ESTIMATION OF VARIANCE DUE TO IMPUTATION IN THE TRANSPORTATION ANNUAL SURVEY (TAS)

Philip M. Steel, U.S. Bureau of the Census, Jun Shao, University of Wisconsin[1]
Philip M. Steel, U.S. Bureau of the Census, Statistical Research Division, Washington DC 20233

Key Words: Variance estimation, imputation, economic survey, survey design

Introduction

The Transportation Annual Survey (TAS) polls trucking and warehousing establishments about various forms of revenue, costs and inventory. As with many of the Census Bureau's surveys, item and establishment non-response is imputed in order to produce a "complete" data set. One of the problems attendant to this approach is an underestimation of variance, as calculated by traditional variance estimators. We present results of estimators that include the contribution of imputation to variance in TAS and discuss some of the problems and benefits encountered in the application.

The interaction between imputation and variance calculation has been an area of active research in recent years. Estimators for specific forms of imputation have been developed. For example, Rao and Shao (1992) develop an estimator for hot deck imputation; Lee, Rancourt and Särndal (1995) derive a estimator of variance for ratio imputation. Shao and Steel (forthcoming) describe a general methodology for deriving variance estimators for data with more than one kind of imputation. TAS employs a wide variety of imputation techniques and provides an opportunity to test the application of this methodology.

Methodology for variance estimation

Non-response is often modeled as an additional stage of sampling, after the usual sample. However we may take the conceptual tack introduced by Fay (1991), where one considers non-response as a characteristic of a portion of the population and that the sampling is 'performed' on the respondent population:

Population ⇨ Complete sample ⇨ Sample with non-respondents

vs.

Population ⇨ Census of respondents ⇨ Sample with non-respondents

So that instead of

$$V(\hat{Y}_I - Y) = E_S[V_\gamma(\hat{Y}_I - Y)] + V_S[E_\gamma(\hat{Y}_I - Y)] \quad (1)$$

we get the somewhat more convenient form

$$V(\hat{Y}_I - Y) = E_\gamma[V_S(\hat{Y}_I)] + V_\gamma[E_S(\hat{Y}_I - Y)] \quad (2)$$

where Y = population total for y
$\hat{Y}_I$ = Horvitz-Thompson estimated total
E_γ, V_γ are the expectation and variance with respect to the probability of response
E_s, V_s are the expectation and variance with respect to the sampling

The inner expectation and variance in (1) is conditional on the sampling, and in (2) is conditional on the response. The first term on the right of equation (2) we designate as v_1, the second as v_2. v_1 contains the naive estimate of variance and v_2 the majority of the variance due to imputation. Their Taylor approximations are of the form:

$$v_1 \approx [\nabla\psi(\hat{T})]' V [\nabla\psi(\hat{T})] \quad (3)$$

$$v_2 \approx [\nabla\varphi(E_\gamma \tilde{T})]' C [\nabla\varphi(E_\gamma \tilde{T})] \quad (4)$$

where V is a KxK matrix whose $(k,l)^{th}$ element is a standard design-based estimator of $\mathrm{Cov}_s(\Sigma_{i\in s} w_i a_{ki} t_{ki}, \Sigma_{i\in s} w_i a_{li} t_{li})$, and C is a (K+1)x(K+1) matrix whose $(k,l)^{th}$ element is $\mathrm{Cov}_\gamma(\Sigma_{i\in P} a_{ki} t_{ki}, \Sigma_{i\in P} a_{li} t_{li})$, $0 \le k, l \le K$, $a_{0i} \equiv 1$ and $t_{0i} \equiv y_i$, and

$$E(\hat{Y}_I) - Y = \varphi(\tilde{T}), \qquad \tilde{T} = (\sum_{i\in P} a_{1i} t_{1i}, \ldots, \sum_{i\in P} a_{Ki} t_{Ki}) \quad (5)$$

$$\hat{Y}_I = \psi(\hat{T}), \qquad \hat{T} = (\sum_{i\in s} w_i a_{1i} t_{1i}, \ldots, \sum_{i\in s} w_i a_{Ki} t_{Ki}) \quad (6)$$

where a_{ki} indicates the use of the k^{th} imputation type on the i^{th} observation, and t_{ki} is the reported or imputed value, w_i the establishment weight. The exact forms of (5) and (6) vary according to the imputation methodology, and will be introduced below.

---

[1] This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and so not necessarily reflect those of the Census Bureau.

The Survey

The Transportation Annual Survey (TAS) is a survey of warehouse and trucking. Two different survey forms are used; for this paper we will consider only those establishments receiving the warehouse form, about 650 establishments. The sample is drawn once every five years from a frame constructed from the Standard Statistical Establishment List (SSEL). The initial sample is stratified by kind of business and size (as determined by payroll or number of employees), stratum boundaries are generally determined by a cumulative square root of F rule. The number of strata varies from 6 to 10. The sampling is simple random sampling within strata. About one third of establishments are selected with certainty. Due to their large size, these may account for half or more of the estimated totals. The sampling rates for the remaining strata vary from 0.5 down to 0.015, most typically around 0.1. The initial sample is supplemented by a birth process, selected from additions to the SSEL. The survey is periodically benchmarked to the economic census. The results here are not benchmarked, and may differ substantially from the published, benchmarked numbers.

Payroll imputation

A variety of imputation techniques are used. The imputations of first resort are based on establishment-level data, e.g., current year business expenses may be imputed by the ratio of current year to prior year payroll applied to prior year business expenses. If there is not sufficient data to perform any "cold deck" imputation, a cell based ratio estimate may be used. The imputation cell for ratio imputation is the entire kind of business. For the payroll variable six different imputations are used (the numbering convention will be used throughout; cy stands for current year, py for prior year):

a1-reported
a2-(cy expenses/py expenses)*py reported payroll
a3-(cy expenses/py expenses)*py admin payroll
a4-(cy admin payroll/py admin payroll)*
py reported payroll
a5-cy admin payroll
a6-SIC level ratio of payroll to expenses*
cy expenses
a7-SIC level ratio of current to prior year reported
payroll*py reported payroll

While all the methods are used on some occasion, methods A4, A5, and A6 are used for all but 2% of cases. A case for suppressing the other methods of imputation can probably be made, but for this variable we implement the full methodology.

Payroll variance estimation

The current variance estimation is done using a random group method. The estimation is done separately for certainty and non-certainty cases. It does not include a finite population correction, perhaps as an intentional overestimation. We include those in estimates in table 1, but need a naive estimate of variance that includes fpc.

For v_1 we use both a repeated random replication estimator (RRR) and a linear estimator, given in equation (7).

$$V_1=\sum_{h=1}^{H}(1-\frac{n_h}{N_h})\sum_{i\in s_h}(\frac{n_h}{n_h-1}\xi_i-\frac{1}{n_h}\sum_{i\in s_h}w_i\xi_i)^2 \qquad (7)$$

For the RRR method the data is reimputed using its replicate weights, yielding different ratios for imputation. The differences arising from the reimputed ratios in the total estimate are negligible for the payroll variable. Administrative and other cold deck values are fixed. Since any random imputation (e.g. ratio) contributes to v_1, we cannot say that v_2 contains all the variance due to imputation, though in the cases examined here the imputation variance effects v_1 only slightly.

The RRR estimate is a replicate method with some similarity to a random group estimate with the number of random groups, G, equal to the minimum stratum size plus one. The function of the G^{th} group is to allow an assignment that results in random groups of uniform size, extra units are placed into the G^{th} group. Rather than making G-1 independent estimates of the total, the i^{th} replicate estimate is made by giving the i^{th} group a high weight and members of all other groups a relatively low weight.
The high weights were around 4 and the low weights around 0.7. G-1 estimates of the total are made for each random group assignment, the process of random group assignment and estimation are repeated until the estimate has "settled", in our case the procedure was repeated 5 times. For each random group assignment there are G random groups but only G-1 estimates are made. The "extra" random group never receives the high weighting and is used to ensure that all of the other random groups have the same number of members from each strata. Since every case needs the opportunity to receive the high weight, the process

must be repeated several times. In the worst case the probability of being assigned to the (extra) G^{th} group approaches 0.5 for a strata of size 2*(G-1)-1. In the best case the G^{th} group is empty, where the stratum divides up evenly into the G-1 group. In the application, G was taken to be 11 for receipts estimation and allowed to vary (upward) for the payroll imputation. Shao and Chen (draft) will addresses this method in detail.

This gives us three estimates of v_1, including the method used in current production. For v_2 we derive an estimator by the method outlined above. To do this we express the payroll variable as:

$$\hat{Y}_I=\sum_{i\in s} w_i(a_{1i}y_i+a_{2i}\frac{z_i\tilde{y}_i}{\tilde{z}_i}+a_{3i}\frac{z_i\tilde{x}_i}{\tilde{z}_i}+a_{4i}\frac{x_i\tilde{y}_i}{\tilde{x}_i}+ a_{5i}x_i+a_{6i}\hat{R}_{\underset{z}{y}}z_i+a_{7i}\hat{R}_{\underset{\tilde{y}}{y}}\tilde{y}_i) \quad (8)$$

where $\sim$ indicates prior year, y is payroll, x is administrative payroll, z is total expenses and R is a cell ratio estimate of proportion. That is, by using indicators a_{ki}, we explicitly include the imputation calculation for nonrespondents. With this expression for Y, we then use (4) to calculate v_2 :

$$V_v[E_s(\hat{Y}_I)]\approx \sum_{2\le k\le 5} p_k(1-p_k)\sum_{i\in P} e_{ki}^2-2\sum_{2\le k\le l\le 5} p_k p_l\sum_{i\in P} e_{ki}e_{li}+ \sum_{6\le k\le 7}(c_k+1)p_k\sum_{i\in P} e_{ki}^2+2\sum_{6\le k\le l\le 7}\frac{p_k p_l}{p_1}\sum_{i\in P} e_{ki}e_{li} \quad (9)$$

with p_k estimated by $\sum_{i\in s}w_i a_{ki}/\sum_{i\in s}w_i$, $c_6\approx p_6/p_1$ and $c_7\approx p_7/p_8$ where p_8 is the probability of observed reported values for y in both the current and prior year. The e_{ki} are the differences between the imputed estimate and the actual value. We then estimate $\sum e_{ki}e_{li}$ by $\sum_{i\in s}w_i a_{ki}e_{ki}e_{li}\sum_{i\in s}w_i/\sum_{i\in s}w_i a_{ki}$ where a_{ki} indicates that the differences between reported and imputed values can be calculated for both e_{ki} and e_{li}.

A comparison of the estimates of v_1 can be found in table 1.

The adminstrative data supplying values for the e_{ki} was subjected to an edit prior to its use. The edit required that the ratio of current-year-administrative payroll to prior-year payroll not exceed 20 nor fall below 0.05. The edit changed the value for a_{ki} from 1 to 0 for 7 observations.

Table 2 gives the v_2's contribution to the total variance ($v_2/(v_1+v_2)$). The results indicates that the naive estimates (v_1) underestimate the true variance, perhaps by as much as 50%.

Receipts imputation

Receipts imputation is seemingly straightforward. If the establishment has prior data, the ratio of prior to current year payroll is applied to the prior year receipts. If it is the first year of the sample and the case was not in the prior sample, or if it is a birth, receipts are imputed by the cell ratio of current year receipts to payroll times the establishment's current year payroll. These techniques are employed regardless of whether or not the values are imputed, and the payroll imputation precedes receipts imputation. In the first year there are 15 possible combinations of payroll and receipts imputation. In the second year there are 30. And so on. Not all combinations can occur in practice, and as noted above, many payroll imputations are infrequently used. We created a partial classification of occurring imputations and their frequence of use. Since each imputation contributes several terms to v_2, we were forced to collapse some of these imputation types together, primarily by ignoring the difference between administrative and reported payroll. For receipts imputation we used the following categories:

a1- reported
a2- py receipts*(cy admin payroll/py admin payroll)
a3- cy admin payroll*(91 reported receipts/ 91 reported payroll)
a4- cy admin payroll*(SIC level ratio of cy reported receipts to reported payroll)
a5- cy admin payroll*(SIC level ratio of 90 reported receipts to reported payroll)

The appearance of the 1990 and 1991 data is due to cancellation of administrative payroll for units that are consistent nonrespondents. Table 3 shows the frequency of the collapsed types and response rates for the receipts variable.

Receipts Variance Estimation

For receipts we have only calculated v_1 using the RRR method described above. To find v_2 we express the receipts variable as:

$$\hat{Y}_I=\sum_{i\in s} w_i(a_{1i}y_i+a_{2i}\frac{z_i\tilde{y}_i}{\tilde{z}_i}+a_{3i}\frac{z_i y_i^{(91)}}{x_i^{(91)}}+ a_{4i}\hat{R}_{\underset{x}{y}}z_i+a_{5i}\hat{R}^{(90)}_{\underset{x}{y}}z_i) \quad (10)$$

Again from (4) we derive v_2 to be:

$$V_v[E_s(\hat{Y}_I)]\approx \sum_{k=2,3} p_k(1-p_k)\sum_{i\in P} e_{ki}^2-2p_k p_l\sum_{i\in P} e_{2i}e_{3i}+ (c_4+1)p_4\sum_{i\in P} e_{4i}^2+[(c_5+1)p_5-2c_5p_{5,6}]\sum_{i\in P} e_{5i}^2- 2\sum_{k=2,3} c_5p_{k,6}\sum_{i\in P} e_{ki}e_{5i}+2c_5(c_4p_{1,6}-p_{4,6})\sum_{i\in P} e_{4i}e_{5i} \quad (11)$$

with p_k is estimated by $\sum_{i\in s} w_i a_{ki} / \sum_{i\in s} w_i$, $c_4 \approx p_4/p_1$ and $c_5 \approx p_5/p_6$ where a_6 is the indicator for an observed value for y in 1990 so that $p_6 = E(a_6)$. We again estimate $\sum e_{ki} e_{li}$ by $\sum_{i\in s} w_i a_{kli} e_{ki} e_{li} \sum_{i\in s} w_i / \sum_{i\in s} w_i a_{kli}$ where a_{kli} is an indicator showing that the differences between reported and imputed values can be calculated for both e_{ki} and e_{li}.

Table 4 gives the values found for v_2 and calculate its contribution to the total variance ($v_2/(v_1+v_2)$). We observe that the naive component comes directly from observed values and the components of the imputation variance depend in part on response rate and imputation methodology. We have no explaination, as yet, for the variability of the contribution of v_2. Some of the same features appeared in an examination of the imputation bias and variability (ie some of the reported data lies far from the imputed value).

Conclusion

We are able to estimate the variance due to imputation for several variables in a survey with complex imputation. The procedure required the derivation of a formula specific to the imputation methodology and the imputation had to be thoroughly analyzed in the data set. The effect of imputation varied, but in the most effected case the confidence interval around the estimate was half again as large as the interval derived from the naive variance. The average effect was around a 10%.

A great deal of our efforts went in to finding a good value for the standard variance. As mentioned previously the currently used estimator has built in overestimation, by ignoring the finite population correction. We reimputed the data set for each replicate, recomputing the ratio estimates. Somewhat surprisingly, this had very little effect, though there is no guarantee that this will always be the case. It would be useful to have conditions under which reimputation can be ignored.

The RRR method needs modification to properly handle the birth process. The method we employed is not appropriate for calculating year to year statistics. Whether or not deaths should be retained is under review and makes a substantial difference in the final variance estimates.

There are several weakness in the application. First, some sort of edit on the administrative data was necessary. It seems clear that if a value would be rejected if used in an imputation, that value ought not be used in measuring the accuracy or variability of the imputation method. Administrative data not utilized by the survey is not currently subjected to edit or review. To utilize this method (or, quite likely, any method) of calculating the variance due to imputation, requires some changes and additions to the processing of the survey in order to review and edit all administrative data. The diversity of the imputation methodology, when imputations can be based on prior imputations, is another problem. The collapsed imputation types we used should give a fair, though not rigorous, approximation. Though many of the remaining variables in the survey follow the same basic pattern as the receipts imputation, the sheer number (40+) poses a formidable obstacle to implementing this for these non-critical variables. A final problem is posed by the post-stratification of trucking firms (SICs in the 41 series) for imputation. The trucking firms are split into "specialty", "general" and "unknown" depending on their response to the survey; this additional classification is used in creating the imputation classes. Hence imputation classes no longer contain strata, which may interfere with the calculation of v_2.

Some remedies may be available. It may be possible to show that the interaction terms in equations (9) and (11) are negligible, certainly we observed no instance in which the interaction term exceeded 10% of v_2. This would simplify the formulas considerably. On the survey side, I note that the imputation design was driven primarily by item nonresponse, when in fact the majority of imputation is for unit nonresponse. Some redesign of the imputation procedure with this in mind would simplify the methodology with little sacrifice to accuracy, e.g. impute directly from payroll rather than from a value that itself was imputed from payroll. Consideration might also be given to explicitly using the current year ratio of identicals, rather than implicitely using the ratio that was current when a nonresponding unit entered the sample.

In addition to eventually providing a publishable, more accurate estimate of the variance, the technique has several more immediate benefits. Preliminary results, or results for critical variables, can inform decisions made for a survey in its design stage, e.g. a decision to oversample in a particular strata. It also forces one to thoroughly review imputation and can easily produce an ongoing analysis of imputation as a byproduct of the variance estimation.

REFERENCES

Fay, B.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.

Lee, H., Rancourt, E. and Särndal, C. E. (1995). Variance estimation in the presence of imputed data for the generalized estimation system. *Proceedings of the Section on Survey Research Methods,* American Statistical Association, 384-389.

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

Shao, J. and Chen, Y. (forthcoming). Balanced half samples for survey data under imputation.

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

Table 1
Standard Variances (v_1); Survey year by Standard Industrial Classification

YEAR	V4221	V4222	V4225	V4226	METHOD
1991	3.30E+13	9.28E+13	7.20E+14	1.25E+15	RRR
1992	4.22E+13	2.90E+14	2.10E+15	2.47E+15	RRR
1993	6.73E+14	4.19E+14	3.93E+15	2.55E+15	RRR
1994	1.05E+14	3.69E+14	2.86E+15	4.24E+15	RRR
1995	8.54E+13	2.87E+14	5.22E+15	7.42E+15	RRR
1991	8.51E+13	1.15E+14	4.49E+14	1.93E+15	CURRENT
1992	1.29E+14	2.20E+14	2.13E+15	3.05E+15	CURRENT
1993	1.74E+14	3.36E+14	4.19E+15	3.90E+15	CURRENT
1994	2.15E+14	4.17E+14	3.22E+15	4.47E+15	CURRENT
1995	3.37E+14	7.22E+14	3.53E+15	5.34E+15	CURRENT
1991	3.50E+13	1.22E+14	6.33E+14	1.28E+15	linear
1992	5.57E+13	2.30E+14	1.75E+15	1.97E+15	linear
1993	8.69E+13	3.59E+14	3.57E+15	2.54E+15	linear
1994	8.78E+13	4.79E+14	3.08E+15	3.52E+15	linear
1995	1.51E+14	7.29E+14	4.09E+15	4.47E+15	linear

Table 2
v_2 contributions to total variance (with RRR as v_1)
Survey year by Standard Industrial Classification

YEAR	V4221	V4222	V4225	V4226	METHOD
1991	29.3%	69.7%	51.9%	2.7%	v2/v1+v2
1992	17.7%	7.4%	32.6%	2.8%	v2/v1+v2
1993	10.0%	6.8%	18.5%	3.4%	v2/v1+v2
1994	13.5%	8.8%	2.8%	2.8%	v2/v1+v2
1995	1.7%	9.4%	25.6%	2.1%	v2/v1+v2

Table 3
Receipts imputation types, collapsed

IMPUTATION TYPE	COUNT	PERCENT
reported	2687	78.2
cy admin payroll*(py receipts/py admin payroll)	223	6.7 (a2)
cy admin payroll*(91 reported receipts/91 reported payroll)	63	1.8 (a3)
cy admin payroll*(SIC level ratio of cy reported receipts to reported payroll)	172	5.0 (a4)
cy admin payroll*(SIC level ratio of 90 reported receipts to reported payroll)	285	8.3 (a5)

Weighted response rates, receipts

SURVEY_Y	4221	4222	4225	4226
1991	73.6	70.9	83.9	67.9
1992	58.2	72.6	65.6	71.7
1993	66.8	67.6	69.1	70.7
1994	73.2	62.2	69.7	74.5
1995	74.3	74.3	80.9	66.0

Table 4

Receipts
Comparison of CVs and v_2's contribution to total variance

CV for v_1

YEAR	CV4221	CV4222	CV4225	CV4226
1991	7.01%	3.93%	6.28%	8.40%
1992	5.16%	4.15%	6.34%	6.72%
1993	7324%	3.52%	5.37%	10.34%
1994	5.39%	5.84%	5.09%	8.60%
1995	7.38%	6.14%	9.76%	7.25%

CV for v_1+v_2

YEAR	CV4221	CV4222	CV4225	CV4226
1991	7.83%	4.11%	6.85%	8.57%
1992	6.88%	4.28%	7.51%	6.98%
1993	9.23%	3.73%	6.39%	10.48%
1994	8.01%	6.00%	5.22%	9.01%
1995	7.47%	6.23%	10.14%	7.75%

$v_2/(v_1+v_2)$

YEAR	4221	4222	4225	4226
1991	19.89%	8.94%	15.76%	3.81%
1992	43.82%	6.18%	28.61%	7.43%
1993	38.52%	10.93%	29.37%	2.68%
1994	54.80%	5.16%	4.98%	8.84%
1995	2.52%	2.63%	7.34%	12.59%