6

Model-Assisted Estimation for Domains

In this chapter, we examine the estimation for population subgroups or domains. Regional areas constructed by administrative criteria, such as county or municipality, are typical domains or *domains of interest.* The population also can be grouped into domains by demographic criteria, such as sex and age group, as in a social survey. In a business survey, enterprises are often grouped into domains according to the type of industry. Further, elements can be assigned into domains by demographic criteria within regional areas. In all these instances, *estimation for domains*, or *domain estimation*, refers to the estimation of population quantities, such as totals, for the desired population subgroups. Estimation of domain totals will be discussed in the context of design-based estimation, which is the main approach of the book. In practice, design-based estimation is mainly used for domains whose sample size is reasonably large. For small domains (with a small sample size in a domain), methods falling under the headline of *small area estimation* are often used. In Section 6.1, we outline the framework and basic principles of domain estimation. We also summarize the operational steps of a domain estimation procedure. Section 6.2 introduces two important concepts, estimator type and model choice, in the context of domain estimation. Selected estimators and models are worked out and illustrated in Section 6.3. Section 6.4 includes an empirical examination of properties of some estimators of domain totals based on Monte Carlo experiments. Summary and further reading is in Section 6.5.

6.1 FRAMEWORK FOR DOMAIN ESTIMATION

We focus on the estimation of population totals for domains in a descriptive survey. The estimation of domain totals is discussed from a design-based perspective, with the use of auxiliary information. According to Särndal *et al.* (1992), the framework

Practical Methods for Design and Analysis of Complex Surveys Risto Lehtonen and Erkki Pahkinen
 ISBN: 0-470-84769-7

is called *model-assisted*. The reason for incorporating auxiliary data in a domain estimation procedure is obvious: with strong auxiliary data it is possible to obtain better accuracy for domain estimates, when compared to an estimation procedure not using auxiliary data. Thus, this chapter extends the treatment of model-assisted estimation introduced in Section 3.3.

Different types of auxiliary data can be used in model-assisted estimation. In Section 3.3, we used population-level aggregates of auxiliary variables. Here, we also employ unit-level auxiliary data for model-assisted estimation for domains. These data are incorporated in a domain estimation procedure by unit-level statistical models. This is possible if we make the following technical assumptions: (1) register data (such as population census register, business register, different administrative registers) are available as frame populations and sources of auxiliary data, (2) registers contain unique identification keys that can be used in merging at micro-level data from registers and sample surveys (see Figure 1.1 in Chapter 1). Obviously, access to micro-merged register and survey data involves much flexibility for a domain estimation procedure. This view has been adopted, for example, in Särndal (2001) and Lehtonen *et al.* (2003). Much of the material of this chapter are based on these sources.

The methods specific to small-area estimation include a variety of model-dependent techniques such as synthetic (SYN) estimators, composite estimators, EBLUP (empirical best linear predictor) estimators and various Bayesian techniques, and techniques developed in the context of demography and disease mapping. The monograph by J.N.K. Rao (2003) provides a comprehensive treatment of model-dependent small-area estimation and discusses design-based methodologies for the estimation for domains as well. Other materials include, for example, Schaible (1996), Lawson *et al.* (1999), and Ghosh (2001), who discusses especially empirical and hierarchical Bayes techniques.

Basic Principles

Let us introduce our basic notation for population quantities and sample-specific quantities in the context of domain estimation. The finite population is again denoted by $U = \{1, 2, \ldots, k, \ldots, N\}$ and, in domain estimation, we consider a set of mutually exhaustive subgroups of the population denoted $U_1, \ldots, U_d, \ldots, U_D$ (note that in this chapter we use exclusively a subscript d for domains of interest). We assume that the population U can be used as a sampling frame. This implies that U is available as a computerized data set, for example, a population register, or a register of business firms. We therefore also assume that the frame population U contains (in addition to the 'labels' k of the population elements) values for certain additional variables for all elements $k \in U$ (where the symbol '$\in$' refers to the inclusion of an element in a set of elements). These variables are unique element-identification (ID) keys, domain membership indicators, stratum membership indicators and the auxiliary z-variables.

Denote by y the variable of interest and by Y_k its unknown population value for unit k. The target parameters are the set of domain totals, $T_d = \sum_{k \in U_d} Y_k$, $d = 1, \ldots, D$, where summation is over all population elements k belonging to domain U_d (for simplicity, we use this notation throughout this chapter). Auxiliary information is essential for building accurate domain estimators, and increasingly so when the sample size of domains get smaller. Let $\mathbf{z}_k = (z_{1k}, \ldots, z_{jk}, \ldots, z_{Jk})'$ be the auxiliary variable vector of dimension $J \geq 1$. The value $\mathbf{z}_k$ is assumed to be known for every element $k \in U$. In a survey on individuals, $\mathbf{z}_k$ may specify known data about a person k, such as age, sex, taxable income and other continuous or qualitative variable values. In a business survey, $\mathbf{z}_k$ may indicate the turnover, or the total number of staff, for business firm k. It is important to emphasize that we assume the auxiliary z-data to be at the micro-level, that is, a value is assigned for each population element in the frame register. This is for flexibility, because the data can be then aggregated at higher levels of the population, such as at the domain or stratum level, if desired. Indeed, for some estimators, it suffices to know the population totals $T_{dz_1}, \ldots, T_{dz_j}$ of the auxiliary variables z_j for each domain of interest. In the model-fitting phase, we often assume that a constant value 1 is assigned as the first element in a vector $\mathbf{z}_k$.

For unique identification of domain membership for each population element, we define $\boldsymbol{\delta}_k = (\delta_{1k}, \ldots, \delta_{dk}, \ldots, \delta_{Dk})'$ to be the *domain indicator vector* for unit k, such that $\delta_{dk} = 1$ for all elements $k \in U_d$, and $\delta_{dk} = 0$ for all elements $k \notin U_d$, $d = 1, \ldots, D$. An indicator vector $\boldsymbol{\tau}_k$ for *stratum identification* for population element k is constructed in a similar manner: $\tau_{hk} = 1$ for all $k \in U_h$, $h = 1, \ldots, H$, and $\tau_{hk} = 0$ otherwise, where U_h refers to stratum h and H is the number of strata. Thus, a total of D domain indicator variables and H stratum indicator variables are assumed in the population frame.

A probability sample s of size n is drawn from U using a sampling design $p(s)$ such that an inclusion probability π_k is assigned to unit k. The corresponding sampling weights are $w_k = 1/\pi_k$. Measurements y_k of the response variable y are obtained for the sampled elements $k \in s$. We assume that a unique element ID key is included in sample s making it possible to micro-merge these data with the frame register U.

The domain samples are $s_d = U_d \cap s$, $d = 1, \ldots, D$. A domain is defined *unplanned*, if the domain sample size n_{s_d} is not fixed in the sampling design. This is the case in which the desired domain structure is not a part of the sampling design. Thus, the *domain sample sizes are random quantities* introducing an increase in the variance estimates of domain estimators. In addition, an extremely small number (even zero) of sample elements in a domain can be realized in this case, if the domain size in the population is small. For *planned* domains on the other hand, the *domain sample sizes are fixed in advance by stratification*. Stratified sampling in connection with a suitable allocation scheme is often used in practical applications.

A certain domain structure for a stratified sample of n elements can be illustrated, for example, as in Table 6.1. In the table setting, an unplanned domain structure

Table 6.1 Planned and unplanned domain structures in a stratified sample of n elements.

Unplanned domains	Strata (planned domains)						
	1	2	...	h	...	H	Sum
1	$n_{s_{11}}$	$n_{s_{12}}$	...	$n_{s_{1h}}$	...	$n_{s_{1H}}$	n_{s_1}
2	$n_{s_{21}}$	$n_{s_{22}}$	...	$n_{s_{2h}}$	...	$n_{s_{2H}}$	n_{s_2}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
d	$n_{s_{d1}}$	$n_{s_{d2}}$	⋯	$n_{s_{dh}}$	⋯	$n_{s_{dH}}$	n_{s_d}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
D	$n_{s_{D1}}$	$n_{s_{D2}}$	...	$n_{s_{Dh}}$	...	$n_{s_{DH}}$	n_{s_D}
Sum	n_1	n_2	...	n_h	...	n_H	n

Sample sizes n_{s_d}, $d = 1, \ldots, D$, for *unplanned* domains are not fixed in advance and thus are random variables.
Stratum sample sizes n_h, $h = 1, \ldots, H$ are fixed in the sampling design. Thus, the strata are defined as *planned* domains.
Cell sample sizes $n_{s_{dh}}$ are random variables in both cases.

cuts across the strata, a situation that is common in practice. In other types of structures, strata and domains can be nested such that a stratum contains several unplanned domains (for example, regional sub-areas within larger areas) or the strata themselves constitute the domains. The latter case represents a planned domain structure. Singh *et al.* (1994) illustrates the benefits of the planned domain approach for domain estimation. They presented compromise sample allocation schemes for the Canadian labour force survey to satisfy reliability requirements at the provincial level as well as at sub-provincial level. However, for practical reasons, it is usually not possible to define all desired domain structures as strata.

For the estimation for domains, it is advisable to apply the planned domains approach when possible, by defining the most important domains of interest as strata and to use a suitable allocation scheme in the sampling design, such as power or Bankier allocation (see the next example). It is also beneficial to use a large overall sample size to avoid small expected domain sample sizes if an unplanned domain approach is used. And in the estimation phase, it is often useful to incorporate strong auxiliary data into the estimation procedure by carefully chosen models and estimators of domain totals (see Example 6.2 and Section 6.4).

Example 6.1

Impact of sampling design in estimation for domains: the cases of unplanned and planned domain structures. Problems may be encountered when working with an unplanned domain structure, because small domain samples can be obtained

for domains with a small population size, if the overall sample size is not large, involving imprecise estimation. For example, if the sample has been drawn with simple random sampling without replacement, then the expected sample size in a domain would be $E(n_{s_d}) = n \times (N_d/N)$, thus corresponding to the proportional allocation in stratified sampling. An alternative is based on the *planned domain* structure, where the domains are defined as strata. Then, more appropriate allocation schemes can be used. In this example, the allocation scheme is based on *power allocation* (see Section 3.1). In power or Bankier allocation, the sample is allocated to the domains on the basis of information on the coefficient of variation of the response variable y in the domains and on the possibly known domain totals T_{dz} of an auxiliary variable z. We use a simplified version of power allocation in a hypothetical situation in which the coefficients of variation $\mathrm{C.V}_{dy} = S_{dy}/\overline{Y}_d$ of the response variable y are known in all domains, where S_{dy} and $\overline{Y}_d$ are the population standard deviation and the population mean of y in domain d, respectively.

In power allocation, the domain sample sizes are given by

$$n_{d,pow} = n \times \frac{T_{dz}^a \times \mathrm{C.V}_{dy}}{\sum_{d=1}^{D} T_{dz}^a \times \mathrm{C.V}_{dy}},$$

where the coefficient a refers to the desired power (typical choices are 0, 0.5 or 1). Here we have chosen $a = 0$ for simplicity. Thus, information on coefficients of variation is only used.

We illustrate the methodology by selecting an SRSWOR sample ($n = 392$ persons) from the Occupational Health Care Survey (OHC) data set ($N = 7841$ persons) and estimating the total number of chronically ill persons in the $D = 30$ domains constructed. In the population, the sizes of the domains vary with a minimum of 81 persons and a maximum of 517 persons. The results for the allocation of the sample by proportional allocation (corresponding to an unplanned domain structure) and by power allocation (corresponding to a planned domain structure) are shown in Table 6.2. The domain totals of the number of chronically ill persons are estimated by a Horvitz–Thompson (HT) estimator $\hat{t}_{dHT} = \sum_{k \in s_d} w_k y_k$. The stability of the estimators is measured by the population coefficient of variation of an estimator of a domain total, given by $\mathrm{C.V}(\hat{t}_{dHT}) = \mathrm{S.E}\,(\hat{t}_{dHT})/T_d$.

The results show that SRSWOR sampling produces a large variation in the expected domain sample size: the average domain sample size is 13, the minimum sample size is 4 and the maximum is 26. On the other hand, power allocation smoothes considerably the variation in domain sample size: the minimum domain sample size is now 10 and the maximum is 17. The percentage coefficient of variation varies much in the case of SRSWOR. For example, the difference between the smallest and largest coefficient of variation is over 60% points. In power

Table 6.2 Allocation schemes for a sample of $n = 392$ elements for $D = 30$ domains of the OHC Survey data set. Calculation of the expected domain sample size $E(n_{s_d})$ under an SRSWOR design and realized domain sample size n_d under a stratified SRSWOR design with power allocation ($a = 0$), and the corresponding coefficients of variation (%) of a Horvitz–Thompson estimator $\hat{t}_{dHT}$.

		Domain sample size		Coefficient of variation C.V (%) of HT estimators of domain totals	
Domain		**Unplanned domain structure** Expected under SRSWOR	**Planned domain structure** Realized under stratified SRSWOR (power allocation)	**Unplanned domain structure** SRSWOR	**Planned domain structure** Stratified SRSWOR (power allocation)
d	N_d	$E(n_{s_d})$	n_d	C.V($\hat{t}_{dHT}$)	C.V($\hat{t}_{dHT}$)
10	81	4	11	84.10	38.88
20	101	5	12	78.41	40.54
18	129	6	13	72.69	42.38
3	133	7	15	81.04	45.63
8	141	7	16	81.03	46.54
30	146	7	15	74.80	45.03
21	153	8	12	62.87	41.15
23	156	8	11	57.65	39.05
16	165	8	13	64.94	43.19
1	181	9	17	75.90	48.78
11	187	9	14	63.52	44.52
6	188	9	13	60.37	43.22
28	194	10	10	50.52	38.69
24	200	10	13	58.68	43.39
22	242	12	10	44.27	38.30
15	252	13	14	55.68	45.50
7	292	15	17	60.34	50.06
4	295	15	15	53.92	47.04
13	305	15	13	46.00	43.04
12	311	16	12	44.50	42.38
5	323	16	16	53.50	48.23
25	339	17	11	40.57	41.03
2	352	18	14	46.80	45.74
26	364	18	11	38.87	40.88
29	365	18	11	38.25	40.45
9	366	18	14	45.99	45.85
17	426	21	12	36.67	41.62
14	447	22	13	37.95	43.37
19	490	24	11	33.60	41.22
27	517	26	10	30.68	39.34
Sum	7841	392	392		

allocation, the difference is reduced to 12% points. Thus, power allocation tends to smooth the variation in the coefficient of variation such that large coefficients are considerably decreased. However, the coefficients of variation of estimated domain totals tend to be quite large; this is mainly due to the small overall sample size.

The progression in coefficients of variation can be illustrated graphically. In Figure 6.1, the coefficients of variation have been plotted against domain size in population. The curve for the HT estimator obtained for coefficients of variation under SRSWOR shows clear decrease with increasing domain size. For power allocation, the curve is clearly stabilized.

To continue the specification of the setting for domain estimation, our further technical assumption is as follows. We assume that after data collection from the selected sample and preparation of the final sample data set, denoted by $s(y)$, the population frame U and the sample measurements $s(y)$ can be micro-merged using the unique element ID keys that are available in both data sources. Completing this procedure we have obtained an enhanced frame register data set that includes the auxiliary z-data and stratum and domain indicator variables for all population elements, amended with y-measurements for the elements belonging to the sample.

We have now completed the technical preparations for conducting an estimation for the domains. The operational steps in a domain estimation procedure, given in general terms, are summarized in Box 6.1.

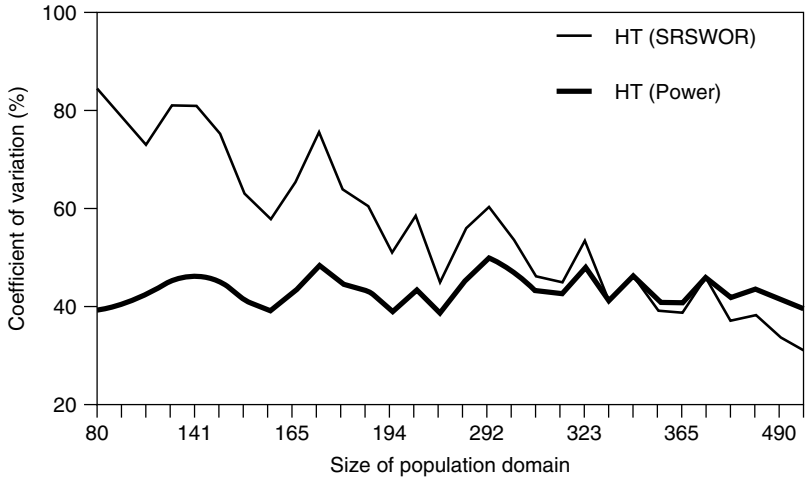

Figure 6.1 Coefficient of variation (%) of Horvitz–Thompson estimator of domain total under SRSWOR sampling (corresponding to the unplanned domain structure) and stratified SRSWOR sampling with power allocation ($a = 0$) (corresponding to the planned domain structure).

BOX 6.1 ***Operational steps in a domain estimation procedure***

Step 1: Construction of frame population Construction of the frame population $U = \{1, 2, \ldots, k, \ldots, N\}$ of N elements containing unique element ID keys, domain indicator vectors $\boldsymbol{\delta}_k$, stratum indicator vectors $\boldsymbol{\tau}_k$, inclusion probabilities π_k for drawing of an n element sample with sampling design $p(s)$, and the vectors $\mathbf{z}_k$ of auxiliary z-data, for all elements k in U.

Step 2: Sampling and measurement Sample selection by using the design $p(s)$ and measurement of the values of the response variable y, and the construction of the sample data set $s(y)$, including the element ID keys, observed values y_k and sampling weights $w_k = 1/\pi_k$, for all elements $k \in s$.

Step 3: Frame population revisited Construction of a combined data set by micro-merging the frame population U and the sample data set $s(y)$ by using the element ID keys.

Step 4: Model choice and model fitting The choice of the model, specification of model parameters and effects, model fitting using the sample data set and model validation and diagnostics. On the basis of the fitted model, calculation of fitted values $\hat{y}_k$ for all population elements $k \in U$ and residuals $\hat{e}_k = y_k - \hat{y}_k$ for all elements $k \in s\,(y)$, the sample data set.

Step 5: Choice of estimator of domain totals and estimation for domains Supply of fitted values, residuals and weights in the chosen estimator for domain totals. Basically, estimators of domain totals labeled 'model-dependent' use the fitted values $\hat{y}_k$, $k \in U$, and the estimators of domain totals labeled 'model-assisted' use the fitted values $\hat{y}_k$, $k \in U$, and in addition, the residuals $\hat{e}_k$ and the weights w_k, $k \in s$.

Step 6: Variance estimation and diagnostics Choice of an appropriate variance estimator. Calculation of standard error estimates and coefficients of variation.

In Table 6.3, we summarize in a hypothetical situation, the progression in the population frame data set that occurs when the operations in Steps 1 to 4 of Box 6.1 are implemented for a domain estimation procedure. Because the vectors $\mathbf{z}_k = (z_{1k}, \ldots, z_{Jk})'$ of auxiliary z-variables are assumed to be known for every population element, including sampled and nonsampled elements, the vector $\mathbf{T}_z = (T_{z_1}, \ldots, T_{z_J})'$ with $T_{z_j} = \sum_{k \in U} z_{jk}$, $j = 1, \ldots, J$, of population totals of auxiliary z-variables is known. Also, domain totals $T_{dz_j} = \sum_{k \in U_d} z_{jk}$, $d = 1, \ldots, D$ and $j = 1, \ldots, J$, can be calculated for each z-variable, because the domain indicators are assumed to be known for all $k \in U$. The sample membership

Table 6.3 Execution of Steps 1, 3 and 4 of Box 6.1 in a domain estimation procedure (hypothetical situation).

	Step 1: Construction of the frame population U				**Step 3**: Merging of the frame population U and the sample data set $s(y)$			**Step 4**: Calculation of fitted y-values and residuals	
Element ID	Domain ID vectors $\boldsymbol{\delta}'_k$	Stratum ID vectors $\boldsymbol{\tau}'_k$	Inclusion probability π_k	Auxiliary z-vectors $\mathbf{z}'_k$	Sampling weight w_k	Sample membership indicator I_k	Study variable y_k	Fitted values $\hat{y}_k$	Residuals $\hat{e}_k$
1	$\boldsymbol{\delta}'_1$	$\boldsymbol{\tau}'_1$	π_1	$\mathbf{z}'_1$	0	0	...	$\hat{y}_1$	...
2	$\boldsymbol{\delta}'_2$	$\boldsymbol{\tau}'_2$	π_2	$\mathbf{z}'_2$	0	0	...	$\hat{y}_2$	...
3	$\boldsymbol{\delta}'_3$	$\boldsymbol{\tau}'_3$	π_3	$\mathbf{z}'_3$	w_3	1	y_3	$\hat{y}_3$	$\hat{e}_3$
4	$\boldsymbol{\delta}'_4$	$\boldsymbol{\tau}'_4$	π_4	$\mathbf{z}'_4$	0	0	...	$\hat{y}_4$	...
5	$\boldsymbol{\delta}'_5$	$\boldsymbol{\tau}'_5$	π_5	$\mathbf{z}'_5$	w_5	1	y_5	$\hat{y}_5$	$\hat{e}_5$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	$\boldsymbol{\delta}'_k$	$\boldsymbol{\tau}'_k$	π_k	$\mathbf{z}'_k$	w_k	1	y_k	$\hat{y}_k$	$\hat{e}_k$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
N	$\boldsymbol{\delta}'_N$	$\boldsymbol{\tau}'_N$	π_N	$\mathbf{z}'_N$	0	0	...	$\hat{y}_N$	...

... Nonsampled element.

indicator variable I is created for the whole population data set such that $I_k = 1$ if $k \in s$, zero otherwise. Obviously, the sum of the indicator variable over the population is n, the sample size. In the model-fitting phase, the fitted values $\hat{y}_k$ are calculated for all N elements $k \in U$. On the other hand, the residuals $\hat{e}_k = y_k - \hat{y}_k$ can be calculated for the sampled elements $k \in s$ only. It is also important to emphasize that the fitted values $\{\hat{y}_k; k \in U\}$ calculated by a given model differ from one model specification to another. This will be apparent in the next section in which models and estimators of domain totals are treated in more detail.

6.2 ESTIMATOR TYPE AND MODEL CHOICE

Important phases in a model-assisted domain estimation procedure are the selection of the type of the estimator of a total, the choice of the auxiliary variables to be used, the formulation of the model for the incorporation of the auxiliary data into the estimation procedure, the model-fitting phase and the derivation of variance estimators for the selected domain total estimators (see Box 6.1). In this section, we consider these phases in a more technical manner.

Estimator Type

We first discuss two concepts, estimator type and model choice, making the basis for the construction of an estimator of the population totals for domains of interest.

The concept *estimator type* refers to the explicit structure of the selected estimator of the domain totals. There are two main types of estimators discussed in this chapter. These are the *generalized regression (GREG) estimator* and the *synthetic (SYN) estimator*. The main conceptual difference in these estimators is that GREG estimators use models as assisting tools, whereas SYN estimators rely exclusively on the model used. Thus, GREG estimators are *model-assisted* and SYN estimators are *model-dependent*. The main consequence of this differing role of a model is that a GREG estimator of a domain total is constructed to be design unbiased (or approximately so) irrespective of the 'truth' of the model. This is a benefit of GREG estimators. However, a GREG estimator can be very unstable if the sample size in a domain becomes small. On the other hand, the bias of a SYN estimator depends heavily on a correct model specification. If the model is severely misspecified, a SYN estimator can involve substantial design bias. If, on the other hand, the model is correctly specified or nearly so, then the bias of a SYN estimator can be small.

In a typical large-scale survey conducted, for example, by a national statistical agency, some domains of interest are large enough, and the auxiliary information strong enough, so that the GREG-type estimators will be sufficiently precise. But for a small domain the variance of a GREG estimator can become unacceptably large, and in this case, the variance of a SYN estimator can be much smaller. Better precision of SYN estimators for small domains favours their use, in particular, for small-area estimation (recall that 'small area' refers to the situation in which the attained sample size in a given domain, or 'area', is small, or very small, even zero).

To summarize the main theoretical properties of the estimator types, GREG estimators are constructed to be design unbiased; the SYN estimators usually are not. Variance of the GREG estimator can be large for a small domain, that is, if the domain sample size is small, causing poor precision. The SYN estimator is usually design biased; its bias does not approach zero with increasing sample size; its variance is usually smaller than that of GREG; this holds especially for small domains. The accuracy, measured by the mean squared error MSE, of a SYN estimator can be poor even in the case of a small variance, if the bias is substantial.

Model Choice

The concept *model choice* refers to the specification of the relationship of the study variable y with the auxiliary predictor variables $z_1, \ldots, z_J$, as reflected by the structure of the constructed model. Model choice has two aspects, the *mathematical form* of the model and the *specification of the parameters and effects* in the model. For example, when working with a continuous study variable, a *linear model formulation* is usually appropriate. For binary or polytomous study variables, one might make a choice for a nonlinear model, such as a binomial or multinomial logistic model. For example, for a binary study variable, a *logistic model formulation* is arguably an improvement on a linear model type, because the fitted y-values

under the former will necessarily fall in to the unit interval, which is not always true for a linear model.

The second aspect of model choice is the specification of the parameters and effects in the model. Some of these may be defined at the fully aggregated population level, others at the level of the domain (domain-specific parameters), yet others at some intermediate level. We will separate a *fixed-effects model formulation* and a *mixed model formulation*. A fixed-effects model can involve population-level or domain-specific fixed effects, or effects specified on an intermediate level. In a mixed model, there are domain-specific *random effects* in addition to the fixed effects. Using a mixed model type, we can introduce stochastic effects that recognize domain differences.

To summarize, the chosen model specifies a hypothetical relationship between the variable of interest, y, and the predictor variables, $z_1, \ldots, z_J$, and makes assumptions about its perhaps complex error structure. Fixed-effects models can often be satisfactory, but mixed models offer additional possibilities for flexible modelling. For every specified model, we can derive one GREG estimator and one SYN estimator, by observing the respective construction principles. However, fixed-effects models have been more common in model-assisted estimators, whereas mixed models have most often been used in model-dependent estimators.

By combining these two aspects of an estimator for domain totals, estimator type and model choice, we get a two-dimensional arrangement of estimators. To illustrate this, we have included in Table 6.4 a number of selected estimators. There are six model-dependent SYN-type estimators and six design-based GREG-type estimators in the table. Each of the six rows corresponds to a different model choice. A *population model* (P-model; rows 1 and 2) is one whose only parameters are fixed effects defined at the population level; it contains no domain-specific parameters. A *domain model* (D-model) is one having at least some of its parameters or effects defined at the domain level. These are fixed effects for rows 3 and 4 and

Table 6.4 Classification of estimators for domain totals by model choice and estimator type.

Model choice			Estimator type	
Specification of model effects	Level of aggregation	Functional form	Model-dependent	Design-based model-assisted
Fixed-effects models	Population models	Linear	SYN-P	GREG-P
		Logistic	LSYN-P	LGREG-P
	Domain models	Linear	SYN-D	GREG-D
		Logistic	LSYN-D	LGREG-D
Mixed models including fixed and random effects	Domain models	Linear	MSYN-D	MGREG-D
		Logistic	MLSYN-D	MLGREG-D

random effects for rows 5 and 6. 'Linear' and 'logistic' refer to the mathematical forms. In Example 6.2 and Section 6.4, we will consider in more detail a number of these estimators.

6.3 CONSTRUCTION OF ESTIMATORS AND MODEL SPECIFICATION

Construction of Estimators of Domain Totals

The estimators of domain totals are constructed in the following three phases (according to Steps 4 and 5 in Box 6.1):

1. The parameters of the designated model are estimated using the sample data set $s(y) = \{(y_k, \mathbf{z}_k); k \in s\}$.
2. Using the estimates of the model parameters and the population vectors $\mathbf{z}_k$, the fitted value $\hat{y}_k$ is computed for every population element k, including elements belonging to the sample and also elements that are not sampled.
3. For obtaining an estimate $\hat{t}_d$ of the total T_d in domain d, the fitted values, $\{\hat{y}_k; k \in U\}$, and the sample observations, $\{y_k; k \in s\}$, are incorporated in the respective formulas for the GREG and SYN estimators.

We will illustrate the domain estimation procedure in the context of linear models. Consider a *fixed-effects linear model* specification such that $y_k = \mathbf{z'}_k\boldsymbol{\beta} + \varepsilon_k$, where $\boldsymbol{\beta}$ is an unknown parameter vector requiring estimation, and ε_k are the residual terms. The model fit yields the estimate $\hat{\boldsymbol{\beta}}$. The supply of fitted values given by $\hat{y}_k = \mathbf{z'}_k\hat{\boldsymbol{\beta}}$ is computed for all elements $k \in U$. Similarly, for a *linear mixed model* involving domain-specific random effects in addition to the fixed effects, the model specification is $y_k = \mathbf{z'}_k(\boldsymbol{\beta} + \mathbf{u}_d) + \varepsilon_k$, where $\mathbf{u}_d$ is a vector of *random effects* defined at the domain level. Using the estimated parameters, fitted values given by $\hat{y}_k = \mathbf{z'}_k(\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}_d)$ are computed for all $k \in U$. In more general terms, models used in the construction of GREG- and SYN-type estimators of domain totals are special cases of *generalized linear mixed models*, such as a mixed linear model and a logistic model (see e.g. McCulloch and Searle 2001; Dempster *et al.* 1981).

The fitted values $\{\hat{y}_k; k \in U\}$ differ from one model specification to another. For a given model specification, an estimator of domain total $T_d = \sum_{k \in U_d} y_k$ has the following structure for the two basic estimator types:

$$\text{Synthetic estimator:} \qquad \hat{t}_{dSYN} = \sum_{k \in U_d} \hat{y}_k \tag{6.1}$$

$$\text{Generalized regression estimator:} \qquad \hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} w_k(y_k - \hat{y}_k) \tag{6.2}$$

where $w_k = 1/\pi_k$, $s_d = s \cap U_d$ is the part of the full sample s that falls in to domain U_d, and $d = 1, \ldots, D$.

Note that $\hat{t}_{dSYN}$ uses the fitted values given by the estimated model, and thus relies on the 'truth' of the model and, therefore, can be biased. On the other hand, $\hat{t}_{dGREG}$ has a second term that aims at protecting against possible model misspecification. Note also that in the case in which there are no sample elements in a domain, $\hat{t}_{dGREG}$ reduces to $\hat{t}_{dSYN}$ for that domain. A Horvitz–Thompson estimator $\hat{t}_{dHT} = \sum_{k \in s_d} w_k y_k$ is often calculated as a reference when assessing the benefits from the more complex estimators.

Model Specification

Let us first discuss *fixed-effects linear models*. Let $\mathbf{z}_k = (1, z_{1k}, \ldots, z_{jk}, \ldots, z_{Jk})'$ be a $(J+1)$-dimensional vector containing the values of $J \geq 1$ predictor variables z_j, $j = 1, \ldots, J$. This vector is used to create the predicted values $\hat{y}_k$, $k \in U$, in the estimators (6.1) and (6.2).

1. *Fixed-effects P-models.* The estimators SYN-P and GREG-P build on the model specification

$$y_k = \beta_0 + \beta_1 z_{1k} + \cdots + \beta_J z_{Jk} + \varepsilon_k = \mathbf{z}'_k \boldsymbol{\beta} + \varepsilon_k \tag{6.3}$$

for $k \in U$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_J)'$ is a vector of fixed effects defined for the whole population. Owing to this property, we call (6.3) the *fixed-effects P-model*. If y-data were observed for the whole population, we could compute the generalized least-squares (GLS) estimator of $\boldsymbol{\beta}$ given by

$$\mathbf{B} = \left(\sum_{k \in U} \mathbf{z}_k \mathbf{z}'_k / c_k \right)^{-1} \sum_{k \in U} \mathbf{z}_k y_k / c_k, \tag{6.4}$$

where the c_k are specified positive weights. With no significant loss of generality, we specify these to be of the form $c_k = \boldsymbol{\lambda}' \mathbf{z}_k$ for $k \in U$, where the $(J+1)$-vector $\boldsymbol{\lambda}$ does not depend on k. As a further simple specification, we can set $c_k = 1$ for all k, and (6.4) reduces to an ordinary least-squares (OLS) estimator. In practice, a weighted least-squares (WLS) estimate for (6.4) is calculated on the observed sample data, yielding

$$\hat{\mathbf{b}} = \left(\sum_{k \in s} w_k \mathbf{z}_k \mathbf{z}'_k \right)^{-1} \sum_{k \in s} w_k \mathbf{z}_k y_k, \tag{6.5}$$

where $w_k = 1/\pi_k$ is the sampling weight of unit k. The resulting predicted values are given by

$$\hat{y}_k = \mathbf{z}'_k \hat{\mathbf{b}}, \quad k \in U. \tag{6.6}$$

By incorporating predicted values $\hat{y}_k$ into (6.1) and (6.2), we obtain the corresponding SYN-P and GREG-P estimators. Note that using a P-model for a given domain d, y-values from other domains also contribute to the predicted values incorporated in an estimator SYN-P and GREG-P for that domain. For this reason, the estimators $\hat{t}_{dSYN-P}$ and $\hat{t}_{dGREG-P}$, using a fixed-effects P-model type, are called *indirect* estimators.

2. *Fixed-effects D-models.* The estimators SYN-D and GREG-D are built with the same predictor vector $\mathbf{z}_k$, but with a different model specification allowing a fixed-effects vector $\boldsymbol{\beta}_d$ separately for every domain, so that

$$y_k = \mathbf{z}'_k\boldsymbol{\beta}_d + \varepsilon_k \tag{6.7}$$

for $k \in U_d$, $d = 1, \ldots, D$, or equivalently,

$$y_k = \sum_{d=1}^{D} \delta_{dk}\mathbf{z}'_k\boldsymbol{\beta}_d + \varepsilon_k \tag{6.8}$$

for $k \in U$, where δ_{dk} is the domain indicator of unit k, defined by $\delta_{dk} = 1$ for all $k \in U_d$, and $\delta_{dk} = 0$ for all $k \notin U_d$, $d = 1, \ldots, D$. Model (6.7) is called the *fixed-effects D-model*. Again, if the model (6.7) could be fitted to the data for the whole subpopulation U_d, the GLS estimator of β_d would be

$$\mathbf{B}_d = \left(\sum_{k\in U_d} \mathbf{z}_k\mathbf{z}'_k/c_k\right)^{-1} \sum_{k\in U_d} \mathbf{z}_k y_k/c_k, \quad d = 1, \ldots, D. \tag{6.9}$$

In practice, the fit must be based on the observed sample data in domain d. Setting again $c_k = 1$ for all k, the following WLS estimator can be used:

$$\hat{\mathbf{b}}_d = \left(\sum_{k\in s_d} w_k\mathbf{z}_k\mathbf{z}'_k\right)^{-1} \sum_{k\in s_d} w_k\mathbf{z}_k y_k, \quad d = 1, \ldots, D. \tag{6.10}$$

The resulting predicted values are given by

$$\hat{y}_k = \mathbf{z}'_k\hat{\mathbf{b}}_d \tag{6.11}$$

for $k \in U_d$; $d = 1, \ldots, D$. By incorporating predicted values $\hat{y}_k$ from (6.11) into (6.1) and (6.2), we obtain the corresponding SYN-D and GREG-D estimators. For a given domain d, y-values are used from that domain only in the model fitting and in the calculation of the predicted values incorporated in an estimator SYN-D and GREG-D in that domain. Thus, the estimators $\hat{t}_{dSYN-D}$ and $\hat{t}_{dGREG-D}$, using a fixed-effects D-model type, are called *direct* estimators. Note that because of the

specification $c_k = \boldsymbol{\lambda}'\mathbf{z}_k = 1$, we have $\sum_{k \in s_d} w_k(y_k - \hat{y}_k) = 0$. Consequently, SYN-D and GREG-D are identical, that is, $\hat{t}_{dSYN-P} = \hat{t}_{dGREG-P}$ for every sample s, when using the fixed-effects D-model specification.

3. *Mixed D-models.* The estimators MSYN-D and MGREG-D build on a two-level linear model, called the *mixed linear D-model*, involving fixed as well as random effects recognizing domain differences,

$$y_k = \beta_0 + u_{0d} + (\beta_1 + u_{1d})z_{1k} + \cdots + (\beta_J + u_{Jd})z_{Jk} + \varepsilon_k = \mathbf{z}'_k(\boldsymbol{\beta} + \mathbf{u}_d) + \varepsilon_k \tag{6.12}$$

for $k \in U_d$, $d = 1, \ldots, D$. Each coefficient is the sum of a fixed component and a domain-specific random component: $\beta_0 + u_{0d}$ for the intercept and $\beta_j + u_{jd}$, $j = 1, \ldots, J$ for the slopes. The components of $\mathbf{u}_d = (u_{0d}, u_{1d}, \ldots, u_{Jd})'$ represent deviations from the coefficients of the fixed-effects part of the model,

$$y_k = \beta_0 + \beta_1 z_{1k} + \cdots . + \beta_J z_{Jk} + \varepsilon_k = \mathbf{z}'_k\boldsymbol{\beta} + \varepsilon_k, \tag{6.13}$$

which agrees with (6.3). More generally, we can have that only some of the coefficients in (6.12) are treated as random, so that, for some j, $u_{jd} = 0$ for every domain d. A simple special case of (6.12), commonly used in practice, is the one that includes domain-specific random intercepts u_{0d} as the only random terms, given by $y_k = \beta_0 + u_{0d} + \beta_1 z_{1k} + \cdots + \beta_J z_{Jk} + \varepsilon_k$. We insert the resulting fitted y-values

$$\hat{y}_k = \mathbf{z}'_k(\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}_d) \tag{6.14}$$

into (6.1) to obtain the two-level MSYN-D estimator. Inserting the fitted values (6.14) into (6.2), we obtain the two-level MGREG-D estimator, introduced by Lehtonen and Veijanen (1999). A two-level D-model (6.12) can be fitted, for example, by estimating the variance components by maximum likelihood (ML) or restricted maximum likelihood (REML) and the fixed effects by GLS given these variance estimates; for details see, for example, Goldstein (2002) and McCulloch and Searle (2001). In estimating a mixed D-model, an assumption is usually made that the random effects follow a joint normal distribution. Note, however, that the assumption of normality is not necessary to obtain approximate unbiasedness for the resulting MGREG-D estimator.

Alternative options are available for the estimation of the design variance for estimators (6.1) and (6.2) of domain totals. When working with *planned domains*, where the domain sample sizes n_d are fixed in the stratified sampling design and, for example, the samples are drawn with SRSWOR in each stratum, approximate variance estimators presented in Section 3.3 for regression estimation can be used separately for each domain. In this setting, a sample of n_d elements is drawn from the population of N_d elements in domain d, and the weights are $w_k = N_d/n_d$ for

all $k \in U_d$. For example, for the GREG estimator (6.2), an approximate variance estimator is given by

$$\hat{v}_{srs}(\hat{t}_{dGREG}) = N_d^2 \left(1 - \frac{n_d}{N_d}\right) \left(\frac{1}{n_d}\right) \sum_{k \in s_d} \frac{(\hat{e}_k - \bar{\hat{e}}_d)^2}{n_d - 1}, \tag{6.15}$$

where the residuals are $\hat{e}_k = y_k - \hat{y}_k$, $k \in s_d$, and $\bar{\hat{e}}_d = \sum_{k \in s_d} \hat{e}_k / n_d$ is the mean of the residuals in domain d, $d = 1, \ldots, D$. It is obvious that in the SRSWOR case in which the weights are constants, for a direct estimator the sum of residuals in each domain is zero. But for other designs, and for an indirect estimator, the sum can differ from zero.

In an *unplanned domain* case, the extra variation due to a random domain sample size n_{s_d} should be accounted for. Let us consider the case of SRSWOR with n elements drawn from the population of N elements. The sampling fraction is n/N and the weights are $w_k = N/n$ for all k. By denoting $y_{dk} = \delta_{dk} y_k$ and $\hat{e}_{dk} = y_{dk} - \hat{y}_k$, $d = 1, \ldots, D$, where the domain membership indicator was given by $\delta_{dk} = 1$ if $k \in U_d$, zero otherwise, we obtain an approximate variance estimator given by

$$\hat{v}_{srs}(\hat{t}_{dGREG}) = N^2 \left(1 - \frac{n}{N}\right) \left(\frac{1}{n}\right) \sum_{k \in s} \frac{(\hat{e}_{dk} - \bar{\hat{e}}_d)^2}{n - 1}. \tag{6.16}$$

Note that also elements outside the domain d contribute to the variance estimate, because $\hat{e}_{dk} = -\hat{y}_k$ for elements $k \notin U_d$ and $k \in s$. An alternative approximate variance estimator is given by

$$\hat{v}_{srs}(\hat{t}_{dGREG}) = N^2 \left(1 - \frac{n}{N}\right) \left(\frac{1}{n}\right) p_d \sum_{k \in s_d} \frac{(\hat{e}_k - \bar{\hat{e}}_d)^2}{n_d - 1} \left(1 + \frac{q_d}{\text{c.v}_{d\hat{e}}^2}\right), \tag{6.17}$$

$d = 1, \ldots, D$, where $p_d = n_d/n$ and $q_d = 1 - p_d$, and $\text{c.v}_{d\hat{e}} = \hat{s}_{d\hat{e}} / \bar{\hat{e}}_d$ is the sample coefficient of variation of residuals in domain d with $\hat{s}_{d\hat{e}}$ as the sample standard deviation of residuals in domain d. The estimator (6.17) corresponds to the variance estimator commonly used under Bernoulli sampling (see Example 2.2).

Let us consider in more detail the choice of a model and the construction of an estimator of the total in the context of ratio estimation and regression estimation for domains. In Section 3.3 the estimation of the total T for the whole population was discussed. There, the auxiliary information assumed to be known at the whole-population level was the total T_z of the auxiliary variable z, and the assisting fixed-effects linear regression model was of the form $y_k = \beta_0 + \beta_1 z_k + \varepsilon_k$, $k \in U$, given by (6.3). The ratio estimator of the population total was given in Section 3.3 by $\hat{t}_{rat} = T_z \times \hat{t}/\hat{t}_z$, and the regression estimator by $\hat{t}_{reg} = \hat{t} + \hat{b}_1(T_z - \hat{t}_z)$, where $\hat{t}$ and $\hat{t}_z$ are SRSWOR estimators of totals T and T_z, respectively and the estimate $\hat{b}_1$ is a sample-based OLS estimate of the finite-population regression coefficient B_1.

For the estimation of domain totals T_d these ratio and regression estimators can be used, but more complex model types can also be introduced, including model types (6.3), (6.7) and (6.12) described above.

Consider a continuous response variable y, whose total T_d is to be estimated for a number of domains of interest U_d, $d = 1, \ldots, D$. Assuming one auxiliary variable z, for example, the following assisting models can be postulated.

1. Fixed-effects P-model for y_k, $k \in U$:
   (1a) $y_k = \beta_0 + \varepsilon_k$ Common intercept model
   (1b) $y_k = \beta_1 z_k + \varepsilon_k$ Common slope model
   (1c) $y_k = \beta_0 + \beta_1 z_k + \varepsilon_k$ Common intercept and slope model.
2. Fixed-effects D-model for y_k, $k \in U_d$, $d = 1, \ldots, D$:
   (2a) $y_k = \beta_{0d} + \varepsilon_k$ Domain-specific intercepts model
   (2b) $y_k = \beta_{1d} z_k + \varepsilon_k$ Domain-specific slopes model
   (2c) $y_k = \beta_{0d} + \beta_{1d} z_k + \varepsilon_k$ Domain-specific intercepts and slopes model.
3. Mixed D-model for y_k, $k \in U_d$, $d = 1, \ldots, D$:
   (3a) $y_k = \beta_{0d} + \varepsilon_k = \beta_0 + u_{0d} + \varepsilon_k$ Domain-specific random intercepts model
   (3b) $y_k = \beta_{0d} + \beta_1 z_k + \varepsilon_k = \beta_0 + u_{0d} + \beta_1 z_k + \varepsilon_k$ Domain-specific random intercepts and common slope model.

Models (1b) and (2b) can be used in ratio estimation for domains and models (1c) and (2c) in regression estimation. It is obvious that indirect SYN and GREG estimators are obtained with model specification (1) and (3), and model type (2) gives direct SYN and GREG estimators.

For example, using the P-model (1b), a SYN estimator (6.1) for domain totals T_d is given by

$$\hat{t}_{dSYN-P} = \sum_{k \in U_d} \hat{y}_k = \sum_{k \in U_d} \hat{b}_1 z_k = T_{dz} \hat{b}_1 = T_{dz} \times \hat{t}_{HT} / \hat{t}_{zHT}, \quad d = 1, \ldots, D, \qquad (6.18)$$

resembling the ratio estimator $\hat{t}_{rat}$ for the whole population, but in $\hat{t}_{dSYN-P}$, domain totals T_{dz} are used instead of the overall total T_z. The estimator for the population slope B_1 is

$$\hat{b}_1 = \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k z_k} = \frac{\hat{t}_{HT}}{\hat{t}_{zHT}},$$

which is the ratio of two HT estimators, $\hat{t}_{HT}$ and $\hat{t}_{zHT}$, of totals of the study variable y and auxiliary variable z respectively. These total estimates are calculated at the whole-population level and, thus, the estimator of domain totals is *indirect*. While using y-values from the whole sample, the estimator $\hat{t}_{dSYN-P}$ aims at *borrowing strength* from the other domains.

A SYN estimator (6.18) using a type (1b) model can be biased. The bias of $\hat{t}_{dSYN-P}$ is approximated by

$$\text{BIAS}(\hat{t}_{dSYN-P}) = E(\hat{t}_{dSYN-P}) - T_d \doteq -T_{dz}(B_{1d} - B_1),$$

where $B_{1d} = \sum_{k\in U_d} y_k / \sum_{k\in U_d} z_k$ is the domain-specific slope, $d = 1, \ldots, D$, and $B_1 = \sum_{k\in U} y_k / \sum_{k\in U} z_k$ is the slope for the whole population. For a given domain, the bias is negligible if the domain slope closely approximates the population slope. But a substantial bias can be encountered if this condition does not hold.

The corresponding indirect GREG estimator (6.2) for domain totals T_d is given by

$$\begin{aligned}\hat{t}_{dGREG-P} &= \sum_{k\in U_d} \hat{y}_k + \sum_{k\in s_d} w_k(y_k - \hat{y}_k) = \hat{t}_{dSYN-P} + \sum_{k\in s_d} w_k(y_k - \hat{b}_1 z_k) \\ &= \hat{t}_{dHT} + \frac{\hat{t}_{HT}}{\hat{t}_{zHT}}(T_{dz} - \hat{t}_{dzHT}) \qquad (6.19)\end{aligned}$$

mimicking the regression estimator for the whole population, but the underlying model is different. Note that an attempt to 'borrow strength' also holds for the indirect GREG estimator.

The *direct* SYN and GREG estimators of type (2b) use y-values from the given domain only. The estimators are obtained by replacing $\hat{b}_1$ by domain-specific counterparts $\hat{b}_{1d}$ given by

$$\hat{b}_{1d} = \frac{\sum_{k\in s_d} w_k y_k}{\sum_{k\in s_d} w_k z_k} = \frac{\hat{t}_{dHT}}{\hat{t}_{dzHT}}, \quad d = 1, \ldots, D,$$

where $\hat{t}_{dHT}$ and $\hat{t}_{dzHT}$ are HT estimators of totals T_d and T_{dz} at the domain level. The direct SYN estimator $\hat{t}_{dSYN-D}$ hence is

$$\hat{t}_{dSYN-D} = \sum_{k\in U_d} \hat{y}_k = \sum_{k\in U_d} \hat{b}_{1d} z_k = T_{dz}\hat{b}_{1d} = T_{dz} \times \hat{t}_{dHT}/\hat{t}_{dzHT}, \quad d = 1, \ldots, D. \quad (6.20)$$

For this model specification, the direct GREG counterpart $\hat{t}_{dGREG-D}$ coincides with the SYN estimator because the second term in GREG estimator (6.2) vanishes.

Let us consider the relative properties of the estimators (6.18) and (6.20) with respect to bias, precision and accuracy. First, the indirect estimator $\hat{t}_{dSYN-P}$ given by (6.18) is biased, and the bias can be substantial if the model assumption does not hold in a given domain. The direct counterpart $\hat{t}_{dSYN-D}$ given by (6.20), which coincides with the GREG estimator $\hat{t}_{dGREG-D}$, is almost design unbiased, irrespective of the validity of the model assumption. The variance of the indirect estimator (6.18) is of the order n^{-1} and thus can be small even in a small domain if the total sample size n is large. On the other hand, the variance of the direct

estimator (6.20) is of the order n_d^{-1} and becomes large when the sample size n_d in domain d is small. Thus, there is a trade-off between bias and precision, depending on the validity of the model assumption and the domain sample size. Using the mean squared error, $\mathrm{MSE}(\hat{t}_d) = V(\hat{t}_d) + \mathrm{BIAS}^2(\hat{t}_d)$, we can conclude the following. In small domains, the indirect estimator (6.18) can be more accurate than the direct counterpart (6.20) because the variance of (6.20) can be very large. But for large domains (with large domain sample size), the direct estimator can be more accurate, because the squared bias of (6.18) can dominate. This holds especially if the model assumption is violated (this trade-off is examined in more detail, for example, in Lehtonen *et al.* 2003).

In Example 6.2, we study selected estimators for domain totals for a single SRSWOR sample drawn from the OHC Survey data set. In Section 6.4, we examine in more detail the relative properties (bias and accuracy) of the synthetic and generalized regression estimators under different model choices. There, the methods are investigated by Monte Carlo simulation techniques, where a large number of independent SRSWOR samples are drawn from a fixed population.

Example 6.2

Estimation of domain totals by design-based methods under SRSWOR sampling. We illustrate the domain estimation methodology by selecting an SRSWOR sample ($n = 1960$ persons) from the OHC Survey data set ($N = 7841$ persons) and estimating the total number of chronically ill persons in the $D = 30$ domains constructed. In the population, the sizes of the domains vary with a minimum of 81 persons and a maximum of 517 persons. The domain proportion of chronically ill persons varies from 18 to 39%, and the overall proportion is 29%. The intra-domain correlation of being chronically ill (binary response) and the age (in years) varies from 0.08 to 0.55; the overall correlation is 0.28.

In the sampling procedure, we consider the domains as unplanned type. Thus, the domain sample sizes are not fixed in the sampling design but are random variates. A Horvitz-Thompson estimator is first calculated. Auxiliary data are then incorporated into the estimation procedure by using the model-assisted GREG estimator given by (6.2). Values of the auxiliary variable z are measurements of age, being available for all persons in the OHC data set, which we, for this example, assume to constitute the population of interest. Therefore, in this hypothetical situation the domain totals T_d of the study variable y also are known for all domains $d = 1, \ldots, D$, and can be used when comparing the estimates of domain totals.

A simple model (1b) from Example 6.2, given by $y_k = \beta \times z_k + \varepsilon_k$, postulates a uniform ratio $R = T/T_z (= 7.778 \times 10^{-3})$ for all domains. Thus, a GREG estimator built on this P-model is of indirect type. On the basis of the SRSWOR sample of $n = 1960$ elements, an estimate of the ratio R is $\hat{r} = \hat{t}_{HT}/\hat{t}_{zHT} = 7.651 \times 10^{-3}$, where $\hat{t}_{HT} (= 2252.3)$ is the HT estimator of the total T of the study variable y and $\hat{t}_{zHT} (= 294357.5)$ is that of the total T_z of the auxiliary variable z. The predicted y-values are calculated by $\hat{y}_k = \hat{r} \times z_k, k = 1, \ldots, 7841$. Alternative

expressions of the estimators are summarized in (6.21). There, the sampling weights are $w_k = N/n = 7841/1960 = 4.001$, T_{dz} are the known domain totals of the auxiliary variable z and $\hat{t}_{dzHT} = \sum_{k \in s_d} w_k z_k$ are the corresponding HT estimates.

$$\begin{aligned} \hat{t}_{dHT} &= \textstyle\sum_{k \in s_d} w_k y_k = N/n \sum_{k \in s_d} y_k \\ \hat{t}_{dGREG-P} &= \textstyle\sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} w_k (y_k - \hat{y}_k) = \hat{t}_{dHT} + \hat{r}(T_{dz} - \hat{t}_{dzHT}), \end{aligned} \tag{6.21}$$

where s_d (with n_d elements) and U_d (with N_d elements) are the sets of the sample and the population elements belonging in domain d respectively and $d = 1, \ldots, D$. Note that the corresponding indirect synthetic estimator is $\hat{t}_{dSYN-P} = \sum_{k \in U_d} \hat{y}_k = T_{dz} \times \hat{r}$, which is based on the same simple model as the GREG estimator.

In the examination of the accuracy, we use the estimated standard error $\text{s.e}(\hat{t}_d)$ and percentage coefficient of variation $\text{c.v}(\hat{t}_d)\% = 100 \times \text{s.e}(\hat{t}_d)/\hat{t}_d$ of an estimator $\hat{t}_d$. The variance estimators used are as follows:

$$\begin{aligned} \hat{v}_{srs}(\hat{t}_{dHT}) &= N^2 \left(1 - \frac{n}{N}\right) \left(\frac{1}{n}\right) p_d \hat{s}^2_{dy} \left(1 + \frac{q_d}{\text{c.v}^2_{dy}}\right), \text{ and} \\ \hat{v}_{srs}(\hat{t}_{dGREG-P}) &= N^2 \left(1 - \frac{n}{N}\right) \left(\frac{1}{n}\right) p_d \hat{s}^2_{d\hat{e}} \left(1 + \frac{q_d}{\text{c.v}^2_{d\hat{e}}}\right), \end{aligned} \tag{6.22}$$

where $p_d = n_d/n, q_d = 1 - p_d$, variance estimators are $\hat{s}^2_{dy} = \sum_{k \in s_d} (y_k - \overline{y}_d)^2/(n_d - 1)$ and $\hat{s}^2_{d\hat{e}} = \sum_{k \in s_d} (\hat{e}_k - \overline{\hat{e}}_d)^2/(n_d - 1)$, estimated coefficients of variation are $\text{c.v}_{dy} = \hat{s}_{dy}/\overline{y}_d$ and $\text{c.v}_{d\hat{e}} = \hat{s}_{d\hat{e}}/\overline{\hat{e}}_d$, where $\overline{y}_d = \sum_{k \in s_d} y_k/n_d$ and $\overline{\hat{e}}_d = \sum_{k \in s_d} \hat{e}_k/n_d$, and residuals are $\hat{e}_k = y_k - \hat{r} \times z_k$.

In the realized sample, domain sample sizes vary from 24 to 132 elements and the mean size is 65. The situation thus is realistic for design-based estimation for domain totals. We first examine the average performance of the Horvitz-Thompson estimator $\hat{t}_{dHT}$ and the indirect GREG estimator $\hat{t}_{dGREG-P}$. In the first part of Table 6.5, a simple average measure $|\overline{\hat{t}} - \overline{T}|/\overline{T}$ of absolute relative difference is calculated in three domain sample size classes, where $\overline{\hat{t}}$ is the mean of the estimated domain totals $\hat{t}_d$ and $\overline{T}$ is the mean of the true values T_d in a given size class. Absolute relative differences of the HT and GREG estimates tend to decrease with increasing domain sample size, and for a given size class, the figures closely coincide. The realized domain sample size and coefficient of variation have a clear association for GREG and HT estimators: sample coefficients of variation tend to decrease with increasing domain sample size, as is indicated in the average coefficient of variation figures given in the second part of Table 6.5. On average, estimated coefficients of variation are smaller for the GREG estimator.

Domain-wise point estimates, standard errors and coefficients of variation for the 30 domains are given in Table 6.6 in which the domains are sorted by the domain sample size. When compared to the HT estimator $\hat{t}_{dHT}$, use of auxiliary information by the model-assisted GREG estimator $\hat{t}_{dGREG-P}$ clearly improves

Table 6.5 Average absolute relative difference and average coefficient of variation of Horvitz–Thompson and GREG estimates by domain sample size class.

	Average absolute relative difference (%)		Average coefficient of variation (%)	
Size class	HT estimator	GREG estimator	HT estimator	GREG estimator
−39	10.6	10.2	30.8	24.7
40–79	2.0	3.4	23.5	19.8
80–	3.2	3.7	16.0	13.6
All	1.8	1.7	23.0	19.0

accuracy. In all 30 domains, estimated standard errors of the GREG estimator are smaller than those of the HT estimator. In most domains, estimated coefficients of variation are smaller for the GREG estimator, as expected.

Let us complete the example by considering briefly the relationship of the GREG estimator and the corresponding model-dependent indirect SYN estimator $\hat{t}_{dSYN-P} = T_{dz} \times \hat{r}$ in the context of the realized sample. By the expression (6.21) for the GREG estimator, we obtain for example in the first domain ($n_1 = 41$):

$$\begin{aligned}\hat{t}_{1GREG-P} &= \sum_{k \in U_1} \hat{y}_k + \sum_{k \in s_1} w_k (y_k - \hat{y}_k) \\ &= 45.43 + 4.001 \times (-0.5974) = 43.04,\end{aligned}$$

where the sum of predicted values $\hat{y}_k$ in the first domain is calculated as $\sum_{k \in U_1} \hat{y}_k = T_{1z} \times \hat{r} = 5937 \times 0.0076515 = 45.43$. This is the synthetic estimate $\hat{t}_{1SYN-P}$ for the first domain. And, for example, for domain $d = 19$ ($n_{19} = 115$) we obtain $\hat{t}_{19GREG-P} = 160.00$ and $\hat{t}_{19SYN-P} = 138.09$, whereas the true value is $T_{19} = 165$. The bias-adjustment term of the GREG estimator thus happens to adjust successfully the bias of the SYN estimator for these domains. But this does not necessarily hold for all domains. In fact, the GREG estimator is more successful than the SYN estimator in 17 out of 30 domains because in several domains, the bias correction affects to a correct direction but too strongly. In the estimation of the accuracy of the SYN estimator, an estimated mean squared error (MSE) should be used because the SYN estimator is not design unbiased. We will consider the relationship of the GREG and SYN estimators for domain totals in more detail in Section 6.4 and further, in the web extension of the book.

6.4 FURTHER COMPARISON OF ESTIMATORS

In this section, we examine further the properties of model-dependent estimators and model-assisted estimators for domain totals using Monte Carlo simulation methods. For this exercise, we again use the OHC Survey data set. To examine empirically the theoretical properties (bias and accuracy) of the different SYN and

Table 6.6 Estimates of the total number of chronically ill persons in domains calculated for an SRSWOR sample ($n = 1960$) from the OHC data set. Domain sample sizes n_d, domain sizes N_d, population totals T_d, and point estimates, standard error estimates and coefficient of variation estimates (%) for HT and GREG estimators, by domain sample size class.

Domain				Estimate of total		Standard error		Coefficient of variation (%)	
d	n_d	N_d	T_d	$\hat{t}_{dHT}$	$\hat{t}_{dGREG}$	s.e$(\hat{t}_{dHT})$	s.e$(\hat{t}_{dGREG})$	c.v$(\hat{t}_{dHT})$	c.v$(\hat{t}_{dGREG})$
Domain sample size n_d < 40									
20	24	101	31	32.0	31.6	9.77	7.13	30.5	22.5
10	26	81	27	32.0	25.6	10.83	8.05	33.8	31.5
18	26	129	36	20.0	27.2	7.60	6.95	38.0	25.5
23	31	156	57	44.0	53.2	10.82	9.10	24.6	17.1
8	35	141	29	24.0	24.5	8.57	7.88	35.7	32.2
30	36	146	34	32.0	33.8	9.86	8.56	30.8	25.3
3	37	133	29	36.0	32.6	10.77	8.73	29.9	26.8
16	37	165	45	52.0	54.8	12.14	9.15	23.3	16.7
Domain sample size 40 ≤ n_d < 80									
1	41	181	33	40.0	43.0	10.80	9.15	27.0	21.3
21	43	153	48	64.0	55.3	14.55	10.93	22.7	19.8
6	45	188	52	24.0	26.6	8.51	7.67	35.5	28.9
28	51	194	74	88.0	85.4	16.61	11.65	18.9	13.6
24	53	200	55	56.0	55.7	13.21	11.06	23.6	19.9
22	57	242	96	112.0	115.0	17.79	13.08	15.9	11.4
15	58	252	61	60.0	66.4	13.20	11.90	22.0	17.9
11	59	187	47	52.0	39.5	13.30	10.89	25.6	27.6
13	69	305	89	80.0	88.5	15.10	12.86	18.9	14.5
12	73	311	95	56.0	65.9	12.85	11.40	22.9	17.3
4	76	295	65	68.0	68.1	14.39	12.17	21.2	17.9
7	78	292	52	40.0	36.3	11.09	10.17	27.7	28.0
Domain sample size n_d ≥ 80									
2	84	352	86	76.0	78.6	14.95	13.49	19.7	17.2
5	86	323	66	76.0	70.5	15.31	13.62	20.1	19.3
26	89	364	124	124.0	126.0	19.07	15.72	15.4	12.5
29	90	365	128	124.0	124.5	19.12	15.10	15.4	12.1
25	91	339	114	112.0	101.6	18.68	14.81	16.7	14.6
17	99	426	139	176.0	183.3	22.11	16.72	12.6	9.1
9	103	366	89	88.0	79.3	16.66	13.82	18.9	17.4
19	115	490	165	152.0	160.0	20.81	17.13	13.7	10.7
14	116	447	130	136.0	128.4	20.31	16.28	14.9	12.7
27	132	517	197	176.0	173.8	22.94	17.51	13.0	10.1
All	1960	7841	2293	2252.3	2254.8	69.42	66.88	3.1	3.0

GREG estimators for domains, we make the following conventions. First, similarly as in Example 6.2, we consider the OHC data set as a frame population of size 7841 elements, such that the necessary auxiliary data are included at micro-level in the data set. Secondly, we construct for the population frame data set a domain structure involving 60 domains in total. This is because we want to consider also domains with a small sample size. Finally, we will draw a large number of independent SRSWOR samples of 1000 elements from the constructed artificial frame population under an unplanned domain structure. We study the bias and accuracy of estimators on the basis of the average figures calculated over the simulated samples.

We assume (according to the principles presented in Box 6.1) that the constructed OHC frame population of $N = 7841$ persons and $D = 60$ domains includes unique identification keys, domain membership indicators, inclusion probabilities for all elements $k \in U$ for a SRSWOR sample of $n = 1000$ elements and values of the auxiliary z-variable age (in years). The binary response variable y to be measured from the sample elements is chronic illness (value 0: No, 1: Yes).

P-models and D-models are used for the indirect SYN and GREG estimators based on linear models of the general form $y_k = \beta_0 + u_{0d} + \beta_1 z_k + \varepsilon_k$. In the mixed D-model case, model parameters are estimated by restricted maximum likelihood (REML) and generalized least squares (GLS), and predictions $\hat{y}_k = \hat{\beta}_0 + \hat{u}_{0d} + \hat{\beta}_1 z_k$, $k \in U$, are calculated. For a fixed-effects P-model, estimation is based on ordinary least squares (OLS), and predictions are calculated as $\hat{y}_k = \hat{b}_0 + \hat{b}_1 z_k$, $k \in U$. Residuals are calculated as $\hat{e}_k = y_k - \hat{y}_k$, $k \in s$, in both cases. By micro-merging these data in the frame population U (see Table 6.3), the data are successfully completed for domain estimation.

Domain totals to be estimated are given by

$$T_d = \sum_{k \in U_d} Y_k, \quad d = 1, \ldots, D.$$

The indirect estimators to be used are the following:

$$\hat{t}_{dSYN} = \sum_{k \in U_d} \hat{y}_k, \quad d = 1, \ldots, D \text{ (synthetic estimator), and}$$

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} w_k (y_k - \hat{y}_k), \; d = 1, \ldots, D$$

(generalized regression estimator).

In these formulas, the predicted values $\hat{y}_k$, $k \in U$, and observed y-data y_k, sampling weights w_k and residuals $\hat{e}_k$, $k \in s$, provide the materials for the calculation of estimates for domain totals. The indirect estimators use fixed-effects P-models and mixed D-models. For the synthetic estimators $\hat{t}_{dSYN-P}$ and

$\hat{t}_{dMSYN-D}$, only the predictions $\hat{y}_k$ are used. And for the GREG estimators $\hat{t}_{dGREG-P}$ and $\hat{t}_{dMGREG-D}$, predicted values $\hat{y}_k$, observed y-data y_k, sampling weights w_k and residuals $\hat{e}_k = y_k - \hat{y}_k$ are used. In the SRSWOR case considered here, the weights $w_k = N/n$ are constants, and the sum of residuals over the whole sample data set is $\sum_{k \in s} \hat{e}_k = 0$. Note that this does not necessarily hold for the domains because we work with indirect estimators of domain totals.

We compare the bias and accuracy of the various estimators by using estimates $\hat{t}_d(s_v)$ from the K repeated Monte Carlo samples $s_v; v = 1, 2, \ldots, K$. For each domain $d = 1, \ldots, D$, the following Monte Carlo summary measures of bias and accuracy are computed. We use two measures of accuracy, the relative root mean squared error (RRMSE) and the median absolute relative error (MdARE), because for a binary response variable there is sometimes a difference in the conclusions drawn from the two measures.

(i) Absolute relative bias (ARB), defined as the ratio of the absolute value of bias to the true value:

$$\left| \frac{1}{K} \sum_{v=1}^{K} \hat{t}_d(s_v) - T_d \right| / T_d.$$

(ii) Relative root mean squared error (RRMSE), defined as the ratio of the root MSE to the true value:

$$\sqrt{\frac{1}{K} \sum_{v=1}^{K} (\hat{t}_d(s_v) - T_d)^2} / T_d.$$

(iii) Median absolute relative error (MdARE) is defined as follows. For each simulated sample $s_v; v = 1, 2, \ldots, K$, the absolute relative error is calculated and a median is taken over the K samples in the simulation:

$$\underset{\text{over } v = 1, \ldots, K}{\text{Median}} \{|\hat{t}_d(s_v) - T_d|/T_d\}.$$

A summary of the features of the experimental design used in this simple exercise is given in Table 6.7.

A summary of the results for the simple models (1a) and (2a) is presented in Part A of Table 6.8 and for the more complex models (1b) and (2b) in Part B of the table. The results indicate that the bias, measured by the average of absolute relative bias ARB, of the GREG estimators GREG-P and MGREG-D is negligible for all models and in all size classes. The bias for the SYN-type estimators varies with the model choice. The bias of SYN-P is substantial for the extremely simple fixed-effects P-model (1a), and the bias decreases when the more realistic fixed-effects model (1b) is used. A similar effect is noticed for the mixed models (2a) and (2b), which provides the smallest bias figures for SYN estimators. Especially

Table 6.7 Summary of technical details of Monte Carlo experiments.

Population:

OHC Survey frame population of size $N = 7841$ persons

Sample size: $n = 1000$ persons

Number of domains: $D = 60$ areas

Number of simulated samples:

$K = 500$ independent SRSWOR samples (unplanned domain structure)

Response variable y:

Chronic illness (binary; $0 = \text{No}, 1 = \text{Yes}$)

Auxiliary z**-data**:

Domain membership indicators

Age (in years)

Models:

(1a) Linear fixed-effects P-model with intercept only:

$y_k = \beta_0 + \varepsilon_k$

(1b) Linear fixed-effects P-model with age as the predictor:

$y_k = \beta_0 + \beta_1 z_k + \varepsilon_k$

(2a) Linear mixed D-model with random intercepts:

$y_k = \beta_0 + u_{0d} + \varepsilon_k$

(2b) Linear mixed D-model with age as the predictor:

$y_k = \beta_0 + u_{0d} + \beta_1 z_k + \varepsilon_k$

Target parameters:

Domain totals T_d of chronically ill people, $d = 1, \ldots, 60$

Estimators of domain totals:

SYN estimators:

$\hat{t}_{dSYN-P}$ using a linear fixed-effects P-model

$\hat{t}_{dMSYN-D}$ using a two-level linear D-model

GREG estimators:

$\hat{t}_{dGREG-P}$ using a linear fixed-effects P-model

$\hat{t}_{dMGREG-D}$ using a two-level linear D-model

Measures of performance:

Averages calculated over domain size classes of:

ARB Absolute relative bias

RRMSE Relative root mean squared error

MdARE Median absolute relative error

in small domains, the accuracy is better for SYN estimators when compared to GREG estimators, in all model types and with both measures RRMSE and MdARE. But as soon as the domain sample size increases, the difference in accuracy tends to decrease.

The results in Table 6.8 also indicate that the model improvement, that is, moving from a 'weak' model towards a 'stronger' model, is much more prominent for SYN-type estimators than for GREG-type estimators. Note that for this estimation exercise we needed an access to the micro-merged frame population and sample data set. An access to these data is provided by the web extension of the book.

6.5 CHAPTER SUMMARY AND FURTHER READING

In this chapter, we concentrated on design-based model-assisted estimation for domains. This approach is frequently used, for example, in the production of official statistics. We made several assumptions for the treatment of estimation for domain totals. In particular, we assumed that in a given statistical infrastructure, registers

Table 6.8 Simulation results for SYN and GREG estimators for domain totals of chronically ill people with different model choices ($K = 500$ independent SRSWOR samples with $n = 1000$ elements in each).

A. Fixed-effects P-model $y_k = \beta_0 + \varepsilon_k$ and mixed D-model $y_k = \beta_0 + u_{0d} + \varepsilon_k$.

		Average over domains of					
Estimator	Domain sample size class	Domain total in population	Estimate of domain total	Absolute relative bias ARB%	Relative root MSE RRMSE%	Median absolute relative error MdARE%	Domain sample size
SYN-P	0–10	17.5	13.7	36.9	37.4	37.0	5.6
	11–20	37.0	34.4	50.3	50.7	50.3	14.1
	21–	62.4	78.8	43.6	44.2	43.6	32.4
	All	38.2	41.2	43.5	44.0	43.5	16.9
MSYN-D	0–10	17.5	14.9	25.1	33.0	27.9	5.6
	11–20	37.0	35.7	22.7	33.3	25.0	14.1
	21–	62.4	66.3	11.6	26.0	17.4	32.4
	All	38.2	38.2	20.0	30.9	23.6	16.9
GREG-P	0–10	17.5	17.5	2.4	55.2	39.5	5.6
	11–20	37.0	37.0	1.6	40.7	27.8	14.1
	21–	62.4	62.4	1.1	31.1	20.8	32.4
	All	38.2	38.2	1.7	42.8	29.7	16.9
MGREG-D	0–10	17.5	17.3	2.6	53.5	38.9	5.6
	11–20	37.0	37.0	1.9	39.5	27.3	14.1
	21–	62.4	62.5	1.1	30.3	20.2	32.4
	All	38.2	38.2	1.9	41.5	29.1	16.9
B. Fixed-effects P-model $y_k = \beta_0 + \beta_1 z_k + \varepsilon_k$ and mixed D-model $y_k = \beta_0 + u_{0d} + \beta_1 z_k + \varepsilon_k$.							
SYN-P	0–10	17.5	18.0	27.0	28.1	27.1	5.6
	11–20	37.0	36.6	19.6	20.8	19.7	14.1
	21–	62.4	62.0	12.1	13.9	12.5	32.4
	All	38.2	38.1	19.8	21.2	20.0	16.9
MSYN-D	0–10	17.5	18.0	25.9	27.5	26.4	5.6
	11–20	37.0	36.6	17.7	20.2	18.5	14.1
	21–	62.4	62.1	9.7	14.4	11.6	32.4
	All	38.2	38.2	18.1	20.9	19.1	16.9
GREG-P	0–10	17.5	17.5	2.7	53.0	38.5	5.6
	11–20	37.0	37.0	1.4	38.9	26.5	14.1
	21–	62.4	62.5	1.1	30.0	20.2	32.4
	All	38.2	38.2	1.8	41.0	28.7	16.9
MGREG-D	0–10	17.5	17.5	2.7	52.8	38.4	5.6
	11–20	37.0	37.0	1.5	38.8	26.4	14.1
	21–	62.4	62.5	1.0	29.8	20.2	32.4
	All	38.2	38.2	1.8	40.8	28.6	16.9

are available as frame populations and sources of micro-level and aggregate-level auxiliary data, and unique identification keys are available in order to merge the data from a sample survey with data from a statistical register. We believe that fulfilling these conditions can provide much flexibility for sampling design and estimation for domains. For example, the data can then be aggregated at higher levels of the population if desired. The use of unit-level data and unit-level modelling can be beneficial for both design-based model-assisted estimation and model-dependent estimation for domains. It appeared that careful and realistic modelling is especially important in model-dependent estimation for domains. This was demonstrated by a small-scale simulation study. The materials discussed in the examples of this chapter will be worked out further in the web extension of the book.

In practice, design-based model-assisted estimation is most often used for domains whose sample size is reasonably large. For small domains, methods of small-area estimation are used instead. For the estimation for domains, it is recommended to define, if possible, the intended domains as strata in the sampling phase, and to use a suitable allocation scheme, such that a reasonably large sample size is attained for all domains. And in the estimation phase it is advisable to incorporate strong auxiliary data into the estimation procedure by using carefully chosen models.

Supplementing the references mentioned earlier in this chapter, design-based model-assisted estimation for domains is discussed, for example, in Estevao *et al.* (1995) and Estevao and Särndal (1999). Lehtonen and Veijanen (1998) discuss nonlinear GREG estimators, such as a multinomial logistic GREG estimator.

In addition to Rao (2003), model-dependent methods for small area estimation are presented in Ghosh and Rao (1994) and Rao (1999). You and Rao (2002) discuss pseudo EBLUP estimators involving survey weights. Underlying models and their features is a prominent theme in recent literature (Ghosh *et al.* 1998; Marker 1999; Moura and Holt 1999; Prasad and Rao 1999; Feder *et al.* 2000). There is extensive recent literature on small area estimation from a Bayesian point of view, including empirical Bayes and hierarchical Bayes techniques (Datta *et al.* 1999; Ghosh and Natarajan 1999; You and Rao 2000). Some recent publications relate frequentist and Bayesian approaches in small area estimation (Singh *et al.* 1998). Valliant *et al.* (2000) discuss small-area estimation under a prediction approach.