VARIANCE ESTIMATION PROCEDURE FOR VALUE OF CONSTRUCTION PUT IN PLACE SURVEY WITH IMPUTED DATA

Carrie Jones and Masato Asanuma, U.S. Bureau of the Census[1]
Carrie Jones, Room 2136 FOB-4, Washington, DC 20233

Key Words: design-based variance, ratio-based imputation, warm-deck imputation

1. INTRODUCTION

Applying a standard variance estimator to survey data with a high level of imputation leads to underestimation of the variance when the imputed values are treated as if they were reported. Many variance estimation methods attempt to correct for this underestimation (Rao and Shao 1992, Rao and Sitter 1995, Rubin 1996, Fay 1996, Rao 1996, and Deville and Sarndal 1994). It is very difficult to directly apply these techniques to complex surveys such as the Census Bureau's Value of Construction Put in Place (VIP) Survey, where the survey item of interest is imputed from three other survey items (imputed or reported) that use warm-deck and ratio-based imputation procedures. Shao and Steel (forthcoming) propose a general decomposition method for constructing variance estimators for single imputation that address these complexities. The purpose of our research is to apply their method and develop a variance estimator for the VIP Survey that reflects variance due to sampling and imputation.

We express the total variance as a sum of three components: a design-based variance given the nonresponse, a variance due to a survey item falling into a correct imputation cell, and a variance due to random nonresponse of the survey items. The very complex variance estimators of the latter two components have been derived, but they are not yet programmed. Therefore, they are not included in this paper. Our objective in this paper is to compare design-based variance estimates when imputed values are treated as reported as opposed to imputed.

We give an overview of the VIP Survey in Section 2 and explain our imputation methodology in Section 3. Section 4 presents an estimator of total value of construction which includes ratio based imputation methods. Section 5 provides a general description of the variance estimator of total value of construction based on the Shao and Steel decomposition method. In this section we also produce numerical results from the design-based variance component for imputed values treated as both imputed and observed values. Section 6 contains conclusions.

2. VIP SURVEY

Each month the Manufacturing and Construction Division (MCD) of the U.S. Bureau of the Census conducts the privately owned Nonresidential VIP Survey to measure the total value of nonresidential construction activity performed in the United States. The sampling frame is a list of construction projects in the United States, excluding Hawaii, valued at $50,000 or more that have started or will start construction within 60 days. MCD purchases the list from the F.W. Dodge Division of McGraw-Hill. We identify the nonresidential construction projects in Hawaii by obtaining building permit notifications from the permit-issuing places of Honolulu, Maui, Kauai, and Hawaii counties, and we select projects not covered by building permit systems by area sampling. On average there are 7,500 projects in the survey at any one time. These include newly selected projects as well as projects carried over from previous months. We stratify sample projects by type of construction and contract value, independently take a systematic sample of projects from each of the 66 strata, and contact sampled cases each month by mail or telephone in an effort to obtain progress reports until completion of the project.

We request owners or builders to report the value of construction (*VIP*) activity performed in the previous month. Some of the other survey data we collect include total construction cost (*RVITM5C*); architectural, engineering, and miscellaneous construction costs (*AE&M*); start date of construction (*STRTDATE*); and completion date of construction. Some of these items may not be reported during the month of selection (*SELDATE*), i.e., the initial month

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

of contact. Two other variables of interest are the project's contract value (*PROJSELV*) and major type of construction (*MTC*). They are provided in the F.W. Dodge sampling frame.

Each month MCD tabulates *VIP* estimates at the national level by *MTC* and publishes preliminary estimates for the current month, revised estimates for one month and two months ago, and a final revision in May of the following year. In the near future MCD will begin issuing monthly and annual publications with newly defined *MTC* and introduce subcategories of the major types of construction in the annual publications. These changes motivate us to develop new variance methods. Since on average about 42% of the preliminary *VIP* estimates, 37% of the revised *VIP* estimates, and 20% of the final *VIP* estimates are imputed, it is very important for the new *VIP* variance estimator to include a component for imputation.

3. IMPUTATION METHODOLOGY

The *VIP* estimate for each sampled project is expressed as the final weight multiplied by the *VIP* for a record. The final weight is a product of the following: (inverse of the probability of selecting the record) x (outlier adjustment factor) x (*AE&M* adjustment factor) x (Dodge duplication factor). Of these, the *AE&M* adjustment factor which prorates the *AE&M* costs over the life of the project, defined as [1+*AE&M/RVITM5C*], contains a ratio of two survey items that may be imputed; therefore, they add imputation variation to the *VIP* variance estimator. *STRTDATE* defines part of the *VIP* imputation cells, so it will add variation if it is not reported. Thus, variation due to imputation of *VIP*, *RVITM5C*, *AE&M*, and *STRTDATE* is included in the *VIP* variance estimator. The methodology for imputing *VIP*, *RVITM5C*, *AE&M*, and *STRTDATE* is as follows:

3.1 VIP

If *VIP* is not reported for a sampled case due to late reporting or refusal, we impute the value depending on the reporting status of *STRTDATE*. We also reimpute projects that have not reported *VIP* for any given month (*VIPMONTH*) in the past 24 month period. Each nonreporting *VIP* sampled project *i* is assigned to the appropriate imputation cell and is imputed as follows:

$$\hat{VIP}_i = \begin{cases} \frac{Y_r}{X_r} \times RVITM5C_i, \text{ if } STRTDATE \text{ is reported,} & (3.1.1) \\ \frac{Y_r}{X_r + X_c} \times RVITM5C_i, \text{ otherwise.} & (3.1.2) \end{cases}$$

where

Y_r represents the sum of *VIP* from cases reporting *VIP* and *RVITM5C*,

X_r represents the sum of *RVITM5C* from cases reporting *VIP* and *RVITM5C*,

X_c represents the sum of *RVITM5C* from cases reporting *RVITM5C* that have completed construction activity.

Y_r, X_r, and X_c are computed for each *VIPMONTH* and *VIP* imputation cell where the *VIP* imputation cells are defined by five *RANGE* groups (difference in months: *VIPMONTH - STRTDATE*) and seven *RVITM5C* value groups.

3.2 RVITM5C and AE&M

The nonreporting *RVITM5C* or *AE&M* sampled project *i* is assigned to the appropriate imputation cell and is imputed as follows:

$$RVIT\hat{M}5C_i = \frac{X_h}{Z_h} \times PROJSELV_i, \qquad (3.2.1)$$

$$AE\hat{\&}\, M_i = \frac{V_h}{Z_h} \times RVITM5C_i. \qquad (3.2.2)$$

where

X_h represents the sum of reported *RVITM5C* from cases that completed construction since 1992,

Z_h represents the sum of *PROJSELV* from cases that completed construction since 1992 and reported *RVITM5C*,

V_h represents the sum of *AE&M* from cases that completed construction since 1992 and reported *RVITM5C* and *AE&M*.

X_h, Z_h, and V_h are updated annually for each item imputation cell which is defined by five *MTC* and six *PROJSELV* value groups.

3.3 STRTDATE

If a sampled project does not report a *STRTDATE*, we use a warm-deck procedure by selecting a donor from a pool of projects in the sample or selected within the past twenty-four months. Each potential donor must have reported a *STRTDATE* and the absolute difference between start and selection date of the donor must be less than or equal to 24 months, i.e., $DIFF_{donor} = |SELDATE_{donor} - STRTDATE_{donor}| \leq 24$.

The nonreporting *STRTDATE* sampled project i is assigned to the appropriate imputation cell as described above for *RVITM5C* and *AE&M*. A donor is randomly selected with replacement, and *STRTDATE* for project i is imputed as follows:

$$STR\hat{T}DATE_i = SELDATE_i - DIFF_{donor}. \tag{3.3.1}$$

Some sampled projects will include imputed data for all four survey items using the imputation methods described above. Each of these four imputation methods must be reflected in the monthly *VIP* estimation process. The *VIP* estimator in the next section illustrates how the reporting status of the survey items and the imputation methods are intertwined in the formula.

4. VIP ESTIMATOR

We construct the monthly estimate of *VIP* by summing over sampling strata. For each sampling stratum m, the estimator of *VIP* is:

$$\hat{Y}_m = \sum_{fr} \sum_{i \in fr} h_{ifr} w_i y_i^{(b)} \left[1 + \frac{v_i^{(c)}}{x_i^{(a)}} \right] \tag{4.1}$$

$$= \sum_{fr} \sum_{i \in fr} h_{ifr} w_i b_i c_i y_i$$

$$+ \sum_{fr} \sum_{i \in fr} h_{ifr} w_i b_i (1 - c_i) y_i$$

$$+ \sum_{fr} \sum_{i \in fr} \left(\frac{h_{ifr} w_i c_i v_i b_i y_i}{x_i} \right)$$

$$+ \sum_{fr} R_{fr_2} \sum_{i \in fr} h_{ifr} w_i a_i (1 - b_i) x_i$$

$$+ \sum_{fr} R_{fr_2} \sum_{i \in fr} h_{ifr} w_i c_i (1 - b_i) v_i$$

$$+ \sum_{fr} R_{fr_2} \sum_{i \in fr} h_{ir} w_i (1 - a_i) d_i z_i \left[\sum_{st} h_{ist} R_{st} \left(\frac{5C}{SV} \right) R_{st} \left(\frac{6}{5C} \right) h_{if|st} \right]$$

$$+ \sum_{fr} R_{fr_2} \sum_{i \in fr} h_{ir} w_i (1 - a_i) d_i z_i \left[\sum_{st} h_{ist} R_{st} \left(\frac{5C}{SV} \right) h_{if|st} \right]$$

$$+ \sum_{fr} R_{fr_3} \sum_{i \in fr} w_i (1 - a_i)(1 - d_i) z_i \left[\sum_{st} h_{ist} R_{st} \left(\frac{5C}{SV} \right) R_{st} \left(\frac{6}{5C} \right) h_{ifr|st} \right]$$

$$+ \sum_{fr} R_{fr_3} \sum_{i \in fr} w_i (1 - a_i)(1 - d_i) z_i \left[\sum_{st} h_{ist} R_{st} \left(\frac{5C}{SV} \right) h_{ifr|st} \right]$$

$$+ \sum_{fr} \sum_{i \in fr} h_{ifr} w_i b_i y_i (1 - c_i) \left[\sum_{st} h_{ist} R_{st} \left(\frac{6}{5C} \right) \right]$$

$$+ \sum_{fr} R_{fr_2} \sum_{i \in fr} h_{ifr} w_i a_i (1 - b_i)(1 - c_i) x_i \left[\sum_{st} h_{ist} R_{st} \left(\frac{6}{5C} \right) \right], \tag{4.2}$$

where the indices, response indicators, imputation factors, and variables are defined below:

Indices

i = a sampled project in a sampling stratum,
k = a completed project in a sampling stratum,
m = the sampling stratum,
s = *PROJSELV* value group for *RVITM5C*, *AE&M*, and *STRTDATE* imputation cells,
t = the *MTC* for *RVITM5C*, *AE&M*, and *STRTDATE* imputation cells,
f = *RVITM5C* value group for *VIP* imputation cells,
r = *RANGE* group for *VIP* imputation cells.

Response Indicators

a_i = 1 if *RVITM5C* value is reported for project i; otherwise a_i=0,
b_i = 1 if *VIP* value is reported for project i; otherwise b_i=0,
c_i = 1 if *AE&M* value is reported for project i; otherwise c_i=0,
d_i = 1 if *STRTDATE* is reported for project i; otherwise d_i=0,
h_{ist} = 1 if project i is in *PROJSELV* value group s and *MTC* t; otherwise, h_{ist}=0,
h_{ifr} = response indicator for project i with reported *STRTDATE* and *RVITM5C*; h_{ifr} = 1 if project i is in *RVITM5C* value group f and *RANGE* r; otherwise, h_{ifr}=0,

$h_{ifr|st}$ = response indicator for project i with imputed *STRTDATE* and *RVITM5C*; $h_{ifr|st}=1$ if project i is in *RVITM5C* value group f and *RANGE* r; otherwise, $h_{ifr|st}=0$.

Imputation Factors

$$R_{fr_2}\left(\frac{VIP}{5C}\right) = \left(\frac{\sum_{fr} h_{ifr} w'_i b_i y_i}{\sum_{fr} h_{ifr} w'_i b_i x_i}\right),$$

VIP imputation factor as described in equation 3.1.1 for *RVITM5C* value group f and *RANGE* r,

$$R_{fr_3}\left(\frac{VIP}{5C}\right) = \left(\frac{\sum_{fr} h_{ifr} w'_i b_i y_i}{\sum_{fr} h_{ifr} w'_i b_i x_i + \sum_{fr} h_{kfr} w'_k b_k x_k}\right),$$

VIP imputation factor as described in equation 3.1.2 for *RVITM5C* value group f and *RANGE* r,

$R_{st}(5C/SV)$ = *RVITM5C* imputation factor as described in equation 3.2.1 for *PROJSELV* value group s and *MTC* t,

$R_{st}(6/5C)$ = *AE&M* imputation factor as described in equation 3.2.2 for *PROJSELV* value group s and *MTC*.

Variables

v_i = the *AE&M* value for project i,
w_i = the adjusted final weight for project i (*AE&M* adjustment factor divided out from the final weight),
w'_i = the final weight for project i,
x_i = the *RVITM5C* value for project i,
y_i = the monthly *VIP* value for project i,
z_i = *PROJSELV* value for project i,

$$x_i^{(a)} = \begin{cases} x_i \text{ if } a_i = 1, \\ R_{st}(5C/SV)z_i \text{ if } a_i = 0, \end{cases}$$

$$y_i^{(b)} = \begin{cases} y_i \text{ if } b_i = 1, \\ R_{fr_2}(VIP/5C)x_i \text{ if } b_i = 0,\ a_i = 1, \text{ and } d_i = 1, \\ R_{fr_3}(VIP/5C)x_i \text{ if } b_i = 0,\ a_i = 1, \text{ and } d_i = 0, \\ R_{fr_2}(VIP/5C)R_{st}(5C/SV)z_i \text{ if } b_i = 0,\ a_i = 0, \text{ and } d_i = 1, \\ R_{fr_3}(VIP/5C)R_{st}(5C/SV)z_i \text{ if } b_i = 0,\ a_i = 0, \text{ and } d_i = 0, \end{cases}$$

$$v_i^{(c)} = \begin{cases} v_i \text{ if } c_i = 1, \\ R_{st}(6/5C)x_i \text{ if } c_i = 0 \text{ and } a_i = 1, \\ R_{st}(6/5C)R_{st}(5C/SV)z_i \text{ if } c_i = 0 \text{ and } a_i = 0. \end{cases}$$

5. VARIANCE ESTIMATOR

As previously mentioned, the imputation rates for preliminary estimates are very high and treating imputed *VIP* as reported in standard variance formulas substantially underestimates the true variance.

Shao and Steel use a sample-response path considered by Fay (1991) which differs from the normal sample-response path.

> normal path: population ⇒ complete sample ⇒ sample with nonrespondents
>
> Fay's path: population ⇒ census with nonrespondents ⇒ sample with nonrespondents

Applying this concept of reversing the order of the sample-response sequence to the VIP Survey, the sample-response path is

> population ⇒ census with nonrespondents (γ) ⇒ sample with nonrespondents (s) ⇒ sample with nonrespondents (*);

where

> γ indicates randomness of responses for reporting *RVITM5C*, *VIP*, *AE&M*, and *STRTDATE*,
>
> s indicates randomness due to sampling,
>
> * indicates randomness of warm-deck imputation of *STRTDATE* falling into the correct *VIP* imputation cell.

We use an alternative sample-response path where we assume that the warm deck process closely simulates donors from the population. Since the random processes are independent, we switch the order of * and s without affecting the values. Conceptually, the final sample-response path becomes

> population ⇒ census with nonrespondents (γ) ⇒ census with nonrespondents (*) ⇒ sample with nonrespondents (s).

Then the variance decomposition is

$$Var_{\gamma *s}(\hat{Y}_m) = E_{\gamma}Var_{*s}(\hat{Y}_m|\gamma) + Var_{\gamma}E_{*s}(\hat{Y}_m|\gamma)$$

$$= E_{\gamma}E_{*}Var_{s}(\hat{Y}_m|\gamma,*) + E_{\gamma}Var_{*}E_{s}(\hat{Y}_m|\gamma,*) + Var_{\gamma}E_{*}E_{s}(\hat{Y}_m|\gamma,*)$$

$$= \quad V_1 \quad + \quad V_2 \quad + \quad V_3 \,,$$

where E_γ and Var_γ represent the expectation and variance with respect to the probability of response. Similarly, E_*, Var_*, E_s, and Var_s represent the expectation and variance in the respective random processes. This can be extended over all strata.

The first component, V_1, is the sample design variance which can be estimated by any standard variance estimator. We adopted a Census Bureau variance software package called VPLX developed by Fay which contains a variety of replication methods. We chose the stratified jackknife method with clusters of size 20 within each sampling stratum. We computed *VIP* variance estimates for preliminary March 1998 data by treating imputed values as reported (V_0) and by using equation 4.2 for each replicate (V_1). The computer processing time was 1 hour and 5 minutes.

Table 1 compares the two variance estimates V_0 and V_1 by *MTC*. Thirty-eight percent of the total value of construction for privately owned Nonresidential projects is imputed at the preliminary estimation stage. The sample design variance estimate for the $8,045,417,000 is underestimated by about 53 percent relative to V_1 if imputation methodology is ignored. On the other hand, the individual types of construction such as Industrial, Office, Commercial and Health Care are underestimated by about 10, 13, 22, and 15 percent, respectively. One reason why the underestimation for the total is much larger than the underestimation for individual types of construction is the fact that the *VIP* imputation cells do not depend on type of construction. Therefore, the imputation ratio factors we use for a specific type of construction are derived using all types of construction. If we combine Hotel and Office, the underestimation relative to V_1 increases to 23 percent. Now suppose we combine Hotel, Office, and Commercial. The underestimation relative to V_1 increases to 36 percent. As we continue to aggregate additional types of construction, the underestimation will continue to increase to the amount of underestimation for the Nonresidential total, 53 percent.

TABLE 1
COMPARISON OF DESIGN-BASED VARIANCE ESTIMATES

			VARIANCE ESTIMATES (Millions^2)		RATIO
MTC	VIP (Millions)	PERCENT IMPUTED	V_0	V_1	$\frac{V_0}{V_1}$
Hotel	709	49	569	732	.78
Office	1,421	36	3,498	4,029	.87
Commercial	2,024	39	3,280	4,207	.78
Health Care	995	37	2,091	2,449	.85
Education	536	29	1,279	1,317	.97
Religious	352	37	505	527	.96
Recreation	385	47	454	495	.92
Transportation	124	33	23	24	.96
Power	71	51	3	4	.77
Industrial	1,314	37	7,289	8,114	.90
Not Elsewhere Classified	112	28	359	362	.99
Total	8,045	38	18,226	38,458	.47

6. CONCLUSION

We believe our current *VIP* variance estimates which treat imputed values as reported are underestimated. We apply a decomposition method proposed by Shao and Steel to our survey which allows for variance due to imputation. We are able to successfully quantify the impact of imputation in the sample design variance component. Our initial results support the fact that in order to correctly estimate the true variance when imputation rates are high, the form of the estimator must reflect the actual imputation procedures used.

Although the derivation of the variance formulas may be complex and time-consuming, Shao and Steel's method has delivered encouraging results so far. It also allows us to implement one component at a time because of the additive nature of the variance estimation formula.

ACKNOWLEDGEMENTS

The authors greatly appreciate Donald Luery of Economic Statistical Methods and Programming Division for his guidance and advice throughout the research. We also thank Bob Fay of the Director's Office and George Train of Statistical Research Division for their assistance with VPLX.

REFERENCES

Deville, J.C. and Sarndal, C.E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, **10**, 381-394.

Fay, Robert E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, **91**, 490-498.

Fay, Robert E. (1991). A design-based perspective on missing data variance. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 429-440.

Rao, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, **91**, 499-506.

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.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, **82**, 453-460.

Rubin, Donald B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, **91**, 473-489.

Shao, J. and Steel, P.M. (forthcoming). Variance estimation for imputed survey data with non-negligible sampling fractions.