Using the Superpopulation Model for Imputations and Variance Computation in Survey Sampling

Petr Novák , Václav Kosina

Abstract:

In present work we study variance computation techniques for population estimates based on survey sampling and imputation. We use the superpopulation regression model, which means that the target variable values for each statistical unit are treated as random realisations of a linear regression model with weighted variance. We focus on models with one auxiliary variable and no intercept, which have many applications in business statistics. Furthermore, we deal with cases where the estimates are not independent and thus the covariance must be computed, and where the auxiliary variables are random variables instead of constants.

Keywords: Survey sampling, Variance, Imputation

1 General assumptions

For computation of survey estimates of business statistics, the CZSO is recently exploring a new approach, in which all data for units that are not in the sample are imputed, instead of computing the population sums or means through weighting. Therefore, new techniques for survey variance computation had to be developed.

We want to estimate the population sum $Y = \sum_{i=1}^{N} y_i$. Suppose that we have sampled only n observations, the $N-n$ remaining values must be estimated. We use the superpopulation model, with following assumptions:

- The data y_i are random variables with $y_i = x_i\beta + e_i$,
- the error terms e_i are independent with distribution $e_i \sim (0, c_i\sigma^2)$,
- x_i and c_i are known constants for all i=1,...,N
- β and σ^2 are unknown parameters.

2 Variance computation with simple regression imputations

Let us derive the formula for variance of the estimated sum of observed variable. We observe n realisations of the variable, which we call the sample *sam*. There are $N - n$ more realised varables, which values we want to estimate with the knowledge of x_i and c_i. This unknown part we call *the imputed part* or simply *imp*. More accurately we want to estimate the sum

$$Y = \sum_{i \in sam} y_i + \sum_{i \in imp} y_i.$$

We use classical regression estimates with one covariate and no intercept (the regression line passes through the origin). We impute in the following way:

$$\hat{y}_i = x_i \hat{\beta} = x_i * \frac{\sum_{sam} w_i x_i y_i / c_i}{\sum_{sam} w_i x_i^2 / c_i},$$

where w_i are some appropriately chosen weights (discussed later). Note, that for $c_i := x_i$ we get the most commonly used fraction

$$\hat{\beta} = \frac{\sum_{sam} w_i y_i}{\sum_{sam} w_i x_i},$$

but we allow different values of c_i because they may differ with each methology. We can easily verify, that

$$E\hat{\beta} = \frac{\sum_{sam} w_i x_i E y_i / c_i}{\sum_{sam} w_i x_i^2 / c_i} = \beta$$

$$var\hat{\beta} = \frac{\sum_{sam} w_i^2 x_i^2 / c_i^2 var y_i}{(\sum_{sam} w_i x_i^2 / c_i)^2} = \frac{\sum_{sam} w_i^2 x_i^2 / c_i}{(\sum_{sam} w_i x_i^2 / c_i)^2} \sigma^2 =: \sigma_\beta^2.$$

We want to estimate the variance of $\hat{Y}$. Because of the superpopulation model, the variables y_i which we estimate are random variables instead of constants, therefore we cannot use the common formula

$$var\hat{Y} = E(\hat{Y} - E\hat{Y})^2.$$

In fact, we are interested in the mean square error of the difference of the real and estimated (predicted) values of the random variables

$$mse(\hat{Y}) = E(\hat{Y} - Y)^2$$

given the realisation of the sample data. We should write $E(\hat{Y} - Y|sam)^2$, but we leave the condition out for space saving reasons. This is the main difference

from the usual theoretical methods in survey sampling, where all data are taken as constants and the randomness is included in the models in form of inclusion indicators. If we take Y as realisations of random variables from the superpopulation model, we can derive the formulas for the variance also in more complex situations.

For the imputed data we have

$$E\hat{y}_i = Ex_i\hat{\beta} = x_i\beta = Ey_i,$$

therefore the estimate is unbiased. For the mse we then get

$$E(\hat{Y} - Y)^2 = E(\hat{Y}_{imp} - Y_{imp})^2 = E(\hat{Y}_{imp} - E\hat{Y}_{imp} - (Y_{imp} - EY_{imp}))^2 =$$

$$= E(\hat{Y}_{imp} - E\hat{Y}_{imp})^2 + E(Y_{imp} - EY_{imp})^2 - 2E(\hat{Y}_{imp} - E\hat{Y}_{imp})(Y_{imp} - EY_{imp}).$$

The cross part will be zero, because it consists of two independent terms, both with a zero mean ($\hat{Y}_{imp}$ is computed form the sample, Y_{imp} is the rest). Therefore

$$mse(\hat{Y}) = var\hat{Y}_{imp} + varY_{imp} = varX_{imp}\hat{\beta} + c_{imp}\sigma^2 = X^2_{imp}\sigma^2_{\beta} + c_{imp}\sigma^2.$$

The constants x_i and c_i are known, for estimating $\widehat{mse}(\hat{Y})$ we only need to use an appropriate estimate for σ^2, i.e.

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{sam}\frac{(y_i - \hat{\beta}x_i)^2}{c_i}$$

or

$$\hat{\sigma}^2 = \frac{1}{\sum_{sam} w_i - \bar{w}_i}\sum_{sam}\frac{w_i(y_i - \hat{\beta}x_i)^2}{c_i},$$

where $\bar{w}_i = \frac{1}{n}\sum_{sam} w_i$.

We see, that the estimate of mse consists of the model paramater estimates on the sample part and of the sums of auxiliary variables on the imputed part of the data.

3 Variance of chain imputations

Suppose we deal with data y_i estimated with the help of an auxiliary variables x_i, which is known only for the units in the sample, elsewhere it is imputed with the help of known constants z_i. We assume the same model as above:

$$y_i|x_i \sim (\beta_y x_i, c_i\sigma_y^2), \qquad x_i \sim (\beta_z z_i, d_i\sigma_z^2).$$

The regression parameters are estimated in a following way:

$$\hat{\beta}_{y|x} = \frac{\sum_{sam} w_i x_i y_i/c_i}{\sum_{sam} w_i x_i^2/c_i}, \qquad \hat{\beta}_x = \frac{\sum_{sam} v_i z_i x_i/d_i}{\sum_{sam} v_i z_i^2/d_i}.$$

The estimates have then similar properties:

$$\hat{\beta}_{y|x} \sim \left(\beta_y, \sigma^2_{\beta_{y|x}} := \frac{\sum_{sam} w_i^2 x_i^2/c_i}{(\sum_{sam} w_i x_i^2/c_i)^2}\sigma_y^2\right), \qquad \hat{\beta}_x \sim (\beta_x, \sigma^2_{\beta_x}).$$

At first, $\hat{x}_i$ are imputed, afterwards we impute $\hat{y}_i$ with their help:

$$\hat{x}_i = \hat{\beta}_x z_i, \qquad \hat{y}_i = \hat{\beta}_{y|x}\hat{x}_i.$$

For the imputed part we have

$$E\hat{y}_i = EE[\hat{y}_i|x_i] = EE[\hat{\beta}_{y|x}\hat{x}_i|x_i] = E\beta_y\hat{x}_i$$

$$= \beta_y E\hat{x}_i = \beta_y\beta_x z_i = EE[y_i|x_i] = Ey_i.$$

We want to compute the mean square error of the prediction of the random variables Y estimated through $\hat{Y}$. With the help of conditional variance decomposition we get

$$\begin{aligned}
mse(\hat{Y}) &= E(\hat{Y}_{imp} - Y_{imp})^2 = E(\hat{Y}_{imp} - E\hat{Y}_{imp})^2 + E(Y_{imp} - EY_{imp})^2 \\
&= var\hat{Y}_{imp} + varY_{imp} \\
&= Evar[\hat{Y}_{imp}|\boldsymbol{X}] + varE[\hat{Y}_{imp}|\boldsymbol{X}] \\
&+ Evar[Y_{imp}|\boldsymbol{X}] + varE[Y_{imp}|\boldsymbol{X}] \\
&= E\hat{X}^2_{imp}\sigma^2_{\beta_{y|x}} + var[\hat{X}_{imp}\beta_y] + Ec_{imp}\sigma_y^2 + var[X_{imp}\beta_y] \\
&= E[\hat{X}^2_{imp}\sigma^2_{\beta_{y|x}} + c_{imp}\sigma_y^2] + \beta_y^2(var\hat{X}_{imp} + varX_{imp}) \\
&= EE[(\hat{Y}_{imp} - Y_{imp})|\boldsymbol{X}]^2 + \beta_y^2 E(\hat{X}_{imp} - X_{imp})^2 \\
&= Emse(\hat{Y}|\boldsymbol{X}) + \beta_y^2 mse(\hat{X}).
\end{aligned}$$

The second term may be estimated with adding $\hat{\beta}_{y|x}$ and $\widehat{mse}\hat{X}$. The computation of the expectation with respect to the distribution of x_i in the first term would be relatively complex, because of the values x_i are in both nominator

and denominator of $\sigma^2_{\beta_{y|x}}$. We need to find an appropriate estimate, we can use instead of $Emse(\hat{Y}|\boldsymbol{X})$

$$\widehat{mse}(\hat{Y}|\hat{\boldsymbol{X}}) = \hat{X}^2_{imp}\hat{\sigma}^2_{\beta_{y|x}} + \hat{c}_{imp}\hat{\sigma}^2_{y|x}.$$

We get $\hat{X}_{imp}$, $\hat{\sigma}^2_{y|x}$ and $\hat{\sigma}^2_{\beta_{y|x}}$ through the estimates of $\hat{x}_i$, the estimate $\hat{c}_{imp}$ follows from the chosen model for the variance, i.e. $c_i := x_i$ or $c_i := x_i^2$. We have

$$\widehat{mse}(\hat{Y}) = \widehat{mse}(\hat{Y}|\hat{\boldsymbol{X}}) + \hat{\beta}^2_{y|x}\widehat{mse}(\hat{X}).$$

When we work with a chain structure having more levels, the first term $\widehat{mse}(\hat{Y}|\hat{\boldsymbol{X}})$ and $\hat{\beta}_{y|x}$ remain the same, because they are conditional estimates given their auxiliary variable. The second term may be obtained through another chain estimation, so we are getting a recurrent formula, which leads so far until it reaches an auxiliary variable which is known for all units (i.e. administrative data sources).

4 Stratification level shifts - - covariance computation

The CZSO works with the stratification approach, where the surveyed enterprises are divided into groups depending on the number of employees, type of economic activity, region etc. The stratification has more levels, going from relatively small groups to larger ones. In each stratum, the regression parameters are estimated separately. When it is not possible to obtain the estimates in given stratum, for example because of a low number of responding units, we use the estimates in the corresponding superior stratum at a higher stratification level.

Consider the non-chained regression from section **2**. Let m be a small stratum where the estimates for β_m and σ^2_m could not be obtained. Let S be its superior stratum (one or more levels higher), with enough units to compute the estimates $\hat{\beta}_S = \frac{\sum_{S_{sam}} w_i x_i y_i / c_i}{\sum_{S_{sam}} w_i x_i^2 / c_i}$. For the variance of the estimate of the sum Y_m we impute $\hat{y}_i = \hat{\beta}_S x_i$ and we get

$$\begin{aligned} mse(\hat{Y}_m) &= var\hat{Y}^m_{imp} + varY^m_{imp} \\ &= var\hat{X}^m_{imp}\hat{\beta}_S + varY^m_{imp} = (X^m_{imp})^2\sigma^2_{\beta_S} + c_{imp}\sigma^2_m. \end{aligned}$$

The estimate for $\sigma^2_{\beta_S}$ is obtained from the superior stratum S, σ^2_m is completely unknown and cannot be estimated from m, therefore we use the estimate for σ^2_S instead.

Suppose we now have one stratum S in a higher level, which consists of two substrata: one small (m) and one good (d), where it is possible to estimate β_d and σ_d^2. We want to obtain the variance for the sum Y for the whole S. Using above given formulas and the independence assumption for e_i, we get

$$\begin{aligned} mse(\hat{Y}) &= var\hat{Y} + varY = var(\hat{Y}_m + \hat{Y}_d) + var(Y_m + Y_d) \\ &= var\hat{Y}_m + var\hat{Y}_d + 2cov(\hat{Y}_m, \hat{Y}_d) + varY_m + varY_d \\ &= mse(\hat{Y}_m) + mse(\hat{Y}_d) + 2cov(\hat{Y}_m, \hat{Y}_d). \end{aligned}$$

The covariance is computed in the following way:

$$\begin{aligned} cov(\hat{Y}_m, \hat{Y}_d) &= cov(X_{imp}^m \hat{\beta}_S, X_{imp}^d \hat{\beta}_d) = X_{imp}^m X_{imp}^d cov(\hat{\beta}_S, \hat{\beta}_d) \\ &= X_{imp}^m X_{imp}^d cov\left(\frac{\sum_{S_{sam}} w_i x_i y_i / c_i}{\sum_{S_{sam}} w_i x_i^2 / c_i}, \frac{\sum_{d_{sam}} w_i x_i y_i / c_i}{\sum_{d_{sam}} w_i x_i^2 / c_i} \right) \\ &= \frac{X_{imp}^m X_{imp}^d}{\sum_{S_{sam}} \frac{w_i x_i^2}{c_i} \sum_{d_{sam}} \frac{w_i x_i^2}{c_i}} cov\left(\sum_{S_{sam}} \frac{w_i x_i y_i}{c_i}, \sum_{d_{sam}} \frac{w_i x_i y_i}{c_i} \right). \end{aligned}$$

The variables y_i belonging to m and d are mutually independent, therefore it is enough to take the sum only through d in the first term of the covariance. Denote as B_S and B_d the sums we have taken out of the parentheses in the denominator:

$$\begin{aligned} &= \frac{X_{imp}^m X_{imp}^d}{B_S B_d} cov\left(\sum_{d_{sam}} w_i x_i y_i / c_i, \sum_{d_{sam}} w_i x_i y_i / c_i \right) \\ &= \frac{X_{imp}^m X_{imp}^d}{B_S B_d} var\left(\sum_{d_{sam}} w_i x_i y_i / c_i \right) \\ &= \frac{X_{imp}^m X_{imp}^d}{B_S B_d} \sum_{d_{sam}} w_i^2 x_i^2 / c_i^2 vary_i \\ &= \frac{X_{imp}^m X_{imp}^d}{B_S B_d} \sum_{d_{sam}} w_i^2 x_i^2 / c_i \sigma_d^2 = X_{imp}^m X_{imp}^d \frac{B_d}{B_S} \sigma_{\beta_d}^2. \end{aligned}$$

If we estimate the parameter $\sigma_{\beta_d}^2$ from the good stratum d, we get the whole variance. In a similar way, the covariance of estimates for any two strata can be obtained, even for s m_1 and m_2, for which the estimates are taken from the strata S_{m_1} a S_{m_2} (clearly $m_i \subseteq S_{m_i}$). Denote $S_d = S_{m_1} \cap S_{m_2}$ and $S = S_{m_1} \cup S_{m_2}$, in our case is S_d the smaller of the sets S_{m_2} and S_{m_1}. For the covariance we get

$$cov(\hat{Y}_{m_1}, \hat{Y}_{m_2}) = \frac{X_{imp}^{m_1} X_{imp}^{m_2}}{B_{S_{m_1}} B_{S_{m_2}}} \sum_{i \in S_d} w_i^2 x_i^2 / c_i \sigma_{S_d}^2 = X_{imp}^{m_1} X_{imp}^{m_2} \frac{B_{S_d}}{B_S} \sigma_{\beta_{S_d}}^2.$$

It cat be further shown, that for a larger stratum S consisting of $d = 1, ..., D$ good and $m = 1, ..., M$ small strata we get

$$mse(\hat{Y}_S) = \sum_{d=1}^{D} mse(\hat{Y}_d) + \sum_{m=1}^{M} mse(\hat{Y}_m) + 2\sum_{m=1}^{M} X_{imp}^{m} \sum_{d=1}^{D} X_{imp}^{d} \frac{B_d}{B_S} \sigma_{\beta_d}^2 + \sum_{m_i \neq m_j} X_{imp}^{m_i} X_{imp}^{m_j} \sigma_{\beta_S}^2.$$

5 Stratification level shifts - - chained imputations

We generalize now the methods used for stratification level shifts for the cases, when the data y_i are imputed with help of estimated auxiliary variables x_i, which are obtained through regression with respect to known constants z_i. In terms of model parameters we have $y_i|x_i \sim (\beta_y x_i, c_i\sigma_y^2)$ and $x_i \sim (\beta_z z_i, d_i\sigma_z^2)$. Let S be a large stratum consisting of substrata m (small) and d (good). Then the mean square error can be decomposed as:

$$\begin{aligned} mse(\hat{Y}_S) &= var\hat{Y}_S + varY_S \\ &= var\hat{Y}_d + var\hat{Y}_m + 2cov(\hat{Y}_d, \hat{Y}_m) + varY_d + varY_m \\ &= mse(\hat{Y}_d) + mse(\hat{Y}_m) + 2cov(\hat{Y}_d, \hat{Y}_m). \end{aligned}$$

Both mse for sums just in strata d and m can be estimated through methods given in section (2):

$$\widehat{mse}(\hat{Y}_d) = \widehat{mse}(\hat{Y}_d|\hat{X}) + \hat{\beta}_{yd|x}^2 \widehat{mse}(\hat{X}_d),$$

$$\widehat{mse}(\hat{Y}_m) = \widehat{mse}(\hat{Y}_m|\hat{X}) + \hat{\beta}_{yS|x}^2 \widehat{mse}(\hat{X}_m).$$

The covariances are computed with help of conditional covariance decomposition:

$$\begin{aligned} cov(\hat{Y}_d, \hat{Y}_m) &= Ecov[\hat{Y}_d, \hat{Y}_m|X] + cov(E[\hat{Y}_d|X], E[\hat{Y}_m|X]) \\ &= Ecov[\hat{Y}_d, \hat{Y}_m|X] + \beta_{yd}\beta_{yS}cov(\hat{X}_d, \hat{X}_m). \end{aligned}$$

The computation of the mean of the first term with respect to X would be rather difficult, we substitute it with the estimate with the help of $\hat{X}$:

$$\widehat{cov}(\hat{Y}_d, \hat{Y}_m) = \widehat{cov}[\hat{Y}_d, \hat{Y}_m|\hat{X}] + \hat{\beta}_{yd}\hat{\beta}_{yS}\widehat{cov}(\hat{X}_d, \hat{X}_m)$$

The coefficients $\hat{\beta}_{yd}$ and $\hat{\beta}_{yS}$ and the first term of the sum can be computed given the estimates $\hat{x}_i$:

$$\widehat{cov}[\hat{Y}_d, \hat{Y}_m|\hat{X}] = \hat{X}^m_{imp}\hat{X}^d_{imp}\frac{\hat{B}^x_d}{\hat{B}^x_S}\hat{\sigma}^2_{\beta_y d},$$

the second covariance may be expressed as

$$\widehat{cov}(\hat{X}_d, \hat{X}_m) = Z^m_{imp}Z^d_{imp}\frac{B^z_d}{B^z_S}\hat{\sigma}^2_{\beta_z d}.$$

We also get a recurrent formula for the covariances, too. If Z would have an auxiliary variable which must be estimated, the estimate of the second term will be chained until it leads to constant covariates.

It can be also shown, that the formula will work also when in the strata m or d are some values y_i imputed, but corresponding values x_i are observed in the sample.

The covariance computation for more than two strata can be generalized in a similar way as in the case with no chain structure.

6 Remarks

6.1 Special cases

The above described techniques are quite general. Often we work simply with $c_i := x_i$, the variance formula is then reduced to

$$mse(\hat{Y}) = X^2_{imp}\sigma^2_\beta + c_{imp}\sigma^2 = X^2_{imp}\frac{\sum_{sam} w_i^2 x_i}{(\sum_{sam} w_i x_i)^2}\sigma^2 + X_{imp}\sigma^2.$$

When the weights are constant, we get

$$mse(\hat{Y}) = X_{imp}(X_{imp}\frac{X_{sam}}{X^2_{sam}} + 1)\sigma^2 = X_{imp}\frac{X_{all}}{X_{sam}}\sigma^2.$$

If no auxiliary information is available, we may use $x_i \equiv 1$, which means that we impute just the sample mean. We obtain

$$mse(\hat{Y}) = (N-n)\frac{N}{n}\sigma^2 = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\sigma^2,$$

which is the commonly used formula for simple random sampling variance.

6.2 Choosing the weights

For getting the population estimates, we use imputations with help of the superpopulation model, rather than the commonly used reweighting techniques. The weights are used only in the estimates $\hat{\beta}$, and therefore they have a different meaning.

If we observe just one stratum alone with no relation to others, it would be appropriate to use constant weights (which may simply be equal to one for that case, because the constant in the numerator and denominator of $\hat{\beta}$ cancels out).

If we apply some outlier-detection methods to point out data that may not fit the model, we can simply put $w_k := 0$ for that units, meaning that they will not influence the parameter estimates in any way.

In the case when we need to use a higher level stratification to obtain the estimates, the weights can be chosen in a way that they reflect the proportion of sampled units in each sub-strata, i.e. $w_k := N_k/n_k$ for sub-stratum k with n_k from N_k units sampled. Therefore the data from the greater strata influence the estimates more than the data from the smaller strata.

7 Conclusions

The superpopulation regression model and all-data imputation presents an alternative approach how to estimate the population totals in survey sampling. It is then easier to report the data with respect to various groupings. We have shown how to compute the mean square error of the estimators, in order to assess data quality. In simple cases, this approach leads to the same results as the commonly used formulas for classic simple random sampling. However, with the help of superpopulation model it is easier to develop variance estimates in more complex cases with sophisticated stratification and chain structure, as we have shown.