Journal of Official Statistics, Vol. 16, No. 2, 2000, pp. 133–138

A Note on Jackknife Variance Estimation for the General Regression Estimator

Pierre Duchesne[1]

We derive in this article explicit jackknife variance estimators of the general regression estimator (*GREG*) using the random group technique. A corrected version is proposed that removes a large part of the positive model bias. A small simulation is presented.

Key words: Confidence interval; jackknife; regression estimator; survey sampling; variance estimation.

1. Introduction

Let $U = \{1, \ldots, N\}$ be a finite population. Suppose that we know the total T_x of an auxiliary variable x of dimension p. A sample s is observed from a πps sampling plan. Let π_k and π_{kl} be the first and second inclusion probabilities, respectively. Our goal is to estimate the total $T_y = \Sigma_U\, y_k$ of a positive variable y with $\{(x_k, y_k),\, k \in s\}$ and T_x.

The general regression estimator (*GREG*) of T_y is given by

$$\hat{T}_{GREG} = \sum_s d_k g_{ks} y_k$$

where

$$g_{ks} = 1 + (T_x - \hat{T}_{x\pi})' \left(\sum_s d_k x_k x_k' / c_k \right)^{-1} x_k / c_k$$

is the 'g-weight', $d_k = \pi_k^{-1}$ is the sampling weight, $\hat{T}_{x\pi} = \Sigma_s x_k / \pi_k$ is the Horvitz-Thompson estimator of T_x, and c_k is chosen by the user. Särndal (1996) discusses the choice of c_k. The asymptotic variance AV for the *GREG* is given by

$$AV(\hat{T}_{GREG}) = \sum\sum_U \Delta_{kl} \check{E}_k \check{E}_l$$

where $E_k = y_k - x_k' B, B = (\Sigma_U x_k x_k' / c_k)^{-1} \Sigma_U x_k y_k / c_k, \Delta_{kl} = \pi_{kl} - \pi_k \pi_l$ and $\check{E}_k = E_k / \pi_k$.

Since the asymptotic variance is an ordinarily unknown quantity, Särndal et al. (1989)

[1] Service de l'enseignement des méthodes quantitatives de gestion, École des Hautes Études Commerciales 3000 chemin côte ste-Catherine, Montreall, Québec, Canada, H3T-2A7. E-mail: pierre.duchesne@hec.ca
Acknowledgments: I would like to thank Professor Carl-Erik Särndal, Professor Bengt Rosén, the Chief Editor, and two anonymous referees for helpful comments.

suggested the following g-weighted variance estimator given by

$$\hat{V}_g = \hat{V}_g(\hat{T}_{GREG}) = \sum\sum_s \check{\Delta}_{kl}(g_{ks}\check{e}_{ks})(g_{ls}\check{e}_{ls}) \tag{1}$$

where $e_{ks} = y_k - x_k'\hat{B}_s$, $\hat{B}_s = (\Sigma_s d_k x_k x_k'/c_k)^{-1}\Sigma_s d_k x_k y_k/c_k$, $\check{\Delta}_{kl} = \Delta_{kl}/\pi_{kl}$ and $\check{e}_{ks} = e_{ks}/\pi_k$. With $\hat{V}_g$, we can construct a confidence interval for T_y given by $\hat{T}_{GREG} \pm z_{1-\alpha/2}[\hat{V}_g]^{1/2}$, that is expected to be valid approximately to the $1-\alpha$ confidence level.

The jackknife technique is another popular method to obtain a variance estimator. That method is described in Wolter (1985) and Särndal et al. (1992, chap. 11). We derive in the next section explicit jackknife variance estimators of the *GREG*. A corrected version is proposed that removes a large part of the positive model bias in Section 3. A small simulation is given in Section 4 to illustrate the proposed estimator. We conclude with a discussion in Section 5.

2. Jackknife Variance Estimation

In this section, we obtain explicit formulas for the jackknife variance estimators of the *GREG*. Let the sample be divided into A groups of size m partitioning the sample, where $Am = n$, where n is the sample size. The two jackknife variance estimators advocated by Särndal et al. (1992) are given by

$$\hat{V}_{JK1} = \frac{A-1}{A}\sum_{a=1}^{A}(\hat{T}_{GREG}(a) - \hat{T}_{GREG,JK})^2$$

$$\hat{V}_{JK2} = \frac{A-1}{A}\sum_{a=1}^{A}(\hat{T}_{GREG}(a) - \hat{T}_{GREG})^2$$

where $\hat{T}_{GREG}(a)$ is the *GREG* calculated without the group a and

$$\hat{T}_{GREG,JK} = \frac{1}{A}\sum_{a=1}^{A}\hat{T}_{GREG}(a)$$

Since the two formulas $\hat{V}_{JK1}$ and $\hat{V}_{JK2}$ are related by the following relation

$$\hat{V}_{JK1} = \hat{V}_{JK2} - (A-1)(\hat{T}_{GREG} - \hat{T}_{GREG,JK})^2 \tag{2}$$

we can conclude that $\hat{V}_{JK1} \leq \hat{V}_{JK2}$ and it is easy to pass from one form to the other. In practice, the two formulas give very similar results.

We consider in the following the maximal number of groups, that is the case $A = n, m = 1$. See the remark 11.5.3 of Särndal et al. (1992, pp. 441–442). With that hypothesis, we now use the random group technique to obtain explicit formulas for $\hat{V}_{JK2}$ and for $\hat{V}_{JK1}$ using Expression (2). Under that technique, we suppose that conditional on s, each group $\{i\}$ is obtained by simple random sampling. In that case, there are n random subsamples $s_{(i)} = s - \{i\}$. The inclusion probability that the unit k will be in the final subsample, denoted $\pi_k(i)$, is $\pi_k(i) = (n-1)/n\pi_k$. Using that technique, we obtain the following result:

PROPOSITION 1. *The jackknife variance estimator $\hat{V}_{JK2}$ of the GREG estimator is* *given by*

$$\hat{V}_{JK2} = \frac{n}{n-1}\sum_s(\tilde{g}_{is}\breve{e}_{is} - n^{-1}\hat{T}_{e\pi})^2 \tag{3}$$

where $\tilde{g}_{is} = (g_{is} - n^{-1}T_x'M_s^{-1}x_i/c_i)/(1-h_i),\ \hat{T}_{e\pi} = \Sigma_s e_{ks}/\pi_k, h_i = d_i x_i' M_s^{-1} x_i/c_i,$
$e_{is} = y_i - x_i'\hat{B}_s,\ M_s = \Sigma_s d_k x_k x_k'/c_k$

Proof. Let $s_{(i)}$ denote the sample s without unit i. Since

$$\begin{aligned}\hat{T}_{GREG} &= \sum_s d_k g_{ks} y_k \\ &= \hat{T}_{y\pi} + (T_x - \hat{T}_{x\pi})'\hat{B}_s\end{aligned}$$

the *GREG* without unit i can be written as

$$\hat{T}_{GREG}(i) = \hat{T}_{y\pi}(i) + (T_x - \hat{T}_{x\pi}(i))'\hat{B}_s(i)$$

where $\hat{T}_{y\pi}(i) = \sum_{s(i)} y_k/\pi_k(i)$ and similarly for $\hat{T}_{x\pi}(i)$, and

$$\hat{B}_s(i) = \left\{\sum_{s(i)} x_k x_k'/(c_k\pi_k(i))\right\}^{-1}\sum_{s(i)} x_k y_k/(c_k\pi_k(i))$$

With some algebra, we can show that

$$\hat{T}_{GREG}(i) = \hat{T}_{GREG} + \frac{1}{n-1}\sum_s e_{ks}/\pi_k - \frac{n}{n-1}(g_{is} - n^{-1}T_x'M_s^{-1}x_i/c_i)\breve{e}_{is}/(1-h_i)$$

Finally, using the relation 2, we obtain the following corollary

COROLLARY 1.

$$\hat{V}_{JK1} = \hat{V}_{JK2} - \frac{1}{n-1}(\hat{T}_{e\pi} - \sum_s \tilde{g}_{is}\breve{e}_{is})^2$$

It is interesting to note that with the exception of the factor $\tilde{g}_{is}$, Formula 3 looks like the simplified variance estimator $\hat{V}_0$ of Särndal et al. (1992, ex. (11.2.6), p. 422), where e_{is} now replaces y_i.

3. Corrected Estimator for the Model Bias

Särndal et al. (1992) note that, with the exception of the Horvitz-Thompson estimator, there are no exact results concerning the properties of the jackknife variance estimator. We study in this section the model bias of the Formula 3. Let the ξ regression model for $y_1, \ldots, y_N$ be given by

$$y_k = x_k'\beta + \epsilon_k$$

where the ϵ_k are independent under the model, and such that $E_\xi(\epsilon_k) = 0, V_\xi(\epsilon_k) = \sigma^2 c_k$, where E_ξ and V_ξ indicate the mean and variance under the model. We assume like Särndal et al. (1989) that for some λ independent of k

$$c_k = \lambda' x_k \tag{4}$$

and we similarly define the *prototype* $\hat{V}^*$ for $\hat{V}_g$ as

$$\hat{V}^* = \sum\sum_s \breve{\Delta}_{kl}(g_{ks}\breve{\epsilon}_k)(g_{ls}\breve{\epsilon}_l)$$

Särndal et al. (1992, p. 232) give several examples of variance structures satisfying Condition (4). Note that under that condition we have $\Sigma_s e_{ks}/\pi_k = 0$. We now recall a result that will be useful in the sequel.

LEMMA 1. *Under the model ξ, for any given realized sample s, the model mean, the model mean squared error and the relative model bias of the prototype $\hat{V}^*$ are given by*

$$(i) \quad E_\xi(\hat{V}^*) = \sigma^2\left(\sum_s g_{ks}^2 c_k/\pi_k^2 - \sum_U g_{ks}c_k\right)$$

$$(ii) \quad MSE_\xi(\hat{T}_{GREG}) = E_\xi(\hat{T}_{GREG} - T_y)^2 = \sigma^2\left(\sum_s g_{ks}^2 c_k/\pi_k^2 - \sum_U c_k\right)$$

$$(iii) \quad RMB_\xi(\hat{V}^*) = \frac{E_\xi(\hat{V}^*) - MSE_\xi(\hat{T}_{GREG})}{MSE_\xi(\hat{T}_{GREG})} = \frac{-\sum_U (g_{ks}-1)c_k}{\sum_s g_{ks}^2 c_k/\pi_k^2 - \sum_U c_k}$$

Proof. See Särndal et al. (1989).

We study properties of the jackknife variance estimator *prototype*. Under Condition (4), it is given by

$$\hat{V}_{JK2}^* = \frac{n}{n-1}\sum_s \tilde{g}_{ks}^2\breve{\epsilon}_k^2$$

Under the model, note that $E_\xi(\hat{V}_{JK2}^*) = n/(n-1)\sigma^2\Sigma_s\tilde{g}_{ks}^2 c_k/\pi_k^2 = A_s$. Suppose that all h_i are negligible (their sample mean is $n^{-1}\Sigma_s h_i = p/n$). Then $g_{ks} \approx \tilde{g}_{ks}$ and A_s will be of the same order that the first term in the right member of (i) in the lemma. We have approximately

$$E_\xi(V_{JK2}^*) - E_\xi(\hat{V}^*) \approx \sigma^2\sum_U g_{ks}c_k$$

suggesting that the jackknife variance estimator overestimates the true variance, which is well-known. The relative model bias of the jackknife variance estimator can also be calculated using (ii) in the lemma:

$$RMB_\xi(\hat{V}_{JK2}^*) \approx \frac{\sum_U c_k}{\sum_s g_{ks}^2 c_k/\pi_k^2 - \sum_U c_k} \tag{5}$$

Looking at the numerator of 5, the relative model bias $RMB_\xi(\hat{V}_{JK2}^*)$ is expected to be more important than for $\hat{V}_g$. It may however be small if the first term in the denominator dominates the second term in Formula (5). It can be seen that under simple random sampling, if the sampling fraction $f = n/N$ is small, then $RMB_\xi(\hat{V}_{JK2}^*)$ can be negligible.

However, since the positive bias may be more important in practice, we consider the following modification:

$$\hat{V}_{JK3}^* = \frac{n}{n-1}\sum_s (1-\pi_k)\tilde{g}_{ks}^2\breve{\epsilon}_k^2 \tag{6}$$

We can justify that modification with the following argument. Under the model, we now obtain

$$E_{\xi}(\hat{V}^{*}_{JK3}) = A_s - \frac{n}{n-1}\sigma^2 \sum_s \tilde{g}^2_{ks} c_k/\pi_k \tag{7}$$

If $g_{ks} \approx \tilde{g}_{ks}$, then the second term of that right member of Expression (7) will be of the same order as that for the second member of (i) in the lemma since $\Sigma_s g^2_{ks} c_k/\pi_k = \Sigma_U g_{ks} c_k$ (see Särndal et al. (1989, Expression 5.6)). Note that in the case of the simple random sampling, $\hat{V}^{*}_{JK3}$ is simply $\hat{V}^{*}_{JK2}$ affected by the finite population correction. However, our analysis is in the more general setting of a πps sampling plan. Wolter (1985) discusses some methods to remove the bias of the jackknife variance estimator in the Horvitz-Thompson case. See also Särndal et al. (1992, pp. 439–440). These ideas are applied here to the *GREG*.

4. Illustration

We consider a small Monte Carlo simulation for the variables $y = RMT85 \times 10^{-4}$, $x_1 = CS82$ and $x_2 = SS82$ for the $MU281$ population (of size $N = 281$) in Särndal et al. (1992). This study is a complement to the one in Särndal et al. (1992, pp. 278–280). Like them, we carried out 5,000 repeated simple random samples, each with size $n = 100$. The main objective of the simulation study is to evaluate coverage properties of confidence intervals at the 95% level

$$\hat{T} \pm 1.96[\hat{V}(\hat{T})]^{1/2}$$

where $\hat{T}$ is the *GREG* estimator, and $\hat{V}(\hat{T})$ is a variance estimator. We consider the *GREG* estimator with only x_1, the *GREG* estimator with only x_2, and the *GREG* estimator with x_1 and x_2. We always included an intercept and let $c_k \equiv 1$ throughout the study. We consider the variance estimator given in Formula (1), the jackknife variance estimators given in Formula (3) and the corrected Version (6). Formula (1) becomes under simple random sampling

$$\hat{V}_g = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\frac{\sum_s g^2_{ks} e^2_k}{n-1}$$

Results are presented in Table 1, where $\bar{\hat{T}}$ and $S^2_{\hat{T}}$ are the sample mean and sample variance of the 5,000 estimates $\hat{T}$; $\bar{\hat{V}}_g$, $\bar{\hat{V}}_{JK2}$ and $\bar{\hat{V}}_{JK3}$ are the sample means of the 5,000 variance estimates, $\hat{V}_g$, $\hat{V}_{JK2}$ and $\hat{V}_{JK3}$, respectively; and ECR_g, ECR_{JK2} and ECR_{JK3} are the respective coverage rates for the *GREG* based on $\hat{V}_g$, $\hat{V}_{JK2}$, $\hat{V}_{JK3}$, respectively. The final column gives the approximate variance for $\hat{T}$ given by

$$AV(\hat{T}) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\frac{\sum_U E^2_k}{N-1}$$

The results in Table 1 show that in our experience, $\hat{V}_{JK3}$ gives good coverage properties and in the three cases the variance of the 5,000 estimates is close to $\bar{\hat{V}}_{JK3}$. In that limited simulation, $\hat{V}_{JK3}$ seems to compare reasonably well with $\hat{V}_g$.

5. Discussion

In this article, explicit jackknife *GREG* variance estimators are exhibited. These formulas give new examples of the well-known rule of thumb that jackknifing leads to

Table 1. Results of the simulation

Estimator	$\bar{\bar{T}}$	$S^2_{\hat{T}}$	$\bar{\hat{V}}_g$	ECR_g	$\bar{\hat{V}}_{JK2}$	ECR_{JK2}	$\bar{\hat{V}}_{JK3}$	ECR_{JK3}	AV
$\hat{T}_{GREG}(x1)$	5.31	0.122	0.115	0.937	0.189	0.977	0.121	0.942	0.116
$\hat{T}_{GREG}(x2)$	5.30	0.124	0.118	0.934	0.191	0.978	0.123	0.941	0.117
$\hat{T}_{GREG}(x1, x2)$	5.31	0.056	0.052	0.929	0.088	0.978	0.057	0.939	0.052

Note: The total T_y is 5.315.

overestimation of the variance. An idea for possible overestimation ''correction'' is presented, leading to a modified estimator. In the numerical illustration, we obtain reasonable properties with the corrected version. We do not claim that the proposed estimator is superior to other estimators. In fact, in a vast majority of situations occurring in practice, $\hat{V}_g$ may be preferable. Jackknife estimators are perhaps applicable to exceptional situations (shortage of time, one-time use, etc). It seems to appear that their chief merit is that they require less programming efforts. For example, there is no need to evaluate all the π_{kl} as in $\hat{V}_g$. See Särndal (1996) for a discussion of this problem. However, if jackknife variance estimators for the *GREG* are needed, it is hoped that Proposition 1 will be useful.

6. References

Särndal, C.E. (1996). Efficient Estimators with Simple Variance in Unequal Probability Sampling. Journal of the American Statistical Association, 91, 1289–1300.

Särndal, C.E., Swensson, B., and Wretman, J.H. (1989). The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of the Finite Population Total. Biometrika, 76, 527–537.

Särndal, C.E., Swensson, B., and Wretman, J.H. (1992). Model Assisted Survey Sampling. New-York: Springer-Verlag.

Wolter, K.M. (1985). Introduction to Variance Estimation. New-York: Springer-Verlag.

Received October 1998
Revised June 1999