Using the DAG Jackknife to Measure the Variance of an Estimator in the Presence of Item Nonresponse

Darcy Miller[1], Phillip Kott[2]
[1]National Agricultural Statistics Service, 3251 Old Lee Highway, Fairfax, VA 20002
[2]RTI International, 6110 Executive Boulevard, Rockville, MD 20852

Abstract
The National Agricultural Statistics Service (NASS) uses the delete-a-group (DAG) jackknife to estimate variances and mean squared errors in many of its surveys. The DAG jackknife provides nearly unbiased estimates under a host of complex designs and processes. It involves breaking the sample of unit respondents into T (T= 15 or 30) groups and then sequentially deleting each group from the sample, leaving T replicate groups. Then T sets of replicate weights are created, and T replicate estimates are calculated using these weights. The DAG jackknife estimator is the sum of the squared differences between the T replicate estimates and the original estimate (entire sample) multiplied by (T-1)/T. Although the DAG jackknife currently accounts for unit nonresponse, item imputations are treated as real reported values. Through a small simulation study, we explore using the principles of DAG jackknife to account for the additional variance in an estimated mean or total due to the imputation process by creating replicate imputations.

Key Words: Variance Estimation, DAG Jackknife, Item Non-Response, Imputation

1. Introduction

Missing data exists. Although significant improvements in survey instruments and data collection have been made, surveys are returned without full responses. Likewise, experiments can be well planned, yet attrition and mechanical failure lead to incomplete data. Despite the difficulty in obtaining complete observations from sampled units, the need remains to perform effective statistical analysis with the data. Without adaptation to the processing and analysis of the data set from traditional methods, the missing values leave room for bias in the estimates. To mitigate this, many statisticians intelligently create values to fill in the "holes" in the data set (imputation). Imputation is widely accepted and practiced. While this may reduce bias in the target estimate, inferences drawn from the data will not be valid if the imputed values are treated as real values. Classical methods to calculate the variance of an estimator which treat imputed values as real are insufficient.

NASS is increasingly using the DAG jackknife to estimate variances and mean squared errors, since it provides nearly unbiased estimates under a host of

complex designs and processes. It is a natural extension to examine the use of the DAG jackknife as a tool to calculate the variance of an estimator. Rao and Shao (1992) originated an approach based on adjusted replication. Further developments have been made by Shao, Chen and Chen (1998) and more. Cohen (2002) outlined a simpler implementation of these methods.

We extend the DAG Jackknife when estimtating a total using imputed values derived on a linear model. Section 2 gives an overview of the DAG jackknife; Section 3 provides the adjusted imputation method. Section 4 lays out the framework for a small simulation study, the results of which are in Section 5. A short conclusion is offered in Section 6.

2. Overview of DAG Jackknife

Observation	Variable	Sampling Weight
1	x	6
2	x	12
3	.	18
4	x	6
5	x	12
6	.	24
7	x	18
8	x	12
9	x	36
10	.	36
11	x	12
12	x	24
13	x	18
14	.	6

Observation	Variable	Replicate Weight
1	x	7
2	x	14
4	x	7
6	.	14
7	x	21
8	x	14
9	x	42
10	.	42
11	x	14
12	x	42
13	x	21
14	.	7

Figure 1: The table on the left represents the sample of unit respondents with non-overlapping groups represented by colors (e.g. blue = group 1, peach = group 2, red = group 3, etc.). The table on the right represents the 3rd replicate (sample of unit respondents with the red = group 3 dropped). The sampling weight has been adjusted (replicate weight) for the absence of the group 3 observations.

The DAG jackknife is one of a host of replication methods that are used to calculate variances of estimators, particularly under complex sampling designs or when estimating a nonlinear target. Replication methods are implemented by "re-sampling" from the sample. In the case of the DAG Jackknife, this is done by creating T non-overlapping groups of the unit respondents and sequentially dropping each group from the sample. What remains of the sample after dropping the tth group (t = 1, 2, ... , T) is called the tth replicate. Sample weights are adjusted to form replicate weights, which account for the loss of observations.

Figure 1 above provides a small scale representation of the process, although there are generally more groups formed (to retain degrees of freedom) and more observations in each group.

An estimate of a total, $\hat{\theta}$, is calculated using the sampling weights. The estimated variance of $\hat{\theta}$ is based on the difference between the estimate $\hat{\theta}$ using sampling weights and estimate $\hat{\theta}$ using replicate weights for each replicate.

More formally, the set of unit respondents from a sample, S, is divided into T groups, $S_1, \ldots, S_T$. The t^{th} replicate is defined to be $S_{(t)} = S - S_t$. The total of item Y can be computed by summing over values for Y in the population, U.

$$\theta = \sum_{i \in U} y_i \qquad (1)$$

An estimate of a total, $\hat{\theta}$, is computed from S using the sampling weights $\{w_k; k \in S\}$

$$\hat{\theta} = \sum_{k \in S} w_k y_k \qquad (2)$$

Similarly, an estimate of a total, $\hat{\theta}^{(t)}$, using the t^{th} replicate can be computed from $S_{(t)}$ using the t^{th} set of replicate weights $\{w_{k(t)}; k \in S_{(t)}\}$.

$$\hat{\theta}^{(t)} = \sum_{k \in S(t)} w_{k(t)} y_k \qquad (3)$$

And, $\hat{\theta}$ has its variance estimated by:

$$\text{Var}\left(\hat{\theta}\right) = \frac{T-1}{T} \sum_{t=1}^{T} \left(\hat{\theta}^{(t)} - \hat{\theta}\right)^2 \qquad (4)$$

3. Adjusted Jackknife Imputation

If imputed values are naively treated as real values, the variance of the estimate will be biased, regardless of the variance estimator used. This includes the jackknife variance estimator. Using replication methods, such as the jackknife, to account for imputation by adjusting imputations was introduced by Rao and Shao (1992). Essentially, they proposed to re-impute each replicate and apply formula (4) with the $\hat{\theta}^{(t)}$ computed based on the t^{th} adjusted replicate.

For mean imputation and random hot deck imputation, Rao and Shao's adjustment for a unit respondent with a missing value for y_k is

$$y^*_{k(t)} = y^*_k + \hat{E}\left(y^*_{k(t)}\right) - \hat{E}\left(y^*_k\right), \qquad (5)$$

where y_k^* is the imputed value for y_k using all sample respondents, $y_{k(t)}^*$ is the imputed value for y_k using the respondents in the t[th] replicate, and $\hat{E}$ is estimated expectation of y_k under the imputation model.

Rao and Shao's adjusted jackknife produces asymptotically unbiased and consistent jackknife variance estimators for means and totals for imputation methods such as mean imputation, random hot deck imputation, ratio or regression imputation, but produces serious overestimation in the case of nearest neighbor imputation.

With the typical delete-1 jackknife, this implementation would be computationally intensive, even with a simple imputation scheme such as mean imputation. Di Zio et al. (2008) suggested the use of the DAGjackknife and the extended-delete-a-group jackknife (EDAGjackknife) with the Rao & Shao adjustment for hot deck imputation to reduce computation. The EDAGjackknife is a small modification to the reweighting step when using DAGjackknife proposed by Kott (2001) to handle the bias introduced by the DAGjackknife when the number of primary sampling units in a strata is small. Simulations where some strata have a small number of primary sampling units by Di Zio et al. showed the EDAGjackknife with Rao & Shao adjustment for hot deck imputation to perform well in terms of precision and computational feasibility.

We propose re-imputing replicates with the DAGjackknife where a prediction model has been used for imputation. That is, when imputing with

$$y_k^* = m(x_k'b) \quad or \quad y_k^* = m(x_k'b) + e_k^*, \tag{7}$$

re-impute each replicate with

$$y_{k(t)}^* = m\left(x_k'b_{(t)}\right) \quad or \quad y_{k(t)}^* = m\left(x_k'b_{(t)}\right) + e_k^* \ , \tag{8}$$

where $b_{(t)}$ is calculated using the data and replicate weights for the t[th] replicate, and the e_k^* remain the same across replicates.

4. Simulation Study Framework

We simulated 100 samples (n = 1200) of farm operations from Illinois using the 2007 Census of Agriculture. We drew a stratified sample with a simple random sample within strata (see Table 1).

Farm Type/Total Value of Production	Small	Medium	Large
Crop	Strata 1	Strata 2	Strata 3
Livestock	Strata 4	Strata 5	Strata 6

Table 1: Sampling design strata. Farm Type is pooled to be defined as either Crop or Livestock. Total Value of Production defines the size of the farm based on the value of its production and is pooled into three groups (small, medium, and large).

Our target is Total Corn Acres Harvested for Illinois, $\theta = \sum_{i \epsilon U} y_i$, where y_i are individual values of Total Corn Acres Harvested in Illinois. The sampled values of Total Acres Harvested, x_k , and Total Corn Acres Harvested, y_k were kept: (x_k, y_k) . We imposed missingness completely at random (MCAR) for 25% of the y_k .

From the sample, S, we formed T= 20 non-overlapping groups, $S_1, \ldots, S_{20}$. The t^{th} replicate is $S_{(t)} = S - S_t$.

Three imputation models were considered in this study. Separate simple linear models were assumed for Crop and Livestock operations. In our notation, R is the set of observations with non-missing values for y_k after we imposed missingness.

4.1 Imputation Method 1

$$y_k^* = \begin{cases} x_k b & y_k \text{ missing} \\ y_k & y_k \text{ observed} \end{cases} , \qquad (9)$$

$$\text{where } b = \frac{\sum_{k \epsilon R} w_k y_k}{\sum_{k \epsilon R} w_k x_k} . \qquad (10)$$

We re-impute replicate with

$$y_{k(t)}^* = \begin{cases} x_k b_{(t)} & y_k \text{ missing} \\ y_k & y_k \text{ observed} \end{cases} , \qquad (11)$$

$$\text{where } b_{(t)} = \frac{\sum_{k \epsilon R \cap S_{(t)}} w_{k(t)} y_k}{\sum_{k \epsilon R \cap S_{(t)}} w_{k(t)} x_k} . \qquad (12)$$

4.2 Imputation Method 2

$$y_k^* = \begin{cases} x_k b + \frac{x_k}{\bar{x}_k} e_k^* & y_k \text{ missing} \\ y_k & y_k \text{ observed} \end{cases} , \qquad (13)$$

where b is calculated using (10), $e_k^* \sim N(0, s_r^2)$, $s_r^2$is the weighted sample variance among respondents for r $= y_k - x_k b$, and the e_k^* are truncated so that $|e_k^*| \leq \bar{x}_k$ to eliminate the possibility of negative imputations.

We re-impute replicate with

$$y_{k(t)}^* = \begin{cases} x_k b_{(t)} + \frac{x_k}{\bar{x}_k} e_k^* & y_k \text{ missing} \\ y_k & y_k \text{ observed} \end{cases} , \qquad (14)$$

where $b_{(t)}$ is calculated using (12), $and\ e_k^*$is described above

4.3 Imputation Method 3

$$y_k^* = \begin{cases} x_k b_k^* & y_k \text{ missing} \\ y_k & y_k \text{ observed} \end{cases} , \qquad (15)$$

where b_k^* is a random draw from the set $\left\{ \frac{y_k}{x_k} : k \in R \cap S_t \right\}$ with probability of selection proportional to $w_k x_k$ for reasons explained in a Kott and Folsom (2010). Note that this effectively sets the $\frac{x_k}{\bar{x}_k} e_k^*$ term in equation (13) equal to$x_k(b_k^* - b)$.

We re-impute replicate with

$$y_{k(t)}^* = \begin{cases} x_k b_k^* \frac{b_{(t)}}{b} & y_k \text{ missing} \\ y_k & y_k \text{ observed} \end{cases} , \qquad (16)$$

which is asymptotically identical to $x_k b_{(t)} + x_k(b_k^* - b)$but always nonnegative.

5. Simulation Study Results

Remember, our target, $\theta = \sum_{i \in U} y_i$, is Total Corn Acres Harvested for Illinois.

Initially, we looked at the empirical relative bias (EmpRBIAS) of each imputation model as well as the sample. Let a = 1, ... 100 be a sample simulation.

$$EmpRBIAS_{*}(\hat{\theta}_{*}^{(a)}) = \frac{100^{-1}\sum_{a=1}^{100}(\hat{\theta}_{*}^{(a)} - \theta)}{\theta} \quad , \qquad (17)$$

where * = {Sample, Model 1, Model 2, Model 3}, $\theta = \sum_{i \epsilon U} y_i$ is the true total corn acres harvested for Illinois and $\hat{\theta}_{*}^{(a)}$ is the estimate for total corn acres harvested for Illinios using sampling weights. EmpRBIAS is summarized in Table 2.

Metric	Sample	Model 1	Model 2	Model 3
EmpRBIAS	<0.001	<0.001	0.055	-0.038

Table 2: Empirical relative bias for the sample simulation, Model 1, Model 2, Model 3.

The DAGjackknife estimates the combined sample and prediction model mean squared error. So, we will compare and empirical relative root mean square error (EmpRRMSE) to an average DAGjackknife variance converted to a relative root mean square error for both the naive DAG jackknife with imputed values treated as real values (NaiveJackRRMSE) and DAG jackknife with replicates re-imputed (JackRRMSE).

Since we have T=20 groups, the DAGjackknife variance estimate was calculated using

$$Var_{jack,*}\left(\hat{\theta}_{*}^{(a)}\right) = \frac{20-1}{20}\sum_{t=1}^{20}\left(\hat{\theta}_{*,t}^{(a)} - \hat{\theta}_{*}^{(a)}\right)^2 \quad , \qquad (18)$$

where * = {Sample, Model 1, Model 2, Model 3}, $\hat{\theta}_{*}^{(a)}$ is the estimate for total corn acres harvested for Illinois, and $\hat{\theta}_{*,t}^{(a)}$ is the estimate for total corn acres harvested for Illinois for the t^{th} replicate using the corresponding replicate weights.

The EmpRRMSE is

$$EmpRRMSE\left(\hat{\theta}_{*}\right) = \frac{\sqrt{\left[100^{-1}\sum_{a=1}^{100}\left(\hat{\theta}_{*}^{(a)} - \theta\right)^2\right]}}{\theta} \quad , \qquad (19)$$

where * = {Sample, Model 1, Model 2, Model 3}, $\hat{\theta}_{*}^{(a)}$ is the estimate for total corn acres harvested for Illinois, and $\theta = \sum_{i \epsilon U} y_i$ is total corn acres harvested for Illinois.

Our NaiveJackRRMSE (imputations treated as real values) and JackRRMSE (replicates re-imputed) are found using

$$NaiveJackRRMSE\left(\hat{\theta}_*\right) \text{or JackRRMSE } \left(\hat{\theta}_*\right) = \frac{\sqrt{\left[100^{-1}\sum_{a=1}^{100}\left(Var_{jack,*}\left(\hat{\theta}_*^{(a)}\right)\right)\right]}}{\theta}, \quad (20)$$

where * = {Sample, Model 1, Model 2, Model 3}, $Var_{jack,*}\left(\hat{\theta}_*^{(a)}\right)$ is the DAGjacknife variance estimate for sample a, and $\theta = \sum_{i\epsilon U} y_i$ is total corn acres harvested for Illinois.

Results of the simulation study are in Table 3 and summarized below.

Metric	Sample	Model 1	Model 2	Model 3
EmpRBIAS	<0.001	<0.001	0.055	-0.038
EmpRRMSE	0.057	0.059	0.074	0.057
NaiveJackRRMSE	0.054	0.052	0.06	0.055
JackRRMSE	0.054	0.056	0.063	0.059

Overall, we see evidence that JackRRMSE performs fairly well, and better than the NaiveJackRRMSE, which always underestimates. We should not, however, be too bold in conclusions drawn from only 100 simulated samples.

6. Conclusion

Building on the work of Rao & Shao (1992) and others through a small simulation study, we see evidence that the DAGjackknife can be used to account for imputation when using a prediction model by re-imputing replicates. Future work includes investigating other response mechanisms, group numbers, other targets beyond totals, domain estimation, and the role of weights.

References

Brick, J.M., Kalton, G., and Kim, J.-K. (2004) "Variance estimation with hot deck imputation using a model," Survey Methodology, 30, 57-66.

Chen, J. and Shao, J. (2001) "Jackknife variance estimation for nearest neighbour imputation," Journal of the American Statistical Association, 96, 453, 260-269.

Cohen, M. (2002) "Implementing Rao-Shao Type Variance Estimation with Replicate Weights," Survey Methodology, Vol. 28, No. 1, 97-101.

Di Zio, M. and Righi, P. et al (2008) "Variance Estimation in Presence of Imputation: an Application to ISTAT Survey Data," European Conference on Quality in Official Statistics, Rome, Italy.

Kott, P.S. (1998) "Using the Delete-a-Group Jackknife Variance Estimator in NASS Surveys," NASS Research Report 98-01 (revised 2011).

Kott, P.S. (2001) "The Delete-a-group Jackknife," Journal of Official Statistics, 17, 521-526.

Kott, P.S. and Folsom, R.E. (2010) "Weights, Double Protection, and Multiple Imputation," In *JSM Proceedings,* Survey Research Methods Section. Alexandria, VA: American Statistical Association.

Rao J.N.K. and Shao, J. (1992) "Jackknife variance estimation with survey data under hot deck imputation," Biometrika, 79 811-822.

Rao, J.N.K. and Sitter, R. (1992) "Jackknife variance estimation under imputation for missing survey data," Technical Report, Laboratory for Research in Statistics and Probability, Caleton University, Ottowa.

Shao, J., Chen, Y., and Chen, Y. (1998) "Balanced repeated replication for stratified multistage survey data under imputation," Journal of the American Statistical Association, 98.

Skinner, C.J. and Rao, J.N.K. (1993) "Jackknife variance estimation for multivariate statistics under hot deck imputation from common donors," Journal of Statistical Planning and Inference, 102, 149 – 167.