OUTLIERS IN SAMPLE SURVEYS

P.D. Ghangurde, Statistics Canada, Ottawa, K1A 0T6, Canada

KEYWORDS: Variance-inflation, outlier robust estimation

1. Introduction

In the literature on regression analysis several approaches for detection and treatment of outliers have been developed. In addition to methods based on the mean-shift and variance-inflation models, estimators based on order statistics such as trimmed and Winsorized means and M-estimators based on robust regression methods are available. Regression diagnostics provide methods for critical examination of models and measures of influence of individual outliers and groups of outliers on estimates of parameters (see Beckman and Cook (1983); Cook and Weisberg (1982)).

The objective of this paper is to develop outlier robust estimators for sample surveys, based on variance-inflation model. This model is a simple extension of superpopulation model often implicitly assumed for the traditional design-based estimators and more explicitly used in the prediction approach. Although these estimators are obtained as optimal estimators of parameters of this model, the results on the bias and variance of these estimators and optimal weight reduction for outliers, presented in this paper, are in the framework of finite population sampling. These outlier robust estimators are not model dependent and have not been evaluated by prediction approach. Outlier robust estimators in finite population sampling based on robust regression and prediction approach have been investigated by Chambers (1986).

The problem of outliers has been considered in the literature on finite population sampling in the context of estimation of mean or total, usually assuming no auxiliary information in estimation. Estimators obtained by methods based on order statistics, such as Winsorization and trimming, and weight reduction, have been investigated by assuming simple random sampling (see e.g. Fuller (1970); Ernst (1980)). It seems that it is not possible to extend methods based on order statistics to sample designs involving stratification and different sampling ratios and non-response rates between strata and unequal probabilities of selection, which result in unequal design weights. Sample surveys are often periodic with rotation samples designed for estimation of changes. Moreover, estimates are needed at several levels such as stratum, group of strata and domains. Because of these features of sample surveys, estimators based on reduction of weights of outliers are more convenient for use in practice.

In Section 2 we introduce the model in which variance is a function of an auxiliary variable x and assume that outliers have inflated variances. The optimal estimators of parameter β are derived by assuming k outliers ($1 \leq k < n$) in a random sample of size n. In Section 3 conditional mean square error of these estimators and optimal weight reduction have been derived by assuming simple random sampling from a finite population. In concluding remarks in Section 4, comments have been made on possible extensions of outlier robust estimation and the problem of estimation of unit variances and covariances of x and y for outliers and non-outliers.

2. Variance-inflation Model

Consider a linear model

$$\underset{n\times 1}{Y} = \underset{n\times p}{X}\ \underset{p\times 1}{\beta} + \underset{n\times 1}{e}, \qquad (2.1)$$

where $e \sim (0, \sigma^2 W)$, σ^2 unknown, W is a diagonal variance-covariance matrix with elements w_i depending on x_i, $i=1, 2, \ldots, n$, Y is a n-vector of responses of y, X is a design matrix of p auxiliary variables each with n observations assumed fixed, β is a p-vector of regression coefficients and e is an error term. Under the model assuming $p=1$, $w_i = x_i^g$, ratio of sample means $\bar{y}/\bar{x}$ and mean of ratios $\bar{r} = \frac{1}{n}\left(\sum_{i=1}^{n} y_i/x_i\right)$ are the best linear unbiased estimators of β for $g=1$ and 2 respectively. The model is appropriate for categorical variables in socio-economic surveys. It is known that values of g for many variables lie in the interval $[1,2]$ and more often closer to 1 than to 2. In practice in multi-purpose sample surveys with several y-variables and an auxiliary variable x possibly used for stratification, ratio estimation, although less than optimal for variables with $g>1$, is often used for all y-variables for convenience of uniform weighting method.

We now consider the variance-inflation model for k outliers ($1 \leq k < n$) which are the last k sample units, without loss of generality. Thus

$$\underset{n\times 1}{Y} = \underset{n\times p}{X}\ \underset{p\times 1}{\beta} + \underset{n\times 1}{e}, \qquad (2.2)$$

where $e \sim (0, \sigma^2 W(k))$ and $W(k)$ is a diagonal variance-covariance matrix with elements w_i, $i=1, 2, \ldots, n-k$ and w_i/w, $i = (n-k+1), \ldots, n$; w is unknown constant ($0 < w \leq 1$). This model is a simple extention of the variance-inflation model considered by Pregibon (1981), Cook, Holschuh and Weisberg (1982) and Thompson (1985).

We consider expression for the best linear unbiased estimator $\hat{\beta}_i(w)$ of β under (2.2) for the case of one outlier, the ith sample unit. Thus

$$\hat{\beta}_i(w) = \hat{\beta} - \frac{(X'W^{-1}X)^{-1}\, X_i'(y_i - \hat{y}_i)(1-w)}{w_i[1 - (1-w)\, V_{ii}]}, \qquad (2.3)$$

where for $p=1$, $(X'W^{-1}X)^{-1}$ and X_i' are scalars $\hat{\beta} = (X'W^{-1}X)^{-1}(X'W^{-1}Y)$ is the estimator of β under (2.1) and $V_{ii} = w_i^{-1}X_i(X'W^{-1}X)^{-1}X_i'$ is the ith diagonal element of variance-covariance matrix $V = V(X\hat{\beta})$, called leverage of X_i. Also, $0 \leq V_{ii} \leq 1$ when $p=1$. For large values of X_i, V_{ii} is close to 1, which makes contribution of i to $\hat{\beta}_i(w)$ very large. The second term on the right hand side of (2.3) shows change due to variance-inflation of ith sample unit. Thus influence of both residual $(y_i - \hat{y}_i)$ and leverage

V_{ij} is reduced due to factor $(1-w)$. For $k(>1)$ outliers and $w_i = x_i$, the estimator $\hat{\beta}_{(i)}(w)$, where (i) represents group of k outliers can also be given in the form which shows weight reduction of outliers, by

$$\hat{R} = \frac{wk\bar{y}_k + (n-k)\,\bar{y}_{n-k}}{wk\bar{x}_k + (n-k)\,\bar{x}_{n-k}}. \tag{2.4}$$

When $w_i = x_i^2$, $\hat{\beta}_{(i)}(w)$ is given by

$$\hat{r} = \frac{wk\bar{r}_k + (n-k)\,\bar{r}_{n-k}}{wk + (n-k)}, \tag{2.5}$$

and when $x_i = 1$ for $i = 1, 2, \ldots, N$, $\hat{\beta}_{(i)}(w)$ is given by

$$\hat{\bar{Y}} = \frac{wk\bar{y}_k + (n-k)\,\bar{y}_{n-k}}{wk + (n-k)}, \tag{2.6}$$

where $\bar{y}_k$ and $\bar{x}_k$ are sample means of outliers, $\bar{y}_{n-k}$ and $\bar{x}_{n-k}$ are sample means of non-outliers, $\bar{r}_k$ and $\bar{r}_{n-k}$ are sample means of ratios $r_i = y_i/x_i$ for outliers and non-outliers respectively. When $w \to 0$, in limit these estimators are $\hat{R} = \bar{y}_{n-k}/\bar{x}_{n-k}$, $\hat{r} = \bar{r}_{n-k}$ and $\hat{\bar{Y}} = \bar{y}_{n-k}$, which can also be obtained as optimal estimators of β under corresponding mean-shift model. The estimator of the population total can be obtained from (2.6) as $N\hat{\bar{Y}}$ and it shows reduction of simple random sampling weight $\frac{N}{n}$ of k outliers by factor w and resulting adjustment of weights of all n units by factor $n/[n-k(1-w)]$. From (2.5) and (2.6) it can be seen that the results for $\hat{r}$ can be obtained from those for $\hat{\bar{Y}}$ by substituting r_i for y_i for all i.

These estimators have been investigated in the following Section 3 using conditional inference given the number of outliers k assuming simple random sampling from a finite population. Although it may be possible to use prediction approach and to obtain model dependent estimators incorporating design weights, the results in this paper have been obtained in the framework of finite population sampling. For detection of outliers in the case of ratio estimation and sample mean Cook's distance seems to have good potential (Cook and Weisberg (1982)).

3. Outlier Robust Estimation

We assume that a finite population of size N contains an unknown proportion P of outliers. The population mean $\bar{X}$ of an auxiliary variable x is assumed known. A simple random sample of size n is drawn without replacement from the population and outliers are identified on the basis of values of some test statistic. Sample units may be identified as outliers if the ith unit values (x_i, y_i) lie in some region of the sample space determined by the test. The test could be based on diagnostics such as Cook's distance (see Ghangurde (1989)).

The outlier robust ratio estimator of the population mean of y's is given by $\hat{\bar{Y}}_R = \hat{R}\bar{X}$, where

$$\hat{R} = \frac{wk\bar{y}_k + (n-k)\,\bar{y}_{n-k}}{wk\bar{x}_k + (n-k)\,\bar{x}_{n-k}} \tag{3.1}$$

is estimator of the population ratio $R = \bar{Y}/\bar{X}$, of population means $\bar{Y}$ and $\bar{X}$. The ratio R can be expressed as

$$R = \frac{P\,\bar{Y}_1 + (1-P)\,\bar{Y}_2}{P\,\bar{X}_1 + (1-P)\,\bar{X}_2}, \tag{3.2}$$

where $\bar{X}_1$ and $\bar{Y}_1$ are unknown population means of outliers, $\bar{X}_2$ and $\bar{Y}_2$ are unknown population means of non-outliers, P is unknown proportion of outliers in the finite population of (x,y). We also assume that σ_{2x}^2, σ_{2y}^2 are unit variances of non-outliers σ_{1x}^2, σ_{1y}^2 are unit variances of outliers, $\sigma_{1x}^2 > \sigma_{2x}^2$ and $\sigma_{1y}^2 > \sigma_{2y}^2$. Also, σ_{1xy} and σ_{2xy} are unit covariances of outliers and non-outliers respectively. By substituting $\bar{Y}_1 = \delta_1\bar{Y}_2$ and $\bar{X}_1 = \delta_2\bar{X}_2$ we have

$$R = \frac{[1 + P(\delta_1 - 1)]\,\bar{Y}_2}{[1 + P(\delta_2 - 1)]\,\bar{X}_2}. \tag{3.3}$$

Since in general, $\delta_1 \neq \delta_2$, $R \neq \bar{Y}_2/\bar{X}_2$. We obtain the conditional bias $B(\hat{R}|k)$ by Taylor series linearization method as

$$B(\hat{R}|k) \doteq \left[\frac{kw\bar{Y}_1 + (n-k)\,\bar{Y}_2}{kw\bar{X}_1 + (n-k)\,\bar{X}_2}\right] - R$$

$$= \frac{(\delta_2 - \delta_1)}{[1-P(1-\delta_2)]}\left\{\frac{P(n-k) - kw(1-P)}{n-k(1-w\delta_2)}\right\}\left(\frac{\bar{Y}_2}{\bar{X}_2}\right). \tag{3.4}$$

When $w=0$, $\hat{R} = \bar{y}_{n-k}/\bar{x}_{n-k}$ and

$$B(\hat{R}|k) = \frac{(\delta_2 - \delta_1)\,P}{(1-P(1-\delta_2))}\left(\frac{\bar{Y}_2}{\bar{X}_2}\right).$$

When $w=1$, $\hat{R} = \bar{y}_n/\bar{x}_n$ with no weight reduction and

$$B(\hat{R}|k) = \frac{(\delta_2 - \delta_1)}{(1-P(1-\delta_2))}\left\{\frac{nP-k}{n-k(1-\delta_2)}\right\}\frac{\bar{Y}_2}{\bar{X}_2}.$$

It can be seen that $f(w) = \frac{P(n-k) - kw(1-P)}{n-k(1-w\delta_2)}$ is a monotonic decreasing function of w in $(0,1)$ and $f(w) \gtrless 0$ according as $P \gtrless \frac{kw}{[n-k(1-w)]}$. Hence, if $P > \frac{k}{n}$, conditional bias of $\hat{R}$ for $w=0$ is in absolute value greater than that of $\hat{R}$ for w in $(0,1)$. If $P < \frac{k}{n}$ absolute conditional bias increases as w increases in $[\frac{P(n-k)}{(1-P)k}, 1]$ and may not be in absolute value lesser than that for $w=0$. This analysis of conditional bias

of $\hat{R}$ is only of theoretical interest; P is not known in practice. By linearization of (3.4) and taking expectation over k, unconditional bias of $\hat{R}$ is given by

$$B(\hat{R}) = \frac{P(\delta_2-\delta_1)}{1+P(\delta_2-1)} \left(\frac{\bar{Y}_2}{\bar{X}_2}\right)(1-P)$$

$$[(1+P) - w(1+P+P\delta_2) + w^2 P\delta_2] + O\left(\frac{1}{n}\right). \quad (3.5)$$

When w=0,

$$B(\hat{R}) = \frac{P(\delta_2-\delta_1)(1-P^2)}{(1+P(\delta_2-1))}\left(\frac{\bar{Y}_2}{\bar{X}_2}\right) \gtrless 0 \text{ according as } \delta_2 \gtrless \delta_1.$$

Also when w=1, $B(\hat{R}) = 0$ to $O(\frac{1}{n})$. The results for outlier robust sample mean $\hat{\bar{Y}}$ given in (2.6) can be obtained from above results by substitution $x_i=1$ for i=1, 2, ..., N, and $\bar{x}_1 = \bar{x}_2 = \delta_2 = 1$.

By Taylor series linearization method

$$V(\hat{R}|k) \doteq \left(\frac{\tilde{R}kw}{\tilde{X}}\right)^2 V(\bar{x}_k|k) + \left(\frac{(n-k)\tilde{R}}{\tilde{X}}\right)^2 V(\bar{x}_{n-k}|k)$$
$$+ \left(\frac{kw}{\tilde{X}}\right)^2 V(\bar{y}_k|k) + \left(\frac{n-k}{\tilde{X}}\right)^2 V(\bar{y}_{n-k}|k)$$
$$- \frac{2\tilde{R}k^2w^2}{\tilde{X}^2} Cov(\bar{x}_k,\bar{y}_k|k)$$
$$- \frac{2\tilde{R}(n-k)^2}{\tilde{X}^2} Cov(\bar{x}_{n-k}, \bar{y}_{n-k}|k), \quad (3.6)$$

where $\tilde{X} = [kw\bar{X}_1+(n-k)\bar{X}_2]$, $\tilde{Y} = [kw\bar{Y}_1+(n-k)\bar{Y}2]$ and $\tilde{R} = \tilde{Y}/\tilde{X}$.

In order to obtain an optimal value of w the conditional mean square error of $\hat{R}$ given k can be minimized as a function of w. However, if conditional bias ratio is small an optimum w obtained by minimizing conditional variance of $\hat{R}$ can be used. By equating derivative of conditional variance with respect to w to zero we have an equation of third degree in w involving several parameters. An analytical solution cannot be obtained unless estimates are substituted for several parameters involved. An alternative is to minimize limiting value of $V(\hat{R}|k)$ for infinite N, which is the same as minimizing superpopulation variance of $\hat{R}$ under the variance-inflation model with μ_x and μ_y, means of x and y respectively. It may be noted that under the model, bias of $\hat{R}$ is zero. The optimum w is given by

$$w = \frac{\sigma_{2x}^2\mu_y^2 + \sigma_{2y}^2\mu_x^2 - 2\sigma_{2xy}\mu_x\mu_y}{\sigma_{1x}^2\mu_y^2 + \sigma_{1y}^2\mu_x^2 - 2\sigma_{1xy}\mu_x\mu_y}. \quad (3.7)$$

This expression for optimum w obtained by assuming infinite N is simple and needs estimation of fewer parameters. The means μ_x and μ_y can be estimated by sample means $\bar{x}_n$ and $\bar{y}_n$ respectively. Although σ_{1x}^2, σ_{1y}^2 and σ_{1xy} can be estimated from sample outliers, more efficient estimation can be done by an extension of Minimum Norm Quadratic Unbiased Estimation (MINQUE) method (see Rao (1970)) to estimation of heteroscedastic variances and covariances when x and y are random.

The optimum w, although obtained under the variance-inflation model, can still be used in finite population sampling and mean square error can be estimated for values of w close to the optimum given by (3.7) to decide on choice of another value of w. However, due to instability of estimates of bias, it may be preferable to use w given in (3.7).

In the case of sample mean $\hat{\bar{Y}}$ the optimum w_o can be obtained by minimizing mean square error and is given by

$$w_o = \frac{(1 - \frac{n-k}{N(1-P)})\sigma_{2y}^2 + (n-k)P(\delta_1-1)^2\bar{Y}_2^2}{(1 - \frac{k}{NP})\sigma_{1y}^2 + k(1-P)(\delta_1-1)^2\bar{Y}_2^2}. \quad (3.8)$$

In the literature several estimators obtained by weight reduction and which can be considered as linear combinations of $\bar{y}_k$ and $\bar{y}_{n-k}$ have been investigated (see e.g. Hidiroglou and Srinath (1981)). Their estimator $\hat{\hat{\bar{Y}}} = \frac{rk}{N}\bar{y}_k + (1 - \frac{rk}{N})\bar{y}_{n-k}$ with optimal $r=r_o$ obtained by minimizing mean square error of $\hat{\hat{\bar{Y}}}$ is the same as $\hat{\bar{Y}}$ with optimal $w=w_o$, since $kw_o/((n-k) + kw_o) = r_ok/N$. The important advantage of estimators suggested above is that the weight reduction can be estimated from survey data and extensions to other sample designs seem possible. Thus w_o can be estimated by

$$\hat{w}_o = \frac{(1 - \frac{n}{N})\hat{\sigma}_{2y}^2 + k(1 - \frac{k}{n})(\bar{y}_k - \bar{y}_{n-k})^2}{(1 - \frac{n}{N})\hat{\sigma}_{1y}^2 + k(1 - \frac{k}{n})(\bar{y}_k - \bar{y}_{n-k})^2}. \quad (3.9)$$

Although, it is possible to obtain estimate $\hat{\sigma}_{1y}^2$ from sample outliers, the estimator is unstable for small values of k. Alternative methods for estimation of heteroscedastic variances have been proposed in the literature (see Kleffe (1977)). The MINQUE was developed for $p \geq 1$ and n unequal variances (see Rao (1970)).

Assuming that the last k units are outliers, estimators of unit variances obtained by MINQUE method are given by

$$\hat{\sigma}_{1y}^2 = \frac{n\sum_{i=n-k+1}^{n}(y_i-\bar{y}_n)^2}{k(n-2)} - \frac{\sum_{i=1}^{n}(y_i-\bar{y}_n)^2}{(n-1)(n-2)},$$

$$\hat{\sigma}_{2y}^2 = \frac{n\sum_{i=1}^{n-k}(y_i-\bar{y}_n)^2}{(n-k)(n-2)} - \frac{\sum_{i=1}^{n}(y_i-\bar{y}_n)^2}{(n-1)(n-2)}. \quad (3.10)$$

These estimators are more efficient than the usual estimators of σ_{1y}^2 and σ_{2y}^2 and also give more efficient estimators of β as compared to those obtained by weighted least squares (Rao and Subrahmaniam (1971)).

4. Concluding Remarks

The variance-inflation model seems to be an appropriate model for outliers in sample surveys. In the optimal estimator based on the model influence of outliers with large residuals and high leverage is reduced due to factor $(1-w)$. Under appropriate assumptions about dependence of variance on x, it reduces to outlier robust estimators $\hat{R}$, $\hat{r}$ and $\hat{\bar{Y}}$, in which weights of outliers are reduced. In the case of ratio estimator, although the problem of determination of optimal weight was simplified by assuming an infinite population, it could still be used in the case of finite population sampling. Although it may be possible to use prediction approach and to obtain model dependent estimators incorporating design weights, conditional inference in finite population sampling framework shows that these outlier robust estimators have desirable properties. Extension of these estimators to stratified sampling incorporating design weights is being investigated.

Implicit in the variance-inflation model for outliers is the assumption that superpopulation is a mixture of distributions with the same mean but different variances. In the case of sampling from mixture distributions belonging to the exponential family, maximum likelihood estimators of parameters have been obtained and sufficient conditions have been established for outliers to occur in samples from these distributions (see e.g. Gather and Kale (1988)). It would be of interest to investigate into the possibility of establishing similar conditions for occurrence of outliers in sampling from finite populations which have mixture distributions.

Although unit variances and covariances for outliers can be estimated from outliers in the sample, the estimator could be unstable for small values of k. In the case of ratio estimation an extension of MINQUE method for estimation of heteroscedastic variances and covariances of x and y is needed. The problem of variance estimation was not discussed since it does not involve any new methodology, once w is treated as a component of weights of outliers.

References

Beckman, R.J. and Cook, R.D. (1983), "Outliers," Technometrics, Vol. 25, No. 2, pp. 119-149.

Chambers, R.L. (1986), "Outlier robust finite population estimation," JASA, Vol. 81, No. 396, pp. 1063-1069.

Cook, R.D. Holschuh, N. and Weisberg, S. (1982), "A note on an alternative outlier model," J.R.S.S., B, 44, No. 3, pp. 370-376.

Cook, R.D. and Weisberg, S. (1982), Residuals and Influence in Regression, Chapman and Hall.

Ernst, L.R. (1980), "Comparisons of estimators of the mean which adjust for large observations," Sankhya, C, 42, pp. 1-16.

Fuller, W.A. (1970), "Simple estimators of the mean of skewed populations," Technical Report, Department of Statistics, Iowa State University.

Gather, U. and Kale B.K. (1988), "Maximum likelihood estimation in the presence of outliers," Communications in Statistics, Theory and Methods, 17 (11), pp. 3767-3784.

Ghangurde P.D. (1989), "Outlier robust estimation in finite population sampling," presented at the Statistical Society of Canada meeting, Ottawa.

Hidiroglou, M.A. and Srinath, K.P. (1981), "Some estimators of a population total from simple random samples containing large units," JASA, Vol. 78, No. 375, pp. 690-695.

Kleffe, J. (1977), "Optimal estimation of variance components," Sankhya, Vol. 39, B, pp. 211-244.

Pregibon, D. (1981)' "Logistic regression diagnostics," The Annals of Statistics, Vol. 9, No. 4, pp. 705-724.

Rao, C.R. (1970), "Estimation of heteroscedastic variances in linear models," JASA, Vol. 65, No. 329, pp. 161-172.

Rao, J.N.K. and Subrahmaniam, K. (1971), "Combining independent estimators and estimation in linear regression with unequal variances," Biometrics, 27, pp. 971-993.

Thompson, R. (1985), "A note on restricted maximum likelihood with an alternative outlier model," J.R.S.S. (B), 47, No. 1, pp. 53-55.