Estimation of Totals and Regression Parameters by Combining Data from Two Independent Surveys

J. N. K. Rao*
Carleton University, Ottawa, Canada jrao@math.carleton.ca

Jae Kwang Kim
Iowa State University, Ames, Iowa, USA jkim@iastate.edu

We consider two different scenarios of combining data from two independent surveys to make inferences on parameters of interest. In scenario 1, a large sample from survey 1 observes only auxiliary variables related to a variable of interest and a much smaller sample from survey 2 observes both the variable of interest and the auxiliary variables. We generate synthetic values of the variable of interest by fitting a working model to survey 2 data and then predicting the variable of interest associated with the auxiliary variables observed in survey 1. A projection estimator of a total or a domain total is simply obtained from survey 1 weights and associated synthetic values reported in survey 1 data file. Replication variance estimators are obtained by augmenting the synthetic data file for survey 1. In scenario 2, regression analysis is studied when some of the predictor variables of interest are observed in a different survey. Instrumental variables and fractional imputation are used to implement statistical matching or data fusion and make inference on the regression parameters from the completed data file.

Key Words: Fractional imputation, instrumental variables, regression analysis, synthetic values.

I. Introduction

We consider two different scenarios of combining data from two independent surveys from the same target population consisting of N elements to make inferences on parameters of interest. In scenario 1, a large sample A_1 from survey 1 collects information on a variable x and a much smaller sample A_2 from survey 2 collects information on both x and a variable of interest y which is more expensive to observe than x. Our primary interest under scenario 1 is to create a single synthetic dataset of proxy variables $\tilde{y}_i$ for the unobserved y_i associated with x_i in survey 1 by fitting a working model to survey 2 data. The proxy data together with the associated survey weights, w_{i1}, of survey 1 are then used to produce projection estimators of the population total of y and domain totals of y. In one application of the synthetic data approach, survey 2 observed both self-reported health measurements, x_i, and clinical measurements from physical examinations, y_i, for a small sample A_2 of individuals, while the much larger survey 1 observed only x_i. Only

the imputed, or synthetic, data and associated survey 1 weights are released to the public (Reiter, 2008). We use a model-assisted approach, based on a working model relating y to x, to generate the synthetic values, and our approach is robust against failure of the working model.

In scenario 2, we have two independent samples A and B. Sample A observes x and y_1 while sample B measures x and another variable y_2. The variables of interest y_1 and y_2 are not jointly observed. Our interest is to create a synthetic value $\tilde{y}_1$ associated with the unobserved y_1 for each element in sample B by finding a statistical match from sample A on the basis of the variable x common to samples A and B. Completed data file with $x, \tilde{y}_1$ and y_2 is used to estimate the parameters of linear regression of y_2 on y_1 and other parameters of interest.

A popular method of statistical matching assumes that the variables y_1 and y_2 not jointly observed are conditionally independent given the common variable x. For example, nearest neighbor (NN) matching assumes conditional independence (CI). The conditional association between y_1 and y_2 given x cannot be estimated from the observed data. However, Rassler (2004) proposed a multiple imputation (MI) method based on explicit models to impute the unobserved y_1 in survey B for different values describing the conditional association. Imputed data sets are then used to estimate the unconditional association between y_1 and y_2. Simulation results indicated that the proposed method performs well in terms of confidence interval coverage unlike NN. In section 4 we propose an alternative method based on an instrumental variable (IV) for the unobserved y_1 in sample B.

2. Scenario 1

2.1 Total and domain totals

Suppose that the total $Y = \sum_{i=1}^{N} y_i$ is the parameter of interest and that the working model for y_i is $E(y_i \mid x_i) = m(x_i, \beta) = m_i$, var $(y_i \mid x_i) = \sigma^2 a(m_i)$ for some known function $a(m_i)$ and cov $(y_i, y_j \mid x_i, x_j) = 0$ for $i \neq j$. Based on $\{(y_i, x_i), i \in A_2\}$ we obtain an estimator $\hat{\beta}$ as the solution of estimating equations $\sum w_{i2}(y_i - m_i)h_i = 0$, where w_{i2} are the survey 2 weights, $h_i = (\partial m_i / \partial \beta) / a(m_i)$ and the summation is over $i \in A_2$. We assume that the first element of the vector h_i is equal to 1. For a continuous variable y and linear regression with $m_i = \beta_0 + \beta_1 x_i$ and $a(m_i) = 1$, we have $h_i = (1, x_i)'$. Similarly, for a binary variable y and logistic regression with $\text{logit}(m_i) = \beta_0 + \beta_1 x_i$ and $a(m_i) = m_i(1 - m_i)$ we have $h_i = (1, x_i)'$. In those cases, the first estimating equation reduces to $\sum w_{i2}(y_i - \hat{m}_i) = 0$ where $\hat{m}_i = m(x_i, \hat{\beta})$.

In the case of a continuous variable y we compute the imputed or synthetic values $\tilde{y}_i = m(x_i, \hat{\beta}) = \hat{m}_i$ for each $x_i \in A_1$ and report them in the survey 1 data file, using $\hat{\beta}$ obtained from survey 2 data. In the case of a binary variable y the survey 1 data file will report binary synthetic values $\tilde{y}_i = 1$ and 0 with associated fractions $\hat{m}_i$ and $1 - \hat{m}_i$ for each $i \in A_1$. The projection estimator of the total Y based on the reported synthetic values in the survey 1 data file is given by

$$\hat{Y}_p = \sum w_{i1} \tilde{y}_i \qquad (1)$$

where the summation is over $i \in A_1$ and w_{i1} is the survey weight associated with $i \in A_1$. Kim and Rao (2012) showed that the projection estimator (1) is asymptotically design unbiased if the first estimating equation above holds or the first element of h_i is equal to 1 for $i \in A_2$. Note that the projection estimator (1) is derived from the working model but our results do not depend on the validity of the working model, although efficiency of estimators may be affected.

For the estimation of a domain total $Y_d = \sum_{i=1}^{N} \delta_i(d) y_i$, the projection estimator is given by $\hat{Y}_{d,p} = \sum w_{i1} \delta_i(d) \tilde{y}_i$ where $\delta_i(d)$ is the indicator variable for domain d taking the value 1 if unit $i \in A_1$ belongs to the domain and 0 otherwise. For domains specified in advance or planned, the working model can be augmented by including the domain indicator and the resulting synthetic values are reported in the survey 1 data file. This ensures asymptotic design unbiasedness of the domain projection estimator $\hat{Y}_{d,p}$ in the case of a linear regression or a logistic regression working model (Kim and Rao 2012). Alternatively if the domain is not planned then the asymptic bias of the domain projection estimator relative to the domain total is negligible if the domain indicator $\delta_i(d)$ is approximately unrelated to the residual $r_i = y_i - m_i$. This will be the case if the working model is correctly specified.

For estimating the variance of the projection estimators, Kim and Rao (2012) proposed a pseudo-replication method that requires the generation of synthetic data $\{\tilde{y}_i^{(k)}, i \in A_1\}$ corresponding to each set of replicate weights $\{w_{i1}^{(k)}, i \in A_1\}$ associated with survey 1 only. This method enables the user to correctly estimate the variance of the projection estimator without access to the data from survey 2. The data file will contain additional columns $\{\tilde{y}_i^{(k)}, i \in A_1\}$ associated with the columns of replicate weights $\{w_{i1}^{(k)}, i \in A_1\}$, $k = 1, ..., L_1$ where L_1 is the number of replicates created from survey 1. Hence the price one pays to use only survey 1 synthetic data is to increase the number of columns in the data file by L_1 for each variable y for which synthetic values are generated. Typically, the number of such variables may be small. The replicate projection estimator

is computed from the additional columns $\tilde{y}_i^{(k)}$ and $w_{i1}^{(k)}$ as $\hat{Y}_p^{(k)} = \sum w_{i1}^{(k)} \tilde{y}_i^{(k)}$ and the resulting replicate variance estimator is of the form

$$v_{\text{rep}}(\hat{Y}_p) = \sum_k c_k (\hat{Y}_p^{(k)} - \hat{Y}_p)^2 \qquad (2)$$

where the factor c_k depends on the replication method used. Kim and Rao (2012) established the design consistency of the variance estimator (2). Replication variance estimator for the domain projection estimator is similarly obtained.

2.2 Distribution function

Suppose we wish to estimate the distribution function $F_N(t) = N^{-1} \sum_{i=1}^{N} I(y_i \le t)$ of a continuous variable y for a given t, where $I(.)$ is the indicator function. If y_i for $i \in A_1$ were observed then a design-consistent estimator of $F_N(t)$ is given by $\hat{F}(t) = \sum \tilde{w}_{i1} I(y_i \le t)$, where $\tilde{w}_{i1} = w_{i1} / \sum w_{i1}$ are the normalized survey 1 weights. In the case only x_i is observed for $i \in A_1$, the projection estimator $\hat{F}_p(t) = \sum \tilde{w}_{i1} I(\hat{m}_i \le t)$ based on the deterministic imputed values $\tilde{y}_i = \hat{m}_i$ will not be design consistent unlike the projection estimator $\hat{Y}_p$ of the total Y. Hence, we cannot make design-based inferences on the distribution function using a working model.

It will be necessary to use a design-model approach assuming that an imputation model holds. In particular, we assume the linear regression model $y_i = \beta_0 + \beta_1 x_i + \sigma \varepsilon_i$, where the ε_i are independent and identically distributed with mean 0 and variance σ^2. We obtain design-weighted estimators $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\sigma}$ from survey 2 data $\{(y_i, x_i), i \in A_2\}$ and calculate the standardized residuals $e_j = \hat{\sigma}^{-1}(y_j - \hat{\beta}_0 - \hat{\beta}_1 x_j)$. We then select n_1 residuals e_i^* with probabilities proportional to w_{j2} and with replacement from the set of standardized residuals e_j, where n_1 is the number of units in A_1. The imputed values of y_i for survey 1 are then given by $y_i^* = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\sigma} e_i^* = \hat{m}_i + \hat{\sigma} e_i^*$ for $i \in A_1$ and the resulting projection estimator of $F_N(t)$ is given by $\tilde{F}_p(t) = \sum \tilde{w}_{i1} I(y_i^* \le t)$. In the context of missing data, Chauvet, Deville and Haziza (2013) proposed a similar method for estimating the distribution function and established its design-model consistency. We are presently studying the properties of the proposed projection estimator $\tilde{F}_p(t)$.

The projection estimator of the total Y is given by $\tilde{Y}_p = \sum w_{i1} y_i^* = \sum w_{i1} \hat{m}_i + \hat{\sigma} \sum w_{i1} e_i^* = \hat{Y}_p + \hat{\sigma} \sum w_{i1} e_i^*$. If the residuals e_i^* are selected by balanced sampling to satisfy $\sum w_{i1} e_i^* = 0$, then $\tilde{Y}_p = \hat{Y}_p$ and, as noted earlier, $\hat{Y}_p$ is asymptotically design unbiased regardless of the validity of the model because

$\sum w_{i2}(y_i - \hat{m}_i) = 0$. Thus a balanced sampling approach would give a projection estimator of total asymptotically design unbiased and at the same time provide a design-model consistent estimator of the distribution function. Properties of such a balanced sampling approach will be studied. Chauvet, Deville and Haziza (2013) proposed balanced sampling to eliminate the imputation variance in the context of missing data.

3. Scenario 2

In scenario 2, we have two independent samples A and B and sample A observes x and y_1 while sample B measures x and another variable y_2. The variables of interest y_1 and y_2 are not jointly observed. We are interested in the parameters β_0 and β_1 of the linear regression model $y_{2i} = \beta_0 + \beta_1 y_{1i} + e_i$ with $e_i \sim (0, \sigma_e^2)$ by creating a synthetic value $\tilde{y}_1$ associated with the unobserved y_1 for each element in sample B on the basis of the common variable x observed in both samples A and B. If y_{1i} associated with y_{2i} were observed in the sample B, then the least squares estimators are denoted by β_0^* and β_1^*.

We now consider a two-step procedure to generate synthetic values $\tilde{y}_{1i}$ for $i \in B$. *Step 1*: Estimate the conditional distribution $f(y_1 \mid x, \eta)$ from sample A by $\hat{f}_a(y_1 \mid x, \hat{\eta})$, where $\hat{\eta}$ is the design -weighted estimator of the parameter η of the normal conditional density. *Step 2*: For each element i in sample B use the x_i value to generate imputed value $\tilde{y}_{1i}$ of y_{1i} from $\hat{f}_a(y_1 \mid x_i, \hat{\eta})$. We then perform linear regression of y_{2i} on $\tilde{y}_{1i}$ for $i \in B$ to obtain least squares estimators $\tilde{\beta}_0$ and $\tilde{\beta}_1$ as proxies to the unknown β_0^* and β_1^*. The estimators $\tilde{\beta}_0$ and $\tilde{\beta}_1$ will be consistent under the CI assumption but biased if the CI assumption is not satisfied. In the case of linear regression models for y_1 on x and y_2 on y_1, it is not necessary to make distributional assumptions and one can use stochastic regression imputation from sample A to generate $\tilde{y}_{1i}$ for the elements i in sample B.

We now turn to the two-stage least squares (2SLS) approach by assuming that x is an instrumental variable for y_1 in the sense that the conditional distribution of y_2 given y_1 and x does not depend on x: $f(y_2 \mid y_1, x) = f(y_2 \mid y_1)$. That is, y_2 and x are conditionally independent given y_1. We replace the unobserved y_{1i} for $i \in B$ by the least squares predictor $\hat{y}_{1i} = \hat{\alpha}_0 + \hat{\alpha}_1 x_i$ obtained from the sample A data $\{(y_{1j}, x_j), j \in A\}$ under the working linear regression model $y_{1j} = \alpha_0 + \alpha_1 x_j + u_j$ with $u_j \sim (0, \sigma_u^2)$ and then perform linear regression of y_{2i} on $\hat{y}_{1i}$ to obtain the 2SLS least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. Under the IV assumption it can be shows that the 2SLS estimators are unbiased with respect to the assumed linear regression model $y_{2i} = \beta_0 + \beta_1 y_{1i} + e_i$ with $e_i \sim (0, \sigma_e^2)$. The 2SLS method is very simple to implement. We have also evaluated the variance of

the 2SLS estimator $\hat{\beta}_1$ which shows that the variance of the estimator can be large if the correlation between x and y_1 is small.

The 2SLS method is designed to estimate only the regression parameters and also it is not directly applicable if the regression model is not linear. For general models, we have developed a parametric fractional imputation method, making distributional assumptions, but it is not discussed in this short paper.

4. Simulation study

We have conducted simulation study by generating samples under the IV assumption. As expected, the CLT-based estimators led to large biases and mean squared errors unlike the 2SLS estimators. Results are not reported in the paper.

References

Chauvet, G., Deville, J. C. V. and Haziza, D. (2013). On balanced random imputation in surveys. *Biometrika*, 100, in press.

Kim, J. K. and Rao, J. N. K. (2012). Combining data from two independent surveys: a mode-assisted approach. *Biometrika*, 99, 83-100.

Rassler, S. (2004). Data fusion: Identification problems, validity, and multiple imputation. *Austrian Journal of Statistics*, 33, 153-171.

Reiter, J. (2008). Multiple imputation when records used for imputation are not used or disseminated for analysis. *Biometrika*, 95, 368-375.