Model-based estimation

10

Harm Jan Boonstra and Bart Buelens

Statistische Methoden (201106)

Statistics Netherlands

The Hague/Heerlen

Explanation of symbols

.	= data not available
*	= provisional figure
**	= revised provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2010–2011	= 2010 to 2011 inclusive
2010/2011	= average of 2010 up to and including 2011
2010/'11	= crop year, financial year, school year etc. beginning in 2010 and ending in 2011
2008/'09–2010/'11	= crop year, financial year, etc. 2008/'09 to 2010/'11 inclusive

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Henri Faasdreef 312
2492 JP The Hague

Prepress
Statistics Netherlands - Grafimedia

Cover
TelDesign, Rotterdam

Information
Telephone +31 88 570 70 70
Telefax +31 70 337 59 94
Via contact form: www.cbs.nl/information

Where to order
E-mail: verkoop@cbs.nl
Telefax +31 45 570 62 68

Internet
www.cbs.nl

60165201106 X-37

Table of Contents

1. Introduction to the theme

1.1 General description and reading guide

The theme of model-based estimation can be interpreted very broadly. Most branches of statistics are model-based in the sense that results are based on *explicit* probability models that relate missing information to available information. The objective of this description is to demonstrate how relatively simple models can be used to make estimates for population variables. We are concentrating on estimates of averages or totals for finite populations or subpopulations.

The theory of survey sampling is usually described from a design-based perspective, which focuses on the probability mechanism that is used to select the sample. However, there are situations in which a design-based approach does not work well or at all. Two such situations are those in which:

1. there is no known random sample design, such as for administrative data from incomplete registers or for specific kinds of internet surveys.
2. there is too little sample data available to make reliable estimates. This is particularly the case if the level of detail for which figures must be produced is high, such that the sample size is small in the various subpopulations.

In these situations, model-based estimation methods can be used. First, we will discuss how population totals can be estimated using linear regression models. These models do not depend explicitly on a sample design and can therefore be used in situation 1. We then state a number of issues that must be paid attention to in the use of these models. The estimators that are derived from the use of these models are also called synthetic estimators. If sufficient suitable auxiliary variables are available as input for the regression model, then synthetic estimators can also be used to make estimates for small areas.

In the second subtheme, we discuss a simple type of model that is suitable for small area estimates, also in situations where synthetic estimates are not adequate. Small areas are subpopulations for which the sample size is too small to make reliable direct (design-based) estimates. The model, which links the various areas with one another, possibly by making use of relevant auxiliary information at area level, provides better estimates. The subtheme of small area estimators thus deals with situation 2.

1.2 Scope and relationship with other themes

We distinguish between two subthemes in the theme of model-based estimation: synthetic estimators and small area estimators. Other themes in the Methods Series, including macro-integration and seasonal correction/time series models, also use model-based estimation methods.

The concept of the synthetic estimator is used for estimators based on regression models with only fixed effects, without random effects. Fixed effects are the common regression coefficients, while random effects can be interpreted as a group of regression coefficients with the prior constraint that they are distributed around 0 according to a multivariate probability distribution. We will only describe synthetic estimators based on linear regression models here. Therefore, synthetic estimators based on, for example, a logistic regression model are not discussed. The subtheme of small area estimators further addresses models with both fixed and random effects, also known as mixed models. Here the random effects correspond to the area indicators, and can explain differences between areas when these differences cannot be explained by the other auxiliary variables used.

The model for small area estimates that is addressed in the second subtheme is formulated at area level. This means that the area averages to be estimated are directly modelled in terms of auxiliary variables at area level. In addition, direct estimates and the associated variance estimates are used as input data for areas. Small area models formulated at the unit level, say, the person level, are not discussed.

In the case of random samples, the synthetic estimator for the population total based on linear regression models will often correspond to the general regression estimator, certainly in sample designs with equal inclusion probabilities. Section 2.5 addresses this in more detail. The methodology of the general regression estimator and the related weighting are addressed in other themes in the Methods Series: ‘Sampling theory’ and ‘Weighting for non-response’. There is also a strong relationship with the method of regression imputation addressed in ‘Imputation’. In this case, missing values are replaced at micro-level by imputed values based on a regression model, and in contrast to synthetic estimators, the estimation of population totals or averages does not have to be a primary goal.

1.3 Place in the statistical process

In the statistical process, ‘Estimation’ follows after ‘Checking and editing the data’. This is not different for synthetic estimation. Here it will relate to, for example, register data that must first be linked to a population register (“backbone”) and then edited.

The small area estimators that are discussed in the second subtheme use direct estimates and the associated variance estimates at area level as input data. These direct estimates can arise from a weighting of the sample data. The Bascula weighting program can calculate both estimates for subpopulations and the associated variance estimates; see Nieuwenbroek and Boonstra (2002).

1.4 Definitions

Concept	Description
Synthetic estimates	Estimates based on a (linear) regression model, for which scores on the target variable for the non-observed units are predicted from the model.
Small areas	Publication cells for which the amount of observed data is so small that an estimate based only on this data would be too inaccurate.
Small area estimates	Estimates for small areas. In order to get accurate estimates, a model is used that relates areas in such a way that data from different areas contributes to the estimate for a particular small area.

1.5 General notation

We assume a target variable *y* for which the population total or mean must be estimated. This variable takes the values of $y_1, \ldots, y_N$ for the population units $U=\{1, \ldots, N\}$. For a subset *s* of n < N unique units, the values of *y* are known. We assume that these values are error-free. The subset *s* can be the response of a sample/random sample or it can be the set of units of an incomplete register, but we will usually call it the sample. We indicate the complement of the sample in the population, thus all of the units for which the target variable is not known, as $r = U \backslash s$. This consists of $N - n$ units.

In addition, we assume the presence of a vector of auxiliary variables *x* which are known for the entire population *U*. The vector *x* has the dimension *p*.

Sample averages are written as $\bar{y}, \bar{x}$ and population totals as t_y, t_x. Therefore $t_y = \sum_{i \in U} y_i$ and $\bar{y} = (1/n)\sum_{i \in s} y_i$, etc. Population averages are obtained by dividing t_y, t_x by *N*, $q_y = t_y / N$ and $q_x = t_x / N$, where the population size *N* is assumed to be known.

For *m* subpopulations or areas, we use an index $d = 1, \ldots, m$. For example, this means that $t_{y;d} = \sum_{i \in U_d} y_i$ is the population total of *y* in area *d*, and $q_{y;d} = t_{y;d} / N_d$ is the population average of *y* in area *d*, with N_d the population size of area *d*. Sample averages are written as $\bar{y}_d, \bar{x}_d$; these are averages of the sample s_d of size n_d which falls in subpopulation *d*. The other, not observed, $N_d - n_d$ units in subpopulation *d* are indicated by r_d.

2. Synthetic estimators

2.1 Brief description

The linear regression model for the target variable *y* given the auxiliary variables *x* is

$$y_i = \beta^T x_i + \varepsilon_i \ , \tag{2.1.1}$$

where β is a *p*-vector of regression coefficients and ε_i a normally distributed error term, independently for *I = 1, …, N*.

The model is fitted using the sample data *(y, x)* for the units in *s*. The units in *r = U\s* are then predicted based on the fitted model. The estimate for the population total of *y* will thus be

$$\hat{t}_y = \sum_{i \in s} y_i + \sum_{i \in r} \hat{y}_i = n\bar{y} + \hat{\beta}^T (t_x - n\bar{x}) \ , \tag{2.1.2}$$

where $\hat{\beta}$ is the vector of estimated regression coefficients. Therefore, in (2.1.2), the observed sample speaks for itself and the rest is predicted according to the fitted model. Note that the population totals t_x must be known, in addition to the sample averages $\bar{x}$. Due to the linear structure of the model, the individual values of *x* for units in *r=U\s* do not have to be available.

Other population variables can be estimated in a similar manner. Totals for subpopulations, for example, are estimated according to

$$\hat{t}_{y;d} = \sum_{i \in s_d} y_i + \sum_{i \in r_d} \hat{y}_i = n_d \bar{y}_d + \hat{\beta}^T (t_{x;d} - n_d \bar{x}_d) \, , \tag{2.1.3}$$

where $t_{x;d}$ are the population totals of *x* in subpopulation *d.*

A more extensive description follows in section 2.3. Additional information about model-based estimation for samples from finite populations can be found in Rubin (1987), Ghosh and Meeden (1997), Vaillant et al (2000) and Rao (2003).

2.2 Applicability

The linear regression model and the resulting synthetic estimators are widely applicable. In particular, these model-based estimation techniques can be used to replace design-based methods to estimate population totals when a known random sample design is not present. This is the case when estimating based on incomplete registers, of which the VAT turnover register is an example. The incompleteness of the data from this register is caused in part by the need to publish at a certain time when not all the returns have been received, and partly because of imperfect matching to the General Business Register (*Algemeen Bedrijvenregister* – ABR).

In addition, the synthetic estimators can be used to estimate totals or averages of subpopulations in which the sample sizes are too small to use direct estimators per

subpopulation. In this situation, we are referring to synthetic estimators as long as *no* area-specific auxiliary variables are included in the vector *x*. As such, the synthetic estimators are a simple type of small area estimators.

The linear regression model assumes a quantitative target variable *y*. Sometimes, a transformation can be carried out on the target variable to make the model more suitable. For a target variable that is always positive, a logarithmic transformation, for example, can be useful. The transformed data can then be used to fit the model. In that case, the prediction of population totals becomes slightly more complicated because the fitted values must be transformed back before adding them up.

A logistic regression model is sometimes used for categorical data with a 0/1 variable *y* per category. A linear regression model can be interesting in these cases as well. An advantage of the linear regression model is that it is easier to fit. As long as the fitted values $\hat{y}_i = \hat{\beta}^T x_i$ remain within the interval [0, 1], except for a few possible exceptions, the linear regression model seems to be a reasonable choice.

When using synthetic estimators for population totals or means, it is important to take possible selection effects into consideration as far as possible. Selection effects are effects that cause systematic differences in the target variable between the sample and the rest of the population. As a result, the model (2.1.1), which is postulated for the entire population, becomes less useful to predict the non-observed part of the population; this can give rise to bias.

To reduce selection effects, it is important to expand the model with auxiliary variables that explain these selection effects as effectively as possible. This can be studied by comparing the sample averages of the auxiliary variables with the population averages. Auxiliary variables that show large differences should be included in the model, unless it is clear that these do not relate to the target variable *y*.

In the case of incomplete registers, selection effects can arise by means of the registration procedures, or through linkage with other sources. These effects are not always easy to identify, but it is important to study the auxiliary variables possibly associated with *y* for potential systematic differences between registered and missing units.

Also in the case of random samples, selection effects almost always arise as a result of non-response. For model-based estimation, the same applies as for weighting sample data: that, to prevent bias, an attempt must be made to include auxiliary variables in the model that explain both the non-response and the target variable.

Planned selectivity arises from a sample design with unequal inclusion probabilities. The design-based methodology corrects for this by weighting with the inverse inclusion probabilities. In model-based methods, in this situation, the inclusion probability variable itself is included as auxiliary variable in a suitable way in the regression model.

When using synthetic estimators for small areas, it is very important that good auxiliary information is available. If the auxiliary information used is not very predictive for the target variable, then the estimates of area means are pulled too much towards the general sample average. If that is the case, then the specific small area methods based on models with random area effects, such as described in the following subtheme, are more suitable.

2.3 Detailed description

2.3.1 The linear regression model

The linear regression model is provided in equation (2.1.1). The error terms ε_i are assumed independently and normally distributed according to

$$\varepsilon_i \sim N(0, v_i \sigma^2) . \tag{2.3.1}$$

The auxiliary variable *v* is also called the variance structure. The model variances $v_i > 0$ give the model more flexibility. As such, it is known that many variables in business statistics have heteroscedasticity, in which the dispersion around the linear predictor $\beta^T x_i$ increases with the size of the companies. In this case, a certain positive power of the size (the number of employees or another measure) may be used for v_i, see also Hedlin et al (2001). The values v_i must be known for the sampling units, and furthermore, the population total t_v must be known to estimate variances. If there is no information that indicates that the model variances are different, then we use $v_i = 1$ for all units, and therefore $t_v = N$.

2.3.2 Fitting the model

The standard estimate for the vector of regression coefficients is

$$\hat{\beta} = \left(\sum_{i \in s} x_i x_i^T / v_i \right)^{-1} \sum_{i \in s} x_i y_i / v_i . \tag{2.3.2}$$

Assuming (2.3.1), this estimate is optimal, in the sense that the expected square error (based on the model) is minimised. Note that $\hat{\beta}$ does not depend on the variance parameter σ^2. This is important for the variance estimates that are discussed in section 2.3.5. An estimate for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i \in s} (y_i - \hat{\beta}^T x_i)^2 / v_i , \tag{2.3.3}$$

where *p* is the dimension of the vector *x*.

2.3.3 Model-based estimates for population totals

The estimate for the population total t_y based on the fitted model and the population totals of the auxiliary variables is

$$\hat{t}_y = n\bar{y} + \hat{\beta}^T (t_x - n\bar{x}), \tag{2.3.4}$$

with $\hat{\beta}$ as in (2.3.2). Sometimes, with the synthetic estimator only the term $\hat{\beta}^T t_x$ is meant. For small sampling fractions *n/N*, this is almost the same as (2.3.4). Under the condition that the variance structure *v* is also included in the vector of auxiliary variables *x*, it can be demonstrated that $\bar{y} - \hat{\beta}^T \bar{x} = 0$ and, in that case, (2.3.4) is exactly equal to $\hat{\beta}^T t_x$.

2.3.4 Synthetic estimators for subpopulations

The estimates for the subpopulation totals $t_{y;d}$ based on the fitted model are given in (2.1.3) with $\hat{\beta}$ as in (2.3.2). For small sampling fractions n_d / N_d, the estimates are closely approximated by $\hat{\beta}^T t_{x;d}$.

2.3.5 Model-based variance estimates

The expected variance of the error $t_y - \hat{t}_y$ based on the model is

$$Var(t_y - \hat{t}_y) = \left((t_x - n\bar{x})^T \left(\sum_{i \in s} x_i x_i^T / v_i \right)^{-1} (t_x - n\bar{x}) + t_v - n\bar{v} \right) \sigma^2, \tag{2.3.5}$$

where t_v is the population total and $\bar{v}$ the sample average of the variable *v*. We obtain a variance estimate of $\hat{t}_y$ by substituting the estimate (2.3.3) for σ^2. The first term of (2.3.5) is the variance resulting from uncertainty in the estimated coefficients $\hat{\beta}$, and the second term, $(t_v - n\bar{v})\sigma^2 = \sum_{i \in r} v_i \sigma^2$, is the prediction variance that is always there, even if β were known.

The error variances of the area total estimates are

$$Var(t_{y;d} - \hat{t}_{y;d}) = \left((t_{x;d} - n_d \bar{x}_d)^T \left(\sum_{i \in s} x_i x_i^T / v_i \right)^{-1} (t_{x;d} - n_d \bar{x}_d) + t_{v;d} - n_d \bar{v}_d \right) \sigma^2,$$

where $t_{v;d}$ and $n_d \bar{v}_d$ are the population and sample total respectively of the variable *v* in area *d*. Estimates are obtained by filling in $\hat{\sigma}^2$ for σ^2. If the differences between areas are not sufficiently explained by the differences in auxiliary variables, then these model-based variance estimates will often be too low. A model with random effects would be better in that case, but the literature also includes a description of alternative design-based MSE estimators, which also contain a term for the bias of the synthetic area estimators, see Rao (2003), section 4.2.4.

2.3.6 Software

The model-based estimates and variance estimates discussed can be calculated in standard statistical software packages relatively easily. In a package with extensive regression and matrix facilities such as R/S-Plus, the necessary programming work is limited.

For a large class of models, the synthetic estimators for population totals (but not for area totals) correspond to the general regression estimators, see section 2.5. These can be calculated with the Bascula weighting program, see Nieuwenbroek and Boonstra (2002) for instructions. However, Bascula does not offer the option to select a variance structure *v*.

2.3.7 Model selection

We limit ourselves here to linear regression models, therefore the selection of a model is mainly based on the selection of a vector with suitable auxiliary variables *x*, and possibly a variance structure *v*. In addition, it is possible to transform the target variable into a variable that is better described by the linear regression model. Here we will further address the selection of auxiliary variables *x*.

The vector of auxiliary variables *x* must consist of variables that are related to the target variable *y*. The better a auxiliary variable correlates with *y*, the more important it is to include this variable in the vector *x*. Furthermore, it should be studied which auxiliary variables in the sample have a different distribution than in the rest of the population; the sample is not representative for these variables. Also if these variables do not correlate very strongly with *y*, it is still better to include them in the vector *x*, to reduce any possible bias. Not all selection effects can be corrected in this way. The lack/presence of units in the sample *s*, despite the addition of many explanatory auxiliary variables, can still be related to the target variable itself. That part of the selection effects cannot be corrected for with the regression model described here.

When expanding the vector of auxiliary variables, it is important how much a new auxiliary variable still *adds* in terms of predictive power for *y* and/or the sample selection mechanism. The number of auxiliary variables may not be too large compared to the sample size *n*, otherwise there will be a danger of "overfitting" which causes the model to lose its predictive power.

Including the variance structure *v* as one of the components of the vector *x* of auxiliary variables provides not only for the simplification of some expressions, but also for a certain robustness against some types of misspecification. For example, in homoscedastic models ($v_i = 1$ for all units), it is normal to have an intercept in the model, corresponding to a constant component in the vector *x*.

To select a reasonable variance structure *v*, the relation of residuals based on the model with $v_i = 1$ with relevant auxiliary variables can be studied. In the literature, various tests for heteroscedasticity are discussed; see for example Greene (1997). In

particular, variance estimates such as (2.3.4) are sensitive to the misspecification of *v*. Various variance estimators that are more robust are discussed in Valliant et al (2000), Chapter 5.

2.4 Example

Using a possible application to the short-term statistics (*Kortetermijnstatistieken* – KS), we hope to clarify the above description.

For the short-term statistics the turnover movements from month to month are estimated. A fixed sample is currently still used for this, but it is the intention that the Tax Administration's VAT turnover registration will be used instead in the near future. The VAT turnover will replace the turnover gathered via primary data collection. Here, we will not address the possible differences between VAT turnover and the actual turnover, but we will take the VAT turnover as the starting point for estimating turnover movements. For reasons of convenience, we will also ignore here the issue concerning annual, quarterly and monthly filers.

The VAT data about the relevant period is not available for all business units in the ABR. This incompleteness is due to a number of causes, including the fact that some VAT returns are not available on time, as well as the presence of matching errors with the ABR. However, there is no known random sampling mechanism that indicates what the chances are that the VAT data is available for the business units. Design-based estimates are therefore not possible. The model-based approach is not dependent on a random sampling mechanism, but tries to model the reasons for the missing data if these correspond to the target variables.

We assume a publication cell: in other words, a subpopulation for which must be published, at two times, *t* and *t* – 1. We indicate the units in the population with U_t and U_{t-1} (subpopulations of the ABR at times *t* and *t* – 1) and the VAT turnover variable with *y* at time *t* and *z* at time *t* – 1. The turnover movement is defined as

$$O_t = \frac{t_y}{t_z} = \frac{\sum_{i \in U_t} y_i}{\sum_{i \in U_{t-1}} z_i} \; . \tag{2.4.1}$$

VAT turnover data from periods *t* – 1 and *t* is available for business units $s_{t-1} \subset U_{t-1}$ and $s_t \subset U_t$, and missing for the other units.

Both population totals in (2.4.1) can be estimated based on the model. Auxiliary information from the ABR can be used for this purpose. An important auxiliary variable is the number of employed persons (*werkzame personen* – WP). A simple model would then be

$$y_i = \beta_1 + \beta_2 \, \mathrm{WP_i} + \varepsilon_i \qquad \text{with} \qquad \varepsilon_i \sim N(0, \mathrm{WP_i}\sigma^2) \, ,$$

and auxiliary variables $x_i = (1, \mathrm{WP_i})^T$. The correlation between turnover and the number of employed persons does not have to be exactly linear, and it is possible to

experiment with various powers of WP, both in the vector *x* and in the variance structure.

Other possible auxiliary variables that can be added to the vector *x* are categorical variables such as a breakdown according to business sectors or legal structure. The ABR variable size class (*grootteklasse* – GK) is derived from WP and does not add anything to the model, provided that the dependence of WP is well specified. This will often not be the case, and then it can actually be useful to add GK.

After fitting the model (separately for *t* and *t* – 1), (2.4.1) is estimated according to

$$\hat{O}_t = \frac{\hat{t}_y}{\hat{t}_z} = \frac{n_t \bar{y} + \hat{\beta}_t^T (t_{x;t} - n_t \bar{x}_t)}{n_{t-1} \bar{z} + \hat{\beta}_{t-1}^T (t_{x;t-1} - n_{t-1} \bar{x}_{t-1})}. \qquad (2.4.2)$$

In this way, population totals are estimated in numerator and denominator each based on its own model, usually with corresponding auxiliary variables. These cross-sectional estimates can potentially be further improved with the use of multivariate or time series models. Variances for the numerator and denominator of (2.4.2) can be estimated using (2.3.5). A variance estimate for the ratio can be obtained using linearisation.

If a publication cell has too little data to properly fit the model, then several (preferably comparable) cells can be combined. The estimates for the separate publication cells are then calculated using formula (2.1.3).

2.5 Characteristics

While this part of the Methods Series was not initially described for the situation of a known sample design, it is still instructive to map out the relation of the synthetic estimator with the design-based general regression estimator for a population total in this situation. As stated earlier, a model-based estimator based on a linear regression model sometimes corresponds to a design-based estimator. The general regression estimator (GREG) for the population total t_y for a sample obtained with inclusion probabilities π_i is (Särndal et al, 1992)

$$\hat{t}_y^{GREG} = \hat{t}_y^{HT} + \hat{\gamma}^T (t_x - \hat{t}_x^{HT}), \qquad (2.5.1)$$

$$\hat{\gamma} = \left(\sum_{i \in s} x_i x_i^T / \pi_i \right)^{-1} \sum_{i \in s} x_i y_i / \pi_i ,$$

where $\hat{t}_y^{HT} = \sum_{i \in s} y_i / \pi_i$ and $\hat{t}_x^{HT} = \sum_{i \in s} x_i / \pi_i$ are the Horvitz-Thompson estimators for the population totals of *y* and *x*. This is also called a model-assisted estimator because implicit use is made of a model, but in such a way that the design-unbiasedness continues to apply in approximation. A sufficient condition for the exact equivalence of the synthetic estimator (2.3.4) and the GREG (2.5.1) is that (1) $v_i = \pi_i$ (for all *i*, up to a constant factor) so that the coefficients $\hat{\beta}$ and $\hat{\gamma}$ are the

same, and (2) $1-\pi_i = c^T x_i$ for a constant *p*-vector *c*, for all *i*. The second condition states that the variable $1-\pi_i$ must be in the vector of auxiliary variables *x*. This is certainly the case if both the constant and the inclusion probability are in *x*. To demonstrate the equivalence, we first multiply both sides of the condition (2) by $x_i^T \hat{\gamma}/\pi_i$ and sum over $i \in s$. This gives us

$$(\hat{t}_x^{HT} - n\bar{x})^T \hat{\gamma} = \sum_{i \in s} c^T x_i y_i / \pi_i = \hat{t}_y^{HT} - n\bar{y} \ ,$$

for which the last equivalent condition (2) is applied one more time. Together with $\hat{\beta} = \hat{\gamma}$, this gives $\hat{t}_y = \hat{t}_y^{GREG}$. See Boonstra (2005) for these and other relationships between design-based and model-based estimators, and for further references to the literature.

One of the implications of this similarity between design-based and model-based estimators is that we can always ensure that a synthetic estimator for a population total is approximatitely design-unbiased for a known random sample design: choose the variance structure $v_i = \pi_i$ and ensure that both the constant and the inclusion probability variable (= variance structure) are in the vector *x* of auxiliary variables. Conversely, it is useful to know which explicit model assumptions form the basis for design-based estimators. A model that poorly describes the actual target variable will not result in a large design bias but could lead to a large design variance. Luckily, the design variance can be kept in check by using a sample size that is sufficiently large.

2.6 Quality indicators

In addition to assessing the model-based variance estimates, which are a measure for the accuracy of the estimates given the model (and therefore are also sensitive to the model selection), the model itself should also be subjected to various tests.

A plot of the residuals $e_i = y_i - \hat{\beta}^T x_i$ against the fitted values $\hat{y}_i = \hat{\beta}^T x_i$ often provides insight into the possible misspecification of the model, and potential improvements. In a good model, the residuals will generally be normally distributed around 0 and there will be no further dependence with the fitted values. Goodall (1983) offers a detailed description of model diagnosis using residuals.

Another strategy is to compare the different models using certain model selection indicators, to ultimately choose the best model. These model selection measures weigh up model fit against the complexity of the model. Models with a relatively large number of coefficients do indeed have small residuals, but the predictive power can drop sharply due to too many coefficients. The predictive power of a model can also be studied in a more direct way by fitting the model on part of the data and then determining the errors in the predictions of the rest of the data. A sensitivity analysis in which a model is changed in various reasonable ways is also a manner of obtaining insight into the quality of the model estimates.

Accepted measures for variable selection in regression models are AIC and BIC, while cross-validation is a direct measure for the predictive power of a model, see, for example, Hastie et al (2003). However, the subject of model selection and evaluation is very broad and fairly complex. Furthermore, this subject has an extra dimension in sampling theory, namely that of selection effects, as described briefly in section 2.3.7. We will not address this further here.

3. Small area estimators

3.1 Short description

In a sample, we refer to small areas when we are talking about subgroups of the population that have a sample size that is too small to make reliable direct estimates. With model-based estimation methods, information from other areas is used to improve the estimate for each small area. A model assumption must be made, and an estimation method must be selected to make estimates using the model.

In this subtheme, we describe the EBLUP estimator (Empirical Best Linear Unbiased Predictor), assuming a linear mixed model in which auxiliary information can be included at area level. This model is known as a Fay-Herriot (FH) model (Fay and Herriot, 1979), and is defined as

$$\begin{aligned} \hat{\theta}_{y;d} &= \theta_{y;d} + \varepsilon_d \\ \theta_{y;d} &= \theta_{x;d}^T \beta + v_d \end{aligned} \qquad (3.1.1)$$

where $\varepsilon_d \sim N(0, \psi_d)$ and $v_d \sim N(0, \sigma_v^2)$ for $d = 1, \ldots, m$ and m is the number of areas. The population average of y for area d is $\theta_{y;d}$. $\hat{\theta}_{y;d}$ is a direct, design-based estimator for $\theta_{y;d}$ with error ε_d. Direct estimates are only based on information from the area itself. We assume that the estimates $\hat{\theta}_{y;d}$ are not biased, with variance estimates ψ_d. The vector $\theta_{x;d}$ consists of area-specific auxiliary variables. The random effects v_d have variance σ_v^2 and are independent of ε_d.

A linear mixed model distinguishes itself from the linear regression models such as they are used in the subtheme synthetic estimators through the presence of so-called random effects, in this case, random area effects. The variations in the area estimates which are not explained by the auxiliary variables or the sampling errors are accounted for by the random effects of model (3.1.1). In most cases, the vector β will contain an intercept μ. The effects $\mu + v_i$ then form a set of area intercepts, with a joint underlying distribution $N(\mu, \sigma^2)$. This gives rise to an alternative, hierarchical or multi-level formulation of the model. Gelman and Hill (2006) and Longford (2005) contain extensive descriptions of hierarchical or multi-level models.

Boonstra et al (2007) study a number of alternative models for small area estimates, including models formulated on unit level versus area level, and linear versus logistic models. Based on research and simulation studies (Boonstra et al., 2007), the approximation using a linear mixed model at area level is initially selected because this retains the balance between simplicity and accuracy. This method is also implemented in a prototype software tool (Buelens, 2007).

In Rao (2003) the methods and formulas are described in more detail.

3.2 Applicability

The goal is to make estimates for small areas. By definition, direct estimates for small areas are unreliable. The model-based method is also interesting for calculating more accurate estimates for small areas. This is usually the case when estimates for small subgroups of the population must be calculated, and when no account was taken of this in the sample design.

The presence of random effects in the model equations (3.1.1) ensures that areas can differ from one another according to the model, apart from variations caused by differences in auxiliary variables. To make good estimates based on the model, it is important that good explanatory variables $\theta_{x;d}$ are available as auxiliary information. If the auxiliary information does not correlate well with the target variable, then the random effects will gain influence, and the model will have less predictive power. It is therefore explicitly expected that the areas are equivalent. If areas differ very strongly, and these differences are not accounted for by the auxiliary information, this results in greater random effects. The selection of good auxiliary variables boils down to selecting a suitable model; this process is known as model selection. This aspect is briefly discussed in sections 2.6 and 3.6.

For model (3.1.1), the auxiliary information only has to be available at area level. If the auxiliary information is available at unit level, then sample averages $\bar{x}_d$ can be used instead of population averages as auxiliary variables in (3.1.1). This allows for better correction for non-response; see Boonstra et al (2007).

3.3 Detailed description

3.3.1 Linear mixed model at area level

We use an FH model, a linear mixed model at area level, as defined in (3.1.1). The statement (3.1.1) can also be written as

$$\hat{\theta}_{y;d} = \theta_{x;d}^T \beta + v_d + \varepsilon_d . \qquad (3.3.1)$$

When fitting this model, we use direct, design-based estimates $\hat{\theta}_{y;d}$ and variance estimates ψ_d. The estimator $\hat{\theta}_{y;d}$ can be a Horvitz-Thompson or a regression estimator, for example.

Because we are dealing with small areas, the variance estimates ψ_d can be unstable. A solution is to pool these estimates. If $\hat{\theta}_{y;d}$ is the sample average area *d*, then the corresponding estimate of the design variance, assuming an unrestricted random sample, is

$$\psi_d = \frac{1}{n_d}(1-\frac{n_d}{N_d})s_d^2,$$

where the sample variance is

$$s_d^2 = \frac{1}{n_d - 1}\sum_{i\in S_d}(y_i - \hat{\theta}_{y;d})^2 ,$$

When pooling, we calculate the sample variance for areas together, or even all the areas together. In this last case, we calculate a pooled sample variance,

$$s_{pooled}^2 = \frac{1}{n-m}\sum_{d=1}^{m}(n_d - 1)s_d^2 ,$$

and use this instead of the area-specific variances s_d^2 in the above expression for ψ_d.

The model variance σ_v^2 is estimated using the Fay-Herriot moment estimator (Rao, 2003). As the starting point of this method, it is noted that

$$E\left(\sum_d \frac{(\hat{\theta}_{y;d} - \theta_{x;d}^T\tilde{\boldsymbol{\beta}})^2}{\psi_d + \sigma_v^2}\right) = E\left(h(\sigma_v^2)\right) = m - p \qquad (3.3.2)$$

where m is the number of small areas, p is the dimension of the vector of auxiliary variables $\theta_{x;d}$, and

$$\tilde{\beta} = \tilde{\beta}(\sigma_v^2) = \left(\sum_d \theta_{x;d}\theta_{x;d}^T/(\psi_d + \sigma_v^2)\right)^{-1}\left(\sum_d \theta_{x;d}\hat{\theta}_{y;d}/(\psi_d + \sigma_v^2)\right). \qquad (3.3.3)$$

The estimate $\hat{\sigma}_v^2$ is obtained by the iterative solving of

$$h(\sigma_v^2) = m - p . \qquad (3.3.4)$$

We use $\sigma_v^{2(a=0)} = 0$ as the starting value, and calculate

$$\sigma_v^{2(a+1)} = \sigma_v^{2(a)} + \frac{1}{h'_*(\sigma_v^{2(a)})}\left(m - p - h(\sigma_v^{2(a)})\right) \qquad (3.3.5)$$

where $h'_*(\sigma_v^2) = -\sum_d \frac{(\hat{\theta}_{y;d} - \theta_{x;d}^T\tilde{\beta})^2}{(\psi_d + \sigma_v^2)^2}$ is an approximation of the derivative of $h(\sigma_v^2)$.

This iterative process converges quickly, usually in less than ten iterations. If no positive solution is found, then $\hat{\sigma}_v^2 = 0$ is used. In this last case, there are no random effects and we will obtain a synthetic estimator at area level.

The bias and the variance of this estimate are indicated by B($\hat{\sigma}_v^2$) and V($\hat{\sigma}_v^2$) respectively, where

$$B(\hat{\sigma}_v^2) = \frac{2\left(m\sum_d (\psi_d + \hat{\sigma}_v^2)^{-2} - \left(\sum_d (\psi_d + \hat{\sigma}_v^2)^{-1} \right)^2 \right)}{\left(\sum_d (\psi_d + \hat{\sigma}_v^2)^{-1} \right)^3}, \tag{3.3.6}$$

$$V(\hat{\sigma}_v^2) = 2m\left(\sum_d (\psi_d + \hat{\sigma}_v^2)^{-1} \right)^{-2}. \tag{3.3.7}$$

3.3.2 Empirical Best Linear Unbiased Predictor (EBLUP)

The Empirical Best Linear Unbiased Predictor (EBLUP) estimator based on model (3.3.1) is indicated by

$$\hat{\theta}_{y;d}^{eblup} = \gamma_d \hat{\theta}_{y;d} + (1-\gamma_d)\theta_{x;d}^T \hat{\beta} \tag{3.3.8}$$

where

$$\hat{\beta} = \tilde{\beta}(\hat{\sigma}_v^2) = \left(\sum_d \gamma_d \theta_{x;d} \theta_{x;d}^T \right)^{-1} \left(\sum_d \gamma_d \theta_{x;d} \hat{\theta}_{y;d} \right) \tag{3.3.9}$$

and

$$\gamma_d = \frac{\hat{\sigma}_v^2}{\psi_d + \hat{\sigma}_v^2}. \tag{3.3.10}$$

The EBLUP estimator (3.3.8) is a weighted combination of a direct estimator $\hat{\theta}_{y;d}$ and a synthetic estimator $\theta_{x;d}^T \hat{\beta}$. The direct estimator is given a large weight γ_d (3.3.10) if the variance ψ_d is small. In other words, the EBLUP estimates are mainly based on the direct estimates when these are accurate, and on the model-based estimates in the other case.

The mean square error (MSE) of the EBLUP estimates (3.3.8) are estimated as

$$mse(\hat{\theta}_{y;d}^{eblup}) = g_{1d}(\hat{\sigma}_v^2) - B(\hat{\sigma}_v^2)(1-\gamma_d)^2 + g_{2d}(\hat{\sigma}_v^2) + 2g_{3d}(\hat{\sigma}_v^2), \tag{3.3.11}$$

where

$$g_{1d}(\hat{\sigma}_v^2) = \gamma_d \psi_d,$$

$$g_{2d}(\hat{\sigma}_v^2) = (1-\gamma_d)^2 \theta_{x;d}^t \left(\sum_d \theta_{x;d} \theta_{x;d}^T / (\psi_d + \hat{\sigma}_v^2) \right)^{-1} \theta_{x;d},$$

$$g_{3d}(\hat{\sigma}_v^2) = \psi_d{}^2(\psi_d + \hat{\sigma}_v^2)^{-3} V(\hat{\sigma}_v^2).$$

The first term, g_{1d}, is the inherent prediction variance that would also be present if β and σ_v^2 were known; g_{2d} and g_{3d} are the contributions from the uncertainty in β and σ_v^2 respectively.

3.3.3 Prototype software SmallAreaEstimator

The model described in section 3.3.1 and the EBLUP estimator described in section 3.3.2 are implemented in a prototype software tool, the *SmallAreaEstimator* (Buelens, 2007). This tool is a plug-in for SPSS and offers the user a graphical user interface that allows small area estimates to be made in the SPSS software environment.

3.4 Example

An example from the Dutch Labour Force Survey (*Enquête-Beroepsbevolking* – EBB) is the estimation of the annual employment figures at municipal level. There is a demand for these figures, but the design of the existing EBB sample does not allow reliable estimates to be made at this level. For many municipalities there are insufficient observations, or for some municipalities even no observations at all.

As an example, we use the CAPI sample of the EBB from 2005. This concerns 86,589 people, and 454 municipalities. We use the *SmallAreaEstimator (SAE)* software tool referred to in section 3.3.3. The direct estimator implemented in the current version of the software is the unweighted sample mean. With SAE, we calculate these direct estimates $\hat{\theta}_{y;d}$ and the accompanying variances ψ_d. The variation coefficient *vc* = $\sqrt{\psi_d} / \hat{\theta}_{y;d}$ can be used as a measure for acceptable accuracy. If we use the maximum value for the *vc* of 0.2 as a criterion, then the direct estimates for only 38 municipalities are sufficiently accurate.

As auxiliary information for the model-based estimates, we use the number of people registered with the Centre for work and income (*Centrum voor Werk en Inkomen*) in each municipality, and the population sizes are split into three age groups. Using SAE, we calculate the EBLUP estimates and the corresponding MSEs. The EBLUP estimates are sufficiently accurate for 437 municipalities (they have a *vc* smaller than than 0.2).

This example shows that higher accuracy can be achieved by using the EBLUP estimator instead of a simple direct design-based estimator.

3.5 Characteristics

Model-based small area estimates have a smoothing effect. The distribution of the estimates will have a smaller spread than the distribution of the actual values. Consequently, high extreme values will often be underestimated, and low extreme

values will be overestimated. It is therefore possible that small area estimates that are generally assumed to be good will not be good at all for specific individual areas. This situation will arise, for example, in atypical areas: areas that, for whatever reason, also differ essentially from all other areas and for which these differences are not explained by the auxiliary variables used.

The EBLUP estimator (3.3.8) is a combination of a direct and a synthetic estimator. Asymptotically, the EBLUP is design unbiased because (3.3.8) for large n_d approaches the direct estimator, which is design unbiased.

For areas that are well represented in the sample, the direct estimates will be accurate, and the weights γ_d (3.3.10) large. In other words, the EBLUP estimator for these areas is mainly, and sometimes almost completely, based on the direct estimator. It is therefore also important to select a good direct estimator, also because the direct estimates are used to fit the model. The prototype described in section 3.3.3 currently still uses the unweighted sample mean as the direct estimator, and therefore does not offer alternative options. It is expected that subsequent versions of this prototype will offer better options, by including weights. In this way, regression estimators will also be able to be used as direct estimators in the software, for example, using weights calculated in Bascula.

3.6 Quality indicators

Initially, one can examine the standard errors and the variation coefficients of the model-based estimates compared to those of the direct estimates. However, the standard errors of the model-based estimates are also estimates themselves based on the model, and thus must be dealt with carefully. Thanks to the random effects in the model and the asymptotic unbiasedness of the design of the EBLUP estimator, they are however reasonably robust.

There are a number of other measures and tests that can indicate whether the selected model and the accompanying estimates are plausible; see section 2.6. The study into these measures and tests for model selection is still being conducted at DMH.

4. References

Boonstra, H.J. (2005), *Model-based estimation of a finite population total: a Bayesian approach*. Statistics Netherlands, Heerlen.

Boonstra, H.J., Buelens, B. and Smeets, M. (2007), *Estimation of municipal unemployment fractions – a simulation study comparing different small area estimators.* Statistics Netherlands, Heerlen.

Buelens, B. (2007), *Methodologie van de kleinedomeinschattingen in het prototype 'SmallAreaEstimator'*. Internal report, Statistics Netherlands, Heerlen.

Fay, R.E. and Herriot, R.A. (1979), Estimation of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association* 74, 268-277.

Gelman, A. and Hill, J. (2006), *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge USA.

Ghosh, M. and Meeden, G. (1997), *Bayesian Methods for Finite Population Sampling*. Chapman & Hall.

Goodall, C. (1983), Examining Residuals. In: *Understanding Robust and Exploratory Data Analysis*, D.C. Hoaglin, F. Mosteller and J.W. Tukey (editors), Wiley, New York.

Greene, W.H. (1997), *Econometric Analysis*. Prentice Hall.

Hastie, T., Tibshirani, R., and Friedman, J.H. (2003), *The Elements of Statistical Learning*. Springer-Verlag.

Hedlin, D., Falvey, H., Chambers, R., and Kokic, P. (2001), Does the Model Matter for GREG Estimation? A Business Survey Example. *Journal of Official Statistics* 17 (4), 527-544.

Longford, N. (2005), *Missing Data and Small-Area Estimation: Modern Analytical Equipment for the Survey Statistician*. Springer-Verlag.

Nieuwenbroek, N. and H.J. Boonstra (2002), *Bascula 4.0 Reference Manual*. Statistics Netherlands, Heerlen.

Rao, J.N.K. (2003), *Small Area Estimation*. Wiley, New York.

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*. Springer-Verlag.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000), *Finite Population Sampling and Inference – A Prediction Approach*. Wiley, New York.

Version history

Version	Date	Description	Authors	Reviewers
Dutch version: Modelmatig schatten / Synthetische schatters en Kleinedomeinschatters				
1.0	18-12-2007	First Dutch version	Harm Jan Boonstra Bart Buelens	Vincent de Heij
1.1	23-01-2008	Minor modifications to layout	Harm Jan Boonstra Bart Buelens	
English version: Model-based estimation / Synthetic estimators and Small area estimators				
1.1E	17-02-2011	First English version	Harm Jan Boonstra Bart Buelens	