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).

The authors thank Drs. D. Brody, A. Looker and V. L. Parsons for providing the NHANES III data discussed here, and for helpful comments on related statistical and substantive issues. This research was supported in part by the U.S. National Center for Health Statistics. The views expressed here are those of the authors and do not necessarily reflect the policies of the U.S. National Center for Health Statistics.

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 νth finite population with N_ν ultimate units is a simple random sample selected with replacement from an infinite superpopulation. Also, assume that population ν has been partitioned into L_ν strata with $N_{\nu h}$ primary sampling units (PSUs). From the hth 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 $p_{\nu hij}$, $j = 1, \ldots, N_{\nu hi}$. 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 νth 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 νth 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 $x_{\nu hij} = \{1, \delta_{\nu hij}(y_1), \ldots, \delta_{\nu hij}(y_k)\}'$ and $w_{\nu hij}$ denote, respectively, the observed vector and the associated survey weight for each sampled unit (h, i, j) in the νth sample, where $\delta_{\nu hij}(y_l) = 1$ if $Y_{\nu hij} \leq y_l$ 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 $\hat{X}_\nu = (\hat{X}_{\nu 1}, \ldots, \hat{X}_{\nu,k+1})' = \sum_{h=1}^{L_\nu} \sum_{i=1}^{n_{\nu h}} \sum_{j=1}^{n_{\nu hi}} w_{\nu hij} x_{\nu hij}$. 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 π_l, the

superpopulation quantile is given by

$$q_{\pi_l} = \inf\{y : F(y) \geq \pi_l\} = F^{-1}(\pi_l).$$

The corresponding sample quantile estimator is defined by

$$\hat{q}_{\pi_l:\nu D} = \inf\{y : \hat{F}_\nu(y) \geq \pi_l\} = \hat{F}_\nu^{-1}(\pi_l)$$

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

$$\hat{q}_{\pi_l:\nu D} = q_{\pi_l} - \{f(q_{\pi_l})\}^{-1}\{\hat{F}_\nu(q_{\pi_l}) - F(q_{\pi_l})\} + o_p(n_\nu^{-1/2}), \quad (1)$$

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

$$n_\nu^{1/2}\{\hat{q}_{\pi:\nu D} - q_\pi\} \xrightarrow{\mathcal{L}} N_k(0, V_Y), \quad (2)$$

where $V_Y = lim_{\nu\to\infty} n_\nu V\{\hat{q}_{\pi:\nu D}\}$ and V_Y is positive definite. We may also show that, as $\nu \to \infty$,

$$n_\nu \left\{\hat{V}(\hat{q}_{\pi:\nu D}) - V(\hat{q}_{\pi:\nu D})\right\} \xrightarrow{p} 0, \quad (3)$$

where

$$\hat{V}(\hat{q}_{\pi:\nu D}) = \hat{D}_\nu \hat{V}\{\hat{F}_\nu(\hat{q}_{\pi:\nu D})\}\hat{D}_\nu. \quad (4)$$

In addition, $\hat{V}\{\hat{F}_\nu(\hat{q}_{\pi:\nu D})\}$ is the estimated covariance matrix of $\hat{F}_\nu(y)$ evaluated at $\hat{q}_{\pi:\nu D}$ such that, as $\nu \to \infty$, $n_\nu \left\{\hat{V}[\hat{F}_\nu(\hat{q}_{\pi:\nu D})] - V[\hat{F}_\nu(q_{\pi:\nu D})]\right\} \xrightarrow{p} 0$. Also, $\hat{D}_\nu$ is the diagonal matrix whose lth diagonal element, given by (Francisco and Fuller, 1991, Theorem 4),

$$\begin{aligned}\hat{d}_{\nu l} &= \left(2z_{1-\frac{\alpha}{2}}\{\hat{V}[\hat{F}_\nu(q_{\pi_l})]\}^{1/2}\right)^{-1} \\ &\times\left\{\hat{F}_\nu^{-1}\left(\pi_l + z_{1-\frac{\alpha}{2}}\{\hat{V}[\hat{F}_\nu(q_{\pi_l})]\}^{1/2}\right)\right. \\ &\left.-\hat{F}_\nu^{-1}\left(\pi_l - z_{1-\frac{\alpha}{2}}\{\hat{V}[\hat{F}_\nu(q_{\pi_l})]\}^{1/2}\right)\right\},\end{aligned}$$

is a design-based estimator of $d_l = \{f(q_{\pi_l})\}^{-1}$ and $z_{1-\frac{\alpha}{2}} = \Phi^{-1}(1-\frac{\alpha}{2})$ such that as $\nu \to \infty$,

$$\hat{d}_{\nu l}^{-1} \xrightarrow{p} f(y_{\pi_l}). \quad (5)$$

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

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 $F(y) = \Phi\left(\frac{y-\mu}{\sigma}\right)$ 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 pth quantile is uniquely expressed as

$$q_p = \mu + \sigma\Phi^{-1}(p). \quad (6)$$

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 $\hat{q}_{\pi_l:\nu D}$ against v_l, where $v_l = \Phi^{-1}(\pi_l)$. Under the normality assumption for the superpopulation model of Y, expression (6) becomes the pth superpopulation quantile. A straightforward application of the least squares method to expression (6) leads to ordinary least squares (OLS) estimators of μ and σ. In general, however, sample quantiles are correlated so that $\hat{V}(\hat{q}_{\pi:\nu D})$ 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, $\hat{D}(\hat{q}_{\pi:\nu D}) = \text{diag}[\hat{V}(\hat{q}_{\pi:\nu D})]$ provides weighted least squares (WLS) estimators. Note that only the weighted least squares estimators would be available if $\hat{V}(\hat{q}_{\pi:\nu D})$ is singular. Let $\hat{\mu}_{\nu M}$ and $\hat{\sigma}_{\nu M}$ 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

$$\hat{q}_{\pi^*:\nu M} = \hat{\mu}_{\nu M} + \hat{\sigma}_{\nu M} z^*,$$

where $z^* = \Phi^{-1}(\pi^*)$. Let $z_\pi = (z_{\pi_1}, \ldots, z_{\pi_k})'$ and let $Z_\pi = [1_k, z_\pi]$, where 1_k is the k-dimensional column vector of ones. Let $\theta = (\mu, \sigma)'$ and $\hat{\theta}_{\nu M} = (\hat{\mu}_{\nu M}, \hat{\sigma}_{\nu M})'$ denote the parameter vector and its estimator associated with the least squares method M. Thus we have the following vector of estimated quantiles for π in the matrix form;

$$\hat{q}_{\pi:\nu M} = Z_\pi \hat{\theta}_{\nu M},$$

where

$$\hat{\theta}_{\nu M} = (Z_\pi' B_{\nu M}^{-1} Z_\pi)^{-1} Z_\pi' B_{\nu M}^{-1} \hat{q}_{\pi:\nu D}$$

and $B_{\nu M} = I$, $\hat{V}(\hat{q}_{\pi:\nu D})$ and $\hat{D}(\hat{q}_{\pi:\nu D})$ 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 $e_{\pi_l:\nu M} = \hat{q}_{\pi_l:\nu D} - \hat{q}_{\pi_l:\nu M}$ be the lth residual of the chosen method M. Then the standardized residual, defined by,

$$e^*_{\pi_l:\nu M} = \frac{e_{\pi_l:\nu M}}{\widehat{\text{se}}(e_{\pi_l:\nu M})},$$

may provide a (standardized) deviation attributed to each fitted quantile, where $\widehat{\text{se}}(e_{\pi_l:\nu M})$ is an estimated standard error of $e_{\pi_l:\nu M}$. Hence, a simple test of the goodness-of-fit of the proposed superpopulation model follows from the statistic $X^2_{GOF}(\hat{q}_{\pi:\nu M}) = \sum_{l=1}^{k} e^{*2}_{\pi_l:\nu M}$. The $k \times 1$ residual vector may be written as $e_{\pi:\nu M} = (e_{\pi_1:\nu M}, \ldots, e_{\pi_k:\nu M})' = R_{\nu M} \hat{q}_{\pi:\nu D}$ for each method M, where

$$R_{\nu M} = I_k - Z_\pi \{Z_\pi' B_{\nu M}{}^{-1} Z_\pi\}^{-1} Z_\pi' B_{\nu M}{}^{-1}.$$

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

$$e_{\pi:\nu M} = R_{\nu M}(\hat{q}_{\pi:\nu D} - q_\pi).$$

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

$$R_{\nu M} \xrightarrow{p} R_M$$

where $R_M = I_k - Z_\pi \{Z_\pi' B_M{}^{-1} Z_\pi\}^{-1} Z_\pi' B_M{}^{-1}$ for $M = OLS, GLS(D)$ and $WLS(D)$, respectively. Thus it follows that as $\nu \to \infty$,

$$n_\nu^{1/2} e_{\pi:\nu M} \xrightarrow{\mathcal{L}} N_k(0, V_M),$$

where $V_M = plim_{\nu\to\infty} n_\nu \hat{V}(e_{\pi:\nu M}) = R_M V_Y R_M'$ and $\hat{V}(e_{\pi:\nu M}) = R_{\nu M} \hat{V}(\hat{q}_{\pi:\nu D}) R_{\nu M}'$ is a consistent estimator of $V(e_{\pi:\nu M})$.

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

$$\begin{aligned} &X^2_{GOF}(\hat{q}_{\pi:\nu M}) \\ &= e_{\pi:\nu M}{}' [\hat{D}(e_{\pi:\nu M})]^{-1} e_{\pi:\nu M} \\ &= [n_\nu^{1/2}(\hat{q}_{\pi:\nu D} - q_\pi)]' (n_\nu^{-1} A_{\nu M}) [n_\nu^{1/2}(\hat{q}_{\pi:\nu D} - q_\pi)], \end{aligned} \tag{7}$$

where $\hat{D}(e_{\pi:\nu M}) = \text{diag}[\hat{V}(e_{\pi:\nu M})]$ and

$$\begin{aligned} A_{\nu M} &= R_{\nu M}' \left\{\text{diag}[\hat{V}(e_{\pi:\nu M})]\right\}^{-1} R_{\nu M} \\ &= R_{\nu M}' \left\{\text{diag}[R_{\nu M} \hat{V}(\hat{q}_{\pi:\nu D}) R_{\nu M}']\right\}^{-1} R_{\nu M}. \end{aligned}$$

This alternative representation is of a quadratic form in $n_\nu^{1/2}(\hat{q}_{\pi:\nu D} - q_\pi)$. Note that $n_\nu^{-1} A_{\nu M} \xrightarrow{p} A_M$ as $\nu \to \infty$, where $A_M = R_M'[\text{diag}(V_M)]^{-1} R_M$. Thus it can be shown that as $\nu \to \infty$,

$$X^2_{GOF}(\hat{q}_{\pi:\nu M}) \xrightarrow{\mathcal{L}} \sum_{l=1}^{k} \lambda_{l:M} X_l^2,$$

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

4.2 Satterthwaite Approximation

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

$$X^2_{RS1}(\hat{\bar{\lambda}}_{\nu M}) = X^2_{GOF}(\hat{q}_{\pi:\nu M}) / \hat{\bar{\lambda}}_{\nu M},$$

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

$1, \ldots, k - r_{\nu M}$. Then a second-order correction is given by

$$X^2_{RS2}(\hat{\bar{\lambda}}_{\nu M}, \hat{a}_{\nu M}) = X^2_{GOF}(\hat{q}_{\pi:\nu M})/\hat{c}_{\nu M},$$

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

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 $\hat{V}\{\hat{F}_\nu(\hat{q}_{\pi:\nu D})\}$, which has at most $n_\nu - L_\nu$ degrees, as noted in Section 2. See, e.g., Korn and Graubard (1990). From expression (4), $\hat{V}_{\nu D}$ 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 $\hat{V}_{\nu D}$ 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 $q_{\pi_l}, l = 1, 2, \ldots, k$, determine k non-overlapping cells of the form $I_l = (q_{\pi_{l-1}}, q_{\pi_l}]$ and the corresponding cell probabilities are given by $\tau_l = \Pr(Y \in I_l) = \pi_l - \pi_{l-1}$, where $\pi_0 = 0$ and $q_{\pi_0} = -\infty$. Note that the open interval (q_{π_k}, ∞) is excluded from consideration. Then one may rewrite $F(q_\pi) = A\tau$, where $\tau = (\tau_1, \ldots, \tau_k)'$ 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 $\hat{F}_{\nu I}(q_\pi)$ of the distribution function vector and $\hat{q}_\pi$ of the associated quantile vectors. Under regularity conditions, as $n_{\nu T} \to \infty$, $n^{1/2}_{\nu T}\{\hat{F}_{\nu I}(q_\pi) - F(q_\pi)\} \xrightarrow{\mathcal{L}} N_k(0, A\Sigma_\tau A')$, where $\Sigma_\tau = \mathrm{diag}(\tau) - \tau\tau'$ (e.g., Agresti, 1990, Section 12.1.5). Using the Bahadur representation of $\hat{q}_{\pi_l:\nu I}$ for IID data (e.g., Serfling 1980, p. 92), we may show that, as $\nu \to \infty$, $n^{1/2}_{\nu T}\{\hat{q}_{\pi:\nu I} - q_\pi\} \xrightarrow{\mathcal{L}} N_k(0, V_I)$, where $V_I = D_f A\Sigma_\tau A' D_f$ and $D_f = \mathrm{diag}\{f(q_{\pi_1})^{-1}, \ldots, f(q_{\pi_k})^{-1}\}$. Note that $\hat{\theta}_{\nu,OLS}$ is a consistent estimator, and the design-based estimator $\hat{V}_{\nu D}$ is not used in computation of $\hat{\theta}_{\nu,OLS}$. Thus a simple estimator of $f(q_{\pi_l}) = f(q_{\pi_l}|\theta)$ follows from direct substitution of $\hat{\theta}_{\nu,OLS}$ for θ and $\hat{q}_{\pi_l:\nu I}$ for q_{π_l}, that is,

$$\hat{d}_{l:\nu I} = \{f(\hat{q}_{\pi_l:\nu I}|\hat{\theta}_{\nu,OLS})\}^{-1}.$$

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

$$\hat{V}^*_{\nu I} = n^{-1}_{\nu T}\hat{D}_{\nu I} A\Sigma_\tau A' \hat{D}_{\nu I},$$

where $\hat{D}_{\nu I} = \mathrm{diag}(\hat{d}_{1:\nu I}, \ldots, \hat{d}_{k:\nu I})$. 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 $\bar{\lambda}_\nu \hat{V}^*_{\nu I}$, where $\hat{\bar{\lambda}}_\nu = k^{-1}\mathrm{tr}(\hat{V}^{*-1}_{\nu I}\hat{V}_{\nu D})$. Therefore, the first-order adjusted model-based covariance matrix estimator is given by

$$\hat{V}_{\nu I} = \hat{\bar{\lambda}} \times \hat{V}^*_{\nu I}.$$

Note that the adjustment factor $\hat{\bar{\lambda}}$ provides a meaurement 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 $\hat{D}_{\nu I} = \mathrm{diag}(\hat{V}_{\nu I})$ for $M = GLS(I)$ 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: Comparsion of Upper Tails of Normal Quantile Plots for $p = 0.75(0.01)0.99$, with Pointwise 95% Design-Based Confidence Intervals for $\hat{q}_{\pi:\nu D}$. log(LEAD) from the Phase 2 of NHANES III Data for Children of All Races Aged 1-5

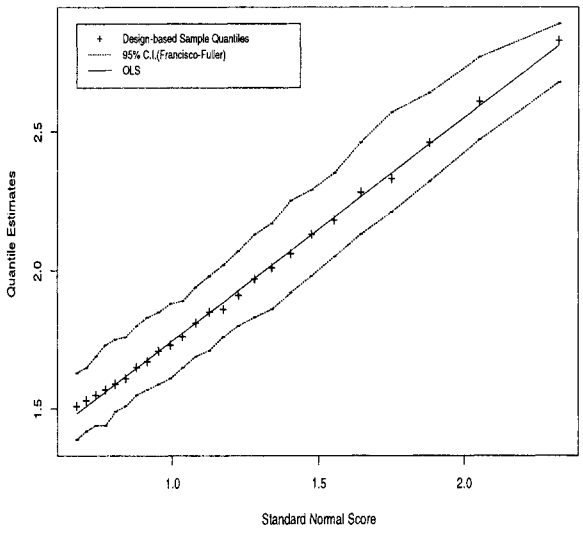
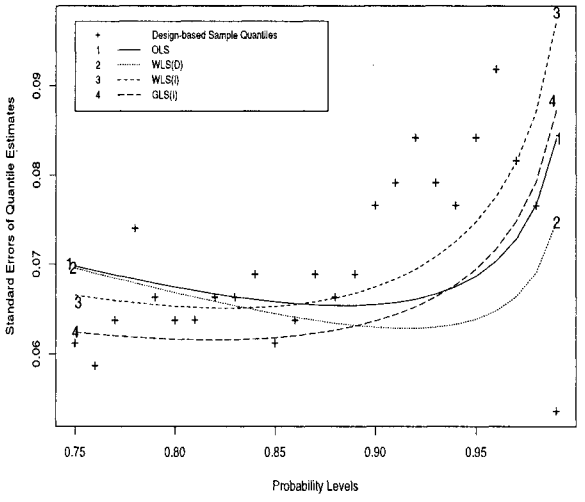


Figure 2: Comparsions of Efficiency in Estimation of Quantiles from Upper Tails of Normal Quantile Plots for $p = 0.75(0.01)0.99$. log(LEAD) from NHANES III Phase 2 Data for Children Aged 1-5



2 (1991-1994) of the U.S. Third National Health and Nutrition Examination Survey (NHANES III). Analyses generally treat the Phase 2 data as arising from a stratified multistage design involving selection of $n_{\nu h} = 2$ PSUs (usually counties) selected with unequal probabilities and with replacement from 23 strata. Additional subsampling was carried out to select area segments (e.g., parts of city or suburban blocks) within a selected PSU, households (or certain types of group quarters) within a selected segment, and persons within a selected household. The data for children aged 1-5 were then collected through a medical examination from approximatedly 2400 participants.

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 clustter. 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 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.

Method	$\hat{\mu}$	$\hat{\sigma}$	$\widehat{\text{se}}(\hat{\mu})$	$\widehat{\text{se}}(\hat{\sigma})$
OLS	0.9405	0.8047	0.0861	0.0468
GLS(D)	-	-	-	-
WLS(D)	0.9403	0.8059	0.0858	0.0425
GLS(I)	0.9844	0.7737	0.0733	0.0437
WLS(I)	0.9545	0.7918	0.0816	0.0517

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.

Table 2: Comparisons of Goodness-of-fit Statistics and the Associated Rao-Scott Type Adjustments for NHANES III (1991-1994) Data on log(LEAD) for Children Aged 1-5.

					Goodness-of-fit Statistics		
Method	$\hat{\bar{\lambda}}_{\nu M}$	$\hat{a}_{\nu M}$	$\hat{c}_{\nu M}$	$\hat{d}_{\nu M}$	Naive	1st Adj.	2nd Adj.
DDO	1.08696	3.95428	18.08296	1.38252	20.23362	18.61493	1.11893
					$(0.62778)^a$	$(0.75297)^a$	$(0.40475)^b$
DDW	1.08696	4.02931	18.73408	1.33447	22.25095	20.47088	1.18773
					(0.50515)	(0.61336)	(0.37354)
IDO	0.74983	4.33829	14.86213	1.16040	11.82175	15.76597	0.79543
					(0.97313)	(0.86510)	(0.42827)
IDW	1.06111	5.20148	29.76988	0.81981	16.89322	15.92032	0.56746
					(0.81446)	(0.85866)	(0.37832)
IIG	0.72633	4.28929	14.08933	1.18569	21.11306	29.06822	1.49851
					(0.57418)	(0.17806)	(0.26819)
IIW	0.75590	3.23294	8.65652	2.00841	13.91843	18.41296	1.60786
					(0.92918)	(0.73472)	(0.44954)

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

Agresti, A. (1990). *Categorical Data Analysis.* New York ; Wiley.

Barnett, V. (1975). Probability Plotting Methods and Order Statistics. *Applied Statistics* **27**, 203–254.

Chernoff, H. and Lieberman, G. J. (1954). Use of Normal Probability Paper. *Journal of the American Statistical Association* **49**, 778–785.

Cran, G. W. (1975). A Note on Chernoff and Lieberman's Generalized Probability Paper. *Journal of the American Statistical Association* **70**, 229–232.

Graybill, F. A. (1976). *Theory and Application of the Linear Model.* Pacific Grove, Calif. ; Wadsworth & Brooks/Cole.

Francisco, C. A. and Fuller, W. A. (1991). Quantile Estimation with a Complex Survey Design. *The Annals of Statistics* **19**, 454–469.

Korn, E. L. and Graubard, B. I. (1990). Simultaneous Testing of Regression Coefficients with Complex Survey Data : Use of Bonferroni t Statistics. *The American Statistician* **44**, 270–275.

Rao, J. N. K. and Scott, A. J. (1981). The Analysis of Categorical Data from Complex Sample Surveys: Chi-squared Tests for Goodness of Fit and Independence in Two-Way Tables. *Journal of the American Statistical Association* **76**, 221–230.

Rao, J. N. K. and Scott, A. J. (1984). On Chi-squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data. *The Annals of Statistics* **12**, 46–60.

Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin* **2**, 110–114.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* New York ; Wiley.

Shao, J. (1996). Resampling Methods in Sample Surveys (with discussion). *Statistics* **27**, 203–254.

Skinner, C. J. (1989). Introduction to Part A. In C. J. Skinner, D. Holt, and T. M. F. Smith (eds.), *Analysis of Complex Surveys*, pp. 23–58. New York: Wiley.

Woodruff, R. S. (1952). Confidence Intervals for Median and Other Position Measures. *Journal of the American Statistical Association* **47**, 635–646.