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

Petr Novák[1] | *Czech Statistical Office, Prague, Czech Republic*
Václav Kosina[2] | *Czech Statistical Office, Prague, Czech Republic*

Abstract

This study is aimed at variance computation techniques for estimates of population characteristics 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 realizations of a linear regression model with weighted variance. We focus on regression models with one auxiliary variable and no intercept, which have many applications and straightforward interpretation in business statistics. Furthermore, we deal with cases where the estimates are not independent and thus the covariance must be computed. We also consider chained regression models with auxiliary variables as random variables instead of constants.

Keywords

Survey sampling, variance estimation, imputation

JEL code

C13

INTRODUCTION

For estimation of population characteristics (mainly totals, means, counts) in business statistics surveys, the Czech Statistical Office (CZSO) has been recently exploring a new approach, in which all data for units that are out of the sample are imputed based on predictions by regression, instead of estimating the population characteristics through weighting. The all-data imputation is based on the superpopulation model (i.e. Cassel et al., 1977, chapter 4). Compared to classical survey methodology (i.e. Hájek, 1960, 1981 or Cochran, 1977), the data are treated as realizations of an infinite population, some of which we know through the survey and some we want to estimate.

Traditional methods, on the other hand, work with the population at hand. All data are treated as fixed constants and the randomness of estimates then comes in form of sample inclusion indicators. The

[1] Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 186 75 Prague 8, Czech Republic. Corresponding author: e-mail: petr.novak@czso.cz, phone: (+420)274052141. Czech Statistical Office, Na padesátém 81, 100 82 Prague 10, Czech Republic.

[2] Czech Statistical Office, Na padesátém 81, 100 82 Prague 10, Czech Republic.

population totals are then estimated by weighting methods, such as the Horwitz-Thompson estimator or the ratio estimator. We show that some of the estimates coincide or are very similar.

The drawback of the superpopulation approach subsists in the fact that it relies heavily on the choice of the regression model and appropriate auxiliary variables. However, the all-data imputation allows to group the data and report the results in any desirable way, because we have a predicted value available for each unit in the population.

It is desirable to assess the quality of the obtained estimators by computing their variance, mean square error or the coefficient of variation. Because of the differences between classic and superpopulation modeling, new techniques for survey error computation had to be explored. At first, we derive the estimator of the standard error computation in simple cases with one auxiliary variable in the regression model. Then, we present extensions of the methods for cases where the population is divided in more strata and where the auxiliary variables used for the regression are themselves imputed and form a chain structure, as explored in Raghunathan et al (2001). We illustrate the methods on simplified examples from business statistics.

1 THE SUPERPOPULATION REGRESSION MODEL

In the superpopulation approach we treat the data as random realizations of an infinite population with some model distribution. Suppose that we have sampled n observations and $N - n$ more values must be estimated in order to cover the population of interest. To find appropriate estimates, we have to choose a suitable regression model, study the dependence between the variable of interest and the covariates on the observed data and use the results to predict the unknown part. First, we consider a simple superpopulation model with one regression variable and following assumptions:

- the data y_i are non-negative 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 positive constants for all i = 1,..., N,
- β and σ^2 are unknown parameters.

By the notation $e_i \sim (0, c_i\sigma^2)$ we mean that the error terms have zero mean and that their variance is equal to $c_i\sigma^2$. Note that we do not assume normality of e_i.

The following methods rely heavily on these assumptions and therefore deviations from the model can make the results inaccurate. The variance scaling constants c_i must be chosen to fit the data well, often it is used $c_i = x_i$ or $c_i \equiv 1$. Methods of assessing the model fit are out of the scope of this paper (see Anscombe, 1961 or Cook and Weisberg, 1983 among others).

We observe n realizations of the variable, which we call the *sample* and denote as *sam*. There are N - n more realized variables, which values we wish to estimate with the knowledge of x_i and c_i. Let us call this unknown part of the population the *imputed* part and denote as *imp*. More accurately we want to estimate the sum:

$$Y = \sum_{i \in sam} y_i + \sum_{i \in imp} y_i , \tag{1}$$

by imputing an estimate for each y_i from the unknown part:

$$\hat{Y} = \sum_{i \in sam} y_i + \sum_{i \in imp} \hat{y}_i . \tag{2}$$

For space saving reasons we will mark the totals with just $\sum_{sam} y_i$ instead of $\sum_{i \in sam} y_i$ etc. We will further use the notation $Y_{sam} = \sum_{sam} y_i$, $Y_{imp} = \sum_{imp} y_i$ and $\hat{Y}_{imp} = \sum_{imp} \hat{y}_i$, similarly for sums of x_i and c_i.

We use classical linear regression model with one covariate and no intercept (the regression line passing through the origin). The estimator of β is obtained using weighted least squares and we use it to impute the data in the following way:

$$\hat{y}_i = x_i \hat{\beta} = x_i \times \frac{\sum_{sam} w_i x_i y_i / c_i}{\sum_{sam} w_i x_i^2 / c_i}, \tag{3}$$

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

$$\hat{\beta} = \frac{\sum_{sam} w_i y_i}{\sum_{sam} w_i x_i}. \tag{4}$$

For constant weights and $c_i := \equiv 1$, we have the classical least-squares estimator:

$\hat{\beta} = \frac{\sum_{sam} x_i y_i}{\sum_{sam} x_i^2}$ and $c_i := x_i^2$ gives the mean ratio $\hat{\beta} = \frac{1}{n} \sum_{sam} \frac{y_i}{x_i}$. It depends on each case, which c_i fits the data best.

We can easily verify regardless of the choice of c_i and w_i, that:

$$E\hat{\beta} = \frac{\sum_{sam} w_i x_i E y_i / c_i}{\sum_{sam} w_i x_i^2 / c_i} = \beta, \tag{5}$$

$$\operatorname{var} \hat{\beta} = \frac{\sum_{sam} w_i^2 x_i^2 / c_i^2 \operatorname{var} y_i}{(\sum_{sam} w_i x_i^2 / c_i)^2} = \frac{\sum_{sam} w_i^2 x_i^2 / c_i}{(\sum_{sam} w_i x_i^2 / c_i)^2} \sigma^2 =: \sigma_\beta^2. \tag{6}$$

Example

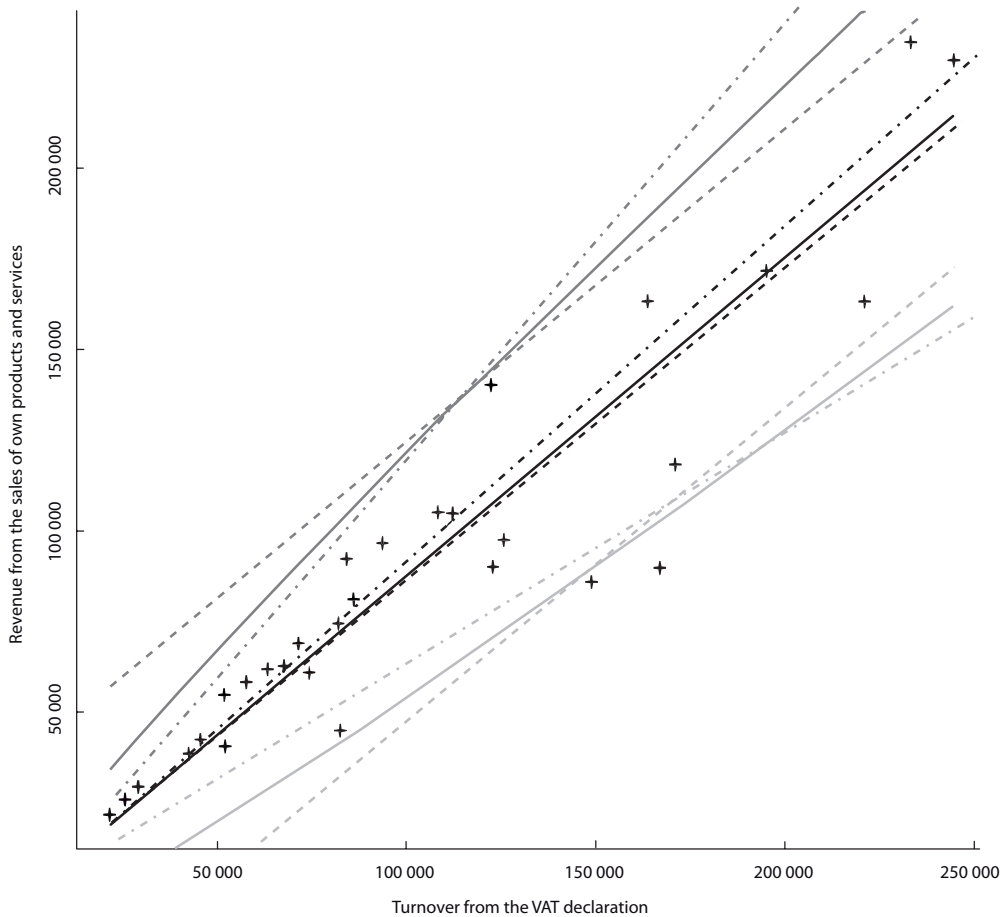
In Figure 1 we see sample data (n = 30) from one particular stratum of the annual structural business survey. We model the dependency of the revenue from the sales of own products and services (y_i) on the turnover given in the VAT declaration (x_i), both given in CZK 1 000. We fitted regression line using $c_i := \equiv 1$ (dashed) $c_i := x_i$ (full) and $c_i := x_i^2$ (dash dot). If the distribution of e_i was Gaussian, we could roughly approximate 95% – confidence bands for the predicted data as $\left(x_i \hat{\beta} - 2\sqrt{c_i}\hat{\sigma}, x_i \hat{\beta} + 2\sqrt{c_i}\hat{\sigma}\right)$, these are marked in gray. The estimated coefficients $\hat{\beta}$, their standard deviations $\hat{\sigma}_\beta$ and the constants $\hat{\sigma}$ are shown in Table 1.

Table 1 Estimated regression parameters

	$c_i := 1$	$c_i := x_i$	$c_i := x^2_i$
$\hat{\beta}$	0.864	0.879	0.923
$\hat{\sigma}$	19 295	53.11	0.143
$\hat{\sigma}_\beta$	0.029	0.03	0.026

Note: $\hat{\beta}$, $\hat{\sigma}$ – estimates of the regression slope β and the standard deviation σ, $\hat{\sigma}_\beta$ – estimated variance of $\hat{\beta}$, c_i – variance scaling.
Source: Simulation – own construction, Czech Statistical Office

Figure 1 Modeling the dependency of the revenue from the sales of own products and services on the turnover given in the VAT declaration, using different variance scaling constants c_i – estimated regression lines with approximate 95% – confidence bands for the data



Source: Czech Statistical Office, data modified to maintain confidentiality

Note that the estimated parameters and therefore also the regression lines are quite similar. Estimators with $c_i := x_i$ and $c_i := x_i^2$ are less sensitive to observations with higher covariate values. The standard deviation parameters $\hat{\sigma}$ differ, because in each case they have a different meaning. The standard deviation of the parameter estimates is again similar in each case. The observations seem to have an increasing deviation from the regression line with higher x_i, which suggests that $c_i := x_i$ or $c_i := x_i^2$ are better choices for the variance scaling than $c_i := 1$.

2 VARIANCE ESTIMATION WITH SIMPLE REGRESSION IMPUTATIONS

Let us derive the formula for the error 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:

$$\operatorname{var} \hat{Y} = E(\hat{Y} - E\hat{Y})^2. \tag{7}$$

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

$$mse\hat{Y} = E(\hat{Y} - Y)^2, \tag{8}$$

given the realization of the sample data. We should write $E(\hat{Y} - Y \mid sam)^2$, 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_i as realizations 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:

$$E\hat{y}_i = Ex_i\hat{\beta} = x_i\beta = Ey_i, \tag{9}$$

therefore $E\hat{Y}_{imp} = EY_{imp}$. For the mse we then get:

$$\begin{aligned} E(\hat{Y} - Y)^2 &= E(\hat{Y}_{imp} - Y_{imp})^2 = E(\hat{Y}_{imp} - E\hat{Y}_{imp} - (Y_{imp} - EY_{imp}))^2 \\ &= E(\hat{Y}_{imp} - E\hat{Y}_{imp})^2 + E(Y_{imp} - EY_{imp})^2 - 2E[(\hat{Y}_{imp} - E\hat{Y}_{imp})(Y_{imp} - EY_{imp})]. \end{aligned} \tag{10}$$

The third (covariance) term 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). Denote $c_{imp} = \sum_{imp} c_i$. Then:

$$\begin{aligned} mse\hat{Y} &= \operatorname{var}\hat{Y}_{imp} + \operatorname{var}Y_{imp} = \operatorname{var}X_{imp}\hat{\beta} + c_{imp}\sigma^2 \\ &= X_{imp}^2\sigma_\beta^2 + c_{imp}\sigma^2. \end{aligned} \tag{11}$$

The constants x_i and c_i are known and $\sigma_\beta^2 = \frac{\sum_{sam} w_i^2 x_i^2 / c_i}{(\sum_{sam} w_i x_i^2 / c_i)^2}\sigma^2$. For establishing the estimate $\hat{mse}\hat{Y}$ we only need to use an appropriate estimate of σ^2, i.e.

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{sam}\frac{(y_i - \hat{\beta}x_i)^2}{c_i}, \tag{12}$$

or

$$\hat{\sigma}^2 = \frac{1}{\sum_{sam} w_i - \overline{w}}\sum_{sam}\frac{w_i(y_i - \hat{\beta}x_i)^2}{c_i}, \tag{13}$$

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

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

It is clear that the more data is in the imputed part, the higher is the mean square error. On the other hand, the more sampled data we have, the more accurately we can estimate $\hat{\beta}$ and therefore σ_β^2 is smaller in the most cases. For example if the weights are constant, then $\sigma_\beta^2 = \frac{1}{\sum_{sam} x_i^2 / c_i}\sigma^2$ is a non-increasing function of *n*.

Example (continued)

In the stratum from the example given in the last section, the revenue from the sales of own products and services was $Y_{sam} = 3\,693\,886$. Suppose we have 50 non-sampled units in the observed stratum. We want to impute the data with the help of known turnover from VAT declaration, for which $X_{imp} = 6\,317\,817$.

We use the same c_i and the estimated regression parameters from above. In Table 2 we see the auxiliary totals c_{imp}, estimated totals $\hat{Y} = Y_{sam} + X_{imp}\hat{\beta}$, the mean square error $m\hat{s}e\hat{Y} = X_{imp}^2\hat{\sigma}_{\beta}^2 + c_{imp}\hat{\sigma}^2$ and the modified coefficient of variation $CV(\hat{Y}) = \frac{\sqrt{m\hat{s}e\hat{Y}}}{\hat{Y}}$ for each choice of c_i.

Table 2 Estimated characteristics

	$c_i := 1$	$c_i := x_i$	$c_i := x^2_i$
$\hat{Y}$	9 154 548	9 247 872	9 527 206
c_{imp}	50	6.31×10^9	1.79×10^{12}
MSE	5.26×10^9	5.42×10^9	6.39×10^9
CV	2.51%	2.52%	2.65%

Note: $\hat{Y}$ – estimated total, MSE – mean square error, CV – coefficient of variation, c_i – variance scaling, c_{imp} – total of c_i – over the imputed part.
Source: Simulation – own construction, primary data: Czech Statistical Office

3 VARIANCE COMPUTATION FOR MORE COMPLEX CASES

By using the superpopulation model, we get closer to linear regression theory and therefore we can derive the variance of the population estimators in various situations where using the classic survey sampling methodology can be overly complicated.

3.1 Variance of chain imputations

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

$$y_i \mid x_i \sim (\beta_y x_i, c_i\sigma_y^2), \qquad x_i \sim (\beta_x z_i, d_i\sigma_x^2), \tag{14}$$

with $y_i \mid x_i$ meaning the conditional distribution of y_i given x_i and d_i being the variance-scaling factors of y_i. The regression parameters are estimated in following way:

$$\hat{\beta}_y = \frac{\sum_{sam} w_i x_i y_i / c_i}{\sum_{sam} w_i x_i^2 / c_i}, \qquad \hat{\beta}_x = \frac{\sum_{sam} v_i z_i x_i / d_i}{\sum_{sam} v_i z_i^2 / d_i}. \tag{15}$$

The estimates have then similar properties:

$$\hat{\beta}_y \sim \left(\beta_y, \sigma^2_{\beta_y} := \frac{\sum_{sam} w_i x_i^2 / c_i}{(\sum_{sam} w_i x_i^2 / c_i)^2}\sigma_y^2 \right), \qquad \hat{\beta}_x \sim (\beta_x, \sigma^2_{\beta_x}). \tag{16}$$

Note that the distribution of $\hat{\beta}_y$ is conditional given the values of x_i, $i = 1,\ldots,n$. At first, $\hat{x}_i$ are imputed, afterwards we impute $\hat{y}_i$ with their help:

$$\hat{x}_i = \hat{\beta}_x z_i, \qquad \hat{y}_i = \hat{\beta}_y \hat{x}_i. \tag{17}$$

Using the conditional expectation, for the imputed part we have:

$$E\hat{y}_i = E[E[\hat{y}_i \mid x_i]] = E[E[\hat{\beta}_y \hat{x}_i \mid x_i]] = E\beta_y \hat{x}_i = \beta_y E\hat{x}_i = \beta_y \beta_x z_i = E[E[y_i \mid x_i]] = Ey_i. \tag{18}$$

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

$$\begin{aligned} mse\hat{Y} &= E(\hat{Y}_{imp} - Y_{imp})^2 = E(\hat{Y}_{imp} - E\hat{Y}_{imp})^2 + E(Y_{imp} - EY_{imp})^2 \\ &= \operatorname{var}\hat{Y}_{imp} + \operatorname{var}Y_{imp} \\ &= E\operatorname{var}[\hat{Y}_{imp} \mid X] + \operatorname{var}E[\hat{Y}_{imp} \mid X] + E\operatorname{var}[Y_{imp} \mid X] + \operatorname{var}E[Y_{imp} \mid X] \\ &= E\hat{X}_{imp}^2\sigma_{\beta_y}^2 + \operatorname{var}[\hat{X}_{imp}\beta_y] + Ec_{imp}\sigma_y^2 + \operatorname{var}X_{imp}\beta_y \\ &= E[\hat{X}_{imp}^2\sigma_{\beta_y}^2 + c_{imp}\sigma_y^2] + \beta_y^2(\operatorname{var}\hat{X}_{imp} + \operatorname{var}X_{imp}) \\ &= EE[(\hat{Y}_{imp} - Y_{imp}) \mid X]^2 + \beta_y^2 E(\hat{X}_{imp} - X_{imp})^2 \\ &= Emse(\hat{Y} \mid X) + \beta_y^2 mse(\hat{X}). \end{aligned} \tag{19}$$

The second term may be estimated by plugging $\hat{\beta}_y$ and $\hat{mse}\hat{X}$ into the formula. The estimation 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 which are in both nominator and denominator of $\sigma_{\beta_y}^2$. We need to find an appropriate estimate, we can use instead of $Emse(\hat{Y} \mid X)$ the term:

$$\hat{mse}(\hat{Y} \mid \hat{X}) = \hat{X}_{imp}^2\hat{\sigma}_{\beta_y}^2 + \hat{c}_{imp}\hat{\sigma}_y^2. \tag{20}$$

We get $\hat{X}_{imp}^2\hat{\sigma}_{\beta_y}^2 + \hat{c}_{imp}\ \hat{X}_{imp}$, $\hat{\sigma}_y^2$ and $\hat{\sigma}_{\beta_y}^2$ through the estimates of $\hat{x}_i$. The estimate $\hat{c}_{imp}$ follows from the chosen model of the variance (i.e. $c_i := x_i$ or $c_i := x_i^2$). We get:

$$\hat{mse}(\hat{Y}) = \hat{mse}(\hat{Y} \mid \hat{X}) + \hat{\beta}_y^2\hat{mse}(\hat{X}). \tag{21}$$

When we work with a chain structure having more levels, the first term $\hat{mse}(\hat{Y} \mid \hat{X})$ and $\hat{\beta}_y$ 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).

3.2 Stratification level shifts – covariance computation

The CZSO works with the stratification approach, where the surveyed enterprises are divided into strata 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, mainly because of a low number of responding units, we use the estimates in the corresponding superior stratum at a higher stratification level.

Let us consider the non-chained regression from section 2. Let m be a small stratum where the estimates for β_m and σ_m^2 could not be obtained. Let S be its superior stratum (one or more levels higher), with enough units to compute the estimates:

$$\hat{\beta}_S = \frac{\sum_{S_{sam}} w_i x_i y_i / c_i}{\sum_{S_{sam}} w_i x_i^2 / c_i}, \tag{22}$$

for the variance of the estimate of the sum Y_m we impute $\hat{y}_i = \hat{\beta}_S x_i$ and we get:

$$\begin{aligned} mse\hat{Y}_m &= \operatorname{var}\hat{Y}_{imp}^m + \operatorname{var}Y_{imp}^m \\ &= \operatorname{var}\hat{X}_{imp}^m\hat{\beta}_S + \operatorname{var}Y_{imp}^m = (X_{imp}^m)^2\sigma_{\beta_S}^2 + c_{imp}\sigma_m^2. \end{aligned} \tag{23}$$

The estimate for $\sigma_{\beta_S}^2$ is obtained from the superior stratum S, σ_m^2 is completely unknown and cannot be estimated from m, therefore we use the estimate for σ_S^2 instead.

Suppose we now have one stratum S in a higher level, which consists of two substrata: one too 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 the above given formulas and the independence assumption for e_i, we get:

$$\begin{aligned} mse\hat{Y} &= \operatorname{var}\hat{Y} + \operatorname{var}Y = \operatorname{var}(\hat{Y}_m + \hat{Y}_d) + \operatorname{var}(Y_m + Y_d) \\ &= \operatorname{var}\hat{Y}_m + \operatorname{var}\hat{Y}_d + 2\operatorname{cov}(\hat{Y}_m, \hat{Y}_d) + \operatorname{var}Y_m + \operatorname{var}Y_d \\ &= mse(\hat{Y}_m) + mse(\hat{Y}_d) + 2\operatorname{cov}(\hat{Y}_m, \hat{Y}_d). \end{aligned} \tag{24}$$

The covariance is computed in the following way:

$$\begin{aligned} \operatorname{cov}(\hat{Y}_m, \hat{Y}_d) &= \operatorname{cov}(X_{imp}^m \hat{\beta}_S, X_{imp}^d \hat{\beta}_d) = X_{imp}^m X_{imp}^d \operatorname{cov}(\hat{\beta}_S, \hat{\beta}_d) \\ &= X_{imp}^m X_{imp}^d \operatorname{cov}\left(\frac{\sum_{S_{sam}} w_i x_i y_i / c_i}{\sum_{S_{sam}} w_i x_i^2 / c_i}, \frac{\sum_{d_{sam}} w_i x_i y_i / c_i}{\sum_{d_{sam}} w_i x_i^2 / c_i} \right) \\ &= \frac{X_{imp}^m X_{imp}^d}{\sum_{S_{sam}} w_i x_i^2 / c_i \sum_{d_{sam}} w_i x_i^2 / c_i} \operatorname{cov}\left(\sum_{S_{sam}} w_i x_i y_i / c_i, \sum_{d_{sam}} w_i x_i y_i / c_i \right). \end{aligned} \tag{25}$$

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:

$$\begin{aligned} &= \frac{X_{imp}^m X_{imp}^d}{B_S B_d} \operatorname{cov}\left(\sum_{d_{sam}} w_i x_i y_i / c_i, \sum_{d_{sam}} w_i x_i y_i / c_i \right) \\ &= \frac{X_{imp}^m X_{imp}^d}{B_S B_d} \operatorname{var} \sum_{d_{sam}} w_i x_i y_i / c_i = \frac{X_{imp}^m X_{imp}^d}{B_S B_d} \sum_{d_{sam}} w_i^2 x_i^2 / c_i^2 \operatorname{var} y_i \\ &= \frac{X_{imp}^m X_{imp}^d}{B_S B_d} \sum_{d_{sam}} w_i^2 x_i^2 / c_i \sigma_d^2 = X_{imp}^m X_{imp}^d \frac{B_d}{B_S} \sigma_{\beta_d}^2. \end{aligned} \tag{26}$$

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. Take m_1 and m_2, for which the estimates are taken from the strata S_{m_1} and S_{m_2}. Denote m_1^{sam} the sampled part of the stratum m_1 etc. If m_1 is a good stratum, then $m_1^{sam} = S_{m_1}^{sam}$, otherwise $m_1^{sam} \subset S_{m_1}^{sam}$. The same for m_2. Suppose that the stratification structure is well ordered, in the way that each substratum is contained in exactly one superior stratum. Denote $S_d^{sam} = S_{m_1}^{sam} \cap S_{m_2}^{sam}$ and $S^{sam} = S_{m_1}^{sam} \cup S_{m_2}^{sam}$. Because of the well-ordered stratification, S_d^{sam} is necessarily either the smaller of the sets $S_{m_1}^{sam}$ and $S_{m_2}^{sam}$ or an empty set if the strata do not overlap. For the covariance we get:

$$\operatorname{cov}(\hat{Y}_{m_1}, \hat{Y}_{m_2}) = \frac{X_{imp}^{m_1} X_{imp}^{m_2}}{B_{S_{m_1}^{sam}} B_{S_{m_2}^{sam}}} \sum_{i \in S_d^{sam}} w_i^2 x_i^2 / c_i \sigma_{S_d^{sam}}^2 = X_{imp}^{m_1} X_{imp}^{m_2} \frac{B_{S_d^{sam}}}{B_{S^{sam}}} \sigma_{\beta_{S_d^{sam}}}^2. \tag{27}$$

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

$$mse(\hat{Y}_S) = \sum_{d=1}^{D} mse\hat{Y}_d + \sum_{m=1}^{M} mse\hat{Y}_m + 2\sum_{m=1}^{M} X_{imp}^m \sum_{d=1}^{D} X_{imp}^d \frac{B_d}{B_S} \sigma_{\beta_d}^2 + \sum_{m_i \neq m_j} X_{imp}^{m_i} X_{imp}^{m_j} \sigma_{\beta_S}^2. \tag{28}$$

3.3 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 $y_i \mid x_i \sim (\beta_y x_i, c_i \sigma_y^2)$, and $x_i \sim (\beta_x z_i, d_i \sigma_x^2)$. Let S be a large stratum consisting of substrata m (small) and d (good). Then the mean square error can be decomposed as:

$$\begin{aligned} mse\hat{Y}_S &= \operatorname{var}\hat{Y}_S + \operatorname{var}Y_S = \operatorname{var}\hat{Y}_m + \operatorname{var}\hat{Y}_m + 2\operatorname{cov}(\hat{Y}_m, \hat{Y}_d) + \operatorname{var}Y_m + \operatorname{var}Y_d \\ &= mse(\hat{Y}_m) + mse(\hat{Y}_d) + 2\operatorname{cov}(\hat{Y}_m, \hat{Y}_d). \end{aligned} \tag{29}$$

Both *mse* of sums just in strata d and m can be estimated through methods given in section (3.1):

$$\hat{mse}(\hat{Y}_d) = \hat{mse}(\hat{Y}_d \mid \hat{X}) + \hat{\beta}_{yd}^2 \hat{mse}(\hat{X}_d), \tag{30}$$

$$\hat{mse}(\hat{Y}_m) = \hat{mse}(\hat{Y}_m \mid \hat{X}) + \hat{\beta}_{yS}^2 \hat{mse}(\hat{X}_m). \tag{31}$$

The covariances are derived with help of conditional covariance decomposition:

$$\begin{aligned} \operatorname{cov}(\hat{Y}_d, \hat{Y}_m) &= E\operatorname{cov}[\hat{Y}_d, \hat{Y}_m \mid X] + \operatorname{cov}(E[\hat{Y}_d \mid X], E[\hat{Y}_m \mid X]) \\ &= E\operatorname{cov}[\hat{Y}_d, \hat{Y}_m \mid X] + \beta_{yd}\beta_{yS}\operatorname{cov}(\hat{X}_d, \hat{X}_m). \end{aligned} \tag{32}$$

The estimation 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 $\hat{X}$:

$$\hat{\operatorname{cov}}(\hat{Y}_d, \hat{Y}_m) = \hat{\operatorname{cov}}[\hat{Y}_d, \hat{Y}_m \mid \hat{X}] + \hat{\beta}_{yd}\hat{\beta}_{yS}\hat{\operatorname{cov}}(\hat{X}_d, \hat{X}_m). \tag{33}$$

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

$$\hat{\operatorname{cov}}[\hat{Y}_d, \hat{Y}_m \mid \hat{X}] = \hat{X}_{imp}^m \hat{X}_{imp}^d \frac{\hat{B}_d^x}{\hat{B}_S^x} \hat{\sigma}_{\beta_y d}^2, \tag{34}$$

the second covariance term may be estimated as:

$$\hat{\operatorname{cov}}(\hat{X}_d, \hat{X}_m) = Z_{imp}^m Z_{imp}^d \frac{B_d^z}{B_S^z} \hat{\sigma}_{\beta_x d}^2. \tag{35}$$

Similarly as for the mean square errors, we now also have a recurrent formula for the covariances. If z_i 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 estimation for more than two strata can be generalized in a similar way as in the case with no chain structure.

4 REMARKS

4.1 Special cases

The above described techniques are quite general. Often we work simply with $c_i := x_i$. The population estimate is then:

$$\hat{Y} = Y_{sam} + X_{imp} \frac{\sum_{sam} w_i y_i}{\sum_{sam} w_i x_i}, \tag{36}$$

which is an analogy to the ratio estimator from the classic survey methodology (i.e. Levy and Lemeshow, 1999),

$$\hat{Y}_R = X_{all} \frac{\sum_{sam} w_i y_i}{\sum_{sam} w_i x_i}. \tag{37}$$

The mean square error then reduces to:

$$mse\hat{Y} = X_{imp}^2 \sigma_\beta^2 + c_{imp}\sigma^2 = X_{imp}^2 \frac{\sum_{sam} w_i^2 x_i}{\left(\sum_{sam} w_i x_i\right)^2} \sigma^2 + X_{imp}\sigma^2. \tag{38}$$

When the weights are constant, we get:

$$\hat{Y} = Y_{sam} + X_{imp}\frac{Y_{sam}}{X_{sam}} = X_{all}\frac{Y_{sam}}{X_{sam}}, \tag{39}$$

which is equal to the ratio estimator. For the error we get:

$$mse\hat{Y} = X_{imp}\left(X_{imp}\frac{X_{sam}}{X_{sam}^2}+1\right)\sigma^2 = X_{imp}\frac{X_{all}}{X_{sam}}\sigma^2. \tag{40}$$

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

$$mse\hat{Y} = (N-n)\frac{N}{n}\sigma^2 = \frac{N^2}{n}\left(1-\frac{n}{N}\right)\sigma^2, \tag{41}$$

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

4.2 Choosing the weights

For getting the population estimates, we use imputations with help of the superpopulation model, rather than the commonly used weighting techniques. The weights are used 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 constants in the numerator and denominator of $\hat{\beta}$ cancel out).

If we apply some outlier-detection methods to identify observations that may not fit the model (see e.g. Grubbs, 1969 or Barnett and Lewis, 1994), 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 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. However, this approach is rather simplified. The proportion of sampled units can be much lower in the studied small stratum than in the neighbouring strata, resulting in overly high weights. Also the dependency of the studied and auxiliary variables may differ between the strata. These considerations open an entire field of Small Area Estimation, which has been extensively studied for example by Rao (2003).

4.3 Multivariate regression

The methods as described in sections 1–3 can be easily generalized to accommodate more regression variables with modeling the data as $y_i = \vec{x}_i^T\vec{\beta} + e_i$, with $e_i \sim (0, c_i\sigma^2)$, covariate vector $\vec{x}_i = (x_{1i},..., x_p)^T$ and the vector of parameters $\vec{\beta} = (\beta_1,..., \beta_p)^T$. We would then have to work with matrix calculus, for the sampled part using $\vec{Y} = (y_1,..., y_n)^T$, a $n \times p$ matrix $X = (\vec{x}_1,..., \vec{x}_n)^T$, vector and an $n \times n$ diagonal variance scaling matrix C with $C_{ii} = \frac{w_i}{c_i}$. Because c_i are one-dimensional, they have to be chosen as a function of one or more of the covariates.

The regression parameters can be than estimated as $\hat{\vec{\beta}} = (X^TCX)^{-1}X^TC\vec{Y}$ with $p \times p$ variance matrix $V = \text{var}\,\hat{\vec{\beta}} = (X^TCX)^{-1}(X^TC^2X)(X^TCX)^{-1}\sigma^2$. We can take:

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{sam}\frac{(y_i - x_i^T\hat{\beta})^2}{c_i}. \tag{42}$$

We estimate the target variable as $\hat{y}_i = \vec{X}_i^T \hat{\vec{\beta}}$ for $i \in imp$. Denote $\vec{X}_{imp} = \left(\sum_{imp} X_{1i}, ..., \sum_{imp} X_{pi} \right)^T$. Then the mean square error of the estimate is:

$$mse\hat{Y} = \vec{X}_{imp}^T V \vec{X}_{imp} + c_{imp}\sigma^2, \tag{43}$$

and the results from section 3 can be generalized similarly. Note that in this way we could include also the intercept term.

Difficulties can arise when the matrix $X^T CX$ is singular or under-determined, which can be the case when there is a linear dependency between the regression variables. It is then impossible to compute the inverse $(X^T CX)^-$, therefore for estimating $\hat{\vec{\beta}}$ one must either omit one or more of the covariates or use some pseudo-inverse matrix $(X^T CX)^-$, such as the Moore-Penrose pseudo-inverse matrix (Penrose, 1955).

5 EXAMPLES

Mean square error estimation by the means of the superpopulation model as shown here has been adapted by the CZSO for business statistics. Larger surveys often have a very detailed stratification structure, with many small strata consisting of only a few units. Also a sequential approach is used, when the most important variables are estimated first and with their help the other ones are imputed, building a chain structure. We show here examples of mean square error and coefficient of variation estimation.

5.1 Revenue from sales of own products and services

First, suppose we want to estimate the aggregate revenue from sales of own products and services in one particular two-digit NACE stratum using the annual structural business statistics survey data from year 2010. The population of enterprises was divided into sampling substrata by size class (1–9, 10–19, 20–49 employees according to the business register) and by three-digit NACE (in this case there are three subgroups, say 1–3). We estimate the regression coefficients for each of the groups separately. If there are less than 15 responding enterprises in one group, we use there the coefficient $\hat{\beta}$ computed over the whole corresponding size class group. As the auxiliary variable x_i, the total turnover from tax declaration was taken. We take again the variance scaling as $c_i \equiv 1$, $c_i = x_i$ and $c_i = x_i^2$ and compare the results. An outlier detection technique based on assessing the influence of each observation on the estimate $\hat{\beta}$ was used.

In Table 3, we see the number of enterprises sampled (*sam*) and non-sampled or non-responding (*imp*) in respective groups. The sample was designed to pay more attention to larger companies. In the higher size classes, all units were sampled and some of them did not respond. There are some strata with relatively few sampled units (enterprises of higher size in 3-digit NACE groups 1 and 3, marked in italics).

Table 3 The number of enterprises in the sampling strata

		NACE3					
		1		2		3	
		Sam	Imp	Sam	Imp	Sam	Imp
Size class	0–9	20	38	86	110	42	82
	10–19	*4*	1	35	4	*14*	0
	20–49	*10*	0	25	1	*12*	1

Note: Sam – sampled part, Imp – imputed part.
Source: Czech Statistical Office

Table 4 The number of enterprises in the imputation groups

		NACE3							
		1		2		3		Total	
		Sam	Imp	Sam	Imp	Sam	Imp	Sam	Imp
Size class	0–9	(2,11,7)	(6,22,10)	(6,61,19)	(11,52,47)	(2,28,12)	(20,35,27)	(10,100,38)	(37,109,84)
	10–19	(1,2,1)	(0,0,1)	(2,21,12)	(1,0,3)	(2,9,3)	0	(5,32,16)	(1,0,4)
	20–49	(0,7,3)	0	(1,16,8)	(0,1,0)	(1,9,2)	(1,0,0)	(2,32,13)	(1,1,0)
	Total	(3,20,11)	(6,22,11)	(9,98,39)	(12,53,50)	*(5,46,17)*	(21,35,27)	(17,164,67)	(39,110,88)

Note: Sam – sampled part, Imp – imputed part.
Source: Czech Statistical Office

The regression coefficient estimates would not be reliable, if taken in these strata separately. Therefore we compute estimates for each whole size class so that the coefficients in smaller NACE groups 1 and 3 are obtained using information also from the group 2. Fortunately, there are no units to estimate in two of the small strata and the other two small strata have both just one non-responding unit.

We estimated $\hat{Y}$, corresponding $\hat{mse}$ and coefficients of variation first for the whole population and then for regional division in which enterprises were divided into three groups by place of residence: i) those residing in the capital city of Prague, ii) in the rest of Bohemia and iii) in Moravia. The number of sampled and non-sampled enterprises in each region can be seen in Table 4 in parentheses (Prague, Bohemia, Moravia).

Table 5 Revenue from sales of own products and services – the whole population

c_i	$\hat{Y}$	MSE	CV
1	11 578 276	5 632 297 044	0.65%
x_i	11 699 438	3 255 484 884	0.49%
x^2_i	11 739 074	7 428 077 251	0.73%

Note: $\hat{Y}$ – estimated total, MSE – mean square error, CV – coefficient of variation, c_i – variance scaling.
Source: Simulation – own construction, primary data: Czech Statistical Office

Table 6 Revenue from sales of own products and services – regions

Region	c_i	$\hat{Y}$	MSE	CV
Prague	1	1 133 291	1 102 787 426	2.93%
	x_i	1 158 584	350 602 637	1.62%
	x^2_i	1 159 533	1 501 735 362	3.34%
Bohemia	1	7 118 493	1 970 034 661	0.62%
	x_i	7 179 980	1 124 227 221	0.47%
	x^2_i	7 202 045	2 714 996 570	0.72%
Moravia	1	3 326 493	1 375 424 879	1.11%
	x_i	3 360 874	644 562 108	0.76%
	x^2_i	3 377 496	1 546 626 660	1.16%

Note: $\hat{Y}$ – estimated total, MSE – mean square error, CV – coefficient of variation, c_i – variance scaling.
Source: Simulation – own construction, primary data: Czech Statistical Office

The mean square error is computed in each of the regions separately, using the coefficients estimated over the sampling strata and the totals of auxiliary data in the region. Note that because the coefficients for small strata are taken from the size-class groups, covariance between estimates has to be computed as shown in section 3.2. We can see the results for each type of variance scaling c_i in Tables 5 and 6.

The estimated totals $\hat{Y}$ using different c_i are similar. The coefficient of variation differs, we can see that $c_i = x_i$ yields more accurate results than $c_i \equiv 1$ or $c_i = x_i^2$ in each case. Generally the estimated coefficients of variations are quite low, which is partly because the sampling ratio was high and the sample focused on larger and more important enterprises and partly also due to good regression fit.

5.2 Revenue from the lease of land

Suppose we want to estimate the total revenue from the lease of land in the same population and the corresponding prediction error. As auxiliary variables x_i, for each enterprise we take the predicted values of the revenue from the sales of own products and services from above. Thus we have a chain structure and therefore it is necessary to use the method described in section 3.2. Because there are some small strata, the covariance has to be computed via the chain structure as shown in section 3.3.

Again, we take the variance scaling as $c_i \equiv 1$, $c_i = x_i$ and $c_i = x_i^2$ and compare the results.

In Tables 7 and 8 we see that the estimated totals are again similar to each choice of c_i. The coefficient of variation of $\hat{Y}$ for the whole population is the lowest with $c_i \equiv 1$. Among the regions it is not so clear, the mean square error is lowest in two cases with $c_i \equiv 1$ and in one case with $c_i = x_i$.

Table 7 Revenue from the lease of land – the whole population

c_i	$\hat{Y}$	MSE	CV
1	31 492	31 291	0.56%
x_i	31 629	53 565	0.73%
x^2_i	31 751	138 821	1.17%

Note: $\hat{Y}$ – estimated total, MSE – mean square error, CV – coefficient of variation, c_i – variance scaling.
Source: Simulation – own construction, primary data: Czech Statistical Office

Table 8 Revenue from the lease of land – regions

Region	c_i	$\hat{Y}$	MSE	CV
Prague	1	15 119	9 898	0.66%
	x_i	15 139	4 542	0.45%
	x^2_i	15 153	13 312	0.76%
Bohemia	1	14 981	16 999	0.87%
	x_i	15 059	38 704	1.31%
	x^2_i	15 123	68 859	1.74%
Moravia	1	1 393	3 307	4.13%
	x_i	1 431	4 909	4.89%
	x^2_i	1 475	33 480	12.41%

Note: $\hat{Y}$ – estimated total, MSE – mean square error, CV – coefficient of variation, c_i – variance scaling.
Source: Simulation – own construction, primary data: Czech Statistical Office

CONCLUSION

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

Because the superpopulation approach is model-based, the results can be inaccurate if the model assumptions are not met. Further research can concern sensitivity analysis on departures from the assumed model, presence of outliers and goodness-of-fit tests.

References

ANSCOMBE, F. J. Examination of Residuals. *Proc. 4th Berkeley Symp.* 1961, 1, pp. 1–36.

BARNETT, V., LEWIS, T. *Outliers in Statistical Data, 3rd edition.* New York: John Wiley & Sons, 1994.

CASSEL, C. M., SÄRNDAL, C. E., WRETMAN, J. H. *Foundations of Inference in Survey Sampling.* New York: John Wiley & Sons, 1977.

COCHRAN, W. G. *Sampling Techniques, 3rd edition.* New York: John Wiley & Sons, 1977.

COOK, R. D., WEISBERG, S. Diagnostics for Heteroscedasticity in Regression. *Biometrika*, 1983, 70, pp. 1–10.

GRUBBS, F. E. Procedures for Detecting Outlying Observations in Samples. *Technometrics,* 1969, 11, pp. 1–21.

HÁJEK, J. *Teorie pravděpodobnostního výběru s aplikacemi na výběrová šetření* (Probability Sampling Theory with Applications to Sample Surveys). Prague: ČSAV (Czechoslovak Academy of Sciences), 1960.

HÁJEK, J. *Sampling from a Finite Population*. New York: Marcel Dekker, 1981

LEVY, P. S., LEMESHOW, S. *Sampling of Populations: Methods and Applications, 3rd edition.* New York: John Wiley & Sons, 1999.

PENROSE, R. A Generalized Inverse For Matrices. *Proceedings of the Cambridge Philosophical Society,* 1955, 51, pp. 406–413.

RAGHUNATHAN, T., LEPKOWSKI, J., HOEWYK, J. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology.* 2001, 27 (1), pp. 85–95.

RAO, J. N. K. *Small Area Estimation.* New York: John Wiley & Sons, 2003.