COMPUTING JACKKNIFE VARIANCE ESTIMATES WITH CERTAINTY AND NONCERTAINTY PSU'S

David A. Marker
Westat, Inc.

Keywords: Jackknife, Replicated Variance, Multi-stage Sampling, Linear Statistics

Introduction

Jackknife variance estimates are computed by dropping one small portion of the data in each replicate and reweighting the remaining data to compensate for those dropped. In the original jackknife procedure developed for a simple random sample (see Efron 1982) of size n, each of the elements is dropped from one of the n replicates with the other (n-1) elements reweighted up to the total.

In practice we do not usually have simple random samples. Instead, a commonly used design is to select two primary sampling units (PSUs) per stratum in a multi-stage sample design. The jackknife variance estimate is then used in its grouped form. One PSU is randomly dropped in each replicate, with the remaining PSU from the same stratum reweighted to represent the entire stratum (see Wolter 1985).

According to the general jackknife theory, the variance estimates should not depend upon which subset of cases is dropped from some PSUs, and which are retained for all replicates. Unfortunately, in the name of cost savings, a short-cut version is often used which violates this basic assumption.

Basic Approach

In the typical paired jackknife design (see Westat, 1989) with I variance strata with J=2 PSU per stratum for a linear estimator X, the i' replicate estimate can be written as

$$X_{i'} = \left(\sum_{i=1}^{I} \sum_{j=1}^{2} x_{ij} \right) + x_{i'j'} - x_{i'j''}$$

where the j' PSU is selected,the j" PSU is dropped, and x_{ij} is the PSU estimate, appropriately weighted, so that

$$X = \sum_{i=1}^{I} \sum_{j=1}^{2} x_{ij}$$

The variance estimate for the full sample estimate X is:

$$V_1(X) = \sum_{i'=1}^{I} (X_{i'} - X)^2 = \sum_{i'=1}^{I} (x_{i'j'} - x_{i'j''})^2$$
$$= \sum_{i'=1}^{I} (x_{i'j''} - x_{i'j'})^2 .$$

Thus, the variance estimate is independent of which PSU is dropped and which is kept. This is also true approximately for nonlinear estimators.

What happens if there are both certainty and noncertainty PSUs? While the above design works for the noncertainty PSUs, it must be modified for certainty PSUs. One possibility is to split each certainty PSU into two parts and randomly drop one part while adjusting the remaining weights. Thus, if there are K certainty PSUs in addition to the 2I noncertainty PSUs, there would be I + K replicates needed to compute variances. (See Rust, 1986 for a discussion of optimally allocating replicates for jackknife variances.)

Common Short-Cut

For many jackknife applications, the cost of computing variances is proportional to the number of replicates used to compute the estimates. (This is true when computer charges are the bulk of the costs.) It is therefore desirable to minimize the number of replicates. A common short-cut is to pair each certainty PSU randomly with a noncertainty variance stratum, split the certainty PSU into two halves, and drop one of these halves in the same replicate that drops one of the noncertainty PSUs from the stratum.

Using this methodology requires I replicates. In each replicate i', there are two subselections being made: a noncertainty PSU with estimate $x_{i'j'}$, and one-half of a certainty PSU with estimate $x_{k'j'}$. The variance estimate now is as follows:

$$V_2(X) = \sum_{i'=1}^{I} \left[(x_{i'j'} - x_{i'j''}) + (x_{k'j'} - x_{k'j''}) \right]^2$$

Unfortunately, this variance estimate **does** depend upon which PSU is selected and which is dropped!

An alternative methodology is to separately use the grouped jackknife to compute a certainty PSU component and a noncertainty PSU component of variance. Assuming the selections were independent, the sum of these two estimates is the estimated variance of the total X. If there are K certainty PSUs, the estimator will require I + K replicates, where

$$V_3(X) = \sum_{i'=1}^{I} (X_{i'} - X)^2 + \sum_{k'=1}^{K} (X_{k'} - X)^2$$
$$= \sum_{k'=1}^{I} (x_{i'j'} - x_{i'j''})^2 + \sum_{k'=1}^{K} (x_{k'j'} - x_{k'j''})^2 .$$

$V_3(X)$, unlike $V_2(X)$, is independent of which PSU is selected and which is dropped. While $E(V_2) = E(V_3)$ (provided that the assignment of j' and j" is random for each i' and k' in the case of V_2), the $Var(V_2 \mid data) > 0$ and $Var(V_3 \mid data) = 0$. In other words, while both methods have the same expected variance, the short-cut method trades a computer cost savings for an added variance of the variance component of error. Unless the cost of the K extra replications is prohibitive, $V_3(X)$ is preferable over $V_2(X)$. (Assuming, of course, that the K certainty PSUs contribute substantially to the estimate X, which is usually the case.)

Example

Assume we have two certainty PSUs and four noncertainty PSUs. The noncertainty PSUs are placed into two variance strata: A and B.

Certainty PSU	Half Estimate	
1	4	6
2	200	400

Variance Stratum	Noncertainty PSU Estimate	
A	10	12
B	100	300

The alternative methodology computes separate estimates for both certainty and noncertainty components of variance,

Certainty component $=$

$$(4 - 6)^2 + (200 - 400)^2 = 40{,}004$$

Noncertainty component $=$

$$(10 - 12)^2 + (100 - 300) = \underline{40{,}004}$$

$$V_3(X) = 80{,}008$$

Using the common short-cut method, there is a 50/50 chance that certainty PSU1 will be paired with variance stratum A or B. In either case, there are four unique variance estimates that can be computed. If certainty PSU1 is paired with variance stratum A, either:

$$V_2(X) = [(4 - 6) + (10 - 12)]^2 + [(200 - 400) + (100 - 300)]^2 = 16 + 160{,}000 = 160{,}016$$

or

$$V_2(X) = [(4 - 6) + (12 - 10)]^2 + [(200 - 400) + (100 - 300)]^2 = 0 + 160{,}000 = 160{,}000$$

or

$$V_2(X) = [(4 - 6) + (10 - 12)]^2 + [(200 - 400) + (300 - 100)]^2 = 16 + 0 = 16$$

or

$$V_2(X) = [(4 - 6) + (12 - 10)]^2 + [(200 - 400) + (300 - 100)]^2 = 0 + 0 = 0$$

The $E(V_2(X) \mid \text{PSU1 with stratum A}) = 80{,}008$.

If certainty PSU1 is paired with variance stratum B:

$$V_2(X) = [(4 - 6) + (100 - 300)]^2 + [(200 - 400) + (10 - 12)]^2 = 40{,}804 + 40{,}804 = 81{,}608$$

or

$$V_2(X) = [(4 - 6) + (300 - 100)]^2 + [(200 - 400) + (10 - 12)]^2 = 39{,}204 + 40{,}804 = 80{,}008$$

or

$$V_2(X) = [(4 - 6) + (100 - 300)]^2 + [(200 - 400) + (12 - 10)]^2 = 40{,}804 + 39{,}204 = 80{,}008$$

or

$$V_2(X) = [(4 - 6) + (300 - 100)]^2 + [(200 - 400) + (12 - 10)]^2 = 39{,}204 + 39{,}204 = 78{,}408$$

The $E(V_2(X) \mid \text{PSU1 with Stratum B}) = 80{,}008$.

So even though $E(V_3(X)) = E(V_2(X)) = 80{,}008$, it is clear that $\text{Var}(V_2(X) \mid \text{sample data}) >> 0 = \text{Var}(V_3(X) \mid \text{sample data})$.

Unconditional Result

A simple corollary to the above result is that

$$\text{Var}(V_2(X)) > \text{Var}(V_3(X)) > 0$$

unconditional on the observed sample data. We know that

(1) $\text{Var}(V_2(X)) = E(\text{Var}(V_2(X) \mid \text{data})) + \text{Var}(E(V_2(X) \text{ data}))$,

and

(2) $\text{Var}(V_3(X)) = E(\text{Var}(V_3(X) \mid \text{data})) + \text{Var}(E(V_3(X) \mid \text{data}))$.

Since the selection of which data to keep and which to drop is conducted independently in the certainty and noncertainty PSUs, we know that $E(V_2(X) \mid \text{data}) = E(V_3(X) \mid \text{data}) > 0$. Hence, the last term on the right-hand side (RHS) of equations (1) and (2) are equal. We have seen earlier that the first term on the RHS of equation (2) is equal to zero, while the comparable term in equation (1) is greater than zero.

Therefore, $\text{Var}(V_2(X)) > \text{Var}(V_3(X)) > 0$.

Acknowledgements

The author would like to thank Morris Hansen, Keith Rust, and the rest of the Statistical Group at Westat for their constructive comments. He would also like to thank Dr. Terrance Connell of the U.S. Department of Housing and Urban Development for raising the questions that led to examining this issue.

References

Efron, Bradley (1982). "The Jackknife, the Bootstrap, and Other Resampling Plans," *Society for Industrial and Applied Mathematics*, Philadelphia, Penn.

Rust, Keith (1986) "Efficient Replicated Variance Estimation", *Proceeding of the Section on Survey Research Methods of the American Statistical Association.*

Westat, Inc. (1989), "The Wesvar Procedure".

Wolter, Kirk, (1985) *"Introduction to Variance Estimation"*, Springer-Verlag: New York.