A Study of Basic Calibration Estimators and their Variance Estimators in Presence of Nonresponse

Yves Thibaudeau, Jun Shao, Jeri Mulrow

U.S. Census Bureau, University of Wisconsin, National Science Foundation
Yves.Thiaudeau@Census.gov, Shao@cs.wisc.edu, JMulrow@nsf.gov

1. Introduction

The Survey of Research and Development in Industry (SRDI), sponsored by the National Science Foundation, is an annual survey involving over 30 thousand sampled companies potentially involved in research and development. Most of the yearly R&D dollar investments in the U.S. are attributable to a relatively small collection of companies, around 5,000. The SRDI surveys most of this core of companies by deterministically selecting them in certainty strata. R&D dollars from these companies are added directly to total estimates, at the state or country level. In 2004, the R&D dollars from the certainty strata accounted for more than 80% of all R&D in the U.S. (table 1). In some states, the R&D dollars from the certainty strata total exceeded 90% of the state totals. The certainty strata totals do not generate a sampling error. But, some companies in the certainty strata do not report their yearly R&D investment. For those, R&D investment is imputed. So, the totals from the certainty strata are subject to nonresponse and imputation errors.

The nonresponse error could be considerable. Table 1 displays the imputed totals for the certainty strata, as well as the totals generated by sampling smaller companies. In the case of California, for example, the imputed R&D is almost as large as the total R&D generated by sampling. So, the nonresponse error is competing with the sampling error size wise.

Currently, a sensible ad-hoc imputation procedure has been used to estimate unreported R&D for the certainty strata. But, and this is the motivation of the paper, no formal statistical assumptions have been developed or presented to support this procedure. Our intention is to research nonresponse compensation procedures germane to the current approach, such as longitudinally-based estimators, and other procedures based on statistical principles, such as calibration estimation.

This paper is set to accomplish two objectives.

1. Identify statistically principled estimators compensating for R&D nonresponse for the dollar amounts of R&D at the state and country level, for the certainty strata. In particular, explore longitudinally-based and calibration estimators. Identify the best of those estimators.
2. Estimate the nonresponse variance for those estimators, proceeding from the statistical principles validating these estimators.

Table 1 - Share of Total Estimated R&D in the Certainty Strata for the SRDI in Eight Prominent States

State	Noncertainty Strata: Number of Companies in Sample	Certainty Strata: Number of Companies in Sample	Noncertainty Strata: Total R&D	Certainty Strata: Total R&D	Certainty Strata: Imputed R&D
U.S.	20016	12046	19947	188353	13005
CA	2747	1632	6151	37376	5551
CT	350	258	223	7839	92.3
IL	888	551	694	9994	410
MA	651	486	1361	11647	792
MI	803	381	396	15309	964
NJ	710	460	379	19569	866
NY	1196	648	741	17754	476
TX	590	1143	844	9350	575

Section 2 presents background on calibration estimators. This class of estimators includes some longitudinal estimation methods germane to the current imputation procedures. In future work, we will make use of the same calibration set-up to expand other types of estimators involving frame information. At the time of this write-up not all the information needed to expand frame-calibrated estimators was available to us. So, we focus on longitudinally-calibrated estimator. Research to compare frame calibration to longitudinal calibration is in progress.

Section 3 describes the specific calibration estimators we consider in the paper, along with the statistical principles motivating them. Numerical results are given for the nonresponse variance of the estimates of total R&D obtained by calibrating. Section 4 discusses other avenues for estimating R&D totals and the nonresponse variance.

2. Calibration Estimation in Presence of Nonresponse

Sarndal et al (2005) propose a general paradigm to illustrate the estimation possibilities in the context of calibration. The calibration operation is instated through a calibration equation. The calibration equation revolves around an auxiliary variable for which comprehensive information is available. Let $\boldsymbol{X}$ be an always observed auxiliary variable. The general form of the calibration equation --or estimation equation-- is

$$\sum_{i \in S_r} w^i \boldsymbol{x^i} = \boldsymbol{X} \qquad (1)$$

This equation must be solved for the calibrating weights $w_i, i \in S_r$, where S_r is the set of subscripts representing the responding survey units. The solution to (1) depends on the specific type of calibration applied to derive the calibration estimator. In the paper, the type of calibration we apply to derive solutions to (1) is based on a linear form (Sarndal Lundstrom p 59). The linear form leads to sensible estimators, but other forms are available in a more general setup (Deville 2000).

Given a solution $w^i, i \in S_r$ for (1), the estimator of the total investment calibrated to the auxiliary information $\left(\left\{x^i; i \in S\right\}, \boldsymbol{X}\right)$ is

$$\hat{Y}_W = \sum_{i \in S_r} w^i y^i \qquad (2)$$

In (1) and (2), $\boldsymbol{X}$ is a calibrator defined at the level of the entire survey universe. In the case of the SRDI, $\boldsymbol{X}$ could be the total payroll of all U.S. companies for each industry type and x^i the payroll for company *i*. In the U.S., every company is required by law to provide this information to the Internal Revenue Service.

In the context of nonresponse, the sample itself is a comprehensive universe relative to the subsample of responding units. So, in addition to the level of the entire survey universe, calibration can be instated at the level of the sample. Under that scheme, $\boldsymbol{X}$ carries information at the level of the sample only. The goal then becomes to use the calibrating information to damper the impact of nonresponse in the form of increased variance. The paper reviews two simple calibration estimators to compete with the current estimator. Both are calibrated to sample-level information.

3. Hot-Decked Unweighted Estimators vs. Calibration Estimators for Estimating R&D

Our geographical unit for conducting analysis is the state. We will aggregate state-level results to obtain statistics descriptive of the entire U.S. We center the attention on the recurring companies in the certainty strata. There is no sampling error involved when constructing an estimator based only of this set of companies.

We consider two types of error when estimating total R&D: the Nonresponse bias and the nonresponse variance. The nonresponse bias is the chronic one-sided departure between the estimator and the true population value it estimates. The nonresponse bias is not observable. But, few believe it is possible to produce estimates completely free of nonresponse bias. The reason is that unrealistic assumptions, such as data missing at random (MAR) or data missing completely at random (MCAR) are usually made to derive estimators in presence of nonresponse. It is the task of the statistician to reduce the nonresponse bias as much as possible.

The nonresponse variance quantifies the uncertainty around the values of the estimator due to the fact that some companies did report R&D, and so R&D had to be imputed. Unlike the bias, it is possible to directly estimate the nonresponse variance. We will use nonresponse variance estimates to asses the accuracy of our calibrated estimators of total R&D investment.

3.1 Current Method

We first look at the current method to impute missing R&D and compensate for nonresponse. For our universe, the set of recurring units, the current method for imputing missing 2004 R&D substitutes the corresponding R&D values from the 2003 survey after adjusting them for industry growth, based on deterministic factors. Problems arise when the 2003 R&D totals are not reported either. Then an earlier value is retrieved, or a de facto mean imputation takes place. We show that, if 2003 R&D was reported for all the 2004 nonresponding companies, total R&D derived from this method is approximately the same as that derived from a calibration procedure we describe in section 3.4.

3.2 Set up and Notation for Calibration Estimation

We will derive two calibration estimators to estimate total R&D for the recurring companies in the certainty strata. We first present the notation. Let S^{2004} and S^{2003} be the sets of the indices representing the companies in the 2004 and 2003 samples respectively and let S represent the overall universe of companies. Since we consider only the recurring cases in the certainty strata, we have $S = S^{2004} = S^{2003}$. Then, let S_r^{2004} and S_r^{2003} be the set of indices representing the companies reporting R&D in 2004 and 2003, and let S_{nr}^{2004} and S_{nr}^{2003} be the set of indices representing the nonresponding companies in the same years. The auxiliary information we will calibrate to, when deriving the first calibration estimator of total R&D, is

$$\alpha = \left(\left\{ I\left(i \in S_r^{2004} \right); i \in S \right\}, N \right)$$

The auxiliary information we will calibrate to, when deriving the second calibration estimator is

$$\beta = \left(\left\{ I\left(i \in S_{nr}^{2004} \cap S_r^{2003} \right);\ i \in S \right\}, N_{nr}^{2004} \right).$$

So, the calibrating variable for the first calibration estimator is the domain-inclusion indicator $I\left(i \in S_r^{2004}\right)$, available for all $i \in S$. The auxiliary information also includes the aggregate *N*, the size of the full sample. Equation

(1) can then be reproduced with $I\left(i \in S_r^{2004}\right)$ and N in lieu of $\boldsymbol{x^i}$ and $\boldsymbol{X}$. Similarly, for the second calibration estimator, the calibrating variable is the indicator $I\left(i \in S_{nr}^{2004} \cap S_r^{2003}\right)$. The aggregate is N_{nr}^{2004}, that is the size of population not reporting R&D in 2004. We describe in more detail how these two calibration estimators are implemented in the next two sections.

3.3 One-Way Calibration on the Stratification for the Self-Representing Units

We define in more details the first calibration estimator for total R&D: the one-way calibration estimator. The reference universe for this estimator is the level of the state. This choice is motivated by the natural tendency for R&D companies to geographically cluster around major cities.

Let $\bar{Y}_r^{2004}$ be the mean of the reported R&D investments (from the responding units) in a state. The estimator calibrated to the auxiliary information α is:

$$\hat{Y} = N\bar{Y}_r^{2004} \qquad (3)$$

where

$$\bar{Y}_r^{2004} = \frac{\sum_{i \in S_r^{2004}} y^{2004,i}}{N_r^{2004}}$$

$y^{2004,i}$ is the analysis variable, that is R&D for company i in 2004. Implicitly, the computation of $\hat{Y}$ involves imputing the nonresponding companies. Imputed R&D is equal to $\bar{Y}_r^{2004}$ for each nonresponding company in the state. The nonresponse variance for $\hat{Y}$ can be estimated analytically. Sarndal (2005) propose the estimator

$$\hat{V}\left(\hat{Y}\right) = \sum_{i \in S_r^{2004}} \left(\frac{N}{N_r^{2004}}\right)\left(\frac{N}{N_r^{2004}} - 1\right)\left(y^{2004,i} - \bar{Y}_r^{2004}\right)^2 \qquad (4)$$

Note, if R&D investments are reported for all the units sampled in the state, $\hat{V}\left(\hat{Y}\right)$ is 0. Table 1 and 2 gives values of $\hat{Y}$ and its variance components for selected states and for the entire U.S.

The assumptions needed to validate the one-way estimator are strong. We must assume that the unreported R&D investments are missing completely at random at the state level. In another words, inclusion to a state completely explains the missing data mechanism. This is likely not true. Other factors, such as the level of R&D investment itself, may contribute to a company's decision not to report it –e.g. if it is near 0. When using this estimator, we should be prepared for a significant nonresponse bias.

3.4 Calibration of R&D Investment to Response Status

The second estimator involves the two-year auxiliary information β. We want to improve on the one-way estimator. To do so we look for a pseudo-strata partition that divides S in homogeneous classes with respect to company nonresponse mechanisms. Ideally our pseudo-strata partition discriminates between company propensity scores. In reality the propensity scores are not available and can only be inferred though variables correlated with propensity.

In our situation we have access to the auxiliary information β. So we can calibrate the analysis variable to the indicator $I\left(i \in S_r^{2003} \cap S_{nr}^{2004}\right)$. Our assumption is that this indicator distinguishes between two classes of

companies, each of them being homogeneous with respect to the propensity to report R&D in 2004. Our calibration estimator is

$$\hat{\hat{Y}} = N_r^{2004}\ \bar{Y}_r^{2004} + N_{nr}^{2004}\left(\bar{Y}_r^{2003/2004} - \bar{Y}_r^{2004}\right) \quad (5)$$

where

$$\bar{Y}_r^{2003/2004} = \frac{\sum_{i \in S_r^{2003} \cap S_{nr}^{2004}} y^{2003,i}}{N_r^{2003/2004}}$$

and

$$N_r^{2003/2004} = \sum_{i \in S_r^{2003} \cap S_{nr}^{2004}} 1$$

Note that the variable of analysis is $y^{2003,i}$, that is 2003 R&D for unit *i*. A nonresponse variance estimator for $\hat{\hat{Y}}$ is

$$\hat{V}\left(\hat{\hat{Y}}\right) = \sum_{i \in S_r^{2003} \cap S_{nr}^{2004}} \left(\frac{N_{nr}^{2004}}{N_r^{2003/2004}}\right)\left(\frac{N_{nr}^{2004}}{N_r^{2003/2004}} - 1\right)\left(y^{2003,i} - \bar{Y}_r^{2003/2004}\right)^2 \quad (6)$$

In (5), $\bar{Y}_r^{2003/2004}$ is the average of the 2003 reported R&D over the $N_r^{2003/2004}$ companies represented in the set $S_r^{2003} \cap S_{nr}^{2004}$. That is $\bar{Y}_r^{2003/2004}$ is the 2003 R&D average for the companies reporting R&D in 2003, but not reporting R&D in 2004. Table 2 compares the current total R&D estimator with $\hat{\hat{Y}}$ and $\hat{Y}$. The nonresponse standard errors (square root of the Nonresponse variance) of $\hat{\hat{Y}}$ and $\hat{Y}$ are also reported. We see $\hat{\hat{Y}}$ is much closer to the current estimator than $\hat{Y}$. Also, table 2 exhibits a nonresponse standard error (root of the nonresponse variance) considerably smaller for $\hat{\hat{Y}}$ than for $\hat{Y}$.

Table 2. Current and New Calibration Estimates (Recurring Units) x 1,000,000

Area	Current Estimator	$\hat{Y}$	$s.d.\left(\hat{Y}\right)$	$\hat{\hat{Y}}$	$s.d.\left(\hat{\hat{Y}}\right)$
U.S.	178442	219147	10985	189790	3256
CA	34274	39667	3871	36003	1788
CT	7614	9403	1483	7595	37.40
IL	9794	12003	1973	10326	385.1
MA	10078	13834	3561	12171	801.6
MI	14999	18486	3825	15777	1030
NJ	18963	23351	2928	18617	221.5
NY	17391	23340	4916	17916	309.5
TX	8464	10347	1477	8546	201.7

4. Discussion and Future Research

The current R&D imputed dollar amount derived from the SRDI is around $13 billions. This dollar amount is approximately two-third of that collected through the probability sample. In this situation, reporting only the sampling variance, as a measure of variability, could be dramatically understating its true value. The paper presents estimation methods that allow for the estimation of the variance of the nonresponse error. This variance does quantify the variability stemming from the fact R&D was unreported and then imputed for a large percentage of companies (table 4.)

The estimators presented in the paper are based on clear statistical assumptions. When these assumptions are valid, or close to valid, (4) and (6) are valid nonresponse variance estimators. The first estimator, the state average, is based on crude, assumptions. Namely it assumes a MCAR process at the state level. The second estimator is based on more adaptive assumptions. Its statistical validity relies on the assumption that the nonrespondents for the 2004 survey form a homogeneous universe with respect to the nonresponse mechanism, regardless of whether or not they reported R&D for the 2003 survey. An ANOVA test (table 3) supports this assumption. The F test points to heterogeneity between 2004 respondents and nonrespondents, and to relative homogeneity within these two clusters.

The second estimator $\hat{\hat{Y}}$ is in fact germane to the current procedure. If the 2004 nonreporting companies all report R&D for the 2003 survey, the current method leads to the same estimator of total R&D as $\hat{\hat{Y}}$, up to a deterministic adjustment factor for changes between 2003 and 2004 at the level of the industry. So, in that case our variance formula can be extended to quantify the nonresponse error of the current method. However, overwhelmingly, companies not reporting R&D in 2004 are also not reporting R&D in 2003. Table 4 shows how the sample universe is divided between the four configurations of reported/not reported R&D in 2003/2004. In practice, it is possible to retrieve older information to compensate for the companies not reporting R&D both in 2004 and 2003. However, the possibility of a bias becomes more real when old data are used to compensate for current data. This argument favors using an estimator of the $\hat{\hat{Y}}$ type to estimate total R&D. More research involving multiyear estimators must be conducted to understand the trade-offs between using estimators of the same type as $\hat{\hat{Y}}$ and estimators involving information recorded at additional points in time.

Beyond allowing for the estimation of the nonresponse variance, $\hat{\hat{Y}}$ naturally leads to formal statistical comparisons across time that were not historically feasible. Provided appropriate variance estimators are derived, it will be possible to statistically assess whether or not there has been growth or decline in R&D at the state and country level from year *t*-1 to year *t* through the statistics $\hat{\hat{Y}}^{t} - \hat{\hat{Y}}^{t-1}$.

Future work will center the attention on the properties of estimators that make use of data collected at additional points in time, when available. In addition, the use of frame information to calibrate cross-section estimators will be explored and evaluated.

Table 3. ANOVA: 2003 Total R&D for 2003 Respondents by 2004 Response Status and SICRCD

Source	D.F.	Mean Square	Sum Squares	F	Significance
2004 Response Status	1	1.74 x 10**17	1.74 x 10**17	3.09	.078
SICRCD	44	9.09 x 10**18	2.06 x 10**17	3.66	< .0001
Full Model	45	9.27 x 10**18	2.06 x 10**17	3.65	< .0001
Model Error	4355	2.46 x 10**20	5.64 x 10**16		

Table 4. Response Status for Recurring Companies in 2003 – 2004.

	2004 R&D Reported	2004 R&D Not Reported
2003 R&D Reported	3525	434
2003 R&D Not Reported	412	843

References

Deville, J. C. (2000). "Generalized Calibration and application to weighting for Non-Response." Proceedings in Computational Statistics, Betlehem and Van Der Heijden eds. pp. 65-76.
Sarndal, C. E., Lundstrom, S. (2005). "Estimation in Surveys with Nonresponse." Wiley.
Request for OMB Review. Survey of Industrial Research and Development (R&D) (Form RD-and RD-1A).