

ESTIMATING FUNCTION RESAMPLING VARIANCE ESTIMATORS UNDER STRATIFIED MULTISTAGE SAMPLING

M. Tausi and J.N.K Rao¹

ABSTRACT

Variance estimation for the poststratified estimator and the generalised regression estimator of a total under stratified multistage sampling is considered. Customary resampling methods (jackknife, balanced repeated replication and bootstrap) require the inversion of a $P \times P$ matrix for each resample, where P is the number of auxiliary variables. This could lead to illconditioned matrices for some of the resamples. We apply the estimating function resampling method of Hu and Kalbfleisch (2000) to obtain variance estimators, using jackknife resampling. This method avoids repeated inverses, and we show that the resulting variance estimator is identical to a jackknife linearization variance estimator of Yung and Rao (1996). We extend the results to cover parameters defined as solutions of census estimating equations.

KEY WORDS: Census estimating equation; Generalised regression estimator; Jackknife.

RÉSUMÉ

L'estimation de la variance d'un estimateur poststratifié et d'un estimateur général de régression sous échantillonnage stratifié à niveaux multiples est considéré. Les méthodes habituelles de rééchantillonnage (jackknife, répétition répétée équilibrée et bootstrap) requièrent l'inversion d'une matrice p fois p pour chaque échantillon, où p représente le nombre de variables auxiliaires. Ceci peut mener à des matrices singulières pour certains échantillons. Nous appliquons la méthode de rééchantillonnage de la fonction d'estimation de Hu et Kalbfleisch (1999) pour obtenir les estimateurs de la variance. Cette méthode évite la répétition d'inverses et nous démontrons que l'estimateur de la variance obtenu est identique à celui obtenu par la méthode de linéarisation du jackknife de Yung et Rao (1996). Nous étendons ces résultats pour couvrir les paramètres définis comme solutions d'équations d'estimation de recensements.

Mots Clé: Équations d'estimation de recensements; estimateur général de régression; jackknife

1. INTRODUCTION

In complex large scale surveys, calibration methods such as poststratification and generalised regression are commonly employed to improve the precision of sample estimates or to benchmark to known population totals. In this paper, we consider variance estimation for the poststratified estimator and the generalised regression estimator of a total under stratified multistage sampling. In particular, we apply the estimating function resampling method of Hu and Kalbfleisch (2000) to the jackknife procedure and we show that the resulting variance estimator is identical to a jackknife linearization variance estimator of Yung and Rao (1996).

Section 2 outlines the stratified multistage sampling design for the estimation of a total, including the

calibration estimator where auxiliary data is used to improve the precision of the sample estimates or to ensure consistency with known population totals. Section 3 compares the jackknife linearization variance estimator of Yung and Rao (1996) with the proposed Estimating Function (EF) jackknife variance estimator. In Section 4, we extend the EF jackknife method to cover parameters defined as solutions of census estimating equations.

2. STRATIFIED MULTISTAGE SAMPLING

In the stratified multistage design, the population under consideration is stratified into L strata and from each stratum h , $n_h \geq 2$ clusters are selected, independently across the strata. We further assume that subsampling within the sampled clusters is

¹ M. Tausi, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6, Canada; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6, Canada.

performed to ensure unbiased estimation of cluster totals, Y_{hi} , $i = 1, \dots, n_h$, $h = 1, \dots, L$.

2.1 Estimation of a total Y

From the specification of the survey design, we obtain the basic design weights, w_{hik} , associated with the sample element $(hik) \in s$, where s denotes the sample of elements. An unbiased estimator of the population total Y is defined by the basic estimator

$$\hat{Y} = \sum_s w_{hik} y_{hik}, \quad (1)$$

where y_{hik} represent the characteristic of interest associated with $(hik) \in s$.

Assuming that the clusters are sampled with replacement, we obtain an estimator of the variance of \hat{Y} , which is given by

$$v(\hat{Y}) = v(y_{hi}) = \sum_h n_h^{-1} (n_h - 1)^{-1} \sum_i (y_{hi} - \bar{y}_h)^2, \quad (2)$$

where $y_{hi} = \sum_k (n_h w_{hik}) y_{hik}$ and $\bar{y}_h = n_h^{-1} \sum_i y_{hi}$.

The operator notation $v(y_{hi})$ indicates that $v(\hat{Y})$ depends only on the y_{hi} 's.

2.2 Calibration

Auxiliary data is commonly used at the estimation stage to ensure consistency with known population totals \mathbf{X} of auxiliary variables \mathbf{x} . The basic design weights, w_{hik} , are then adjusted by the calibration procedure to obtain calibration weights, $w_{hik}^* = w_{hik} a_{hik}$, where a_{hik} is the adjustment factor, with the restriction that $\sum_s w_{hik}^* \mathbf{x}_{hik} = \mathbf{X}$. The calibration estimator of the population total Y is given by

$$\hat{Y}_r = \sum_s w_{hik}^* y_{hik}. \quad (3)$$

Two commonly used calibration procedures are the poststratified estimator and the Generalised Regression (GREG) estimator. For the poststratified estimator, we assume that the population is partitioned into C nonoverlapping poststrata with known population

counts ${}_c M$, $c = 1, \dots, C$. The adjustment factor is then defined by

$$a_{hik} = \frac{{}_c M}{\sum_{{}_c s} w_{hik}} \text{ if } (hik) \in {}_c s,$$

where ${}_c s$ is the sample of elements from poststratum c . If we have two or more poststratifiers with known marginal totals \mathbf{X} , we can apply the GREG estimator of Y by using P -dimensional auxiliary indicator variables \mathbf{x}_{hik} and defining the adjustment factor by $a_{hik} = \mathbf{X}^T \hat{\mathbf{A}}^{-1} \mathbf{x}_{hik}$ where $\hat{\mathbf{A}} = \sum_s w_{hik} \mathbf{x}_{hik} \mathbf{x}_{hik}^T$. This leads to the projection estimator $\hat{Y}_r = \mathbf{X}^T \hat{\mathbf{B}}$, where $\hat{\mathbf{B}} = \hat{\mathbf{A}}^{-1} (\sum_s w_{hik} \mathbf{x}_{hik} y_{hik})$ is the vector of estimated regression coefficients.

3. RESAMPLING VARIANCE ESTIMATORS

We present the EF jackknife variance estimator of \hat{Y}_r and compare it with the jackknife linearization variance estimator of Yung and Rao (1996). In order to apply the jackknife method in stratified multistage sampling, we delete one cluster at a time.

3.1 Delete one cluster jackknife

To calculate the jackknife variance estimator of \hat{Y}_r , we first define the basic jackknife weights, $w_{hik(gj)} = w_{hik} b_{gj}$, where

$$b_{gj} = \begin{cases} 0 & \text{if } hi = gj \\ \frac{n_g}{(n_g - 1)} & \text{if } h = g; i \neq j, \\ 1 & \text{if } h \neq g \end{cases}$$

when the (gj) -th sample cluster is deleted. The calibration jackknife weights are defined by $w_{hik(gj)}^* = w_{hik(gj)} a_{hik(gj)}$ and the calibration jackknife estimator of the population total Y is given by

$$\hat{Y}_{r(gj)} = \sum_s w_{hik(gj)}^* y_{hik}. \quad (4)$$

A customary jackknife variance estimator of \hat{Y}_r is defined by

$$v_J(\hat{Y}_r) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_{r(gj)} - \hat{Y}_r)^2. \quad (5)$$

For the calibration procedures given in section 2.2, we note that we will need to recalculate $\hat{\mathbf{A}}_{(gj)}^{-1}$, and therefore the adjustment factor, for every cluster (gj) that is deleted.

3.2 Jackknife linearization

By linearizing (5) Yung and Rao (1996) obtained a jackknife variance estimator of \hat{Y}_r given by

$$v_{JL}(\hat{Y}_r) = v(e_{hi}^*), \quad (6)$$

where $e_{hi}^* = \sum_k (n_h w_{hik}^*) e_{hik}$ and $e_{hik} = y_{hik} - \mathbf{x}_{hik}^T \hat{\mathbf{B}}$.

3.3 Estimating function jackknife

Let us consider the projection estimator $\hat{Y}_r = \mathbf{X}^T \hat{\mathbf{B}}$. To calculate $\hat{\mathbf{B}}$ we need to solve the sample estimating function $\hat{\mathbf{S}}(\mathbf{B}) = \sum_s w_{hik} \mathbf{x}_{hik} (y_{hik} - \mathbf{x}_{hik}^T \mathbf{B}) = \mathbf{0}$.

Note that $\sum_s w_{hik} \mathbf{x}_{hik} e_{hik} = \mathbf{0}$.

We now apply the estimating function method, proposed by Hu and Kalbfleisch (2000), to the jackknife procedure. First we define

$$\begin{aligned} \hat{\mathbf{S}}_{(gj)}(\hat{\mathbf{B}}) &= \sum_s w_{hik(gj)} \mathbf{x}_{hik} (y_{hik} - \mathbf{x}_{hik}^T \hat{\mathbf{B}}) \\ &= \sum_s w_{hik(gj)} \mathbf{x}_{hik} e_{hik} \end{aligned}$$

and solve $\hat{\mathbf{S}}(\mathbf{B}) = \hat{\mathbf{S}}_{(gj)}(\hat{\mathbf{B}})$ for \mathbf{B} to obtain the EF jackknife estimator $\tilde{\mathbf{B}}_{(gj)}$. An EF calibration jackknife

estimator of Y is defined by $\tilde{Y}_{r(gj)} = \mathbf{X}^T \tilde{\mathbf{B}}_{(gj)}$ and an EF jackknife variance estimator of \hat{Y}_r is given by

$$v_{EFJ}(\hat{Y}_r) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\tilde{Y}_{r(gj)} - \hat{Y}_r)^2. \quad (7)$$

Theorem 1 *The EF jackknife variance estimator of \hat{Y}_r is identical to the jackknife linearization variance*

estimator.

$$v_{EFJ}(\hat{Y}_r) = v(e_{hi}^*) = v_{JL}(\hat{Y}_r) \quad (8)$$

A sketch of the proof of Theorem 1 is given in Appendix I.

There are two important points to highlight in the EF calibration jackknife weights, defined by

$$\begin{aligned} \tilde{w}_{hik(gj)} &= w_{hik}^* - w_{hik(gj)} a_{hik} \\ &+ w_{hik} \left(\sum_s w_{hik(gj)} a_{hik} \mathbf{x}_{hik}^T \right) \hat{\mathbf{A}}^{-1} \mathbf{x}_{hik}. \end{aligned}$$

First the weights may preserve confidentiality in the data since we do not get zero weights for the (gj) -th cluster, and second we only have to invert the full matrix $\hat{\mathbf{A}}$ thereby avoiding the inversion of possibly illconditioned matrices $\hat{\mathbf{A}}_{(gj)}$ that may be obtained from the ordinary jackknife method. The latter seems to occur often in the case of EF bootstrap, which will be studied in a separate paper.

4. GENERAL PARAMETERS θ

The calibration estimator, $\hat{\theta}_r$, of the census parameter θ_N is obtained by solving $\hat{\mathbf{S}}_r(\theta) = \sum_s w_{hik}^* \mathbf{u}_{hik}(\theta) = \mathbf{0}$, which is an estimator of the population estimating function $\mathbf{S}(\theta) = \sum_U \mathbf{u}_{hik}(\theta)$, where U denotes the population of elements. For example, for the linear regression model of y on \mathbf{z} , we have $\mathbf{u}_{hik}(\theta) = \mathbf{z}_{hik} (y_{hik} - \mathbf{z}_{hik}^T \theta)$ where \mathbf{z}_{hik} is a Q -dimensional vector of explanatory variables. Other examples are given in Hidiroglou et al. (1999). A jackknife linearization variance estimator, $v_{JL}(\hat{\theta}_r)$, is also derived in Hidiroglou et al. (1999).

4.1 EF Jackknife

To apply the EF jackknife method, we define $\tilde{\mathbf{S}}_{r(gj)}(\theta) = \sum_s \tilde{w}_{hik(gj)} \mathbf{u}_{hik}(\theta)$ as an EF jackknife estimator of $\mathbf{S}(\theta)$, where $\tilde{w}_{hik(gj)}$ are the EF calibration jackknife weights defined in section 3.3. Then an EF jackknife estimator of θ_N , denoted by $\tilde{\theta}_{r(gj)}$, can be obtained by solving $\tilde{\mathbf{S}}_{r(gj)}(\theta) = \mathbf{0}$ for

θ . The EF jackknife estimator of $\text{cov}(\hat{\theta}_r)$ is given by

$$v_{EFJ}(\hat{\theta}_r) = \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\tilde{\theta}_{r(gj)} - \hat{\theta}_r)(\tilde{\theta}_{r(gj)} - \hat{\theta}_r)^T. \quad (9)$$

Now to obtain $\tilde{\theta}_{r(gj)}$ we use the calibration estimator, $\hat{\theta}_r$, as the starting point and perform only one step of the Newton-Raphson algorithm (Lipsitz, Dear, Zhao, 1994). That is,

$$\tilde{\theta}_{r(gj)} \approx \hat{\theta}_r + \tilde{\mathbf{J}}_{(gj)}^{-1}(\hat{\theta}_r) \tilde{\mathbf{S}}_{r(gj)}(\hat{\theta}_r) \quad (10)$$

where

$$\tilde{\mathbf{J}}_{(gj)}(\theta) = \frac{-\partial \tilde{\mathbf{S}}_{r(gj)}(\theta)}{\partial \theta^T} = \sum_s \tilde{w}_{hik(gj)} \frac{-\partial \mathbf{u}_{hik}(\theta)}{\partial \theta^T}$$

is evaluated at $\theta = \hat{\theta}_r$. Now $\tilde{\mathbf{J}}_{(gj)}^{-1}(\hat{\theta}_r) \approx \mathbf{J}^{-1}(\hat{\theta}_r)$ and (10) becomes

$$\tilde{\theta}_{r(gj)} \approx \hat{\theta}_r + \mathbf{J}^{-1}(\hat{\theta}_r) \tilde{\mathbf{S}}_{r(gj)}(\hat{\theta}_r), \quad (11)$$

where

$$\mathbf{J}(\theta) = \frac{-\partial \hat{\mathbf{S}}_r(\theta)}{\partial \theta^T} = \sum_s w_{hik}^* \frac{-\partial \mathbf{u}_{hik}(\theta)}{\partial \theta^T}$$

is evaluated at $\theta = \hat{\theta}_r$. Noting that $\hat{\mathbf{S}}_r(\hat{\theta}_r) = \mathbf{0}$ we obtain the EF jackknife estimator of $\text{cov}(\hat{\theta}_r)$, which is given in the following theorem.

Theorem 2 *The EF jackknife estimator of $\text{cov}(\hat{\theta}_r)$ is approximately equal to the jackknife linearization variance estimator. That is,*

$$v_{EFJ}(\hat{\theta}_r) \approx v(\mathbf{J}^{-1}(\hat{\theta}_r) \mathbf{e}_{hi(u)}^*) = v_{JL}(\hat{\theta}_r), \quad (12)$$

where the q^{th} component of $\mathbf{e}_{hi(u)}^*$ is defined as

$$e_{hiq(u)}^* = \sum_k (n_h w_{hik}^*) e_{hikq}(\hat{\theta}_r),$$

$$e_{hikq}(\hat{\theta}_r) = u_{hikq}(\hat{\theta}_r) - \mathbf{x}_{hik}^T \hat{\mathbf{B}}_q(\hat{\theta}_r),$$

$\hat{\mathbf{B}}_q(\hat{\theta}_r) = \hat{\mathbf{A}}^{-1}(\sum_s w_{hik} \mathbf{x}_{hik} u_{hikq}(\hat{\theta}_r))$. A sketch of the proof of Theorem 2 is given in Appendix II.

5. CONCLUSIONS

We have discussed the EF jackknife method for the

estimation of a total and extended it to cover parameters which are defined as solutions of census estimating equations. This method has the advantages of being simpler to implement, avoiding the inversion of possibly illconditioned matrices and potentially useful in preserving confidentiality of micro data. Similar results for the bootstrap and the balanced repeated replication (BRR) methods will be reported elsewhere. Extensions to dual frame surveys and adjustment for unit nonresponse are under investigation.

Appendix I: Sketch of the Proof of Theorem 1

Define

$$\begin{aligned} \tilde{Y}_{r(gj)} - \hat{Y}_r &= -\sum_s w_{hik(gj)} a_{hik} e_{hik} \\ &= -\sum_s a_{hik} e_{hik} (w_{hik(gj)} - w_{hik}) \end{aligned}$$

by noting that $\sum_s w_{hik} \mathbf{x}_{hik} e_{hik} = \mathbf{0}$, which implies

that $\sum_s (\mathbf{X}^T \hat{\mathbf{A}}^{-1} \mathbf{x}_{hik}) w_{hik} e_{hik} = \mathbf{0}$. We then have

$$\begin{aligned} \tilde{Y}_{r(gj)} - \hat{Y}_r &= -\left\{ \frac{1}{n_g - 1} \sum_{i \neq j} \sum_k w_{gik} a_{gik} e_{gik} - \sum_k w_{gik} a_{gik} e_{gik} \right\} \\ &= \frac{1}{n_g - 1} (e_{gj}^* - \bar{e}_g^*) \end{aligned}$$

which proves (8), by substituting $\tilde{Y}_{r(gj)} - \hat{Y}_r$ into (7).

Appendix II: Sketch of Proof of Theorem 2

Define the jackknife estimator of $\text{cov}(\hat{\theta}_r)$ by

$$\begin{aligned} v_{EFJ}(\hat{\theta}_r) &= \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\tilde{\theta}_{r(gj)} - \hat{\theta}_r)(\tilde{\theta}_{r(gj)} - \hat{\theta}_r)^T \\ &\approx \mathbf{J}^{-1}(\hat{\theta}_r) \sum_{g=1}^L \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\tilde{\mathbf{S}}_{r(gj)}(\hat{\theta}_r) - \hat{\mathbf{S}}_r(\hat{\theta}_r)) \times \\ &\quad (\tilde{\mathbf{S}}_{r(gj)}(\hat{\theta}_r) - \hat{\mathbf{S}}_r(\hat{\theta}_r))^T \mathbf{J}^{-1}(\hat{\theta}_r) \\ &= \mathbf{J}^{-1}(\hat{\theta}_r) v_{EFJ}(\hat{\mathbf{S}}_r(\hat{\theta}_r)) \mathbf{J}^{-1}(\hat{\theta}_r). \end{aligned} \quad (13)$$

We can then replace y_{hik} in Appendix I by a Q -dimensional vector $\mathbf{u}_{hik}(\hat{\theta}_r)$ and obtain

$$\begin{aligned}
\tilde{\mathbf{S}}_{r(gj)}(\hat{\theta}_r) - \hat{\mathbf{S}}_r(\hat{\theta}_r) &= -\sum_s w_{hik(gj)} a_{hik} \mathbf{e}_{hik}(\hat{\theta}_r) \\
&= -\sum_s a_{hik} \mathbf{e}_{hik}(\hat{\theta}_r) (w_{hik(gj)} - w_{hik}) \\
&= \frac{1}{n_g - 1} (\mathbf{e}_{gj(u)}^* - \bar{\mathbf{e}}_{g(u)}^*) \quad (14)
\end{aligned}$$

where $\mathbf{e}_{gj(u)}^* = \sum_k (n_g w_{gjk}^*) \mathbf{e}_{gjk}(\hat{\theta}_r)$ and $\bar{\mathbf{e}}_{g(u)}^* = n_g^{-1} \sum_j \mathbf{e}_{gj(u)}^*$. Hence substituting

$$v_{EFJ}(\hat{\mathbf{S}}_r(\hat{\theta}_r)) = v(\mathbf{e}_{hi(u)}^*)$$

into (13) yields $v_{JL}(\hat{\theta}_r)$, which is the jackknife linearization estimator of $\text{cov}(\hat{\theta}_r)$ given in Hidioglou et al. (1999).

REFERENCES

- Hidioglou, M.A., Rao, J.N.K. and Yung, W. (1999). « Variance Computation for Complex Surveys using Estimating Equations ». *Proceedings of the Survey Methods Section, Statistical Society of Canada*.
- Hu, F. and Kalbfleisch, J.D. (2000). « The Estimating Function Bootstrap ». *Canadian Journal of Statistics*, 28, (in press).
- Lipsitz, S.R., Dear, K.B.G., and Zhao, L. (1994). « Jackknife estimators of variance for parameter estimates from estimating equations with applications to clustered survival data ». *Biometrika*, 50, 842-846.
- Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer, New York.
- Yung, W. and Rao, J.N.K. (1996). « Jackknife linearization variance estimators under stratified multistage sampling ». *Survey Methodology*, 22, 23-3.