FITTING COMPLEX SURVEY DATA TO THE TAIL OF A
PARAMETRIC DISTRIBUTION

I. Park, Texas A&M University
J. L. Eltinge, Bureau of Labor Statistics and Texas A&M University
I. Park, Department of Statistics, Texas A&M University, College Station, TX 77843-3143
(parkinho@stat.tamu.edu)

Key Words: Misspecification effect matrix; Design effect matrix; Design-based inference; Normal distribution; Quantile plots; Stratified multistage sample design; Superpopulation model; U.S. Third National Health and Nutrition Examination Survey (NHANES III).

1. Introduction

For simple random samples, one sometimes estimates the mean and standard deviation of a normal population from regression of observed quantiles on the corresponding standard normal quantiles. For some general discussion of parameter estimation using least squares methods applied to quantile plots, see, e.g., Chernoff and Lieberman (1954), Barnett (1975), Cran (1975) and references cited therein.

This paper considers an extension of this idea to data collected through a stratified multistage sample survey. Principal attention is devoted to fitting a parametric model to the tail of an underlying superpopulation distribution. With design-based quantile estimators and associated covariance matrix estimators proposed by Francisco and Fuller (1991), direct application of ordinary least squares and generalized least squares methods lead to point estimators of the superpopulation parameters and quantiles; associated variance estimators; and related goodness-of-fit test statistics. We also consider methods based on constrained misspecification effect matrices, extending covariance matrix approximation ideas used in, e.g., Rao and Scott (1981, 1984). The proposed methods are applied to medical examination data from the U.S. Third National Health and Nutrition Examination Survey (NHANES III).

2. Quantile Estimation and Related Variance Estimation for Complex Survey Data

Consider a sequence of finite populations indexed by \( \nu = 1, 2, \ldots \) Suppose that the \( \nu \)th finite population with \( N_\nu \) ultimate units is a simple random sample selected with replacement from an infinite superpopulation. Also, assume that population \( \nu \) has been partitioned into \( L_\nu \) strata with \( N_{\nu h} \) primary sampling units (PSUs). From the \( h \)th stratum, \( n_{\nu h} \geq 2 \) PSUs are selected with replacement using possibly unequal per-draw probabilities \( p_{\nu hi} \) independently across strata, \( i = 1, \ldots , n_{\nu h} \). Furthermore, PSU \((h,i)\) contains \( N_{\nu hi} \) ultimate units, among which \( n_{\nu hi} \) units are also selected by using possibly unequal probabilities \( p_{\nu hj} \), \( j = 1, \ldots , N_{\nu hi} \). Note that we have \( M_{\nu h} = \sum_{i=1}^{n_{\nu h}} N_{\nu hi} \) ultimate units in stratum \( h \) and \( N_\nu = \sum_{h=1}^{L_\nu} M_{\nu h} \) in the \( \nu \)th finite population. Also, \( n_\nu = \sum_{h=1}^{L_\nu} n_{\nu h} \) PSUs and \( n_{\nu T} = \sum_{h=1}^{L_\nu} \sum_{i=1}^{n_{\nu h}} n_{\nu hi} \) ultimate units are selected for the \( \nu \)th sample in the sequence.

For the survey variable \( Y \), assume that the associated superpopulation distribution function \( F(\cdot) \) is continuous. For a fixed vector \( y = (y_1, \ldots , y_k)' \), let \( x_{\nu hj} = \{1, \delta_{\nu hj}(y_1), \ldots , \delta_{\nu hj}(y_k)\}' \) and \( w_{\nu hj} \) denote, respectively, the observed vector and the associated survey weight for each sampled unit \((h,i,j)\) in the \( \nu \)th sample, where \( \delta_{\nu hj}(y_l) = 1 \) if \( Y_{\nu hj} \leq y_l \) or 0 otherwise. Then a set of \( k \) customary estimators of the distribution functions evaluated at \( y_l, l = 1, \ldots , k \), are a vector of \( k \) ratio estimators given by

\[
\hat{F}_\nu(y) = \{\hat{F}_\nu(y_1), \ldots , \hat{F}_\nu(y_k)\}',
\]

where \( \hat{F}_\nu(y_l) = \hat{X}_{\nu,l+1}/\hat{X}_{\nu 1} \) and \( \hat{X}_\nu = (\hat{X}_\nu 1, \ldots , \hat{X}_\nu,k+1)' = \sum_{h=1}^{L_\nu} \sum_{i=1}^{n_{\nu h}} \sum_{j=1}^{n_{\nu hi}} w_{\nu hj} x_{\nu hj} \).

Asymptotic results regarding \( \hat{F}_\nu(y) \) are available in Shao (1996, pp. 209-211) and Francisco and Fuller (1991, Theorem 2).

Now we consider a set of \( k \) prespecified probability values \( 0 < \pi_1 < \ldots < \pi_k < 1 \). For each \( \pi_l \), the...
superpopulation quantile is given by

\[ q_{\pi} = \inf\{y : F(y) \geq \pi\} = F^{-1}(\pi). \]

The corresponding sample quantile estimator is defined by

\[ \hat{q}_{\pi;D} = \inf\{y : \hat{F}_v(y) \geq \pi\} = \hat{F}_v^{-1}(\pi) \]

For stratified single stage sampling, Francisco and Fuller (1991, Theorem 3) demonstrated under regularity conditions that the Bahadur representation of the sample quantile \( \hat{q}_{\pi;D} \) can be given as follows; as \( v \to \infty \),

\[ \hat{q}_{\pi;D} = q_{\pi} - \{f(q_{\pi})\}^{-1}\{\hat{F}_v(q_{\pi}) - F(q_{\pi})\} + o_p(n_v^{-1/2}), \]  

where \( f(y) = F'(y) \) is the density function of \( Y \) and the total number of selected PSUs, \( n_v \), is assumed to increase without bound. Using the above representation and the asymptotic normality of \( \hat{F}_v(y) \), Francisco and Fuller (1991, Theorem 4) established the asymptotic multivariate normality of the sample quantile vector of fixed dimension. Let \( q_\pi \) and \( q_{\pi;D} \) denote the vectors of the superpopulation quantiles and the \( u \)th sample quantiles, respectively, for \( \pi = (\pi_1, \ldots, \pi_k)' \). Under suitable restrictions on the subsampling design, as \( v \to \infty \),

\[ n_v^{1/2}\{q_{\pi;D} - q_{\pi}\} \xrightarrow{D} N_k(0, V_Y), \]

where \( V_Y = \lim_{v \to \infty} n_v V\{\hat{q}_{\pi;D}\} \) and \( V_Y \) is positive definite. We may also show that, as \( v \to \infty \),

\[ n_v \left\{ \hat{V}(\hat{q}_{\pi;D}) - V(\hat{q}_{\pi;D}) \right\} \xrightarrow{p} 0, \]

where

\[ \hat{V}(\hat{q}_{\pi;D}) = \hat{D}_v \hat{V}(\hat{F}_v(\hat{q}_{\pi;D})) \hat{D}_v. \]

In addition, \( \hat{V}(\hat{F}_v(\hat{q}_{\pi;D})) \) is the estimated covariance matrix of \( \hat{F}_v(y) \) evaluated at \( \hat{q}_{\pi;D} \) such that, as \( v \to \infty \), \( n_v \left\{ \hat{V}[\hat{F}_v(\hat{q}_{\pi;D})] - V[\hat{F}_v(q_{\pi;D})] \right\} \xrightarrow{p} 0 \).

Also, \( \hat{D}_v \) is the diagonal matrix whose \( i \)th diagonal element, given by (Francisco and Fuller, 1991, Theorem 4),

\[
\hat{d}_{vi} = \left(2z_{1-\frac{\pi_i}{2}} \{\hat{V}[\hat{F}_v(q_{\pi_i})]\}^{1/2}\right)^{-1}
\times \left\{ \hat{F}_v^{-1}(\pi_i + z_{1-\frac{\pi_i}{2}} \{\hat{V}[\hat{F}_v(q_{\pi_i})]\}^{1/2}) - \hat{F}_v^{-1}(\pi_i - z_{1-\frac{\pi_i}{2}} \{\hat{V}[\hat{F}_v(q_{\pi_i})]\}^{1/2}) \right\},
\]

is a design-based estimator of \( d_i = \{f(q_{\pi_i})\}^{-1} \) and \( z_{1-\frac{\pi_i}{2}} = \Phi^{-1}(1 - \frac{\pi_i}{2}) \) such that as \( v \to \infty \),

\[ \hat{d}_{vi} \xrightarrow{p} f(q_{\pi_i}). \]  

Note that \( \hat{V}\{\hat{F}_v(q_{\pi})\} \) is a linearization estimator of the covariance matrix of the asymptotic distribution of \( \{\hat{F}_v(q_{\pi}) - F(q_{\pi})\} \). See, e.g., Francisco and Fuller (1991, p. 459); and Shao (1996, pp. 207-8). Consequently, its degrees of freedom is at most \( \sum_{h=1}^{L_v}(n_{vh} - 1) = n_v - L_v \).

3. Model-based Estimation of the Superpopulation Parameters and Quantiles from Tails of Quantile Plots

For the remainder of this paper, we assume that the true distribution function may be written in the form \( F(y) = \Phi \left( \frac{v-y}{\sigma} \right) \) for \( y \in \mathbb{R} \), where \( \Phi(\cdot) \) is the standard normal distribution function. Since \( \Phi(x) \) is a strictly increasing continuous function of \( x \), the \( p \)th quantile is uniquely expressed as

\[ q_p = \mu + \sigma \Phi^{-1}(p). \]

For this discussion, we consider a set of \( k \) probability values in one tail, say, \( \pi_l = 0.75(0.01)0.99, \ l = 1, \ldots, k, \) instead of considering the entire interval \((0, 1)\). Given complex survey data, a normal quantile plot can then be constructed by plotting \( \hat{q}_{\pi;D} \) against \( v_l \), where \( v_l = \Phi^{-1}(\pi_l) \). Under the normality assumption for the superpopulation model of \( Y \), expression (6) becomes the \( p \)th superpopulation quantile. A straightforward application of the least squares method to expression (6) leads to ordinary least squares (OLS) estimators of \( \mu \) and \( \sigma \). In general, however, sample quantiles are correlated so that \( V(\hat{q}_{\pi;D}) \) is not a scalar multiple of the identity matrix. If the full covariance matrix is used to obtain a regression line, the resulting generalized least squares (GLS) estimators are anticipated to be more efficient. In addition, use of the diagonal elements, say, \( \hat{D}(\hat{q}_{\pi;D}) = \text{diag}[\hat{V}(\hat{q}_{\pi;D})] \) provides weighted least squares (WLS) estimators. Note that only the weighted least squares estimators would be available if \( \hat{V}(\hat{q}_{\pi;D}) \) is singular. Let \( \hat{\mu}_M \) and \( \hat{\sigma}_M \) denote the least squares estimators based on each method \( M = OLS, GLS(D) \) and \( WLS(D) \). For any \( \pi^* \in (0, 1) \), the associated quantile estimator is then given by

\[ \hat{q}_{\pi^*;M} = \hat{\mu}_M + \hat{\sigma}_M z^*, \]

where \( z^* = \Phi^{-1}(1 - \frac{\pi^*}{2}) \).
where \( z^* = \Phi^{-1}(\pi^*) \). Let \( z = (z_1, \ldots, z_n)' \) and let \( Z = [1, z, \ldots, z] \), where \( 1_k \) is the \( k \)-dimensional column vector of ones. Let \( \theta = (\mu, \sigma)' \) and \( \hat{\theta}_M = (\hat{\beta}_M, \hat{\sigma}_M)' \) denote the parameter vector and its estimator associated with the least squares method \( M \). Thus we have the following vector of estimated quantiles for \( \pi \) in the matrix form;

\[
\hat{q}_{\pi; M} = Z \hat{\theta}_M,
\]

where

\[
\hat{\theta}_M = (Z^*_\pi B^{-1}_\pi Z^*_\pi)^{-1} Z^*_\pi B^{-1}_\pi \hat{q}_{\pi; D}.
\]

and \( B_M = I \), \( \hat{V}(\hat{q}_{\pi; D}) \) and \( \hat{D}(\hat{q}_{\pi; D}) \) for \( M = OLS, GLS(D) \) and \( WLS(D) \), respectively.

4. Goodness-of-Fit Test Statistic

4.1 Asymptotic Distribution

Having obtained fitted quantiles, we can assess the goodness-of-fit of the parametric superpopulation model by comparing the fitted quantiles to the corresponding design-based quantile estimators. Let \( e_{\pi; w; M} = \hat{q}_{\pi; w; D} - \hat{q}_{\pi; w; M} \) be the \( l \)th residual of the chosen method \( M \). Then the standardized residual, defined by,

\[
e_{\pi; w; M}^* = \frac{e_{\pi; w; M}}{\text{se}(e_{\pi; w; M})},
\]

may provide a (standardized) deviation attributed to each fitted quantile, where \( \text{se}(e_{\pi; w; M}) \) is an estimated standard error of \( e_{\pi; w; M} \). Hence, a simple test of the goodness-of-fit of the proposed superpopulation model follows from the statistic

\[
X^2_{GOF}(\hat{q}_{\pi; w; M}) = \sum_{l=1}^{k} e_{\pi; w; M}^* (n-1) \text{se}(e_{\pi; w; M})^{-2}.
\]

The \( k \times 1 \) residual vector may be written as \( e_{\pi; w; M} = (e_{\pi; w; M_1}, \ldots, e_{\pi; w; M_k})' = R_{w; M} \hat{q}_{\pi; w; D} \) for each method \( M \), where

\[
R_{w; M} = I_k - Z^*_\pi \{Z^*_\pi B^{-1}_\pi Z^*_\pi\}^{-1} Z^*_\pi B^{-1}_\pi.
\]

Under the normality assumption for the superpopulation distribution of \( Y \), expression (6) leads to

\[
e_{\pi; w; M} = R_{w; M} (\hat{q}_{\pi; w; D} - q_\pi).
\]

In addition, results (2) and (3) give that under regularity conditions, as \( \nu \to \infty \),

\[
R_{w; M} \to R_M
\]

where \( R_M = I_k - Z^*_\pi \{Z^*_\pi B^{-1}_\pi Z^*_\pi\}^{-1} Z^*_\pi B^{-1}_\pi \) for \( M = OLS, GLS(D) \) and \( WLS(D) \), respectively. Thus it follows that as \( \nu \to \infty \),

\[
n_{\nu}^{1/2} e_{\pi; w; M} \overset{d}{\to} N_k(0, V_M),
\]

where \( V_M = \lim_{\nu \to \infty} n_{\nu} \hat{V}(e_{\pi; w; M}) = R_M V_Y R'_M \) and \( \hat{V}(e_{\pi; w; M}) = R_M \hat{V}(\hat{q}_{\pi; w; D}) R'_M \) is a consistent estimator of \( V(e_{\pi; w; M}) \).

The goodness-of-fit test statistic can be represented as

\[
X^2_{GOF}(\hat{q}_{\pi; w; M}) = e_{\pi; w; M}' \{\hat{D}(\hat{e}_{\pi; w; M})\}^{-1} e_{\pi; w; M} = [n_{\nu}^{1/2} (\hat{q}_{\pi; w; D} - q_\pi)]' [n_{\nu}^{1/2} A_{\nu; M}][n_{\nu}^{1/2} (\hat{q}_{\pi; w; D} - q_\pi)],
\]

(7)

where \( \hat{D}(\hat{e}_{\pi; w; M}) = \text{diag}[\hat{V}(e_{\pi; w; M})] \) and

\[
A_{\nu; M} = R'_{\nu; M} \left\{ \text{diag} \{\hat{V}(e_{\pi; w; M})\} \right\}^{-1} R_{\nu; M}
\]

The goodness-of-fit test statistic can be represented as

\[
X^2_{GOF}(\hat{q}_{\pi; w; M}) \overset{d}{\to} \sum_{l=1}^{k} \lambda_{l; M} X^2_l,
\]

where \( \lambda_{l; M} \geq \ldots \geq \lambda_{k; M} \geq 0 \) are the eigenvalues of the matrix \( V_{Y}^{1/2} A_{\nu; M} V_{Y}^{1/2} \), \( V_{Y}^{1/2} \) is the symmetric square root of \( V_y \), and \( X^2_l \) are \( k \) independent \( \chi^2(1) \) random variables. See, e.g., Graybill (1976, Theorem 4.4.4). Observe that \( A_{\nu; M} \hat{V}(\hat{q}_{\pi; w; D}) \overset{p}{\to} A_M V_Y \) as \( \nu \to \infty \). Thus, in practice, the weights \( \lambda_{l; M} \) can be replaced by consistent estimators \( \hat{\lambda}_{l; M} \), which are the eigenvalues of \( \hat{V}(\hat{q}_{\pi; w; D})^{1/2} A_M \hat{V}(\hat{q}_{\pi; w; D})^{1/2} \).

4.2 Satterthwaite Approximation

In parallel with Rao and Scott (1981, 1984), we may consider some corrections to \( X^2_{GOF}(\hat{q}_{\pi; w; M}) \). Suppose that the matrix \( A_{\nu; M} \hat{V}(\hat{q}_{\pi; w; D}) \) is of rank \( k - r_{\nu; M} \). In addition, suppose that the associated positive eigenvalues \( \hat{\lambda}_{l; M} \) are close to each other so that \( \hat{\lambda}_{l; M} / \hat{\lambda}_{r_{\nu; M}} \approx 1 \), where \( \hat{\lambda}_{r_{\nu; M}} \) denotes the mean of positive eigenvalues \( \hat{\lambda}_{l; D} \) for \( l = 1, \ldots, k - r_{\nu; M} \). Then a first-order correction is

\[
X^2_{GOF}(\hat{q}_{\pi; w; M}) \overset{d}{\to} \sum_{l=1}^{k} \hat{\lambda}_{l; M} X^2_l,
\]

where the asymptotic first moment of which approximately equals that of \( \chi^2(k - r_{\nu; M}) \). If the variation among the positive eigenvalues \( \hat{\lambda}_{l; M} \) is not negligible, the Satterthwaite (1946) procedure may provide a more accurate correction to \( X^2_{GOF}(\hat{q}_{\pi; w; M}) \). Let \( \hat{a}_{\nu; M} \) denote the coefficient of variation of \( \hat{\lambda}_{l; M} \) for \( l =
1, ..., k - r_\nu M. Then a second-order correction is given by

\[ X^2_{\nu M} (\hat{\lambda}_M, \hat{\alpha}_M) = X^2_{\nu M} (\tilde{q}_\nu M) / \hat{c}_\nu M, \]

the asymptotic first and second moments of which are approximately equal those of \( \chi^2 (d_{\nu M}) \), where \( d_{\nu M} = (k - r_\nu M) / (1 + \hat{\alpha}_\nu M) \) and \( \hat{c}_\nu M = \hat{\lambda}_M (1 + \hat{\alpha}_\nu M) \).

## 5. An Alternative Covariance Matrix Estimator Using A Constrained Misspecification Effect Matrix

As addressed in Section 3, our attention will focus on the tail quantiles of a continuous survey variable, e.g., \( p = 0.75(0.01)0.99 \). Customary design-based estimation procedures including those of Woodruff (1952) and Francisco and Fuller (1991) may provide point estimators with a satisfactory level of precision when they are applied to the central region such as \([q_{0.25}, q_{0.75}]\). This is because in many applications, observations are relatively dense in the central region of the distribution. For some survey applications, however, the performance of standard design-based methods may be somewhat less satisfactory. In addition, the presence of relatively few PSUs may result in poor performance of inference based on use of a design-based variance estimator such as a linearization estimator \( V \{ \hat{F}_\nu (\tilde{q}_\nu) \} \), which has at most \( n_\nu - L_\nu \) degrees, as noted in Section 2. See, e.g., Korn and Graubard (1990). From expression (4), \( V_D \) has the same degrees of freedom term. Also, the presence of relatively few PSUs in the survey data (e.g., \( n_\nu - L_\nu < k \)) will result in a singular \( V_D \) so that the generalized least squares method is not feasible. Although one may face possible model misspecification issues, use of a parametric model can reduce the number of parameters to be estimated, which in turn may make the generalized least squares method feasible. These considerations motivate use of the model-based approach in estimation for the tail of the distribution.

Following the ideas used in Rao and Scott (1981, 1984), consider a model-based alternative covariance matrix estimator for the vector of estimated quantiles. The superpopulation quantiles \( q_{n_\nu i} \) determine \( k \) non-overlapping cells of the form \( I_i = (q_{n_\nu i-1}, q_{n_\nu i}) \) and the corresponding cell probabilities are given by \( \pi_i = \Pr (Y \in I_i) = \pi_i - \pi_{i-1} \), where \( \pi_0 = 0 \) and \( q_{n_\nu} = -\infty \). Note that the open interval \( (q_{n_\nu}, \infty) \) is excluded from consideration. Then one may rewrite \( F(q_\tau) = A \tau \), where \( \tau = (\tau_1, ..., \tau_k) \) and \( A \) is a \( k \times k \) lower triangular matrix of ones. Next, suppose for the moment that a hypothetical with-replacement simple random sample of \( n_\nu T \) ultimate units led to alternative estimators \( \hat{F}_\nu (q_\tau) \) of the distribution function vector and \( \tilde{q}_\tau \) of the associated quantile vectors. Under regularity conditions, as \( n_\nu T \to \infty \), \( n_\nu T \{ \hat{F}_\nu (q_\tau) - F(q_\tau) \} \xrightarrow{\nu} N_k (0, A \Sigma \tau A') \), where \( \Sigma \tau = \text{diag} (\tau - \tau') \) (e.g., Agresti, 1990, Section 12.1.5). Using the Bahadur representation of \( q_{n_\nu i} \) for IID data (e.g., Serfling, 1980, p. 92), we may show that, as \( \nu \to \infty \), \( n_\nu T \{ \tilde{q}_{n_\nu i} - q_\tau \} \xrightarrow{\nu} N_k (0, V_i) \), where \( V_i = D_f A \Sigma \tau A' D_f \) and \( D_f = \text{diag} (f(q_{n_\nu i})^{-1}, ..., f(q_{n_\nu k})^{-1}) \). Note that \( \tilde{q}_{n_\nu OLS} \) is a consistent estimator, and the design-based estimator \( V_{\nu D} \) is not used in computation of \( \hat{\theta}_{\nu, OLS} \). Thus a simple estimator of \( f(q_{n_\nu i}) = f(q_{n_\nu i} \theta) \) follows from direct substitution of \( \hat{\theta}_{\nu, OLS} \) for \( \theta \) and \( \hat{q}_{n_\nu i} \) for \( q_{n_\nu i} \), that is,

\[ \hat{d}_{n_\nu i} = \{ f(\hat{q}_{n_\nu i} | \hat{\theta}_{\nu, OLS}) \}^{-1}. \]

Consequently, an estimator of \( V_{\nu I} = V(\tilde{q}_{\nu i}) \) under the specified IID assumption is given by

\[ \hat{V}_{\nu i} = n_\nu^{-2} \hat{D}_{\nu i} A \Sigma \tau A' \hat{D}_{\nu i}, \]

where \( \hat{D}_{\nu i} = \text{diag}(\hat{d}_{n_\nu i}, ..., \hat{d}_{n_\nu k}) \). Then following, e.g., Skinner (1989, Section 2.11) and Rao and Scott (1981), we consider the first-order variance approximation of \( \hat{V}_{\nu D} \) by \( \hat{\lambda}_\nu \hat{V}_{\nu i} \), where \( \hat{\lambda}_\nu = k^{-1} \text{tr}(\hat{V}_{\nu i}^{-1} \hat{V}_{\nu D}) \). Therefore, the first-order adjusted model-based covariance matrix estimator is given by

\[ \hat{V}_{\nu i} = \hat{\lambda} \times \hat{V}_{\nu i}. \]

Note that the adjustment factor \( \hat{\lambda} \) provides a measurement of the effect of model misspecification. Thus, additional least squares fits are provided using \( \hat{V}_{\nu i} \) to form the weighting matrix, denoted by \( B_{\nu M} = \hat{V}_{\nu i} \) and \( D_{\nu i} = \text{diag}(\hat{V}_{\nu i}) \) for \( M = \text{GLS}(I) \) and \( WLS(I) \).

Our proposed methods using tail-fitting quantile estimation procedures are applied to data from Phase 2 of NHANES III in the next section.

## 6. Application to Medical Examination Survey Data

### 6.1 NHANES III Data

This discussion will focus on analysis of the natural logarithm of blood lead, log(LEAD), measured for children of all races aged 1-5 covered by the Phase
Figure 1: Comparison of Upper Tails of Normal Quantile Plots for $p = 0.75(0.01)0.99$, with Pointwise 95% Design-Based Confidence Intervals for $q_{\nu;iD \cdot \log(\text{LEAD})}$ from the Phase 2 of NHANES III Data for Children of All Races Aged 1-5.

Figure 2: Comparisons of Efficiency in Estimation of Quantiles from Upper Tails of Normal Quantile Plots for $p = 0.75(0.01)0.99 \cdot \log(\text{LEAD})$ from NHANES III Phase 2 Data for Children Aged 1-5.

Table 1: Parameter Estimates from Upper Tails of Normal Quantile Plots for $p = 0.75(0.01)0.99$ from the Phase 2 of NHANES III Data for Children of All Races Aged 1-5.

<table>
<thead>
<tr>
<th>Method</th>
<th>$\hat{\mu}$</th>
<th>$\hat{\sigma}$</th>
<th>$\hat{\text{se}}(\hat{\mu})$</th>
<th>$\hat{\text{se}}(\hat{\sigma})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>0.9405</td>
<td>0.8047</td>
<td>0.0861</td>
<td>0.0468</td>
</tr>
<tr>
<td>GLS(D)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WLS(D)</td>
<td>0.9403</td>
<td>0.8059</td>
<td>0.0858</td>
<td>0.0425</td>
</tr>
<tr>
<td>GLS(I)</td>
<td>0.9844</td>
<td>0.7737</td>
<td>0.0733</td>
<td>0.0437</td>
</tr>
<tr>
<td>WLS(I)</td>
<td>0.9545</td>
<td>0.7918</td>
<td>0.0816</td>
<td>0.0517</td>
</tr>
</tbody>
</table>

Note: The generalized least squares method for estimation of sample quantiles using the design-based covariance matrix estimate is not feasible since only 23 strata are available with the Phase 2 data.

6.2 Comparisons of Estimators and Goodness-of-Fit Statistics

Figure 1 presents a plot of the upper 25 design-based sample quantiles for $p = 0.75(0.01)0.99$ against the associated standard normal quantiles accompanied by connected pointwise design-based 95% confidence intervals and the ordinary least squares fitted line. Lines fit by the $WLS(D)$, $WLS(I)$ and $GLS(I)$ methods were very similar to the OLS line and are omitted here to avoid graphical clutter. The close fit of the OLS line indicates that data are consistent with the baseline normality assumption on the superpopulation distribution. Related estimation results for the superpopulation parameters are given in Table 1. The weighted least squares results appear to yield more efficient estimators compared to the OLS method. Figure 2 displays standard errors of design-based sample quantiles and all associated least squares quantile estimates at the corresponding probability values. The goodness-of-fit test statistics reported in Table 2 again suggest that the upper-tail data are consistent with a normal superpopulation model.
<table>
<thead>
<tr>
<th>Method</th>
<th>$\hat{\lambda}_M$</th>
<th>$\hat{a}_M$</th>
<th>$\hat{c}_M$</th>
<th>$\hat{d}_M$</th>
<th>Naive</th>
<th>1st Adj.</th>
<th>2nd Adj.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDO</td>
<td>1.08696</td>
<td>3.95428</td>
<td>18.08296</td>
<td>1.38252</td>
<td>20.23362</td>
<td>18.61493</td>
<td>1.11893</td>
</tr>
<tr>
<td></td>
<td>(0.62778)</td>
<td>(0.75297)</td>
<td>(0.40475)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DDW</td>
<td>1.08696</td>
<td>4.02931</td>
<td>18.73408</td>
<td>1.33447</td>
<td>22.25095</td>
<td>20.47088</td>
<td>1.18773</td>
</tr>
<tr>
<td></td>
<td>(0.50515)</td>
<td>(0.61336)</td>
<td>(0.37354)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IDO</td>
<td>0.74983</td>
<td>4.33829</td>
<td>14.86213</td>
<td>1.16040</td>
<td>11.82175</td>
<td>15.76597</td>
<td>0.79543</td>
</tr>
<tr>
<td></td>
<td>(0.97313)</td>
<td>(0.86510)</td>
<td>(0.42827)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IDW</td>
<td>1.06111</td>
<td>5.20148</td>
<td>29.76988</td>
<td>0.81981</td>
<td>16.89322</td>
<td>15.92032</td>
<td>0.56746</td>
</tr>
<tr>
<td></td>
<td>(0.81446)</td>
<td>(0.85866)</td>
<td>(0.37832)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IIG</td>
<td>0.72633</td>
<td>4.28929</td>
<td>14.08933</td>
<td>1.18569</td>
<td>21.11306</td>
<td>29.06822</td>
<td>1.49851</td>
</tr>
<tr>
<td></td>
<td>(0.57418)</td>
<td>(0.17806)</td>
<td>(0.26819)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IIW</td>
<td>0.75590</td>
<td>3.23294</td>
<td>8.65652</td>
<td>2.00841</td>
<td>13.91843</td>
<td>18.41296</td>
<td>1.60786</td>
</tr>
<tr>
<td></td>
<td>(0.92918)</td>
<td>(0.73472)</td>
<td>(0.44954)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTES: Each goodness-of-fit test statistic is the sum of standardized residuals, as defined in Section 4.1. The first two letters in the method column represent the variance estimation methods for the numerator and denominator, respectively. The third letter represents the least squares method used for quantile estimation. The labels D and I represent the design-based and model/IID-based methods, respectively. The letters O, G and W represent OLS, GLS and WLS, respectively. The superscripts a and b indicate that the p-value calculation is based on 23 and $d_M$ degrees of freedom, respectively.

References


