij}^T \beta)) = P(g_\lambda(y_{ij}) - g_\lambda(x_{ij}^T \beta) > 0) = P_\varepsilon(\varepsilon_{ij} > 0) = 1/2.$$

Consequently, the regression model (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). 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). In the next section, expressions for the estimating equations and sandwich variance estimator are derived.

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 logarithm of the probability density function of $g_{\lambda_2}(g_{\lambda_1}(y_{ij}))$ is given by

$$\log f(y_{ij} | x_{ij}, \beta, \sigma^2, \lambda) \propto -\frac{1}{2} \sum_{i=1}^n m_i \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n \sum_{j=1}^{m_i} (\omega_{ij} - \mu_{ij})^2 + \ln \left(\prod_{i=1}^n \prod_{j=1}^{m_i} J(y_{ij}, \lambda) \right), \quad (3)$$

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, $f(y_{ij}|x_{ij}, \beta, \sigma^2, \lambda) = (2\pi\sigma^2)^{-1/2} \exp\{-(\omega_{ij} - \mu_{ij})^2/(2\sigma^2)\} \mathcal{J}(y_{ij}, \lambda)$. With rescaled sampling weights, δ_{ij} , the pseudo-log-likelihood is given by

$$\ell_{\pi}(\beta, \sigma^2, \lambda) = \sum_{i=1}^n \sum_{j=1}^{m_i} \delta_{ij} \log f(y_{ij}|x_{ij}, \beta, \sigma^2, \lambda), \quad (4)$$

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 (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 maximum likelihood estimate $\hat{\beta}$ of the pseudo-log-likelihood function (4) is the same as solving the weighted estimating equation

$$S(\beta) = \frac{\partial}{\partial \beta} \ell_{\pi}(\beta, \sigma^2, \lambda) = \sum_{i=1}^n \sum_{j=1}^{m_i} S_{ij}(\beta) = -\frac{1}{\sigma^2} \sum_{i=1}^n \sum_{j=1}^{m_i} \delta_{ij} \frac{\partial \mu_{ij}}{\partial \beta} (\omega_{ij} - \mu_{ij}) = 0, \quad (5)$$

where

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

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

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

Note that M is the covariance matrix of the estimating equation and B is the Hessian matrix.

We now derive expressions for M and B under the naive likelihood model. Using the expression derived for $S_{ij}(\beta)$ in (5) the matrix M is easily obtained and is given by

$$M = \frac{1}{\sigma^4} \sum_{i=1}^n X_i^T D_{\nu} X_i,$$

with

$$v_i = \delta_i \circ g'_{\lambda_2}(\eta_i) \circ g'_{\lambda_1}(\tau_i) \circ (\omega_i - \mu_i) \text{ and } D_{\nu} = \text{Diag}(v_i^2),$$

where operator $\circ$ denotes the Hadamard product and δ_i , η_i , τ_i , ω_i , μ_i are vectors corresponding to scalars δ_{ij} , η_{ij} , τ_{ij} , ω_{ij} , μ_{ij} . Next we obtain B by taking the second partial derivative of the pseudo-log-likelihood function (3) with respect to β

$$S'(\beta) = \frac{\partial}{\partial \beta \partial \beta^T} \ell_{\pi}(\beta, \sigma^2, \lambda) = -\frac{1}{\sigma^2} \sum_{i=1}^n \sum_{j=1}^{m_i} \delta_{ij} \left[(\omega_{ij} - \mu_{ij}) \frac{\partial^2 \mu_{ij}}{\partial \beta \partial \beta^T} - \frac{\partial \mu_{ij}}{\partial \beta} \left(\frac{\partial \mu_{ij}}{\partial \beta} \right)^T \right]. \quad (6)$$

Hence

$$B = -\frac{1}{\sigma^2} \mathbf{X}^T D_{\alpha} \mathbf{X},$$

where

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

$D_{\alpha} = \text{Diag}(\alpha)$ and $\delta, \eta, \tau, \omega, \mu$ are vectors corresponding to vectors $\delta_i, \eta_i, \tau_i, \omega_i, \mu_i$.

Therefore the sandwich estimate of variance is given by

$$\hat{V}_{\hat{\beta}} = \mathbf{X}^T \hat{D}_{\alpha} \mathbf{X} \left(\sum_{i=1}^n \mathbf{X}_i^T \hat{D}_{\nu} \mathbf{X}_i \right) \mathbf{X}^T \hat{D}_{\alpha} \mathbf{X} / \hat{\sigma}^6, \quad (7)$$

where all the parameters are estimated using maximum likelihood estimation. Next, we evaluate the performance of the DTBS estimator in finite samples.

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 considered a design consisting of three discrete and continuous covariates with various combinations of different correlations (0.01, 0.05 or 0.10), sample sizes (600 or 6000), number of clusters (30 or 60) and cluster sizes (10, 20, 100 or 200). For each cluster, we simulated multivariate normal observations with exchangeable correlation. 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 $x_{ij3} = 1$ with probability 0.5 and $x_{ij3} = -1$ otherwise. For the gamma distribution $x_{ij2} \sim TN(0, 1, -2, 2)$ and for the Pareto distribution $x_{ij1} \sim U[1, 5]$. The notation $TN(0, 1, a, b)$ represents a standard normal distribution truncated at a and b . The transformation is given by $y_{ij} = F^{-1}(U_{ij})$ and $U_{ij} = \Phi(Z_{ij})$ such that $Z_{ij} = (\alpha_i + \varepsilon_{ij}) / (1 + \sigma_{\alpha}^2)^{1/2}$, $\varepsilon_{ij} \sim N(0, 1)$, $\alpha_i \sim N(0, \sigma_{\alpha}^2)$, $\sigma_{\alpha}^2 = \rho / (1 - \rho)$ 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 1–3 and Web Tables 1–2 indicate that the proposed DTBS method yields estimates that are relatively 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 still small for both the MR and DTBS models and were at most -18.9% , -12.9% for the gamma distribution and 5.7% , 6.2% for the Pareto distribution, respectively (Web Table 2; Table 3). In contrast, the TBS model yields biased estimates that were large compared with the MR and DTBS models for the gamma (with bias as large as -30.7%) and Pareto (with bias as large as 68.9%) distributions, indicating that a single transformation may not be sufficient for extremely skewed distributions. For the log-normal, exponential and Weibull distributions (Tables 1–2; Web Table 1) in which the DTBS and TBS appear to be almost unbiased, the mean squared error of the DTBS and TBS are similar, suggesting that the additional parameter estimated in the DTBS model does not increase the variance of the estimator. In addition, 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). Finally, for the extremely skewed, heavy-tailed Pareto distribution, the mean squared error of the DTBS is discernibly smaller than for the TBS across all simulation configurations. For the extremely skewed, heavy-tailed gamma distribution, the results are more equivocal when comparing DTBS to TBS. There is a clear bias-variance tradeoff in this particular setting: the TBS is more efficient, with relative efficiency ranging from 70% to 99%. However, in this setting, the TBS is also more biased for smaller sample sizes and generally has poorer coverage probabilities.

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 $\mu\text{g L}^{-1}$, 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 4 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 4 is 16 for all models.

Note that the estimated coefficients for TBS and DTBS in Table 4 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 Tables 1–3 and Web Tables 1–2, 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 1(b)). In addition, the QQ plot for ordinary least squares regression strongly indicates a violation of the assumption of normal errors (Figure 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 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.

6. Discussion

As a viable alternative to the existing standard median regression method for complex sample survey, we present a theoretically sound method where a consistent estimator of the standard errors can be conveniently computed. One advantage of our model is that it allows skewness as well as heteroscedasticity of the response because there is a relationship between the median regression function $\mu(x) = x^T\beta$ and the variance $\text{Var}(Y|x)$ of the original response

$$\text{Var}[Y|x] \approx \frac{\sigma_\varepsilon^2}{\left[h'_\lambda(h_\lambda^{-1}(\mu))\right]^2}, \quad (8)$$

where $h_\lambda(\cdot) = g_{\lambda_2}(g_{\lambda_1}(\cdot))$ with h_λ^{-1} being its inverse, even though the after-transformation error ε_{ij} in (1) has common variance σ_ε^2 (See Web Appendix A for the derivation of equation 8). One key difference between our transform-both-sides method and standard median regression based on LAD is that our method assumes that the error ε_{ij} of our model follows a parametric density (at least approximately) whereas standard median regression makes no parametric distributional assumption. We also note that even if the moments of the original response distribution are not defined as in the case of the Cauchy density, our method is still valid as long as the estimating equation of (5) based on our double-transformation is unbiased for the underlying distribution. Based on our simulation study, the proposed DTBS

estimator is found to be relatively robust to presence of 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 and is applicable to any multi-stage complex sampling design.

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 with ν degrees of freedom can be expressed as a scale mixture of normals by letting $\varepsilon_i \sim N(0, u_i^{-1}\sigma^2)$ with $u_i \sim Ga(\nu/2, \nu/2)$. The maximum likelihood estimate of β under the t_ν model has estimating equation

$$\sum_i \sum_j \eta_{ij} \delta_{ij} \frac{\partial \mu_{ij}}{\partial \beta} (\omega_{ij} - \mu_{ij}) / \sigma^2 = 0$$
, where $\eta_{ij} = (\nu + 1) / (\nu + \theta_{ij})$ is a weight corresponding to each observation, $\theta_{ij} = (\omega_{ij} - \mu_{ij})^2 / \sigma^2$ and δ_{ij} is the sampling weight. 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.

One possible reason for very limited use of existing 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 obtained from a median response of a finite population even when we know the population size with covariate value x . Existing quantile regression tools only focus on estimating one pre-determined quantile at each analysis. However, our method can simultaneously produce estimates of all quantile functions using only one estimating equation, that is, a single analysis. The τ -th quantile, for any $0 < \tau < 1$, of response y given x in our model is

$$Q_\tau(y|x) = h_\lambda^{-1} \left\{ h_\lambda(x^T \beta) + F_\varepsilon^{-1}(\tau) \right\} \quad (9)$$

where $h_\lambda^{-1}(u)$ is the inverse of double-transformation $h_\lambda(\cdot) = g_{\lambda_2}(g_{\lambda_1}(\cdot))$ in (1). This quantile, $Q_\tau(y|x)$ for any $0 < \tau < 1$ can be estimated as

$$\hat{Q}_\tau(y|x) = h_{\hat{\lambda}}^{-1} \left\{ h_{\hat{\lambda}}(x^T \hat{\beta}) + \hat{\sigma}_\varepsilon \Phi^{-1}(\tau) \right\} \quad (10)$$

where $\hat{\sigma}_\epsilon \Phi^{-1}(\tau)$ is a parametric estimate of the τ -th quantile of the Gaussian distribution of ϵ_{ij} . The estimates of parameters $\lambda = (\lambda_1, \lambda_2)$ and β are obtained from the single estimating equation of (5) and then the estimator $\hat{\sigma}$ is obtained from using Gaussian distribution of residuals of transformed responses. Alternatively, $Q_\tau(y | x)$ can be estimated by replacing $\hat{\sigma}_\epsilon \Phi^{-1}(\tau)$ in (10) with the empirical quantile F_ϵ^{-1} of the residuals. Hence, using a single 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are grateful for the support provided by the following grants from the US National Institutes of Health: AI 60373, GM 29745, CA 74015, CA 70101, and CA 68484.

References

- Bassett G, Koenker R. An empirical quantile function for linear models with iid errors. *Journal of the American Statistical Association*. 1982; 77:407–415.
- Bickel PJ, Doksum KA. An analysis of transformations revisited. *Journal of the American Statistical Association*. 1981; 76:296–311.
- Binder D. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*. 1983; 51:279–292.
- Carroll RJ, Ruppert D. Power transformations when fitting theoretical models to data. *Journal of the American Statistical Association*. 1984; 79:321–328.
- Carroll, RJ.; Ruppert, D. Transformation and weighting in regression. Vol. 30. CRC Press; 1988.
- Chen Q, Garabrant DH, Hedgeman E, Little RJ, Elliott MR, Gillespie B, Hong B, Lee S-Y, Lepkowski JM, Franzblau A, et al. Estimation of background serum 2, 3, 7, 8-tcdd concentrations by using quantile regression in the umdes and nhanes populations. *Epidemiology*. 2010; 21:S51–S57. [PubMed: 20220524]
- Cheng R, Iles T. Corrected maximum likelihood in non-regular problems. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1987:95–101.
- Feldt-Rasmussen U. Iodine and cancer. *Thyroid*. 2001; 11:483–486. [PubMed: 11396706]
- Fitzmaurice GM, Lipsitz SR, Parzen M. Approximate median regression via the box-cox transformation. *The American Statistician*. 2007; 61:233–238.
- Geraci M. Estimation of regression quantiles in complex surveys with data missing at random: An application to birthweight determinants. *Statistical methods in medical research*. 2013
- He X, Hu F. Markov chain marginal bootstrap. *Journal of the American Statistical Association*. 2002; 97:783–795.
- Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. 1967; 1:221–33.
- Lange K, Sinsheimer JS. Normal/independent distributions and their applications in robust regression. *Journal of Computational and Graphical Statistics*. 1993; 2:175–198.
- Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73:13–22.
- Lohr, S. Sampling: design and analysis. Cengage Learning; 2009.
- Presnell, B.; Booth, JG. Technical Report 470. Gainesville, FL: Department of Statistics, University of Florida; 1994. Resampling methods for sample surveys.

- Shao J. Invited discussion paper resampling methods in sample surveys. *Statistics: A Journal of Theoretical and Applied Statistics*. 1996; 27:203–237.
- Shao J, et al. Impact of the bootstrap on sample surveys. *Statistical Science*. 2003; 18:191–198.
- Stadel B. Dietary iodine and risk of breast, endometrial, and ovarian cancer. *The Lancet*. 1976; 307:890–891.
- Taylor JM. Power transformations to symmetry. *Biometrika*. 1985; 72:145–152.
- Wang JC, Opsomer JD. On asymptotic normality and variance estimation for nondifferentiable survey estimators. *Biometrika*. 2011; 98:91–106.
- Wang N, Ruppert D. Nonparametric estimation of the transformation in the transform-both-sides regression model. *Journal of the American Statistical Association*. 1995; 90:522–534.
- White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*. 1980:817–838.
- Yeo I-K, Johnson RA. A new family of power transformations to improve normality or symmetry. *Biometrika*. 2000; 87:954–959.

(a) DTBS

(b) OLS

(c) DTBS

(d) OLS

Figure 1.

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

Table 1

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.

Kendall's τ	Sample Size	No. of Clusters	Cluster Size	MR [†]			DTBS			TBS			
				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	0.469	3.163	0.609	-0.780	1.817	-0.708
				Mean Squared Error	0.111	0.652	0.665	0.069	0.351	0.372	0.064	0.342	0.361
				Coverage Probability	0.948	0.930	0.943	0.964	0.954	0.975	0.934	0.930	0.945
		60	10	Relative Bias (%)	-1.174	2.120	-0.730	0.439	2.453	1.468	-0.765	1.221	0.217
				Mean Squared Error	0.112	0.573	0.619	0.066	0.339	0.390	0.063	0.328	0.379
				Coverage Probability	0.938	0.951	0.960	0.955	0.947	0.965	0.943	0.937	0.955
0.05	600	30	20	Relative Bias (%)	0.071	1.419	-1.126	0.758	3.207	0.645	-0.480	1.852	-0.718
				Mean Squared Error	0.120	0.676	0.661	0.077	0.352	0.385	0.073	0.341	0.374
				Coverage Probability	0.940	0.936	0.936	0.963	0.954	0.970	0.936	0.933	0.945
		60	10	Relative Bias (%)	-1.440	3.035	-0.269	0.393	2.304	1.477	-0.826	0.945	0.281
				Mean Squared Error	0.119	0.593	0.641	0.070	0.337	0.398	0.068	0.326	0.384
				Coverage Probability	0.932	0.948	0.943	0.957	0.944	0.960	0.939	0.933	0.944
0.10	600	30	20	Relative Bias (%)	0.221	2.853	-1.608	1.025	3.407	0.866	-0.190	1.597	-0.712
				Mean Squared Error	0.130	0.650	0.667	0.087	0.356	0.402	0.084	0.343	0.389
				Coverage Probability	0.923	0.950	0.943	0.963	0.957	0.971	0.934	0.937	0.946
		60	10	Relative Bias (%)	-1.459	2.378	-0.394	0.391	2.207	1.449	-0.776	0.776	0.098
				Mean Squared Error	0.124	0.579	0.670	0.076	0.334	0.402	0.073	0.324	0.391
				Coverage Probability	0.934	0.948	0.925	0.956	0.938	0.959	0.937	0.935	0.946
0.01	6000	30	200	Relative Bias (%)	-0.096	-1.036	0.470	0.457	-0.295	0.852	-0.584	3.032	4.210
				Mean Squared Error	0.012	0.068	0.070	0.009	0.035	0.038	0.007	0.036	0.047
				Coverage Probability	0.940	0.927	0.935	0.955	0.954	0.959	0.956	0.939	0.940
		60	100	Relative Bias (%)	-0.100	-0.812	0.189	0.524	0.085	0.726	1.988	2.303	3.249
				Mean Squared Error	0.011	0.060	0.066	0.007	0.032	0.038	0.005	0.032	0.043
				Coverage Probability	0.951	0.949	0.945	0.952	0.951	0.950	0.972	0.951	0.940
0.05	6000	30	200	Relative Bias (%)	0.143	-1.062	0.899	0.664	0.001	1.082	-0.195	3.180	4.201
				Mean Squared Error	0.020	0.072	0.075	0.017	0.037	0.046	0.021	0.037	0.050

Kendall's τ	Sample Size	No. of Clusters	Cluster Size	MR [†]						DTBS			TBS		
				β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
0.10	6000	60	100	Coverage Probability	0.860	0.923	0.929	0.958	0.952	0.959	0.905	0.940	0.956	0.956	0.956
				Relative Bias (%)	-0.069	-1.016	0.402	0.513	-0.024	0.757	-0.597	1.695	2.949	2.949	2.949
				Mean Squared Error	0.016	0.062	0.072	0.011	0.033	0.043	0.012	0.033	0.047	0.047	0.047
		30	200	Coverage Probability	0.898	0.938	0.933	0.947	0.951	0.946	0.932	0.950	0.948	0.948	0.948
				Relative Bias (%)	0.302	-0.227	1.092	0.906	0.314	1.290	0.159	2.747	3.719	3.719	3.719
				Mean Squared Error	0.030	0.076	0.083	0.028	0.039	0.056	0.037	0.039	0.058	0.058	0.058
	100	60	100	Coverage Probability	0.779	0.919	0.912	0.950	0.951	0.960	0.895	0.937	0.951	0.951	0.951
				Relative Bias (%)	-0.092	-1.090	0.045	0.532	-0.126	0.775	-0.464	1.668	2.995	2.995	2.995
				Mean Squared Error	0.021	0.065	0.076	0.017	0.034	0.049	0.020	0.035	0.051	0.051	0.051
		30	200	Coverage Probability	0.846	0.931	0.919	0.948	0.955	0.945	0.914	0.952	0.957	0.957	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

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.

Kendall's τ	Sample Size	No. of Clusters	Cluster Size	MR [†]			DTBS			TBS		
				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	-4.498	-1.908	-4.535	-4.372	-4.260
				Mean Squared Error	0.147	0.847	0.860	0.086	0.422	0.440	0.086	0.446
				Coverage Probability	0.946	0.931	0.942	0.957	0.960	0.969	0.929	0.947
		60	10	Relative Bias (%)	0.883	-4.068	0.341	-4.590	-1.780	-3.352	-4.507	-3.349
				Mean Squared Error	0.147	0.758	0.820	0.084	0.404	0.451	0.079	0.409
				Coverage Probability	0.937	0.950	0.955	0.950	0.952	0.9630	0.938	0.953
0.05	600	30	20	Relative Bias (%)	-0.389	-2.821	1.325	-4.235	-1.482	-4.379	-3.912	-4.325
				Mean Squared Error	0.158	0.880	0.837	0.098	0.422	0.456	0.092	0.429
				Coverage Probability	0.935	0.940	0.944	0.954	0.957	0.973	0.933	0.944
		60	10	Relative Bias (%)	1.229	-5.623	0.269	-4.678	-1.854	-3.421	-4.525	-3.354
				Mean Squared Error	0.158	0.779	0.857	0.092	0.398	0.458	0.084	0.404
				Coverage Probability	0.937	0.948	0.944	0.941	0.950	0.961	0.931	0.938
0.10	600	30	20	Relative Bias (%)	-0.388	-4.178	2.031	-3.926	-1.292	-4.174	-3.671	-4.104
				Mean Squared Error	0.167	0.854	0.861	0.109	0.424	0.473	0.107	0.429
				Coverage Probability	0.924	0.944	0.941	0.952	0.960	0.973	0.924	0.939
		60	10	Relative Bias (%)	1.442	-4.390	0.092	-4.669	-1.986	-3.457	-4.420	-3.433
				Mean Squared Error	0.164	0.773	0.898	0.098	0.396	0.466	0.091	0.399
				Coverage Probability	0.933	0.939	0.927	0.940	0.948	0.958	0.929	0.947
0.01	6000	30	200	Relative Bias (%)	0.108	0.960	-0.544	-4.192	-1.978	-1.032	-4.223	-1.829
				Mean Squared Error	0.016	0.089	0.093	0.015	0.033	0.033	0.015	0.043
				Coverage Probability	0.943	0.924	0.935	0.936	0.966	0.965	0.885	0.938
		60	100	Relative Bias (%)	0.075	0.731	-0.177	-4.345	-2.095	-1.619	-4.213	-2.283
				Mean Squared Error	0.015	0.079	0.088	0.014	0.040	0.042	0.013	0.047
				Coverage Probability	0.945	0.953	0.945	0.908	0.958	0.955	0.914	0.943
0.05	6000	30	200	Relative Bias (%)	0.028	1.038	-0.854	-3.217	-4.073	-2.779	-4.224	-2.402
				Mean Squared Error	0.027	0.093	0.099	0.025	0.048	0.059	0.033	0.054

Kendall's τ	Sample Size	No. of Clusters	Cluster Size	MR [‡]						DTBS			TBS		
				β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3
				Coverage Probability	Relative Bias (%)	Mean Squared Error	Coverage Probability	Relative Bias (%)	Mean Squared Error	Coverage Probability	Relative Bias (%)	Mean Squared Error	Coverage Probability	Relative Bias (%)	Mean Squared Error
0.10	6000	60	100	0.864	0.924	0.927	0.941	0.960	0.954	0.846	0.922	0.936	0.846	0.922	0.936
				0.150	0.999	-0.278	-3.400	-4.118	-3.125	-4.059	-1.668	-2.477	-4.059	-1.668	-2.477
				0.021	0.084	0.097	0.019	0.043	0.056	0.021	0.046	0.057	0.021	0.046	0.057
		30	200	0.902	0.938	0.941	0.911	0.962	0.949	0.875	0.935	0.936	0.875	0.935	0.936
				0.093	0.330	-0.826	-3.017	-3.690	-2.556	-4.181	-2.278	-3.059	-4.181	-2.278	-3.059
				0.039	0.098	0.108	0.037	0.050	0.071	0.053	0.053	0.064	0.053	0.053	0.064
	100	60	100	0.781	0.923	0.915	0.947	0.959	0.956	0.840	0.909	0.929	0.840	0.909	0.929
				0.311	1.227	0.276	-3.411	-4.227	-3.093	-3.888	-1.776	-2.631	-3.888	-1.776	-2.631
				0.028	0.088	0.102	0.025	0.045	0.063	0.031	0.048	0.061	0.031	0.048	0.061
		30	200	0.853	0.930	0.919	0.916	0.961	0.942	0.872	0.936	0.929	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 3

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

Kendall's τ	Sample Size	No. of Clusters	Cluster Size	MR [†]			DTBS			TBS			
				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 (%)	-0.178	-1.605	1.011	3.764	-0.805	2.534	35.090	5.141	58.024
			Mean Squared Error	2.292	2.352	3.633	0.720	0.589	0.917	3.488	1.709	7.197	
			Coverage Probability	0.944	0.929	0.906	0.981	0.954	0.984	0.885	0.816	0.924	
0.05	600	60	10	Relative Bias (%)	-3.323	-4.252	-1.073	-0.230	-3.122	1.690	33.303	4.479	55.713
			Mean Squared Error	2.150	2.273	3.602	0.626	0.611	1.075	3.382	1.979	7.799	
			Coverage Probability	0.952	0.925	0.924	0.982	0.960	0.972	0.899	0.809	0.920	
0.10	600	30	20	Relative Bias (%)	1.994	3.139	2.295	2.898	-2.169	-0.519	38.824	7.755	62.258
			Mean Squared Error	2.603	2.531	3.967	1.021	0.878	1.221	4.416	1.862	8.562	
			Coverage Probability	0.932	0.917	0.909	0.978	0.956	0.984	0.919	0.853	0.929	
0.01	6000	60	10	Relative Bias (%)	-0.114	-5.695	3.091	1.641	-6.052	-0.687	32.633	-5.978	55.180
			Mean Squared Error	2.390	2.478	4.053	1.016	0.866	1.320	3.684	1.955	8.299	
			Coverage Probability	0.941	0.931	0.909	0.958	0.949	0.971	0.912	0.816	0.923	
0.05	6000	30	20	Relative Bias (%)	3.206	1.292	5.406	2.600	0.238	5.416	42.861	9.221	68.860
			Mean Squared Error	3.005	2.814	4.317	0.989	0.728	1.373	5.620	2.016	10.976	
			Coverage Probability	0.915	0.908	0.892	0.968	0.952	0.968	0.923	0.890	0.931	
0.10	6000	60	10	Relative Bias (%)	1.472	-4.564	2.004	0.934	-4.155	4.092	37.842	-0.861	61.038
			Mean Squared Error	2.619	2.633	4.211	0.883	0.743	1.281	4.255	2.141	9.611	
			Coverage Probability	0.916	0.927	0.906	0.976	0.951	0.972	0.928	0.839	0.923	
0.01	6000	30	200	Relative Bias (%)	0.343	-1.040	0.144	5.507	-0.466	0.631	31.841	-5.483	42.847
			Mean Squared Error	0.256	0.243	0.372	0.083	0.044	0.068	0.734	0.114	1.367	
			Coverage Probability	0.941	0.937	0.901	0.974	0.984	0.987	0.839	0.926	0.858	
0.05	6000	60	100	Relative Bias (%)	0.524	-0.676	1.336	4.887	0.152	0.953	40.688	2.824	51.552
			Mean Squared Error	0.243	0.271	0.329	0.067	0.040	0.064	1.143	0.152	1.766	
			Coverage Probability	0.948	0.918	0.921	0.987	0.986	0.989	0.718	0.938	0.783	
0.10	6000	30	200	Relative Bias (%)	2.304	0.604	2.542	6.099	0.899	1.332	48.683	7.633	59.695
			Mean Squared Error	0.417	0.282	0.575	0.170	0.061	0.112	2.345	0.302	3.197	

Kendall's τ	Sample Size	No. of Clusters	Cluster Size	MR [†]						DTBS						TBS					
					β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3	β_1	β_2	β_3		
0.10	6000	60	100	Coverage Probability	0.877	0.923	0.844	0.943	0.970	0.970	0.928	0.939	0.944								
				Relative Bias (%)	1.586	-0.950	2.369	0.761	-1.157	1.126	43.055	4.407	54.624								
				Mean Squared Error	0.313	0.288	0.404	0.122	0.071	0.135	1.479	0.188	2.381								
				Coverage Probability	0.918	0.921	0.883	0.951	0.969	0.966	0.883	0.947	0.901								
				Relative Bias (%)	5.184	2.539	5.111	2.540	-1.010	2.042	52.766	8.826	64.813								
				Mean Squared Error	0.669	0.337	0.820	0.347	0.099	0.264	3.201	0.322	4.617								
	100	60	100	Coverage Probability	0.798	0.901	0.788	0.900	0.980	0.949	0.870	0.935	0.871								
				Relative Bias (%)	2.712	-1.485	3.709	6.193	-0.039	1.895	44.123	4.311	56.236								
				Mean Squared Error	0.388	0.309	0.524	0.157	0.048	0.109	1.610	0.175	2.611								
				Coverage Probability	0.890	0.912	0.852	0.946	0.976	0.967	0.956	0.944	0.963								

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

Table 4

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 consisting of 6,802 individuals.

Variable	TBS [†]			DTBS [‡]			MR			OLS		
	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												

Variable	TBS [†]			DTBS [‡]			MR			OLS		
	Est.	SE	t	Est.	SE [#]	t	Est.	SE	t	Est.	SE	t
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 were computed using balanced repeated replication (BRR).

^{*} P 0.05,

^{**} P 0.01,

^{***} P 0.001,

^{****} P 0.0001.