Survey Methodology, December 1995
Vol. 21, No. 2, pp. 137–145
Statistics Canada

Estimating Some Measures of Income Inequality from Survey Data: An Application of the Estimating Equations Approach

DAVID A. BINDER and MILORAD S. KOVACEVIC[1]

ABSTRACT

We summarize some salient aspects of the theory of estimation functions for finite populations. In particular, we discuss the problem of estimation of means and totals and extend this theory to estimating functions. We then apply this estimating functions framework to the problem of estimating measures of income inequality. The resulting statistics are nonlinear functions of the observations. Some of them depend on the order of observations or quantiles. Consequently, the mean squared errors of these estimates are inexpressible by simple formulae and cannot be estimated by conventional variance estimation methods. We show that within the estimating function framework this problem can be resolved using the Taylor linearization method. Finally, we illustrate the proposed methodology using income data from Canadian Survey of Consumer Finance and comparing it to the 'delete-one-cluster' jackknifing method.

KEY WORDS: Complex survey design; Gini family coefficient; Lorenz curve ordinate; Low income measure; Quantile share.

1. INTRODUCTION

The measurement and analysis of economic inequality are well covered in econometrics literature from both, theoretical and applied aspects, although the theoretical issues prevail. Estimation of inequality measures and the impact of the design of sample surveys have gotten less attention. Variance estimation, unavoidable in statistical inference based on these measures, is seldom an issue in the relevant econometric literature. It is usually addressed under very strong assumptions and under unsustainable simplifications of the design or the formulae for the approximate variance. In this paper we present a method that can handle with ease both the estimation of the measures of income inequality and the variance estimation of the resulting non-linear statistics. This method is applicable under a variety of sampling designs.

In general, a population distribution can be described by its cumulative distribution function, $F(y) = \Pr\{Y \leq y\}$, where Y is the random variable corresponding to selecting one population unit at random. Throughout this paper, we assume that Y is non-negative. If Y represents income then we are interested in the properties of an income distribution, such as income concentration, income shares for different population shares, low income proportions, *etc.* We are also interested in the quantile function $\xi(p) = F^{-1}(p) = \inf\{y \mid F(y) \geq p\}$.

The Lorenz curve, for example, depicts the cumulative income against the population share. The formal definition of the ordinate of the Lorenz curve corresponding to the 100 p-th percentile of the population is

$$L(p) = \frac{\int_0^{\xi_p} y dF(y)}{\mu_Y}, \tag{1.1}$$

where

$$\int_0^{\xi_p} dF(y) = p, \quad \text{and} \quad \int_0^{\infty} y dF(y) = \mu_Y.$$

The finite population form of the expression (1.1), more familiar to survey statisticians, is given by

$$L(p) = \sum_U Y_i I\{Y_i \leq \xi_p\} \Big/ \sum_U Y_i,$$

where U represents a finite population and $I\{.\}$ is an indicator function.

The income (quantile) share is defined as the percentage of total income shared by the population allocated to the certain income quantile interval $[\xi_{p_1}, \xi_{p_2}]$, $p_1 \leq p_2$. It is equal to the difference of Lorenz curve ordinates

$$Q(p_1,p_2) = L(p_2) - L(p_1).$$

In Figure 1 we give a graph of the Lorenz curve for the Weibull distribution with shape parameter $\alpha = 1.6$, along with the 45° axis. For example, one can read from the graph that not more than 25% of the total income is allocated to the poor half of population, or that the richest 10% of the population earn 20% of the total available income.

[1] David A. Binder, Director, Business Survey Methods Division, and Milorad S. Kovacevic, Senior Methodologist, Household Survey Methods Division, Statistics Canada, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

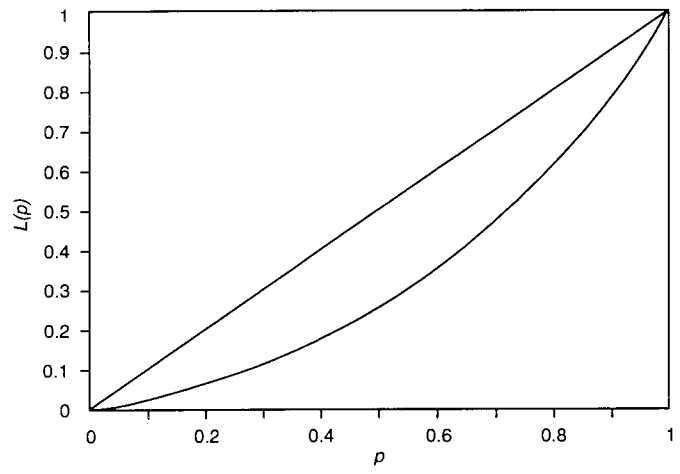

Figure 1. Lorenz Curve for the Weibull Distribution with Shape Parameter $\alpha = 1.6$.

The Gini coefficient measures the degree of the inequality in income distribution. One definition of the Gini coefficient is a linear function of the area between the Lorenz curve and the 45° axis, normalized to lie between 0 and 1. The Gini coefficient in Figure 1 is 0.35. The formal definition of the Gini coefficient (Nygård and Sandström 1981) is

$$G = 1 - 2\int_0^1 L(p)dp = \frac{1}{\mu}\int_0^\infty [2F(y) - 1]ydF(y).$$

A more general family of Gini coefficients, given in Nygård and Sandström (1981) is

$$G_J = \frac{1}{\mu_Y}\int_0^\infty J[F(y)]ydF(y), \tag{1.2}$$

where J is a bounded and continuous function. For the usual Gini coefficient, $J(p) = 2p - 1$.

Another measure of income inequality used by some economists is the Low Income Measure. This is defined as the proportion of the population units whose income is less than half the median income for the population. Formally, this is

$$\Theta = \int_0^{M/2} dF(y), \tag{1.3a}$$

where M is the median defined by

$$\int_0^M dF(y) = \frac{1}{2}. \tag{1.3b}$$

For all these measures, we can express the parameter of interest, Θ, as the solution to the equation

$$\int u(y,\Theta)dF(y) = 0,$$

where $u(y,\Theta)$ is the kernel of the estimating equation. This estimating equation formulation will be discussed in Section 2. In Sections 3, 4, and 5 we give the estimating equations for the above measures along with the approximation of their mean squared error estimates. In Section 6 we present estimators of these measures based on the complex sample design. Section 7 contains an illustration based on the Canadian Survey of Consumer Finance data.

2. USE OF ESTIMATING EQUATIONS FOR FINITE POPULATIONS

The theory for estimating means and totals from finite populations is now well established in the statistical literature. A formulation which encompasses most estimators used in practice is given in Särndal, Swensson, and Wretman (1992). In this section, we briefly review this theory and show how it can be applied to more complex statistics through the use of estimating equations, as described by Binder (1991) and Binder and Patak (1994).

We begin the exposition of the main idea by reviewing the estimation of the population total T_Y and the finite population distribution function $F(y)$. The estimation of the population total is the core of the estimation equations approach of Binder (1991) and Binder and Patak (1994). Let the population total of the variable Y, be defined as

$$T_Y = N\int ydF(y).$$

Note here that $F(y)$ is a step function corresponding to the distribution function for the finite population. We consider estimators of the form:

$$\hat{T}_Y = \sum_{i\in s} w_i(s)y_i = \sum_{i=1}^{N} w_i(s)Y_i, \tag{2.1}$$

where $w_i(s)$ is zero whenever the i-th unit is not in the sample. Expression (2.1) gives, for example, the Horvitz-Thompson (HT) unbiased estimator if

$$w_i(s) = \begin{cases} 1/\pi_i, & i\in s, \\ 0, & i\notin s, \end{cases}$$

or the generalized regression estimator if

$$w_i(s) = \begin{cases} [1 + (T_X - \hat{T}_X)x_i/\hat{T}_{X2}]/\pi_i, & i\in s, \\ 0, & i\notin s, \end{cases}$$

where T_X is the population total of X, and $\hat{T}_X$ and $\hat{T}_{X^2}$ are the HT estimates of the totals of X and X^2 variables, respectively.

Similarly, an estimator for the distribution function is given by

$$N\hat{F}(y) = \sum_{i\in s} w_i(s) I\{y_i \le y\},$$

where

$$I\{y_i \le y\} = \begin{cases} 1 & \text{if } y_i \le y, \\ 0 & \text{if } y_i > y. \end{cases}$$

We note that $\hat{F}(y)$ is uniformly and asymptotically design consistent for $F(y)$, but it is not necessarily a true distribution function, unless

$$\sum_{i\in s} w_i(s) = N.$$

In general, and under certain regularity conditions for complex designs (Francisco and Fuller 1991),

$$\hat{F}(y) - F(y) \to_p 0, \text{ for any } y.$$

That is, the finite population distribution function, $F(y)$, allows a consistent estimator, $\hat{F}(y)$. This property of the $\hat{F}(y)$ will be used later in proving the consistency of the linearized variance estimators for different income statistics.

Now, we review the application of the estimating equations theory to the estimation of any finite population parameter Θ_o that can be expressed as the solution to

$$\int u(y,\Theta_o) dF(y) = 0.$$

We define the estimating equation estimate for Θ_o as that value of $\hat{\Theta}$ for which

$$\int \hat{u}(y,\hat{\Theta}) d\hat{F}(y) = 0, \tag{2.2}$$

where $\hat{u}(y,\Theta)$ is an estimate of $u(y,\Theta)$.

We can rewrite (2.2) as

$$\begin{aligned} 0 &= \int \hat{u}(y,\hat{\Theta}) d\hat{F}(y) \\ &= \int [\hat{u}(y,\hat{\Theta}) - u(y,\Theta_o)] dF(y) + \int u(y,\Theta_o) d\hat{F}(y) + R, \end{aligned} \tag{2.3}$$

where

$$R = \int [\hat{u}(y,\hat{\Theta}) - u(y,\Theta_o)] [d\hat{F}(y) - dF(y)].$$

The decomposition in (2.3) is the basic starting point for all the derivations of variance in the paper. For each parameter considered we will prove that the remainder term, R, is asymptotically negligible.

Binder (1983) considered the case where $\hat{u}(y,\Theta) = u(y,\Theta)$ and where, for large samples,

$$\begin{aligned} &\int [u(y,\hat{\Theta}) - u(y,\Theta_o)] dF(y) \\ &= (\hat{\Theta} - \Theta_o) \left. \frac{\partial \mathrm{E}\{u(y,\Theta)\}}{\partial \Theta} \right|_{\Theta=\Theta_o} + o_p(|\hat{\Theta} - \Theta_o|). \end{aligned}$$

Note that the remainder term R from the decomposition (2.3) should be of order $o_p(|\hat{\Theta} - \Theta_o|)$ to be considered as asymptotically negligible.

For most applications $u(y,\Theta)$ does not need to be estimated by $\hat{u}(y,\Theta)$. However, for some applications such as the Gini coefficient, the function $u(y,\Theta)$ is estimated so that formula (2.2) allows for these cases in general.

Using these approximations, we have

$$\begin{aligned} \hat{\Theta} - \Theta_o \approx &- \left[\left. \frac{\partial \mathrm{E}\{u(y,\Theta)\}}{\partial \Theta} \right|_{\Theta=\Theta_o} \right]^{-1} \\ &\times \int u(y,\Theta_o) d\hat{F}(y) = \int u^*(y) d\hat{F}(y), \end{aligned} \tag{2.4}$$

where

$$u^*(y) = - \left[\left. \frac{\partial \mathrm{E}\{u(y,\Theta)\}}{\partial \Theta} \right|_{\Theta=\Theta_o} \right]^{-1} u(y,\Theta_o).$$

Once we have obtained the expression for $u^*(y)$, the derivation of the variance of $\hat{\Theta}$ becomes straightforward. Since we have approximated $\hat{\Theta} - \Theta_o$ as an estimator of a population total of $u^*(y_i)$'s, we can use the mean squared error calculations for the estimate of total to obtain the variance estimate of $\hat{\Theta}$.

For example, for Θ_o equal to the ratio, T_Y/T_X, we have

$$u = y - \Theta_0 x,$$

$$u^* = \frac{1}{\mu_X} (y - \Theta_0 x).$$

The remainder term in this case is

$$R = \int [y - \hat{\Theta}x - (y - \Theta_0 x)][d\hat{F}(y) - dF(y)].$$

Therefore,

$$-\frac{R}{\hat{\Theta} - \Theta_o} = [\hat{F}(y) - F(y)]x \to_p 0,$$

for any y and any finite x.

Similarly, for population quantiles, we have

$$\begin{aligned} u &= I\{y \le \Theta_0\} - p, \\ u^* &= -\frac{1}{f(\Theta_0)}[I\{y \le \Theta_0\} - p], \end{aligned} \tag{2.5}$$

where $f(\Theta_0)$ is the value of the density function at Θ_0. The second expression in (2.5) is an extension of the Bahadur representation for sample quantiles, as described by Francisco and Fuller (1991). Result (2.5) will be used for the ordinates of the Lorenz curve and for the Low Income Measure, which are discussed in Sections 4 and 5.

The remainder term R in this case reduces to $R = \hat{F}(\hat{\Theta}) - \hat{F}(\Theta_o) - F(\hat{\Theta}) + F(\Theta_o)$. In the case of the simple random sample design, Randles (1982) showed that $R = o_p(n^{-1/2})$. For the complex design situation, under some regularity conditions, Shao and Rao (1994) established a similar asymptotic result: first they showed that $\hat{\Theta} - \Theta_o = O_p(n^{-1/2})$, then that $R = o_p(n^{-1/2})$, and therefore $R = o_p(|\hat{\Theta} - \Theta_o|)$.

3. GINI FAMILY COEFFICIENT

For the Gini family coefficient, given by (1.2), we can use

$$u(y,G_J) = J[F(y)]y - G_J y.$$

Binder's (1983) approach cannot handle the variance estimation of the Gini coefficient. For the Gini coefficient, rather than deriving the variances by breaking the problem into two parts - one for the ratio estimator and the other for the variance of the numerator - we use the estimating equations approach to solve the problem in one step.

Ignoring the remainder term in (2.3), we have the following approximation:

$$\begin{aligned} 0 &= \int \{J[\hat{F}(y)]y - \hat{G}_J y\} d\hat{F}(y) \\ &\approx \int \{J[\hat{F}(y)] - J[F(y)]\} y dF(y) \\ &\quad - (\hat{G}_J - G_J)\int y dF(y) + \int \{J[F(y)]y - G_J y\} d\hat{F}(y). \end{aligned}$$

Letting

$$\int \{J[\hat{F}(y)] - J[F(y)]\} y dF(y) \approx \int [\hat{F}(y) - F(y)] J'[F(y)] y dF(y),$$

and

$$\begin{aligned} \int \hat{F}(y) J'[F(y)] y dF(y) &= \int\int_0^y J'[F(y)] y d\hat{F}(x) dF(y) \\ &= \int \left[\int_y^\infty J'[F(x)] x dF(x)\right] d\hat{F}(y), \end{aligned}$$

we have that

$$\hat{G}_J - G_J \approx \int u^*(y) d\hat{F}(y),$$

where

$$u^* = \frac{1}{\mu_Y}\left[\int_{F(y)}^1 J'(p)F^{-1}(p)dp + J[F(y)]y - G_J y - E\{F(y)J'[F(y)]y\}\right]. \tag{3.1}$$

For the case of independent and identically distributed observations, this yields the same variance result as described by Glasser (1962) and Sendler (1979). To estimate the variance, it is necessary to use estimates for μ_Y, $F(y)$, and G_J in the expression for u^*.

We investigate the asymptotic behaviour of the remainder term R for the usual Gini coefficient G. The remainder is

$$R = \int \{2y[\hat{F}(y) - F(y)] - y(\hat{G} - G)\} \times [d\hat{F}(y) - dF(y)].$$

Denoting the difference $\hat{F}(y) - F(y)$ by $\hat{D}(y)$, the remainder can be expressed as a sum of two integrals

$$R = \int 2y\hat{D}(y)d\hat{D}(y) - \int (\hat{G} - G)yd\hat{D}(y).$$

The first integral is reduced to zero by the integration by parts, so that the remainder is approximated by

$$R \approx -(\hat{G} - G)(\hat{\mu}_Y - \mu_Y)$$
$$= -(\hat{G} - G)o_p(n^{-\frac{1}{2}+\delta}), \quad 0 < \delta < \tfrac{1}{2}.$$

Therefore, we can say that $R = o_p(|\hat{G} - G|)$.

4. LORENZ CURVE ORDINATE AND QUANTILE SHARE

The ordinate of the Lorenz curve was defined in (1.1). In terms of estimating equations, the following two equations are required:

$$u_1(y,L(p)) = I\{y \le \xi_p\}y - L(p)y,$$
$$u_2(y) = I\{y \le \xi_p\} - p.$$

The second equation defines the $100p$-th percentile of the distribution; whereas the first equation defines the ordinate of the Lorenz curve in terms of the $100p$-th percentile. Ignoring the remainder term in (2.3), we have the following approximation:

$$0 = \int [I\{y \le \hat{\xi}_p\} - \hat{L}(p)]yd\hat{F}(y)$$
$$\approx \int_{\xi_p}^{\hat{\xi}_p} ydF(y) - [\hat{L}(p) - L(p)]\int ydF(y)$$
$$+ \int [I\{y \le \xi_p\} - L(p)]yd\hat{F}(y).$$

The first term of this expression can be further approximated as

$$\int_{\xi_p}^{\hat{\xi}_p} ydF(y) \approx (\hat{\xi}_p - \xi_p)\xi_p f(\xi_p),$$

and from (2.5) we see that

$$\hat{\xi}_p - \xi_p \approx -\int \frac{1}{f(\xi_p)} [I\{y \le \xi_p\} - p]d\hat{F}(y), \quad (4.1)$$

so that

$$(\hat{\xi}_p - \xi_p)\xi_p f(\xi_p) \approx -\int \xi_p [I\{y \le \xi_p\} - p]d\hat{F}(y).$$

Therefore, to estimate the variance of the ordinate of the Lorenz curve, the appropriate linearization is given by using

$$u^*(y) = \frac{1}{\mu_Y}[(y - \xi_p)I\{y \le \xi_p\} + p\xi_p - yL(p)].$$

This yields the same result as described by Beach and Davidson (1983) for variances and covariances of ordinates of the Lorenz curve in the case of independent and identically distributed random variables. To estimate the variance it is necessary to use $\hat{\xi}_p$ and $\hat{L}(p)$ in the expression for $u^*(y)$.

To estimate the quantile share $Q(p_1, p_2)$ we need three equations

$$u_1(y,Q(p_1,p_2)) = I\{\xi_{p_1} < y \le \xi_{p_2}\}y - Q(p_1,p_2)y,$$
$$u_2(y) = I\{y \le \xi_{p_1}\} - p_1,$$
$$u_3(y) = I\{y \le \xi_{p_2}\} - p_2.$$

Using the same arguments as before, we arrive at

$$u^*(y) = \frac{1}{\mu_Y}[(y - \xi_{p_2})I\{y \le \xi_{p_2}\}$$
$$- (y - \xi_{p_1})I\{y \le \xi_{p_1}\}$$
$$+ p_2\xi_{p_2} - p_1\xi_{p_1} - yQ(p_1,p_2)].$$

5. LOW INCOME MEASURE

The Low Income Measure was defined in (1.3). In terms of estimating equations, the following two equations are required:

$$u_1(y,\Theta) = I\left\{y \le \frac{M}{2}\right\} - \Theta,$$
$$u_2(y) = I\{y \le M\} - \frac{1}{2},$$

where M denotes the median of the distribution defined by the second equation, whereas the first equation defines the Low Income Measure in terms of the median. Ignoring the remainder term in (2.3), we have the following approximation:

$$0 = \int \left(I\left\{ y \le \frac{\hat{M}}{2} \right\} - \hat{\Theta} \right) d\hat{F}(y)$$

$$\approx \frac{1}{2}(\hat{M} - M) f\left(\frac{M}{2}\right) - (\hat{\Theta} - \Theta) + \int \left(I\left\{ y \le \frac{M}{2} \right\} - \Theta \right) d\hat{F}(y).$$

Using result (4.1) to substitute for $\hat{M} - M$, and solving for $\hat{\Theta} - \Theta$, we obtain

$$\hat{\Theta} - \Theta \approx \int u^*(y) d\hat{F}(y),$$

where

$$u^* = -\frac{f\left(\frac{M}{2}\right)}{2f(M)} \left(I\{y \le M\} - \frac{1}{2} \right) + I\left\{ y \le \frac{M}{2} \right\} - \Theta. \quad (5.1)$$

The problem with applying this result to estimate the variance of the estimated Low Income Measure is that it is necessary to estimate $f(M)$ and $f(M/2)$. To accomplish this, we could use

$$\hat{f}(\xi) = \frac{\hat{F}\left(\xi + \frac{h}{2}\right) - \hat{F}\left(\xi - \frac{h}{2}\right)}{h},$$

for some suitably small h. Alternatively, we could perform the following calculations, as suggested by Francisco and Fuller (1991) for another problem. For a given value of ξ, we estimate the corresponding percentile, $100p$. We then construct the Woodruff interval for that percentile. This is determined by first solving for h_1 and h_2 in

$$\inf_{h_1} \left[\frac{\int [I\{y \le \xi - h_1\} - p] d\hat{F}(y)}{\left[\text{mse}\left\{ \int [I\{y \le \xi\} - p] d\hat{F}(y) \right\} \right]^{1/2}} \le -z_{1-\alpha/2} \right],$$

$$\inf_{h_2} \left[\frac{\int [I\{y \le \xi + h_2\} - p] d\hat{F}(y)}{\left[\text{mse}\left\{ \int [I\{y \le \xi\} - p] d\hat{F}(y) \right\} \right]^{1/2}} \ge z_{1-\alpha/2} \right],$$

where $z_{1-\alpha/2}$ is the $100(1 - \alpha/2)$-th percentile from the standard normal distribution. Then we compute

$$\hat{f}(\xi) = \frac{2z_{1-\alpha/2}\left[\text{mse}\left\{ \int [I\{y \le \xi\} - p] d\hat{F}(y) \right\} \right]^{1/2}}{h_1 + h_2}. \quad (5.2)$$

This calculation uses the asymptotic equivalence of $\hat{\xi} - \xi$ and the estimated sum of the $u^*(y)$'s given by (2.5).

We see that the estimated variance for the Low Income Measure may be somewhat complex to compute. The estimating functions framework has however provided us with the appropriate formulae.

The discussion about the remainder term in the decomposition (2.3) of the low income measure is analogous to that made for the case of the quantile estimation (2.5).

6. ESTIMATION WITH A COMPLEX SURVEY

Let us assume a stratified multistage design with a large number of strata, H, with a few primary sampling units (clusters), $n_h (\ge 2)$, sampled from each stratum. For example, in the Canadian Survey of Consumer Finance (SCF) which uses the Labour Force Survey (LFS) vehicle, the number of strata is several hundreds and the number of clusters per stratum is on average less than six. Let w_{hci} be the normalized weight attached to the i-th ultimate unit in the c-th cluster of the h-th stratum such that the appropriate estimator of mean and the consistent estimator of its mean squared error are

$$\hat{\mu} = \sum_s w_{hci} y_{hci}$$

$$\text{mse}(\hat{\mu}) = \sum_h \frac{n_h}{n_h - 1} \sum_c (u^*_{hc} - \bar{u}^*_h)^2 \quad (6.1)$$

where $u^*_{hc} = \sum_i w_{hci}(y_{hci} - \hat{\mu})$ and $\bar{u}^*_h = 1/n_h \sum_c u^*_{hc}$. We use $\sum_s = \sum_h \sum_c \sum_i$ to denote summation over all ultimate units in the sample incorporating all stages of sampling. We assume that PSU's are selected with replacement.

This paper is not concerned with the efficiency of the estimators but rather the properties of commonly used estimators. An analysis of more complex estimators found in the econometric literature is beyond the scope of our study.

An estimator of the finite population distribution function is

$$\hat{F}(y) = \sum_s w_{hci} I\{y_{hci} \le y\}.$$

A consistent estimator of the approximation of the mean squared error of the distribution function estimated in y takes the form (6.1) where $u^*_{hc} = \sum_i w_{hci}[I\{y_{hci} \leq y\} - \hat{F}(y)]$.

The usual estimate of the finite population quantile is the sample quantile

$$\hat{\xi}_p = \inf\{y_{hci} \epsilon s : \hat{F}(y_{hci}) \geq p\}$$

which is the solution of the estimating equation

$$\sum_s w_{hci}[I\{y_{hci} \leq \hat{\xi}_p\} - p] = 0.$$

Accordingly, using result (2.5), the estimator of the mean squared error of the p-th quantile has the form (6.1) with

$$u^*_{hc} = \frac{1}{[\hat{f}(\hat{\xi}_p)]^2} \sum_i w_{hci}[I\{y_{hci} \leq \hat{\xi}_p\} - p].$$

If the expression (5.2) is used for the estimation of the density function $f(\xi)$, the MSE estimate of the quantile $\hat{\xi}_p$ becomes

$$\text{mse}_\alpha(\hat{\xi}_p) = \left(\frac{D_\alpha(\hat{\xi}_p)}{z_{1-\alpha/2}}\right)^2 \qquad (6.2)$$

where $D_\alpha(\hat{\xi}_p) = (h_1 + h_2)/2 = (\hat{\xi}_U - \hat{\xi}_L)/2$ is the half length of the $100(1 - \alpha)\%$ confidence interval for $\hat{\xi}_p$. In a complex sample design, h_1 and h_2 are obtained as solutions of

$$\hat{\xi}_L = \hat{\xi}_p - h_1 = \inf\{y_{hci} \epsilon s : \hat{F}(y_{hci}) \geq p - z_{1-\alpha/2}\sqrt{\text{mse}[\hat{F}(\hat{\xi}_p)]}\}$$

$$\hat{\xi}_U = \hat{\xi}_p + h_2 = \inf\{y_{hci} \epsilon s : \hat{F}(y_{hci}) \geq p + z_{1-\alpha/2}\sqrt{\text{mse}[\hat{F}(\hat{\xi}_p)]}\}.$$

The estimator (6.2) was also used by Francisco and Fuller (1991). Generally speaking the motivation for (5.2) and consequently for (6.2) comes from Woodruff's (1952) confidence interval for individual quantiles. Francisco and Fuller (1986) and Rao and Wu (1987) used these intervals to derive variance estimators. Although the estimator depends on the confidence coefficient, they showed that it is asymptotically consistent for any significance level α. Rao and Wu (1987) studied the standard errors of quantiles for the cluster samples estimated in this manner. Their Monte Carlo results suggest that 95% confidence interval works well as a basis for extracting the standard error. Binder and Patak (1994) obtained a similar form of the variance estimator by using the estimating equations approach.

The estimate of the usual Gini coefficient is the solution of the following estimating equation

$$\sum_s w_{hci}\{[2\hat{F}(y_{hci}) - 1]y_{hci} - \hat{G}y_{hci}\} = 0$$

and takes the form

$$\hat{G} = \frac{2}{\hat{\mu}} \sum_s w_{hci}\hat{F}(y_{hci})y_{hci} - 1$$

where $\hat{\mu} = \sum_s w_{hci}\, y_{hci}$.

The estimate of the MSE of the Gini coefficient can be computed using expression (6.1) by replacing u^*_{hc}, originally defined by (3.1), with its complex survey form. After some algebraic manipulation we obtain the following expression:

$$u^*_{hc} = \frac{2}{\hat{\mu}} \sum_i w_{hci}\left[A(y_{hci})y_{hci} + B(y_{hci}) - \frac{\hat{\mu}}{2}(\hat{G} + 1)\right]$$

where

$$A(y) = \hat{F}(y) - \frac{\hat{G} + 1}{2}$$

and

$$B(y) = \sum_s w_{hci}\, y_{hci}\, I\{y_{hci} \geq y\}.$$

The Lorenz curve ordinates could be obtained by solving a system of estimating equations

$$\sum_s w_{hci}[I\{y_{hci} \leq \hat{\xi}_p\}y_{hci} - \hat{L}(p)y_{hci}] = 0$$

$$\sum_s w_{hci}[I\{y_{hci} \leq \hat{\xi}_p\} - p] = 0.$$

The resulting estimate is

$$\hat{L}(p) = \frac{1}{\hat{\mu}} \sum_s w_{hci}\, y_{hci}\, I\{y_{hci} \leq \hat{\xi}_p\}.$$

To estimate the mean squared error of the Lorenz curve ordinates we simply use the values of u^*_{hc} defined by (6.3) in (6.1)

$$u^*_{hc} = \frac{1}{\hat{\mu}} \sum_i w_{hci}\big[(y_{hci} - \hat{\xi}_p)I\{y_{hci} \leq \hat{\xi}_p\} + p\hat{\xi}_p - y_{hci}\hat{L}(p)\big]. \qquad (6.3)$$

Similarly, the mse of the quantile share

$$\hat{Q}(p_1,p_2) = \frac{1}{\hat{\mu}} \sum_s w_{hci}\, y_{hci}\, I\{\hat{\xi}_{p_1} < y_{hci} \leq \hat{\xi}_{p_2}\}$$

is approximated by (6.1) using

$$u^*_{hc} = \frac{1}{\hat{\mu}} \sum_i w_{hci}\big[(y_{hci} - \hat{\xi}_{p_2})I\{y_{hci} \leq \hat{\xi}_{p_2}\} - (y_{hci} - \hat{\xi}_{p_1})I\{y_{hci} \leq \hat{\xi}_{p_1}\} + p_2\hat{\xi}_{p_2} - p_1\hat{\xi}_{p_1} - y_{hci}\,\hat{Q}(p_1,p_2)\big].$$

The Low Income Measure defined by (1.3) is estimated as

$$\hat{\Theta} = \hat{F}(\hat{M}/2) = \sum_s w_{hci}\, I\{y_{hci} \leq \hat{M}/2\}.$$

The mean squared error of the low income measure can be estimated approximately by the expression (6.1), where, (from the equation (5.1)):

$$u^*_{hc} = -\frac{\hat{f}(\hat{M}/2)}{2\hat{f}(\hat{M})} \sum_i w_{hci}[I\{y_{hci} \leq \hat{M}\} - 1/2] + \sum_i w_{hci}[I\{y_{hci} \leq \hat{M}/2\} - \hat{\Theta}].$$

7. ILLUSTRATION

The methodology above is illustrated with an application to the family income data collected in the Canadian Survey of Consumer Finance (SCF). We use the file on the Disposable Income of Economic Families obtained for the province of Ontario in 1988. Disposable income is defined as total income after tax reported in the survey. The SCF uses the framework of the Canadian Labour Force Survey which is based on a stratified, multistage design. For more details on the sample design see Singh *et al.* (1990).

We estimated the median M, the Gini coefficient G, the Low Income Measure Θ, Lorenz Curve Ordinates and quintile shares $Q(0,.2), Q(.2,.4), Q(.4,.6), Q(.6,.8), Q(.8,.1.0)$. Their standard errors are obtained using the proposed methodology and the jackknife 'delete-one-cluster' method.

We present a brief description of the jackknife 'delete-one-cluster' method used for this illustration. First, we assume that the estimate of the unknown parameter Θ can be expressed as $\hat{\Theta} = \mathcal{L}(\hat{F})$, where $\hat{F}$ is the estimated distribution function. The estimate of the distribution function $\hat{F}_{(hj)}$ obtained from the sample after removing the j-th sampled cluster of the h-th stratum ($j = 1, \ldots, n_h$, $h = 1, \ldots, H$) is

$$\hat{F}_{(gj)}(y) = \sum_s A_{hci}(g,j)\, w_{hci} I\{y_{hci} \leq y\}$$

where $$A_{hci}(g,j) = \begin{cases} 1, & h \neq g; \\ \dfrac{n_g}{n_g - 1}, & h = g,\ c \neq j; \\ 0, & h = g,\ c = j. \end{cases}$$

Then $\hat{\Theta}_{(gj)} = \mathcal{L}(\hat{F}_{(gj)})$ and the resulting 'delete-one-cluster' jackknife estimator of the variance of $\hat{\Theta} = \mathcal{L}(\hat{F})$ is

$$\mathrm{var}_J(\hat{\Theta}) = \sum_{g=1}^{H} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{\Theta}_{(gj)} - \hat{\Theta})^2.$$

It is known that the jackknife variance estimator performs poorly for quantiles due to its inconsistency (Kovar *et al.* 1988). There are some recent results (Shao and Wu 1989, Rao, Wu and Yue 1992) that suggest that the 'delete d' jackknife and 'delete-one-cluster', under certain conditions, may have desirable asymptotic properties for the variance estimation of non-smooth statistics like quantiles or the low income measure. On the other hand, for statistics like the Gini coefficient the jackknife estimator of the asymptotic variance is consistent (Shao 1993).

Unlike jackknifing, the estimating equations approach is not computationally intensive. It is simple, explicit and incorporates the sample design. It provides formulae for the asymptotic variance that are easy to program despite their complicated form.

Realizing the limitations imposed by using a single sample to make an objective comparison between different methods, the purpose of this example is to point out differences in the standard errors obtained by the estimating equations approach and a computationally intensive method like the jackknifing. Results are summarized in the table below. The direction of the difference in the estimated standard errors confirms the overall conservativeness of the jackknifing method. The difference can be attributed to the upward bias of the jackknifing method in the case of the median, although the 'delete-one-cluster' jackknife is preferable to the 'delete-1' jackknife. For the quantile shares it can be partly explained by the fact that upper quantile shares may not cut over all primary sampling units but rather perform as separated classes which may affect the jackknifing more than the estimating equations method.

Table 1
Measures of Income Inequality and Their Standard Errors

Measure	Estimate	Standard Error	
		Estimating Equations Approach	Jackknifing 'Delete-One-Cluster'
Median	31705	303.3	569.8
Gini	0.3482	0.005	0.005
Low Income Measure	0.1980	0.00586	0.00613
Lorenz Curve Ordinates			
L(0.2)	0.0561	0.00137	0.00175
L(0.4)	0.1745	0.00166	0.00194
L(0.6)	0.3522	0.00246	0.00285
L(0.8)	0.5982	0.00317	0.00393
Quintile Shares			
Q(0, 0.2)	0.0561	0.00137	0.00167
Q(0.2, 0.4)	0.1186	0.00159	0.00221
Q(0.4, 0.6)	0.1775	0.00157	0.00282
Q(0.6, 0.8)	0.2461	0.00158	0.00337
Q(0.8, 1.0)	0.4017	0.00395	0.00451

8. SUMMARY

The problem of estimating the variance of complex statistics, such as measures of income inequality, have eluded statisticians for years. Replication methods such as the jackknife are often suggested for estimation. The advantage of the linearization approach is that it can be used under a wide class of sampling designs and does not suffer from the need for intensive computations which methods such as the bootstrap entail. Through the method of estimating functions and the decomposition given in (2.3), we find that some difficult problems can be solved more easily. A discussion about the order of the remainder term for some of these measures is given as well. A more rigorous proof for a complex sample design can be established along the lines given in Shao and Rao (1994).

REFERENCES

BEACH, C.M., and DAVIDSON, R. (1983). Distribution-free statistical inference with Lorenz curves and income shares. *Review of Economic Studies*, 50, 723-735.

BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

BINDER, D.A. (1991). Use of estimating functions for interval estimation from complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 34-42.

BINDER, D.A., and PATAK, Z. (1994). Use of estimating functions for interval estimation from complex surveys. *Journal of the American Statistical Association*, 89, 1035-1044.

FRANCISCO, C.A., and FULLER, W.A. (1986). Estimation of the Distribution function with a complex survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 37-45.

FRANCISCO, C.A., and FULLER, W.A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, 19, 454-469.

GLASSER, G.J. (1962). Variance formulas for the mean difference and coefficient of concentration. *Journal of the American Statistical Association*, 57, 648-654.

KOVAR, J.G., RAO, J.N.K., and WU, C.F.J. (1988). Bootstrap and other methods to measure errors in survey estimates. *The Canadian Journal of Statistics*, 16, 25-45.

NYGÅRD, F., and SANDSTRÖM, A. (1981). *Measuring Income Inequality*. Stockholm: Almqvist and Wiksell International.

RANDLES, R.H. (1982). On the Asymptotic Normality of Statistics with Estimated Parameters. *Annals of Statistics*, 10, 462-474.

RAO, J.N.K., and WU, C.F.J. (1987). Methods for Standard Errors and Confidence Intervals from Survey Data: Some Recent Work. *Proceedings of the 46th session, International Statistical Institute*, 3, 5-19.

RAO, J.N.K., WU, C.F.J., and YUE, K. (1992). Some Recent Work on Resampling Methods for Complex Surveys. *Survey Methodology*, 18, 209-217.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SENDLER, W. (1979). On statistical inference in concentration measurement. *Metrika*, 26, 109-122.

SHAO, J., and RAO, J.N.K. (1994). Standard Errors for Low Income Proportions Estimated from Stratified Multi-Stage Samples. *Sankhyā, B*, (to appear).

SHAO, J. and WU, C.W.J. (1989). A general Theory for Jackknife Variance Estimation. *Annals of Statistics*, 17,1176-1197.

SHAO, J. (1993). Inferences Based on *L*-statistics in Survey Problems: Lorenz Curve, Gini Family and Poverty Proportion. In *Proceedings of the Workshop on Statistical Issues in Public Policy Analysis*, Carleton University and University of Ottawa.

SINGH, M.P., DREW, J.D., GAMBINO, J.G., and MAYDA, F. (1990). *Methodology of the Canadian Labour Force Survey*, Catalogue No. 71-526, Statistics Canada.

WOODRUFF, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association*, 47,635-646.