Outlier Resistant Alternatives to the Ratio Estimator

JEAN-PHILIPPE GWET and LOUIS-PAUL RIVEST*

Many techniques have been suggested to lower the impact of outliers on sample survey estimates. Outliers can be downweighted by winsorization; that is, by replacing extreme data points by a data-dependent or a predetermined value before calculating estimates. Another approach is to reduce the weight of outliers, from the inverse sampling fraction to 1, in the estimation of population characteristics. In this article the problem of estimating the population mean using auxiliary information in the presence of outliers is considered. A resistant version of the ratio estimator is introduced. It is constructed with a M or a GM estimator of the slope of the regression model through the origin, which is implicitly called on when considering the ratio estimator. The asymptotic biases and asymptotic variances of the proposed alternatives to the ratio estimator are calculated with respect to the randomization of the sampling plan. The selection of a resistant estimator is seen to involve a trade-off between bias and variance. Often, some bias is the price paid to reduce the variance. A mean squared error estimator is proposed. A model-based estimator proposed by Chambers, reducing the weights given to extreme observations to 1, is also studied. A conditional investigation of the bias, given the proportion of outliers in the sample, is carried out. It reveals that the unconditional unbiasedness of the ratio estimator is, in the presence of outliers, deceptive. Its conditional bias varies substantially with the difference between the sample proportion and the population proportion of outliers. It can be severe if the proportion of outliers in the sample is much larger than in the population. The conditional bias of resistant estimators is, on the other hand, more stable. It does not depend as much on the proportion of outliers in the sample. Monte Carlo comparisons of the ratio estimator with resistant alternatives are presented for two populations. These similations show that in the presence of outliers, the mean squared error of resistant estimators can be substantially smaller than that of the ratio estimator. They also show that resistant confidence intervals are interesting alternatives to intervals based on the ratio estimator.

KEY WORDS: Conditional inference; GM estimator; M estimator; Mean squared error; Robust estimation; Sample survey.

Chambers (1986) distinguished two types of outliers in survey data. The first type, a representative outlier, is a sample unit that has been correctly measured. Some of its values differ substantially from most sample values. Such a unit is not unique; the unsampled part of the population may contain similar units, with values markedly different from most unsampled units. The second type, an unrepresentative outlier, is a sample unit whose values have been recorded with errors. The detection and treatment of such errors is done at the editing stage of the survey process. An unrepresentative outlier can also correspond to a unique unit that is unrepresentative of the unsampled part of the population. Unique units can be excluded from the estimation process by giving them a weight of 1 (Rao 1971). Only representative outliers are considered in this article; statistical methods are proposed to lower their impact on the estimation of population characteristics.

When should a data point be called an outlier? The answer to this question depends on the way in which population characteristics are estimated. A sample unit may, for instance, look similar to others when the population mean is estimated by the sample mean but be an outlier when the ratio estimate is used. This is exemplified in the small population of Sukhatme and Sukhatme (1970, p. 185); see also Som (1973, p. 73), which is presented in Figure 1. The y value of the pair (845,663) is similar to the others; however, it is much larger than what would be expected for an x value of 663. When (845,663) is sampled, the ratio estimate is much larger than when it is not sampled. This unit is, therefore, an outlier for the ratio estimator. In general, a unit is an outlier with respect to the regression model that is called on implicitly for constructing estimators of population characteristics. Thus the detection and treatment of outliers have to be model-assisted.

Winsorization, that is, replacing large measurements by a predetermined, or a data-dependent, value, is a popular method to downweight outliers in the estimation of the population mean when the underlying model is the simple location model, $y_i = \mu + \varepsilon_i$. It was investigated by Ernst (1980), Fuller (1991), and Searls (1966), who showed that for skewed populations, the mean squared error of the winsorized sample mean is smaller than that of the sample mean, even if the winsorized mean is a biased estimator of the population mean. Hidiroglou and Srinath (1981) considered various methods for reducing the weights of outliers once they have been identified. They also proposed to study the performance of estimators by conditioning on the proportion of outliers in the sample (see also Rao 1985).

In sample surveys, the word robust is often used to qualify a procedure that is design consistent in the presence of discrepancies to an assumed probability model. In this article "robust" conveys insensitivity to outliers in the sample. The term "resistant," or "outlier resistant," is also used for this purpose (Mosteller and Tukey 1977, p. 203). The proposed outlier robust procedures contrast with procedures robust to model misspecification in that they are not design consistent.

This article introduces outlier-resistant alternatives to the ratio estimator that are derived from M and GM estimators. M estimators were introduced by Huber (1964); see also Huber (1981) and Hampel, Ronchetti, Rousseeuw, and Stahel (1986). Asymptotic biases and asymptotic variances, with respect to the randomization of the sample design, of the resistant estimators are derived in Section 2, where a mean squared error estimator is proposed. In survey sampling, regression M estimators were first considered by Chambers

* Jean-Philippe Gwet is Statistician, Social Survey Method Division of Statistics Canada, Ottawa, Ontario K1A 0T6, and Louis-Paul Rivest is Professor at the Département de mathématiques et de statistiques, Université Laval, Cité Universitaire, Québec G1K 7P4, Canada. This research was supported by NSERC Grant A5244 and by a grant from le Fond pour la formation des chercheurs et l'aide à la recherche from the Province of Québec. The authors gratefully acknowledge the constructive comments of a referee.

Journal of the American Statistical Association
December 1992, Vol. 87, No. 420, Theory and Methods

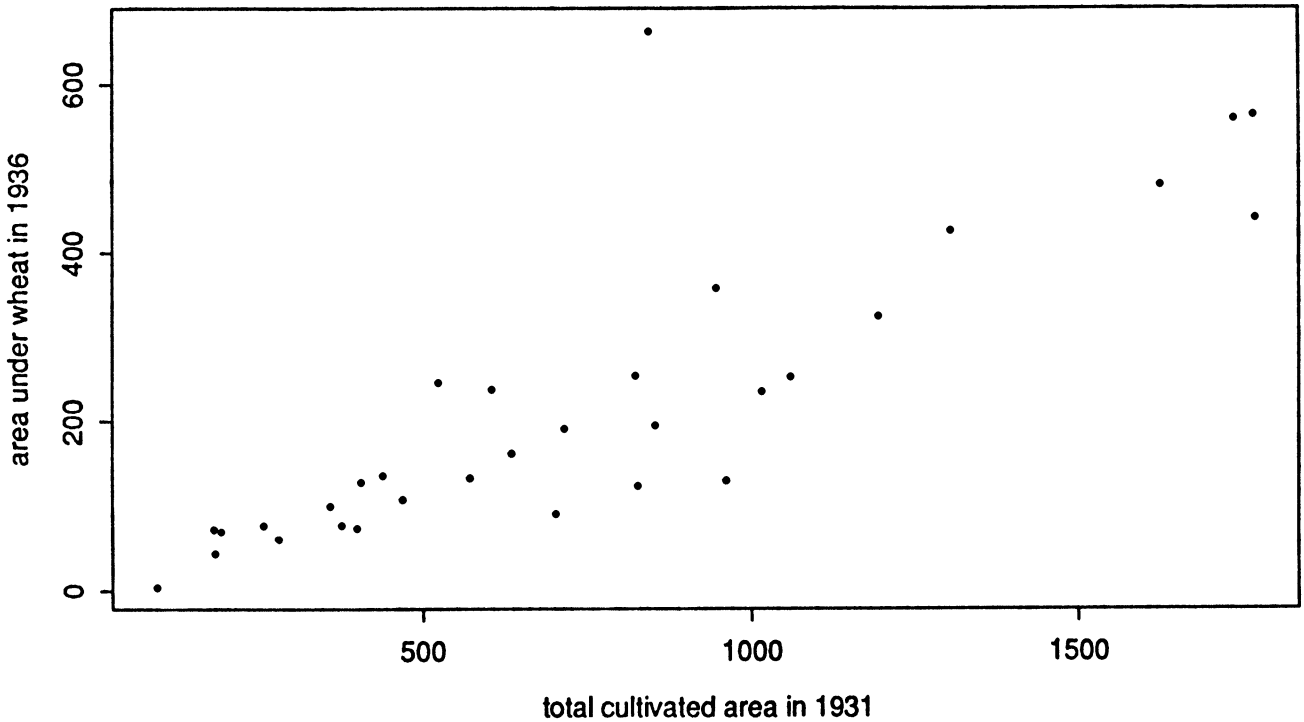

Figure 1. Population 1.

(1986). His investigation was carried in the model-based framework of Royall and Cumberland (1981); this is discussed in Section 3. A conditional investigation of the bias, given the number of outliers in the sample, is presented in Section 4. It shows that when the proportion of outliers in the sample is larger than the corresponding proportion in the population, the ratio estimator may have a substantial bias, larger than that of its resistant competitors. Simulation results are presented in Section 5.

1. RESISTANT ALTERNATIVES TO THE RATIO ESTIMATOR

Suppose that for all individuals in the population $U = \{1, \ldots, i, \ldots, N\}$ the positive auxiliary variable x is known. Let s denote a simple random sample of size n drawn from U. The variable of interest y is observed only for the units contained in s. The y population mean, $\bar{y}$, has to be estimated. In many instances, the model

$$y_i = \beta x_i + \varepsilon_i, \tag{1.1}$$

where $\{\varepsilon_i\}$ is a set of independent random variables with mean 0 and variances proportional to x_i, provides a reasonable description of the relationship between x and y. In such cases the ratio estimator, $\bar{y}_r = \bar{x}\bar{y}_s/\bar{x}_s$, where $\bar{x}$ is the population mean of the variable x and $\bar{y}_s$ and $\bar{x}_s$ are the sample means of the variables y and x, provides a good estimation of $\bar{y}$. (The subscript s always denotes a sample average.) A well-known theoretical justification for $\bar{y}_r$ is that under model (1.1), the ratio

$$\hat{\beta} = \bar{y}_s/\bar{x}_s \tag{1.2}$$

is the weighted least squares estimator of β.

If, as in Figure 1, there are outliers with respect to model (1.1), then it is of interest to consider resistant alternatives to $\hat{\beta}$. This article investigates estimators of $\bar{y}$ calculated using GM estimators of the parameter β of model (1.1). Several types of GM estimators have been proposed in the literature. (See Hampel et al. 1986, p. 315 for a discussion of the various proposals.) Comparisons of various alternatives to the ratio estimators were presented in Rivest and Rouillard (1991), who suggested using the Schweppe version of the regression GM estimators. This GM estimator, denoted by $\hat{\beta}_g$, is defined implicitly as the solution to the equation

$$\sum_s \frac{\sqrt{x_i}}{h(x_i)} \psi\left(\frac{(y_i - \beta x_i)h(x_i)}{K\sqrt{x_i}}\right) = 0, \tag{1.3}$$

where

- $\psi(t)$ is an odd function approximately proportional to t for small values of t
- $h(t)$ is a positive nondecreasing function of t
- K is a positive scaling constant for the residuals of model (1.1)

When $\psi(t) = t$, the solution of (1.3) is given by (1.2), the usual least squares estimator. When $\psi(t)$ is bounded and $h(t)$ is constant, the solution of (1.3) is a regression M estimator. This estimator is not sensitive to outliers among the residuals; however, sample units, with very large values of x, may make regression M estimators unrepresentative of the relationship between x and y for most sample units. (See Hampel et al. 1986, p. 314 for an example of this phenomenon.) To deal simultaneously with outliers among the residuals and among the values of x, one takes a bounded $\psi(t)$ function and $h(t)$ equal to $(t/\bar{x})^{1/2}$. For this choice of ψ and

h, the estimating function for β_g has an appealing simple form:

$$\sum_s \psi\left(\frac{y_i - \beta x_i}{K\sqrt{\bar{x}}}\right) = 0.$$

One can formally write $\hat{\beta}_g$ as a weighted ratio estimator:

$$\hat{\beta}_g = \frac{\sum w_i y_i}{\sum w_i x_i}, \tag{1.4}$$

where the weights w_i are given by

$$w_i = \frac{\psi((y_i - \hat{\beta}_g x_i)h(x_i)/(K\sqrt{x_i}))}{(y_i - \hat{\beta}_g x_i)h(x_i)/(K\sqrt{x_i})}. \tag{1.5}$$

Formula (1.4) is used in Section 5 for the iterative computation of $\hat{\beta}_g$. It shows that in the calculation of $\hat{\beta}_g$, the weight of a sample unit is reduced if it has a large residual.

A wide variety of ψ functions has been proposed for M and GM estimators. They are of two types. The first consists of bounded increasing functions; a famous proposal is that of Huber: $\psi_H(t) = \text{sgn}(t)\min(1, |t|)$. Bounded increasing ψ functions limit the impact of outliers on the estimate. If one wants extreme outliers to have absolutely no impact on the estimate (i.e., the values of the estimate calculated with and without the outliers are the same), one uses what are known as redescending ψ functions. This second type of function is approximately proportional to t when t is around 0, and it converges to 0 as t goes to $\pm\infty$. A typical member of this class is the Cauchy function: $\psi_c(t) = t/(1 + t^2)$.

The determination of the scaling constant K depends on the choice of ψ. By taking K arbitrarily large, one can make $\hat{\beta}_g$ equal to the least squares estimator $\hat{\beta}$. In most applications of robust regression, K is equal to $c\sigma$, where c is a robustness constant chosen by the statistician and σ is the residual scale parameter of model (1.1), which is usually unknown. For our theoretical investigations, we assume that σ is known.

The general form of the proposed resistant alternatives to the ratio estimator is given by $\bar{y}_g = \bar{x}\hat{\beta}_g$. Its asymptotic properties, with respect to the randomization of the sampling plan, are investigated in the next section.

2. ASYMPTOTIC PROPERTIES OF $\bar{y}_g$

The formal framework for investigating asymptotic properties of regression estimators in survey sampling consists of a nested sequence of populations where samples of increasing sizes are taken (see, for instance, Isaki and Fuller 1982; Scott and Wu 1981). To simplify the presentation, heuristic derivations of the main results are given in this section. Some proofs, in the setting of a nested sequence of populations, are given in the Appendix. In this section all expectations are taken with respect to the randomization of the sample design. The only design considered is simple random sampling without replacement.

The first step of the investigation is to determine, in terms of the population under study, the parameter estimated by $\hat{\beta}_g$. For any positive number β, define $\hat{g}(\beta)$ by

$$\hat{g}(\beta) = \frac{1}{n}\sum_s \frac{\sqrt{x_i}}{h(x_i)}\psi\left(\frac{(y_i - \beta x_i)h(x_i)}{K\sqrt{x_i}}\right) \tag{2.1}$$

and note that $\hat{\beta}_g$ is the solution of the equation $\hat{g}(\beta) = 0$. (We assume that ψ is increasing so that (2.1) has a unique solution.) The expectation of $\hat{g}(\beta)$ is given by

$$g(\beta) = \frac{1}{N}\sum_U \frac{\sqrt{x_i}}{h(x_i)}\psi\left(\frac{(y_i - \beta x_i)h(x_i)}{K\sqrt{x_i}}\right). \tag{2.2}$$

If β_g is the solution of $g(\beta) = 0$, then it is intuitively clear that $\hat{\beta}_g$ is estimating β_g. This is stated in the next proposition, which is proved in the Appendix.

Proposition 2.1. If the function ψ is increasing and if x_i is positive for each i, then $\hat{\beta}_g$ is an asymptotically design-consistent estimator (in the sense of Wright 1983) of β_g; that is, $\hat{\beta}_g - \beta_g$ is $o_p(1)$.

Estimators of $\hat{\beta}_g$ based on nonincreasing ψ functions, such as Cauchy function ψ_c, are also consistent under suitable regularity assumptions. Conditions ensuring consistency in infinite populations are given by Huber (1981, p. 127).

Proposition 2.1 shows that $\bar{y}_g$ is not an asymptotically design-consistent estimator of $\bar{y}$, unless $\beta_g = \bar{y}/\bar{x}$. Its asymptotic bias is given by $\beta_g\bar{x} - \bar{y} = -\bar{E}_g$, where $\bar{E}_g$ is the population mean of the theoretical resistant residuals

$$E_{gi} = y_i - \beta_g x_i. \tag{2.3}$$

Correcting $\bar{y}_g$ for its bias by, for instance, subtracting $\bar{e}_{gs}$, the sample mean of the observed residuals $e_{gi} = y_i - \hat{\beta}_g x_i$, for $i = 1, \ldots, n$, yields an estimator sensitive to outliers. It is not possible to have an estimator that is both asymptotically design-consistent and resistant to outliers except for populations where the distribution of the theoretical residuals is symmetric around 0. For such populations, $\beta_g = \bar{y}/\bar{x}$ and the bias is null. In practice one expects the distribution of the E_{gi}'s to be symmetric enough for the bias of $\bar{y}_g$ to be small.

To calculate the asymptotic variance of $\bar{y}_g$, an asymptotic linearization of $\hat{\beta}_g$ is needed. Using a Taylor series argument,

$$\hat{g}(\hat{\beta}_g) - \hat{g}(\beta_g) \approx (\hat{\beta}_g - \beta_g)\frac{d}{d\beta}\hat{g}(\beta_g).$$

Therefore, one can write

$$(\hat{\beta}_g - \beta_g) \approx -\left(\frac{d}{d\beta}g(\beta_g)\right)^{-1}\hat{g}(\beta_g) + o_p(1/\sqrt{n}).$$

This leads to the following result:

Proposition 2.2. An asymptotic expansion for $\hat{\beta}_g$ is given by

$$\hat{\beta}_g = \beta_g + \frac{1}{n}\sum_s \text{IC}_i + o_p(1/\sqrt{n}), \tag{2.4}$$

where IC_i is given by

$$\text{IC}_i = \frac{\dfrac{K\sqrt{x_i}}{h(x_i)}\psi\left(\dfrac{E_{gi}h(x_i)}{K\sqrt{x_i}}\right)}{\dfrac{1}{N}\sum_U x_i\psi'\left(\dfrac{E_{gi}h(x_i)}{K\sqrt{x_i}}\right)}, \tag{2.5}$$

where ψ' is the derivative of ψ and E_{gi} is defined by (2.3).

The quantity IC_i can be seen as a finite population generalization of the influence curve (see Hampel 1974). Proposition 2.2 is the sample survey analogue of the classical linearization of M estimators (see Huber 1981, p. 39). Regularity conditions ensuring the validity of (2.4) in infinite populations were given by Boos and Serfling (1980) and Fernholz (1983). Note that for the ratio estimator, that is, when $\psi(t) = t$, (2.5) simplifies to

$$\mathrm{IC}_i = \frac{y_i - \bar{y}x_i/\bar{x}}{\bar{x}}.$$

The asymptotic properties of $\bar{y}_g$ as an estimator of $\bar{y}$ are summarized in the next proposition.

Proposition 2.3. The asymptotic bias of $\bar{y}_g$ is equal to $-\bar{E}_g$, where $\bar{E}_g$ is the mean of the theoretical residuals defined by (2.3). Its asymptotic variance is given by

$$V(\bar{y}_g) = \frac{1-f}{n}\bar{x}^2 \sum_U \frac{\mathrm{IC}_i^2}{N-1}.$$

When $\psi(t) = t$, or when the tuning constant K is large enough, $\bar{y}_g$ is equal to the ratio estimator $\bar{y}_r$, its asymptotic bias is null, and the expression for its asymptotic variance given in Proposition 2.3 reduces to that for the ratio estimator. Thus when outliers are present, the choice of K involves a trade-off between bias and variance. Selecting large values of K results in a low bias with a possibly large variance, whereas selecting small or moderate values of K might reduce the variance at the cost of an increased bias.

To construct an estimator $v(\bar{y}_g)$ of the mean squared error of $\bar{y}_g$, one needs to estimate the squared bias. A simple bias estimator is $(\bar{y}_g - \bar{y}_r)^2$. It is biased because

$$E((\bar{y}_g - \bar{y}_r)^2) = E^2(\bar{y}_g - \bar{y}_r) + V(\bar{y}_g - \bar{y}_r).$$

The most important component of this bias seems to be the variance term (Rivest and Rouillard 1991). Using Proposition 2.2, $V(\bar{y}_g - \bar{y}_r)$ can be estimated by

$$\bar{x}^2 \frac{1-f}{n} \sum_s \left(\frac{y_i - \bar{y}_r x_i}{\bar{x}_s} - \widehat{\mathrm{IC}}_i\right)^2,$$

where $\widehat{\mathrm{IC}}_i$ is the estimate of IC_i defined by formula (2.5):

$$\widehat{\mathrm{IC}}_i = \frac{\dfrac{K\sqrt{x_i}}{h(x_i)}\psi\left(\dfrac{e_{gi}h(x_i)}{K\sqrt{x_i}}\right)}{\dfrac{1}{n}\sum_s x_i \psi'\left(\dfrac{e_{gi}h(x_i)}{K\sqrt{x_i}}\right)}$$

and the e_{gi}'s are the sample residuals. The proposed squared bias estimator is given by

$$b_2(\bar{y}_g) = \max\left(0, (\bar{y}_g - \bar{y}_r)^2 - \bar{x}^2 \frac{1-f}{n} \sum_s \left(\frac{y_i - \bar{y}_r x_i}{\bar{x}_s} - \widehat{\mathrm{IC}}_i\right)^2\right).$$

This yields the following estimator of the mean squared error of $\bar{y}_g$;

$$v(\bar{y}_g) = b_2(\bar{y}_g) + \frac{1-f}{n(n-1)}\bar{x}^2 \sum_s \widehat{\mathrm{IC}}_i^2. \tag{2.6}$$

When $\psi(t) = t$, formula (2.6) reduces to a classical estimator of $V(\bar{y}_r)$ sometimes known as v_2 (see Wu and Deng 1983).

3. A MODEL-BASED ESTIMATOR

If the y_i's are considered as independent realizations of model (1.1), then one only needs predictions of the y values that are not in the sample to estimate $\bar{y}$. This leads to estimator $\bar{y}_m$ (subscript m stands for model), defined by

$$\bar{y}_m = \frac{1}{N}\left(\sum_s y_i + \sum_{U-s} x_i\hat{\beta}_g\right). \tag{3.1}$$

An estimator similar to (3.1) is discussed by Fuller (1991) and Bellhouse (1987). When a redescending ψ function is used in calculating $\hat{\beta}_g$, this estimator is considered by Chambers (1986) as an outlier rejection strategy, because outliers have no impact on the predictions for the unsampled part of the population. Their weights in the estimation of the population total are equal to 1. As a way to bring outliers into the estimation of the population total, Chambers considered estimator $\bar{y}_{m1}$ obtained by replacing $\hat{\beta}_g$ in formula (3.1) by $\hat{\beta}_{g1}$, where

$$\hat{\beta}_{g1} = \hat{\beta}_g + \frac{\sum_s K\sqrt{x_i}\,\psi_1[(y_i - \hat{\beta}_g x_i)/(K\sqrt{x_i})]}{\sum_s x_i}.$$

In Chambers's notation, ψ_1 is the "prediction" ψ function that increases the weights given to outliers in the estimation of $\bar{y}$, and the ψ function used for calculating $\hat{\beta}_g$ is the "estimation" ψ function. Estimator $\hat{\beta}_{g1}$ is obtained after one iteration of the modified residual algorithm (Huber 1981, p. 181), with $\hat{\beta}_g$ as the starting value. In his Monte Carlo investigations, Chambers considered estimator $\hat{\beta}_{g1}$ with prediction ψ function given by

$$\psi(t) = t\exp(-.25(|t| - 6)^2). \tag{3.2}$$

The asymptotic properties, with respect to the randomization of the sample design, of estimator $\bar{y}_m$ are investigated in the Appendix. This estimator's asymptotic bias is $-(1-f)\bar{E}_g$, and its asymptotic variance is equal to

$$V(\bar{y}_m) = \frac{1-f}{n}\frac{1}{N-1}\sum_U (f(E_{gi} - \bar{E}_g) + (1-f)\bar{x}\,\mathrm{IC}_i)^2.$$

The bias, with respect to the sample design, of model-based estimator $\bar{y}_m$ is smaller than that of $\bar{y}_g$. This bias reduction is achieved at the expense of the design-based variance of $\bar{y}_m$, which is sensitive to extreme observations. Another strategy for bringing in outliers and reducing the bias of $\bar{y}_g$ is to increase the scaling constant K. This would reduce the bias while keeping a bound on the influence of sample units. Ways of increasing K (possibly by making it a function of the sample size and of the sampling fraction) need to be investigated.

4. A CONDITIONAL ANALYSIS OF THE BIAS

The asymptotic design consistency of the ratio estimator is an attractive property that is lacking in its resistant competitors. It means that under repeated sampling of the same population, the discrepancies between the ratio estimates and $\bar{y}$ average out. In practice, however, the statistician is usually

interested in one realized sample. Many authors (for instance, Rao [1985]), argue that the properties of estimators should then be studied conditionally on the realization of variables having an important impact on the estimator, such as the sample proportion of outliers. Thus the vector of conditional biases obtained for the most probable sample proportions of outliers provides a better assessment of the real discrepancy between an estimator and $\bar{y}$ than does the unconditional bias.

Conditional biases in the presence of outliers were considered by Hidiroglou and Srinath (1981) and Rao (1985). They are investigated by dividing population U into two strata, U_1 and U_2, where U_1 contains N_1 outliers with respect to model (1.1) and U_2 contains the remaining $N_2 = N - N_1$ units, which are well described by model (1.1). The population proportion of outliers, $W_1 = N_1/N$, is assumed to be small (say, less than 10%). In a conditional setting, one considers that simple random samples of size n_1 and n_2 ($n_1 + n_2 = n$) are taken out of U_1 and U_2. Let $W_{1s} = n_1/n$ be the sample proportion of outliers and let $B(.\,|\,W_{1s})$ be the conditional bias, given that the sample proportion of outliers is W_{1s}. Thus

$$B(\bar{y}_g|\,W_{1s}) = E(\hat{\beta}_g|\,W_{1s})\bar{x} - \bar{y},$$

where $E(.\,|\,W_{1s})$ denotes the expectation with respect to the stratified sampling plan defined earlier. It is impossible to get an exact analytical expression for the conditional bias. To get asymptotic approximations, one can argue as in Section 2. Asymptotically, $E(\hat{\beta}_g|\,W_{1s})$ is equal to the parameter estimated by $\hat{\beta}_g$; that is, to the solution of

$$E(\hat{g}(\beta)|\,W_{1s}) = 0, \tag{4.1}$$

where $\hat{g}(\beta)$ is defined by (2.1). Straightforward calculations show that $E(\hat{g}(\beta)|\,W_{1s})$ is proportional to

$$g(\beta) + \frac{W_{1s} - W_1}{1 - W_{1s}} \frac{1}{N_1} \sum_{U_1} \frac{\sqrt{x_i}}{h(\sqrt{x_i})} \psi\left(\frac{(y_i - \beta x_i)h(\sqrt{x_i})}{K\sqrt{x_i}}\right). \tag{4.2}$$

It is not generally possible to get an explicit form for the solution of (4.1). But because the second term of (4.2) is small in large samples, one can get good asymptotic approximations. By definition, $g(\beta_g) = 0$. To improve on β_g as an approximation to the solution of (4.1), one can take one iteration of the Newton–Raphson algorithm with β_g as the starting value. This yields

$$\beta_g + \frac{W_{1s} - W_1}{1 - W_{1s}} \frac{1}{N_1} \sum_{U_1} \mathrm{IC}_i.$$

Thus, a first-order approximation to the conditional bias of $\bar{y}_g$ is given by

$$B(\bar{y}_g|\,W_{1s}) = -\bar{E}_g + \frac{W_{1s} - W_1}{1 - W_{1s}} \frac{\bar{x}}{N_1} \sum_{U_1} \mathrm{IC}_i. \tag{4.3}$$

The conditional bias is not too sensitive to the actual number of outliers, because the coefficient of $(W_{1s} - W_1)/(1 - W_{1s})$ is bounded. Writing $\bar{E}_g$ as $W_1\bar{E}_{g1} + W_2\bar{E}_{g2}$, where $\bar{E}_{g1}$ and $\bar{E}_{g2}$ are the averages of the theoretical resistant residuals in the outlier and the nonoutlier strata, shows that the dominant term of (4.3) is $W_1\bar{E}_{g1}$.

For the ratio estimator, equation (4.1) has an explicit solution, which is given by

$$\left[\bar{y} + \frac{W_{1s} - W_1}{1 - W_{1s}} \bar{y}_1\right] \Big/ \left[\bar{x} + \frac{W_{1s} - W_1}{1 - W_{1s}} \bar{x}_1\right],$$

where $\bar{x}_1$ and $\bar{y}_1$ are the averages of variables x and y in the outlier stratum. Thus, the asymptotic conditional bias of $\bar{y}_r$ is equal to

$$B(\bar{y}_r|\,W_{1s}) = \frac{W_{1s} - W_1}{1 - W_{1s}} \bar{E}_{1r} \Big/ \left[1 + \frac{W_{1s} - W_1}{1 - W_{1s}} \frac{\bar{x}_1}{\bar{x}}\right],$$

where $\bar{E}_{1r}$ is the average of the least squares residuals in the outlier stratum. The conditional bias of the ratio estimator depends heavily on the difference between W_{1s} and W_1. It is severe in cases where W_{1s} is much larger than W_1. Such cases may happen, with a relatively high probability, when W_1 is small and n is small.

The previous discussion reveals an important difference between the ratio estimator and its resistant competitors. The unconditional biases of resistant estimators can be seen as a premium paid for being protected against occasional wild samples (with W_{1s} much larger than W_1). On the other hand, the unbiasedness of the ratio estimator is deceptive, because it hides a severe conditional bias that obtains when the sample proportion of outliers is much larger than the population proportion. This is exemplified in the next section.

5. EMPIRICAL INVESTIGATIONS

Five resistant estimators were compared to the ratio estimator by simulations. We also included in our study estimator WINS, defined as $\bar{x}$ multiplied by $\bar{y}_s^1/\bar{x}_s^1$, where $\bar{x}_s^1$ and $\bar{y}_s^1$ are the 1-winsorized mean of the x and the y sample. The 1-winsorized mean is obtained by replacing the largest observation with the second largest observation before taking the sample mean. The 1-winsorized mean has been shown by Fuller (1991) to be an interesting alternative to the sample mean in the presence of outliers. Estimator WINS provides an instance of outlier treatment that is not based on model (1.1). The estimators under investigation are presented in Table 1.

The scaling constant K is equal to $c\sigma$, where σ is the scale parameter of the errors in model (1.1). Parameter σ was

Table 1. Estimators in the Monte Carlo Study

Code	$h(t)$	$\Psi(t)$	K	Type of estimator
RAE	1	t	—	Ratio estimator
HUM	1	$\Psi_H(t)$	1.345σ	g[a]
HUGM	$sqrt(t/\bar{x})$	$\Psi_H(t)$	2.56σ	g
CAM	1	$\Psi_c(t)$	2.385σ	m[b]
CHAM	1	$\Psi_c(t)$	2.385σ	Chambers[c]
CAGM	$sqrt(t/\bar{x})$	$\Psi_c(t)$	4.54σ	m
WINS	—	—		Ratio of 1-winsorized mean

[a] g stands for estimator $\bar{y}_g$ defined in Section 1.
[b] m stands for model-based estimtor $\bar{y}_m$ defined by (3.1).
[c] Chambers denotes estimator $\bar{y}_{m1}$ obtained with estimating Ψ function Ψ_c and prediction Ψ function Ψ_1 defined by (3.2).

considered to be unknown. It was estimated for each Monte Carlo sample by the median absolute deviation of the residuals (MAD) multiplied by 1.483 (Hampel et al. 1986, p. 105). For M estimators, the robustness constant c was chosen such that the large sample efficiency of $\hat{\beta}_g$ is 95% when the errors of model (1.1) are normally distributed. For the GM estimator based on Huber's ψ function, c is the value proposed by Hampel et al. (1986, p. 333). It provides an efficiency of 95% when the errors are normally distributed and when x follows a χ_1^2 distribution. No exact robustness constants c were available for Cauchy GM estimators; Cauchy GM robustness constant c was taken as Huber constant 2.56, multiplied by the ratio of Cauchy constant 2.385, over Huber constant 1.345, for M estimators.

To calculate the M and GM estimates, the iteratively reweighted least squares algorithm (Beaton and Tukey 1974) was used, with the least squares estimate $\bar{y}_s/\bar{x}_s$ as the starting value. To update the current value β_j, one calculates the weights w_i using β_j and computes a weighted ratio estimate β_{j+1} using formulas (1.4) and (1.5). For the M and the GM Huber estimates appearing in the definition of estimators HUM and HUGM of Table 1, the scale was reestimated at each iteration, using MAD of the current residuals. This algorithm converged smoothly. For Cauchy's estimates appearing in the definition of estimators CAM, CAGM, and CHAM, the simultaneous estimation of location and scale can make the estimates ill-defined (Rivest 1989). To avoid numerical problems, Huber estimates were used as starting values, and the scale parameter was not reestimated. The scale value obtained with the Huber $M(GM)$ estimate was used for calculating the Cauchy $M(GM)$ estimate. All our simulation results are based on 2,000 Monte Carlo repetitions.

5.1 Populations Under Study

Two populations were investigated. The first one, presented in Figure 1, has 34 units and an outlier for a moderate value of x. This population was considered by Rao (1969) in a small-sample Monte Carlo investigation of alternatives to the ratio estimator. Samples of size 4, 8, and 12 were drawn repeatedly from this population. The second population has $N = 235$ and was obtained by perturbing the population presented in Appendix E of Kish (1965): the units with $x \leq 2$ were discarded and the y values of 13 units were changed to create outliers. This population appears in Figure 2; a special printing character highlights the units in the outlier stratum. Samples of size 20, 30, and 40 were drawn from this population.

5.2. Results

Table 2 presents the unconditional coefficient of variation (CV), defined as the square root of the mean squared error divided by $\bar{y}$, and the relative bias (RB), defined as the bias divided by $\bar{y}$ for the seven estimators of Table 1 in six simulations. Table 2 reveals that the ratio of 1-winsorized means WINS is uniformly worse than the ratio estimator RAE; this shows the need to base outlier treatment on model (1.1). The coefficients of variation of resistant estimators are between 10% and 20% lower than the ones of the ratio estimator. The optimal estimators depend on the population under study. For Population 1, M estimators (HUM, CAM, and CHAM) generally have smaller coefficients of variation than do GM estimators (HUGM and CAGM); the opposite is true for Population 2. This was predictable. In Population 2, GM estimators are needed to deal adequately with the outliers, which have large x values. The model-based esti-

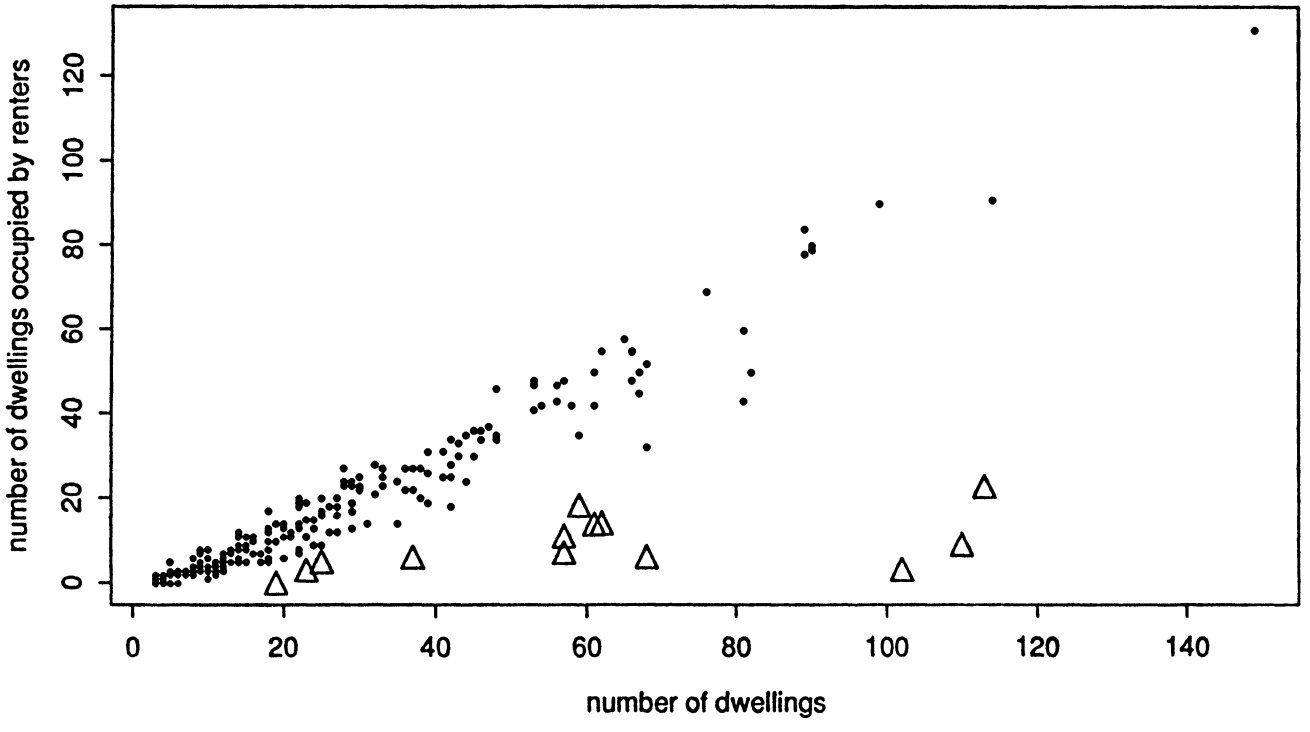

Figure 2. Population 2.

Table 2. Unconditional Simulation Results for Two Populations and Three Sample Sizes

		RAE		HUM		HUGM		CAM		CHAM		CAGM		WINS	
Pop.	n	RB	CV	RB	CV	RB	CV	RB	CV	RB	CV	RB	CV	RB	CV
1	4	.9	23.3	−3.8	16.6	−2.1	18.9	−3.7	16.7	−2.6	17.9	−2.2	18.7	−6.3	24.2
1	8	−.1	13.9	−4.4	10.5	−3.6	10.6	−3.9	10.5	−3.0	10.9	−3.1	11.0	−3.1	15.3
1	12	.4	10.4	−3.8	8.1	−3.2	8.0	−2.8	8.1	−2.1	8.4	−2.2	8.4	−1.1	11.1
2	20	−.2	12.2	2.2	11.0	1.8	10.6	2.2	11.1	2.1	11.1	1.2	10.2	−2.5	12.8
2	30	−.4	10.3	2.5	9.2	2.0	8.8	2.4	9.3	2.2	9.3	1.3	8.6	−2.0	10.8
2	40	−.1	8.2	3.2	7.7	2.8	7.4	2.9	7.7	2.8	7.7	1.8	7.0	−1.4	8.6

NOTE: For each estimator the relative bias (RB) and the coefficient of variation (CV) are presented as percentages.

mators (CAM, CAGM, and CHAM) do not differ much from HUM and HUGM. As could be predicted from the results of Section 3, $\bar{y}_m$ has smaller biases than $\bar{y}_g$.

In these simulations the biases are not important components of the mean squared errors. This contrasts with the simulations of Chambers (1986), where outlier rejection led to estimators having substantial biases. The large sample size ($n = 55$) chosen by Chambers might be the cause of this phenomenon. For the populations of this Monte Carlo study, small biases obtain even if outliers are given weights close to 1. This explains why the performances of Chambers estimator CHAM and CAM are so similar.

The biases of the resistant estimators are much larger than those of the ratio estimator. This result is deceptive. The absolute values of the relative conditional biases of the seven estimators under study are presented in Figure 3 for samples of size 30 drawn out of Population 2. Figure 3 confirms the findings of Section 3: When many outliers are sampled, the bias of the ratio estimator can be quite large, larger than that of its resistant competitors. Noteworthy is the conditional bias of the winsorized estimator WINS, which is slightly larger than the conditional bias of the ratio estimator when two outliers or more are sampled. Simulations not reported here show that when $n = 4$, the bias of the ratio estimator when the outlier is sampled is 40%. Thus from a conditional point of view, the ratio estimator RAE and the winsorized estimator WINS are the most biased estimators.

Confidence intervals for $\bar{y}$ constructed with estimators HUG and HUGM and mean squared error estimator given by formula (2.6) are compared with intervals built with the ratio estimator RAE and variance estimator v_2 in Table 3. The real coverage (CO) of the intervals with nominal coverage rate of 95% and their relative half-length (RHL), defined as the half length divided by $\bar{y}$, are tabulated. The confidence intervals under consideration were constructed using critical values from the t distribution with $n - 1$ degrees of

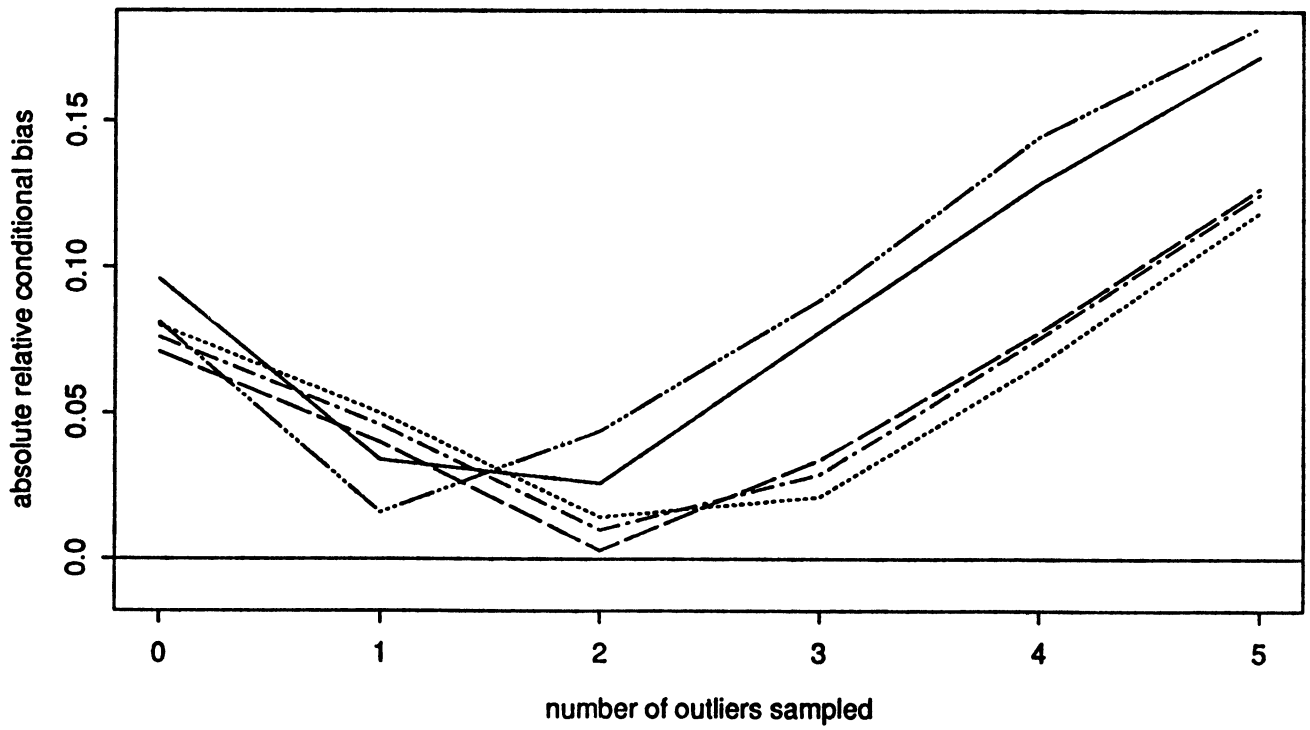

Figure 3. Absolute Relative Conditional Biases for Samples of Size 30 Drawn Out of Population 2. The unconditional probabilities of getting 0, 1, 2, 3, 4, and 5 outliers are given by .19, .32, .26, .14, .06, and .02: ———, RAE; -----, HUM-CAM-CHAM; — - — -, HUGM; — —, CAGM; — --- —, WINS.

Table 3. Coverage (CO) and Relative Half-Length (RHL) of 95% Confidence Intervals Constructed With the Ratio Estimator and Estimator $\hat{\beta}_g$, Presented in Percentages

		RAE		HUM		HUGM	
Pop.	n	CO	RHL	CO	RHL	CO	RHL
1	4	91.1	49.9	88.1	47.4	90.2	48.1
1	8	92.2	26.9	90.3	23.5	91.6	24.2
1	12	91.5	19.7	90.5	17.3	91.8	17.5
2	20	89.1	21.8	91.6	22.6	92.0	21.6
2	30	90.7	17.9	91.9	17.2	92.1	17.2
2	40	92.8	15.6	92.6	15.1	93.0	14.5

freedom. The resistant confidence intervals constructed using HUGM are slightly superior, in terms of coverage and relative half length, to the interval based on v_2, especially in Population 2. Note, however, that all the coverage rates are substantially below their nominal 95% level. The relatively good performance of resistant confidence intervals comes about because in the six simulations under consideration, the bias of resistant estimators accounts for a small percentage of the mean squared error. Rivest and Rouillard (1991) showed that when the squared bias accounts for a large proportion (say more than 20%) of the mean squared error, resistant confidence intervals have poor coverage properties.

6. DISCUSSION

This article has proposed a class of outlier-resistant alternatives to the ratio estimator together with a mean squared error estimator. It has focused on cases when a regression model through the origin provides a good description of the relationship between x and y for most population units. Extensions to multivariate regression models are relatively straightforward.

The proposed resistant estimator is obtained by assigning a weight of $w_i \Sigma_U x_i / \Sigma_s w_i x_i$ to sample unit i, where w_i is defined by (1.6). The resistant weights depend on the variable of interest, because $w_i = w_i(y)$. This is troublesome for multipurpose surveys, where the same set of weights is used for all variables of interest. A possible solution is to calculate resistant sample weight using, for each sample unit, the minimum of the $w_i(y)$'s. This method needs further investigation. The extension of resistant methods to unequal multistage probability sampling is another challenging research problem.

APPENDIX A: PROOF OF PROPOSITION 2.1

Let $\{U_t\}$ be a sequence of finite populations of sizes $\{N_t\}$, with $U_t \supset U_{t-1}$. Let n_t denote the size of a simple random sample s_t taken out of U_t and suppose that n_t goes to ∞. Let $\hat{g}_t(\beta)$ and $g_t(\beta)$ denote expressions (2.1) and (2.2), calculated using sample s_t and population U_t. The estimator $\hat{\beta}_t$ and the parameter β_t are defined as the solutions of equations $\hat{g}_t(\beta) = 0$ and $g_t(\beta) = 0$. (The function ψ is assumed to be continuous and increasing, so that $\hat{\beta}_t$ and β_t are uniquely defined.)

Suppose that the sequence $\{U_t\}$ is such that, as t goes to ∞, (a) $\{\beta_t\}$ converges to some finite limit β_0, (b) $g_t(\beta)$ converges pointwise to a function $g_0(\beta)$ and (c) $g_0(\beta)$ is strictly decreasing around β_0 and satisfies $g_0(\beta_0) = 0$. To prove Proposition 2.1 under these assumptions, consider $P(\hat{\beta}_t - \beta_t > \delta)$ for some $\delta > 0$. Because $\hat{g}_t(\beta)$ is decreasing in β and $\hat{g}_t(\hat{\beta}_t) = 0$, this probability is equal to $P(\hat{g}_t(\beta_t + \delta) > 0)$. It can be expressed as

$$P(\hat{g}_t(\beta_t + \delta) - g_t(\beta_t + \delta) > -g_t(\beta_t + \delta)). \qquad \text{(A1)}$$

Because $E(\hat{g}_t(\beta_t + \delta)) = g_t(\beta_t + \delta)$ and $-g_t(\beta_t + \delta)$ is positive, one can use Chebychev's inequality to show that (A1) is less than or equal to $V(\hat{g}_t(\beta_t + \delta))/g_t(\beta_t + \delta)^2$. As t goes to ∞, $V(\hat{g}_t(\beta_t + \delta))$ is $O(1/n_t)$ as the variance of a sample mean and $\{g_t(\beta_t + \delta)\}$ converges to $g_0(\beta_0 + \delta)$ which is strictly negative according to assumption (c). Thus $\{P(\hat{\beta}_t - \beta_t > \delta)\}$ converges to 0 as t goes to infinity. The proof that $\{P(\hat{\beta}_t - \beta_t < -\delta)\}$ converges to 0 is similar. Thus $\hat{\beta}_t$ is a convergent estimator of β_t.

Note that this proof is a finite population adaptation of the proof of Proposition 2.1 of Huber (1981, p. 48).

APPENDIX B: DERIVATION OF THE BIAS AND THE VARIANCE OF $\bar{y}_m$

The model-based estimator $\bar{y}_m$ defined in Section 2 can be written as $(1 - f)\bar{y}_g + f\bar{y}_d$, where $\bar{y}_d = \bar{x}\hat{\beta}_g + \Sigma_s e_{gi}/n$. Note that $\bar{y}_d$ is equal to $\bar{y}_g$ plus a correction for bias. Thus $\bar{y}_d$ is a design-unbiased estimator of $\bar{y}$, and the bias of $\bar{y}_m$ is equal to $(1 - f)$ times the bias of $\bar{y}_g$. To calculate the large sample variance of $\bar{y}_m$, one can linearize $\bar{y}_d$ using the generalized regression technique of Särndal (1980, 1982). Let $\bar{y}_{d0} = \bar{x}\beta_g + \Sigma_s E_{gi}/n$ be a linear random variable, where E_{gi} is the theoretical resistant residual defined by (2.3). Note that $\bar{y}_d - \bar{y}_{d0} = (\bar{x} - \bar{x}_s)(\hat{\beta}_g - \beta_g)$, which is an $O_p(1/n)$ quantity. Therefore, $\bar{y}_m$ and $(1 - f)\bar{y}_g + f\bar{y}_{d0}$ have the same asymptotic properties. The large sample variance of $(1 - f)\bar{y}g + f\bar{y}_{d0}$ can now be obtained using the asymptotic linearization for $\bar{y}_g$ given in Proposition 2.2.

[*Received July 1990. Revised November 1991.*]

REFERENCES

Beaton, A. E., and Tukey, J. W. (1974), "The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data," *Technometrics*, 16, 147–185.

Bellhouse, D. (1987), "Model-Based Estimation in Finite Population Sampling," *The American Statistician*, 41, 260–262.

Boos, D. D., and Serfling, R. J. (1980), "A Note on Differentials and the CLT and LIL for Statistical Functions, With Applications to *M*-Estimates," *The Annals of Statistics*, 8, 618–624.

Chambers, R. L. (1986), "Outlier Robust Finite Population Estimation," *Journal of the American Statistical Association*, 81, 1063–1069.

Ernst, L. R. (1980), "Comparison of Estimators of the Mean Which Adjust for Large Observations," *Sankhyā, Ser. C*, 42, 1–16.

Fernholz, L. T. (1983), *Von Mises Calculus for Statistical Functionals*, (Lecture Notes in Statistics 19), New York: Springer.

Fuller, W. A. (1991), "Simple Estimators for the Mean of Skewed Populations," *Statistica Sinica*, 1, 137–158.

Hampel, F. R. (1974), "The Influence Curve and its Role in Robust Estimation," *Journal of the American Statistical Association*, 69, 383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. E. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: John Wiley.

Hidiroglou, M. H., and Srinath, K. P. (1981), "Some Estimators of the Population Total from Simple Random Samples Containing Large Units," *Journal of the American Statistical Association*, 76, 690–695.

Huber, P. J. (1964), "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, 35, 73–101.

——— (1981), *Robust Statistics*, New York: John Wiley.

Isaki, C. T., and Fuller, W. A. (1982), "Survey Design Under the Regression Superpopulation Model," *Journal of the American Statistical Association*, 77, 89–96.

Kish, L. (1965), *Survey Sampling*, New York: John Wiley.

Mosteller, F., and Tukey, J. W. (1977), *Data Analysis and Regression*, Reading, MA: Addison-Wesley.

Rao, C. R. (1971), "Some Aspects of Statistical Inference in Problems of Sampling from Finite Populations," in *Foundations of Statistical Inference*, eds. V. P. Godambe and D. A. Sprott, Toronto: Holt, Rinehart & Winston.

Rao, J. N. K. (1969), "Ratio and Regression Estimators," in *New Developments in Survey Sampling,* eds. N. L. Johnson and H. Smith, New York: John Wiley, pp. 213–234.

——— (1985), "Conditional Inferences in Survey Sampling," *Survey Methodology,* 11, 15–31.

Rivest, L. P. (1989), "De l'Unicité des Estimateurs Robustes en Régression Quand le Paramètre d'Échelle et le Paramètre de la Régression Sont Estimés Simultanement," *Canadian Journal of Statistics,* 17, 141–153.

Rivest, L. P., and Rouillard, E. (1991), "*M*-Estimators and Outlier Resistant Alternatives to the Ratio Estimator," in *Proceedings of the 1990 Symposium of Statistics Canada,* pp. 271–285.

Royall, R. M., and Cumberland, W. G. (1981), "An Empirical Study of the Ratio Estimator and Estimators of its Variance," *Journal of the American Statistical Association,* 76, 66–82.

Särndal, C. E. (1980), "On π-Inverse Weighting Versus Best Linear Unbiased Weighting in Probability Sampling," *Biometrika,* 67, 639–650.

——— (1982), "Implications of Survey Design for Generalized Regression Estimation of Linear Functions," *Journal of Statistical Planning and Inference,* 7, 155–170.

Scott, A., and Wu, C. J. (1981), "On Asymptotic Distributions of Ratio and Regression Estimators," *Journal of the American Statistical Association,* 76, 98–102.

Searls, D. T. (1966), "An Estimator Which Reduces Large True Observations," *Journal of the American Statistical Association,* 61, 1200–1204.

Som, R. K. (1973), *A Manual of Sampling Techniques,* London: Heinemann.

Sukhatme, P. V., and Sukhatme, B. V. (1970), *Sampling Theory of Surveys With Applications* (2nd ed.), Rome: Food and Agriculture Organization.

Wright, R. L. (1983), "Finite Population Sampling with Multivariate Auxiliary Information," *Journal of the American Statistical Association,* 78, 879–884.

Wu, C. F., and Deng, L. Y. (1983), "Estimation of the Variance of the Ratio Estimator: An Empirical Study," in *Scientific Inference, Data Analysis, and Robustness,* eds. G. E. P. Box, T. Leonard, and C. F. Wu, New York: Academic Press, pp. 245–277.