2. Mathematical Background

To explain some methods for combining two data sets consider the following situation: Let A and B be two sample surveys. The number of observations are different and not all variables in the two data sets are the same. Moreover some of the variables are observed in both surveys, some are observed in the sample survey A und some other variables are only available in the sample survey B. The idea is now to estimate the missing variables in one survey, lets say in A. Assume that this variable of interest has been observed in the survey B. Many different possibilities to perform this estimation exists. In this chapter the following kind of methods are considered:

1. Regression Models including selected unit-level Small Area Methods
2. Statistical Matching

For (1) linear regression, robust linear regression, logistic regression, linear mixed models and generalized linear mixed models (logistic mixed models) are evaluated.

For (2) random hot deck, sequential random hot deck and weighted random hot deck methods are regarded.

Before the structure of the data sets and the setting is explained, some details on the methods used and evaluated are given in the following.

2.1. Definitions

Since the equivalised household income and the at-risk-of-poverty rate are of crucial importance, they are described in detail in the following.

- *equivalised household income*: It can be seen as the income standardized on a single-person household. Following the description from Statistics Austria on `http://www.statistik.at/web_en/statistics/social_statistics/poverty_and_social_inclusion/index.html` it "is obtained by dividing the available household income by the number of consumption equivalents in the household. It is assumed that, as the size of the household increases and depending on the age of the children, cost savings are achieved in the household through joint budgeting (economies of scale). For weighting purposes, the EU scale (modified OECD scale) is used to calculate a household's resource requirements. An adult living on his or her own is taken as the reference point (= consumption equivalent), with an allocated weighting of 1. For each additional adult, the assumed resource requirement increases by 0.5 consumption equivalents. Each child under the age of 14 is weighted with a consumption equivalent of 0.3. So a household comprising a father, mother and child would have a calculated consumption equivalent of 1.8 compared with a single-person household."

V. Puchner, *Evaluation of Statistical Matching and Selected SAE Methods,* BestMasters,
DOI 10.1007/978-3-658-08224-6_2, © Springer Fachmedien Wiesbaden 2015

- *risk-of-poverty*: Again Statistics Austria defines: "The at-risk-of-poverty is calculated on the basis of the equivalised household income. People are considered to be at-risk-of-poverty or affected by the risk of poverty if their equivalised household income is below an at-risk-of-poverty threshold of 60% of the median." (see `http://www.statistik.at/web_en/statistics/social_statistics/poverty_and_social_inclusion/index.html`).

  Of particular interest is the at-risk-of-poverty rate, i.e. the portion of all households that are considered to be at-risk-of-poverty.

  So in a mathematical notation the estimation of the at-risk-of-poverty rate from a sample is defined as [see, e.g., Alfons et al., 2013]

$$arpr := \frac{\sum_{i \in I_{<arpt}} w_i}{\sum_{i=1}^{n} w_i} \cdot 100 \quad ,$$

  where $I_{<arpt} := \{i \in \{1, \ldots, n\} : x_i < arpt\}$, $\mathbf{x} := (x_1, \ldots, x_n)^T$ with $x_1 \leq \ldots \leq x_n$, is the equivalised household income, $\mathbf{w} := (w_1, \ldots, w_n)^T$ are the corresponding sample weights, n the number of observations and $arpt$ is the estimated at-risk-of-poverty threshold, $arpt = 0.6 \cdot \hat{q}_{0.5}$, where $\hat{q}_{0.5}$ is the weighted median defined as

$$\hat{q}_{0.5} = \hat{q}_{0.5}(\mathbf{x}, \mathbf{w}) = \begin{cases} \frac{1}{2}(x_j + x_{j+1}) & \text{if } \sum_{i=1}^{j} w_i = 0.5 \cdot \sum_{i=1}^{n} w_i \\ x_{j+1} & \text{if } \sum_{i=1}^{j} w_i < 0.5 \cdot \sum_{i=1}^{n} w_i < \sum_{i=1}^{j+1} w_i \end{cases} \quad .$$

  The threshold in 2010 was at a equivalised income of 12371 euros per year (or about 1031 euros a month (12 times)) for a single-person household [see Glaser and Heuberger, 2012].

2.2. Distance Measures

Distance measures are of great importance for statistical matching methods. Therefore some popular distance measures get described, because they are mentioned and used later on (see also Chapter 2.4).

Let $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}$ be p-dimensional vectors. In general a real-valued distance function d fulfills the following properties [see, e.g., D'Orazio et al., 2006]:

1. for any two vectors $\mathbf{x_k}$ and $\mathbf{x_l}$ the function d is symmetric, i.e. $d(\mathbf{x_k}, \mathbf{x_l}) = d(\mathbf{x_l}, \mathbf{x_k})$,
2. for any two vectors $\mathbf{x_k}$ and $\mathbf{x_l}$ the function d is nonnegative, i.e. $d(\mathbf{x_k}, \mathbf{x_l}) \geq 0$, and
3. for any vector $\mathbf{x_k}$ holds, that $d(\mathbf{x_k}, \mathbf{x_k}) = 0$ (property of identity).

Looking at a data set $\mathbf{X}$ with n observations and p variables for each observation, in order to compute the distances it is necessary to distinguish between the several types of the variables, i.e. to consider if the variables are continuous, categorical (maybe binary) or semi-continuous.

Minkowsky distance
A common class of distance measures is based on the Minkowsky distance, which is defined in D'Orazio et al. [2006] as

$$d(\mathbf{x_k}, \mathbf{x_l}) = \left[\sum_{j=1}^{p} c_j^{\lambda} |x_{kj} - x_{lj}|^{\lambda} \right]^{\frac{1}{\lambda}} \quad ,$$

with $\lambda \geq 1$ and c_j a scaling factor for the jth entry. Note that the Minkowsky distance is often defined without the scaling factor c_j, that is $c_j = 1$. It is used for continuous variables.

Manhattan distance
A representative of the Minkowsky distance is for example the Manhattan distance. The parameter λ is defined as 1 and this distance function looks like [see, e.g., D'Orazio et al., 2006]

$$d(\mathbf{x_k}, \mathbf{x_l}) = \sum_{j=1}^{p} c_j |x_{kj} - x_{lj}| \quad .$$

Euclidean distance
Another distance function based on the Minkowsky distance is the Euclidean distance. λ is set to 2, so the distance function is defined as [see, e.g., D'Orazio et al., 2006]

$$d(\mathbf{x_k}, \mathbf{x_l}) = \sqrt{\sum_{j=1}^{p} c_j^2 (x_{kj} - x_{lj})^2} \quad .$$

Maximum distance
The last mentioned representative of the Minkowsky distance is the maximum distance, also called Chebyshev distance. It results letting λ converge to infinity and therefore looks like [see, e.g., D'Orazio et al., 2006]

$$d(\mathbf{x_k}, \mathbf{x_l}) = \max_{j \in \{1,\ldots,p\}} \{c_j |x_{kj} - x_{lj}|\} \quad .$$

Mahalanobis distance
A distance measure including the covariance matrix of the vectors, $\mathbf{\Sigma_{XX}}$, and also used for continuous data, is the Mahalanobis distance. It is defined as [see, e.g., D'Orazio et al., 2006]

$$d(\mathbf{x_k}, \mathbf{x_l}) = (\mathbf{x_k} - \mathbf{x_l})' \mathbf{\Sigma_{XX}}^{-1} (\mathbf{x_k} - \mathbf{x_l}) \quad .$$

Gower distance
An adequate distance measure for mixed type variables is the Gower distance. It looks like [see, e.g., D'Orazio et al., 2006]

$$d(\mathbf{x_k}, \mathbf{x_l}) = \frac{1}{p} \sum_{j=1}^{p} c_j d(x_{kj}, x_{lj}) \quad ,$$

where c_j is set to 1 for binary variables and $c_j = \frac{1}{R_j}$, with R_j defined as the range of the jth variable, $R_j = \max_k\{x_{kj}\} - \min_k\{x_{kj}\}$, for continuous and categorical ordinal variables. $d(x_{kj}, x_{lj})$ is usually defined as $|x_{kj} - x_{lj}|$, see D'Orazio et al. [2006].

2.3. Regression Models Including Selected Small Area Methods

2.3.1. Linear Regression

Regression analysis is used to predict some values of one or more so-called dependent variables with known (independent) variables. In linear regession a linear relationship between the independent and dependent variables is tried to be found.

In the rest of the chapter, only the case of one dependent variable is described.

Multiple Linear Regression

Multiple linear regression means that one dependent variable $\mathbf{Y}$, the response, is tried to get explained with one or more (independent) predictor variables $\mathbf{x_1}, \ldots, \mathbf{x_q}$. $q + 1$ is equal to the number of variables in the model matrix (see also the next page for further explanations). For instance the dependent variable Y could be the income and the independent variables could be the age, the state, the highest completed level of education and the occupational status.

In the multiple linear model $\mathbf{Y}$ is a linear combination of the $\mathbf{x_1}, \ldots, \mathbf{x_q}$ including maybe a constant term, referred to as the intercept, and a random error $\boldsymbol{\varepsilon}$. The values of the predictors $\mathbf{x_1}, \ldots, \mathbf{x_q}$ are fixed. The error $\boldsymbol{\varepsilon}$ includes the influence of other latent variables missing in the model. It is treated as a random variable with special properties, hence also $\mathbf{Y}$ is random. So now given n independent observations of Y and the associated $x_1, \ldots, x_q$ the linear regression model looks like [see, e.g., Johnson and Wichern, 1998]

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 x_{11} + \ldots + \beta_q x_{1q} + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 x_{21} + \ldots + \beta_q x_{2q} + \varepsilon_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 x_{n1} + \ldots + \beta_q x_{nq} + \varepsilon_n \quad , \end{aligned}$$

or using matrix notation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad , \tag{1}$$

whereas $\mathbf{Y}$ is a random vector of dimension n, $\mathbf{X}$ the matrix of dimension $n \times (q + 1)$ with ones in the first column, $\boldsymbol{\beta}$ the $(q+1)$-vector $(\beta_0, \beta_1, \ldots, \beta_q)$ and $\boldsymbol{\varepsilon}$ a random vector of dimension n. In the following $\mathbf{Y}$ stands for the random vector and $\mathbf{y}$ stands for the vector with concrete realisations.

There are some assumptions to the random vector $\boldsymbol{\varepsilon}$: The expected value of ε_i is 0 $\forall i = 1, \ldots, n$, the variance of the errors is equal and constant $\forall i = 1, \ldots, n$ (this is called "homoscedasticity") and the errors are uncorrelated. This can be written as [see, e.g.,

Johnson and Wichern, 1998]

$$\begin{aligned} &1.\ \mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0} \ \text{ and} \\ &2.\ Cov(\boldsymbol{\varepsilon}) = \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \sigma^2 \mathbf{I}_n \end{aligned} \tag{2}$$

In case of ordinary least squares (OLS) the normal assumption of $(\mathbf{Y}, \mathbf{X})$ is necessary to test hypothesis or to estimate confidence intervals when using the classical method.

Ordinary Least Squares (OLS) Estimation

In general, the parameters $\boldsymbol{\beta}$ (i.e. $\beta_0, \beta_1, \ldots, \beta_q$) and σ^2 are unknown and they have to be estimated to predict the response with given predictor variables. There are lots of possibilities to perform the estimation. One of the best known methods is the Ordinary Least Quares (in the following also denoted by OLS). The idea is to minimize the sum of the squared differences $y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_q x_{iq}$. These differences are called residuals and are denoted by ε_i.

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_q x_{iq})^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \to \min$$

The obtained OLS estimates of $\boldsymbol{\beta}$ is denoted by $\hat{\boldsymbol{\beta}}$ and it is expressed by the matrix multiplication $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ [see, e.g., Johnson and Wichern, 1998]. If $\mathbf{X}$ hasn't full rank $q+1$, for the inverse $(\mathbf{X}'\mathbf{X})^{-1}$ a generalized inverse of $\mathbf{X}'\mathbf{X}$ is used. The so-called fitted values of $\mathbf{y}$ are given as the estimation of $\mathbf{y}$: $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Also the estimated residuals can be calculated using only simple matrix operations of $\mathbf{X}$ and $\mathbf{y}$, $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}$ [see, e.g., Johnson and Wichern, 1998].

In this context it is important to declare a model matrix. It has ones in the first column for the intercept. The rest of the matrix is built by the predictor variables. In the case of a continuous variable the corresponding column of the model matrix contains simply the several values of the observations. In the case of a factor variable the group membership is distinguished by more than one column in the model matrix. So in the case of k groups for a factor, the model matrix gets added $k-1$ columns, whereas the first group is the reference group. If an observation belongs to the first group, the $k-1$ added columns will have the value 0 (influence included in the intercept) and if an observation belongs to one of the other $k-1$ groups, the corresponding column will have a 1 as entry and the remaining columns again 0. [see, e.g., Sachs and Hedderich, 2009]

Quality measure

A common measure for the quality of the estimation is the coefficient of determination

R^2, which is given as [see, e.g., Johnson and Wichern, 1998]

$$R^2 = 1 - \frac{\sum_{j=1}^{n} \hat{\varepsilon}_j^2}{\sum_{j=1}^{n} (y_j - \overline{y})^2} = \frac{\sum_{j=1}^{n} (\hat{y}_j - \overline{y})^2}{\sum_{j=1}^{n} (y_j - \overline{y})^2}$$

whereas $\overline{y}$ is the arithmetic mean of the y_i, so $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. It describes the (by the predictor variables) explained proportion of the total variance of the y_i. R^2 takes values from 0 to 1. A value of 1 means that all estimated residuals are 0 (perfect linear relationship) and on the other side a value of 0 means, that all regression coefficients except the intercept β_0 are 0, so the predictor variables $x_1, \ldots, x_q$ have no bearing on the response (no linear relationship). Generally the higher the value, the better the fit. But caution is required, because sometimes it leads to false conclusions. With increasing number of predictor variables also R^2 takes a higher value. In this context the so-called adjusted R-squared has to be mentioned, the number of predictor variables are taken into account here [see, e.g., Sachs and Hedderich, 2009].

Some properties

The estimations $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}$ of the classical regression model (1) with the assumptions (2) have some desirable properties [see, e.g., Johnson and Wichern, 1998]:

1. $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$
2. $\mathbb{E}(\hat{\boldsymbol{\varepsilon}}) = (0)$ and $Cov(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$
3. $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}$ are uncorrelated.
4. The OLS-fit of $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE)(when $\mathbf{X}$ has full rank).

Normal distributed errors

Under the assumption of normal distribution of the residuals $\boldsymbol{\varepsilon}$ with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I_n}$, e.g. $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I_n})$, in case of the classical linear regression model (1) where $\mathbf{X}$ has full rank $q+1$, the OLS estimator $\hat{\boldsymbol{\beta}}$ is distributed as $\mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$ [see, e.g., Johnson and Wichern, 1998].

Weighted linear regression (WLS)

In the case of heterogeneity of variance of the residuals (i.e. if the second assumption of (2) of the classical linear regression model is violated), the OLS estimator will no longer be the "BLUE", because there is loss of efficiency [see, e.g., Heeringa et al., 2010]. But

there is the possibility to put things right if "weights" will be used. The properties

$$\begin{aligned} &1.\ \mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0} \ \text{ and} \\ &2.\ Cov(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{\Sigma} \quad , \end{aligned}$$

whereas $\boldsymbol{\Sigma}$ is not the identity matrix, but a $n \times n$ diagonal matrix, are considered. The diagonal of $\sigma^2\boldsymbol{\Sigma}$ contains the reciprocal values of the weights, what are the residuals variances. A transformation of the original model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

to

$$(\boldsymbol{\Sigma}^{-1/2}\mathbf{Y}) = (\boldsymbol{\Sigma}^{-1/2}\mathbf{X})\boldsymbol{\beta} + (\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon})$$

results in a model that fulfills the "classical" assumptions (2), because $Cov(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}^{-1/2}Cov(\boldsymbol{\varepsilon})\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{-1/2}\sigma^2\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1/2} = \sigma^2\mathbf{I}_n$ [see, e.g., Heeringa et al., 2010]. So the WLS estimator, which is at the same time the OLS estimator of this transformed model, is BLUE again and looks like [see, e.g., Heeringa et al., 2010]

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= ((\boldsymbol{\Sigma}^{-1/2}\mathbf{X})'\boldsymbol{\Sigma}^{-1/2}\mathbf{X})^{-1}(\boldsymbol{\Sigma}^{-1/2}\mathbf{X})'\boldsymbol{\Sigma}^{-1/2}\mathbf{y} = \\ &= (\mathbf{X}'(\boldsymbol{\Sigma}^{-1/2})'\boldsymbol{\Sigma}^{-1/2}\mathbf{X})^{-1}\mathbf{X}'(\boldsymbol{\Sigma}^{-1/2})'\boldsymbol{\Sigma}^{-1/2}\mathbf{y} = \\ &= (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y} \quad . \end{aligned}$$

So with the denotation of $\mathbf{W}$ for the matrix of the weights, that is $\mathbf{W} = \sigma^2\boldsymbol{\Sigma}^{-1}$, the WLS estimator is written as $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$ [see, e.g., Heeringa et al., 2010].

Working with complex sample survey data it is necessary to use weights in order to include the differences in sample inclusion probabilities, unit nonresponse and so on. The weight of a unit specifies the number of people that are represented by the particular observation. These sampling weights can be integrated in the regression model via the WLS estimation. So the weighted least squares estimation results in the formula $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$ for the regression parameters, where $\mathbf{W}$ is the diagonal matrix of the weights.

2.3.2. Robust Linear Regression

By reason of the high sensibility to outliers of the linear regression method, the interest is in methods that can deal properly with data containing outliers. The so called "breakdown point" is a measure for the robustness, which indicates the minimum proportion of the data contaminated with outliers that can make the estimator "useless" (e.g. the estimator takes arbitrarily large values) [see, e.g., Rousseeuw and Leroy, 2003]. OLS estimators has a breakdown point of $\frac{1}{n}$. $\frac{1}{n}$ converges to zero for increasing sample size n and hence it can be said that OLS estimation has a breakdown point of 0%.

There are several methods to increase the breakdown point of regression estimators, but in this contribution only the idea of so called "M"-estimators is explained. Instead of

minimizing the sum of the squared residuals, the aim is to minimize the sum of another function ρ of the residuals [see, e.g., Rousseeuw and Leroy, 2003],

$$\sum_{i=1}^{n} \rho(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_q x_{iq}) = \sum_{i=1}^{n} \rho(\varepsilon_i) \rightarrow \min$$

For logical reasons this function ρ is a symmetric function, i.e. $\rho(t) = \rho(-t)\ \forall t$, and it has a unique minimum at zero [see, e.g., Rousseeuw and Leroy, 2003]. A well known version of M-estimators is Huber's M-estimator, where the equation

$$\sum_{i=1}^{n} \min(c, \max(\frac{\varepsilon_i}{\hat{\sigma}}, -c))\mathbf{x_i} = \mathbf{0}$$

with $\mathbf{x_i} = (x_{i1}, \ldots, x_{ip})$ and $\mathbf{0} = (0, \ldots 0)$ has to be solved. Details can be found in the book "Robust Regression and Outlier Detection" of Rousseeuw and Leroy [2003].

2.3.3. Logistic Regression

The intention is to model and predict a binary variable. Due to the fact that linear regression models require a continuous response variable, they can't be used for the estimation. A popular technique to consider binary responses is logistic regression.

The solution is to work with the posterior probabilities $\mathbb{P}(Y = 0|\mathbf{X} = \mathbf{x})$ and $\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x})$, so the probability that the response of one observation belongs to category 0 or 1 given the corresponding predictor variables. The idea is that a transformation of the posterior probabilities is linear in $\mathbf{x}$. The model looks like [see, e.g., Hastie et al., 2009]

$$\log \frac{\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x})}{\mathbb{P}(Y = 0|\mathbf{X} = \mathbf{x})} = \beta_0 + \beta_1 x_1 + \ldots + \beta_q x_q \quad .$$

One can show that

$$\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_q x_q)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_q x_q)} \quad \text{and}$$

$$\mathbb{P}(Y = 0|\mathbf{X} = \mathbf{x}) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_q x_q)} \quad .$$

The used transformation is the monotone logit-transformation: $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$. Note that the posterior probabilities sum up to 1.

The estimation of the parameter vector β is usually made with the maximum likelihood method and the solving is done iteratively [see, e.g., Hastie et al., 2009]. There is no closed form of the solution as it is the case in linear regression models.

Again it is necessary to work with weights if it is handled with real sample survey data. The weights get incorporated in the logistic regression model in the weighted (pseudo-)(log-)likelihood function used for the calculation of the parameters. For details on the weighted logistic regression see Heeringa et al. [2010].

2.3.4. Linear Mixed Regression

The idea of mixed regression models is, in addition to fixed effects, the inclusion of so called "random effects". While the regression coefficients of the previous sections are now noted as fixed effects, random effects are added to the model. Every unit of a sample survey belongs to a certain domain. A domain is a subset of the population U such as for example the people of a federal state, another geographical area population or also a class defined by age and gender. If it was worked as before only with fixed effects so as to incorporate these differences between the classes or domains, there would be too many parameters that would have to be estimated. So now random effects get added in order that possible differences between the different domains get considered in a model.

A mixed effects model can be formulated as [see, e.g., Christensen, 2011]

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \quad ,$$

where $\mathbf{u}$ is a vector of domain-specific random effects $u_d \sim \mathcal{N}(\mathbf{0}, \sigma_u^2)$ for domain U_d and $\mathbf{Z}$ a matrix. For a single observation $k \in U_d \subset U$ the model for domain-specific random intercepts is given by

$$Y_k = \mathbf{x_k'}\boldsymbol{\beta} + u_d + \varepsilon_k \quad ,$$

with $\varepsilon_k \sim \mathcal{N}(\mathbf{0}, \sigma^2)$.

First, the values of $\boldsymbol{\beta}$, σ_u^2 and σ^2 have to be estimated. Afterwards the values of the random effects are estimated.

For estimated samples from complex surveys, weights should be included in the model. Details can be found in the resources of the AMELI project 2011 [see, e.g., Lehtonen et al., 2011].

2.3.5. Logistic Mixed Regression

Analogous to linear mixed effects regression models, random effects can be included in the logistic mixed effects model as well. Hence the mixed logistic model is given by [see, e.g., Lehtonen et al., 2011]

$$\mathbb{P}(Y_k = 1|u_d) = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_q x_q + u_d)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_q x_q + u_d)} \quad ,$$

with unit $k \in U_d$, $\boldsymbol{\beta}$ the vector of the fixed effects und u_d the domain-specific random effect.

Again weights can be included in the model working with sample survey data [see Lehtonen et al., 2011].

2.3.6. The Transmission of the Model

The procedure is to use regression methods for the estimation of the variable of interest in sample survey A. For that reason the model observed for sample survey B is transmitted

to the sample survey A. For this purpose a regression model is built in the sample survey B using predictor variables that are observed in both sample surveys A and B, and using the variable of interest as response, i.e., regarding for example linear regression, $\hat{\boldsymbol{\beta}}_{\boldsymbol{B}} = (\mathbf{X_B}'\mathbf{X_B})^{-1}\mathbf{X_B}'\mathbf{y_B}$ with $\mathbf{X_B}$ the model matrix built by intercept and predictor variables from survey B (here: EU-SILC) and $\mathbf{y_B}$ the response variable in survey B.

The estimated parameter vector $\hat{\boldsymbol{\beta}}_{\boldsymbol{B}}$ is now used as parameter vector in the model for A, more precisely, in the model $\hat{\mathbf{y}}_{\mathbf{A}} = \mathbf{X_A}\hat{\boldsymbol{\beta}}_{\boldsymbol{B}}$, where in $\hat{\mathbf{y}}_{\mathbf{A}}$ are the estimated values for the variable of interest, that is initially missing in A, and $\mathbf{X_A}$ is the model matrix of survey A, including ones in the first column and the corresponding interaction and contrasts for the common variables of A and B (used for the model estimation in survey B) in the other columns. Note that the (common) variables in $\mathbf{X_A}$ have to be in the same order as in $\mathbf{X_B}$, the model matrix (with ones in the first column) used in the estimated model for survey B.

2.4. Statistical Matching

In general two approaches in statistical matching exists: the micro and the macro approach. In the following only the micro approach is considered. This is the version where in the survey A an additional variable is simulated, i.e. values (of the variable observed in survey B and not observed in survey A) get imputed, so that a "synthetic" data set results. The survey A is called "recipient" file and survey B the "donor" file. [see, e.g., D'Orazio et al., 2006]

Table 1: Survey A and B in one data set: the variables X_j, $j = 1, \ldots, f$, are observed only in survey A, the variables Z_j, $j = 1, \ldots, h$, are observed only in survey B and the variables Y_j, $j = 1, \ldots, g$ are observed both in A and B.

survey	X_1	X_2	$\ldots$	X_f	Y_1	Y_2	$\ldots$	Y_g	Z_1	Z_2	$\ldots$	Z_h
A	x^A_{11}	x^A_{12}	$\ldots$	x^A_{1f}	y^A_{11}	y^A_{12}	$\ldots$	y^A_{1g}				
	x^A_{21}	x^A_{22}	$\ldots$	x^A_{2f}	y^A_{21}	y^A_{22}	$\ldots$	y^A_{2g}				
	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$				
	$x^A_{n_A1}$	$x^A_{n_A2}$	$\ldots$	$x^A_{n_Af}$	$y^A_{n_A1}$	$y^A_{n_A2}$	$\ldots$	$y^A_{n_Ag}$				
B					y^B_{11}	y^B_{12}	$\ldots$	y^B_{1g}	z^B_{11}	z^B_{12}	$\ldots$	z^B_{1h}
					y^B_{21}	y^B_{22}	$\ldots$	y^B_{2g}	z^B_{21}	z^B_{22}	$\ldots$	z^B_{2h}
					$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
					$y^B_{n_B1}$	$y^B_{n_B2}$	$\ldots$	$y^B_{n_Bg}$	$z^B_{n_B1}$	$z^B_{n_B2}$	$\ldots$	$z^B_{n_Bh}$

Different possibilities to perform the imputation are available, for example, the non-parametric techniques "nearest neighbor distance hot deck imputation" (in the following

http://www.springer.com/978-3-658-08223-9

Evaluation of Statistical Matching and Selected SAE Methods
Using Micro Census and EU-SILC Data
Gissing, V.
2015, XIII, 101 p. 6 illus., Softcover
ISBN: 978-3-658-08223-9