Lehtonen-Pahkinen (2004) Practical Methods for Design and Analysis of Complex Surveys. Wiley.

3.3 MODEL-ASSISTED ESTIMATION

Introduction

In the techniques discussed so far, auxiliary information of the population elements is used in the sampling phase to attain an efficient sampling design. We now turn to a different way of utilizing auxiliary information. Our aim is to introduce estimators that can be used for the selected sample to obtain better estimates of the parameters of interest, relative to the estimates calculated with estimators based on the sampling design used.

Let us assume that appropriate auxiliary data are available from the population as a set of auxiliary variables. Of these variables, some might be categorical and some continuous. Some auxiliary data are perhaps used for the sampling procedure. Others can be used for improving efficiency; a way to do this is, for example, to use an auxiliary variable z, which is related to our study variable y, for a reduction of the design variance of the original estimator of the population total of y. In Särndal *et al.* (1992), these techniques are discussed in the context of *model-assisted design-based estimation. Model-assisted* estimation refers to the property of the estimators that models such as linear regression are used in incorporating the auxiliary information in the estimation procedure for the finite-population parameters of interest, such as totals. Model-assisted estimation should be distinguished from the multivariate survey analysis methods to be discussed in Chapter 8. There, models are also used but for multivariate survey analysis purposes.

In the following text, a brief review is given on model-assisted estimation. More specifically, *poststratification, ratio estimation* and *regression estimation* are considered. The methods are special cases of so-called *generalized regression estimators*. All these methods are aimed at improving the estimation from a given sample by using available auxiliary information from the population. This can result in estimates closer to the true population value and a reduction in the design variance of an estimator calculated from the sampled data.

In model-assisted estimation, an auxiliary variable z, which is related to the study variable y, is required. If this variable is categorical, the target population U can be partitioned into subpopulations $U_1, \ldots, U_g, \ldots, U_G$ according to some classification principle. In poststratification, these subpopulations are called *post-strata*. If the poststrata are internally homogeneous, this partitioning can capture a great deal of the total variance of the study variable y, resulting in a decrease in the design-based variance of an estimator. Moreover, poststratification can be used to obtain more accurate point estimates and reduce the bias of sample estimates caused by nonresponse.

The auxiliary variable z is often continuous. If it correlates strongly with the study variable y, a linear regression model can be assumed with y as the dependent variable and z as the predictor. This regression can be estimated from the observed sample and used in the estimation of the original target parameter. For this, ratio estimation and regression estimation can be used. By these methods, substantial gains in efficiency and increased accuracy are often achieved.

To construct a model-assisted estimator, two kinds of weights are considered. The preliminary weights are the usual sampling design weights w_k, which generally are the inverses of the inclusion probabilities π_k; these weights are extensively used in this book. The other type of weights are called *g weights* and their values g_k depend both on the selected sample and on the chosen estimator. The product $w_k^* = g_k w_k$ gives new weights known as *calibrated* weights, which are used in the model-assisted estimators. Thus, using calibrated weights, a model-assisted estimator can be written as $\hat{t}_{cal} = \sum_{k=1}^{n} w_k^* y_k$. A property of the calibrated weights is that for example for ratio estimation, the estimator $\hat{t}_{z,cal} = \sum_{k=1}^{n} w_k^* z_k$ of the total of the auxiliary z-variable reproduces exactly the known population total T_z. The g weights and calibrated weights will be explicitly given for poststratification, ratio estimation and regression estimation.

The basic principles of model-assisted estimation are most conveniently introduced for SRSWOR, although natural applications in practical situations are often under more complex designs. A further simplification is that only one auxiliary variable is assumed. Also, this assumption can be relaxed if multiple auxiliary variables are available as is assumed in discussing regression estimation. The concept of *estimation strategy* will be used referring to a combination of the sampling design and the appropriate estimator. The model-assisted strategies to be discussed are shown in Table 3.12. In the design-based reference strategies, no auxiliary information is used.

Poststratification

Poststratification can be used for improvement of efficiency of an estimator if a discrete auxiliary variable is available. This variable is used to stratify the sample data set after the sample has been selected. Recall from Section 3.1 that

Table 3.12 Estimation strategies for population total.

Strategy		Auxiliary information	Assisting model
		Design-based strategies	
SRSWOR		Not used	None
SRSWR		Not used	None
		Model-assisted strategies	
Poststratification	SRS*pos	Discrete	ANOVA
Ratio estimation	SRS*rat	Continuous	Regression (no intercept)
Regression estimation	SRS*reg	Continuous	Regression

stratification of the element population as part of the sampling design often gave a gain in efficiency. This was achieved by an appropriate choice of the stratification variables so that the variation in the study variable y within the strata would be small. Poststratification has a similar aim. To avoid confusion with the usual (pre)stratification, the population is partitioned into G groups that are called *poststrata*.

To carry out poststratification, the sample data are first combined with the appropriate auxiliary data obtained perhaps from administrative registers or official statistics. Combining the sampled data with poststratum information and the corresponding selection probabilities, we can proceed with the estimation in basically the same way as if it were being done by ordinary (pre)stratification. Certain differences exist, however. Because we are stratifying after the sample selection or, more usually, after the data collection, we cannot assume any specific allocation scheme. The sample size n is fixed but how it is allocated to the different strata is not known until the sample is drawn. This property causes no harm to the estimation of, for example, the total, but estimating of the variance of the total estimator requires more attention.

The *poststratified estimator* for the total T of y is given by

$$\hat{t}_{pos} = \sum_{g=1}^{G} \hat{t}_g = \sum_{g=1}^{G} \sum_{k=1}^{n_g} w_{gk}^* y_{gk}, \tag{3.20}$$

where $\hat{t}_g = N_g \overline{y}_g$ is an estimator of the poststratum total T_g and N_g is the size of the poststratum g. The poststratum weights are $w_{gk}^* = g_{gk} w_{gk}$, where the g weights are $g_{gk} = N_g / \hat{N}_g$ with the *estimated poststratum sizes* in the denominator, and w_{gk} are the original sampling weights. The calculation of w_{gk}^* will be illustrated in Example 3.9. The variance of $\hat{t}_{pos}$ can be determined in various ways, depending on how one uses the configuration of the observed sample. The configuration

refers to how the actual poststratum sample sizes n_g are distributed, and if this is taken as given, the *conditional variance* is simply the same as the usual variance for stratified samples:

$$V_{srs,con}(\hat{t}_{pos}|n_1,\ldots,n_g,\ldots,n_G) = \sum_{g=1}^{G} N_g^2 \left(1-\frac{n_g}{N_g}\right)\frac{S_g^2}{n_g}, \tag{3.21}$$

where the poststratum variances are given by $S_g^2 = \sum_{k=1}^{N_g}(Y_{gk}-\overline{Y}_g)^2/(N_g-1)$. By averaging (3.21) over all possible configurations of n, the *unconditional variance* is obtained. This gives an alternative variance formula,

$$V_{srs,unc}(\hat{t}_{pos}) = \sum_{g=1}^{G} N_g^2 \left(1-\frac{E(n_g)}{N_g}\right)\frac{S_g^2}{E(n_g)}, \tag{3.22}$$

where $E(n_g)$ is the expected poststratum sample size. This variance can be approximated in various ways. One of the approximations is

$$V_{srs,unc}(\hat{t}_{pos}) \doteq N^2\left(1-\frac{n}{N}\right)\left(\frac{1}{n}\right)\left[\sum_{g=1}^{G}\left(\frac{N_g}{N}\right)S_g^2 + \left(\frac{1}{n}\right)\sum_{g=1}^{G}\left(1-\frac{N_g}{N}\right)S_g^2\right]. \tag{3.23}$$

The difference between the conditional and unconditional variances could be considerable if the sample size is small. The corresponding variance estimators $\hat{v}_{srs,con}(\hat{t}_{pos})$ and $\hat{v}_{srs,unc}(\hat{t}_{pos})$ are obtained by inserting $\hat{s}_g^2$ for S_g^2, where $\hat{s}_g^2 = \sum_{k=1}^{n_g}(y_{gk}-\overline{y}_g)^2/(n_g-1)$. For illustrative purposes, both variances $V_{srs,con}$ and $V_{srs,unc}$ are estimated in the next example.

Example 3.10

Estimation with poststratification. The sample used is drawn with SRSWOR from the *Province'91* population in Section 2.3 (see Example 2.1). The sample is poststratified according to administrative division of the municipalities into urban and rural municipalities. The target population contains $N_1 = 7$ urban and $N_2 = 25$ rural municipalities. The two poststrata have the value 1 for urban and 2 for rural municipalities.

In Table 3.13, the sample information used for the estimation with poststratification is displayed.

Let us consider more closely the estimation of the total T. The poststratum totals of UE91 estimated from the table are $\hat{t}_1 = N_1\overline{y}_1 = 7\times 1868 = 13\,076$ and $\hat{t}_2 = N_2\overline{y}_2 = 25\times 201.2 = 5030$. Using these estimates, the poststratified estimate for T is $\hat{t}_{pos} = \hat{t}_1 + \hat{t}_2 = 18\,106$.

Alternatively, the total estimate $\hat{t}_{pos}$ can be calculated using the poststratum weights w_k^*. To calculate w_k^*, the original sampling weights w_k should be adjusted by the sample dependent g_k weights. For this, first the estimate of the poststratum size is determined. Denoting by w_{gk} the original element weight of a sample element that belongs to the poststratum g, an estimate for poststratum size $\hat{N}_g$ is given by summing up these original weights. Then, the corresponding g weight for an element k in poststratum g is simply $g_{gk} = N_g/\hat{N}_g$, where N_g is the exact size of the poststratum g. For example, in Table 3.13, the original sampling weight under SRS is $w_k = 4$, or a constant for each population element. In the first poststratum, the poststratum size is $N_1 = 7$ and its estimated size is $\hat{N}_1 = 4 + 4 + 4 = 12$, because there are three sampled elements in the first poststratum. Thus, the corresponding g weight is $g_{1k} = N_1/\hat{N}_1 = 7/12 = 0.5833$. Finally, the poststratum weights are given for the first poststratum by $w_{1k}^* = g_{1k} \times w_{1k} = 0.5833 \times 4 = 2.3333$. This value turns out to be the same for all the sampled elements for the first poststratum (urban municipalities). Using the poststratum weights, the estimate $\hat{t}_{pos}$ will be equal to that previously calculated.

Estimation results for all the estimators are displayed in Table 3.14. The original setting of sample identifiers remains, say STR = 1 and CLU = ID, but the element weights are to be replaced by the poststratum weights, and the sampling rate is 0.43 for the first poststratum and 0.20 for the second poststratum. Original sampling weights are used and the sampling rate is 0.25 for both poststrata for estimation of unconditional variance. Note that this procedure roughly approximates the formula given in (3.23). For comparison, the design-based estimates $\hat{t}$, $\hat{r}$ and $\hat{m}$ obtained under SRSWOR are included.

Table 3.13 A simple random sample drawn without replacement from the *Province'91* population with poststratum weights.

Sample design identifiers			Element	Study variables		Poststratification		
							g	Post.
STR	CLU	WGHT	LABEL	UE91	LAB91	POSTSTR	WGHT	WGHT
1	1	4	Jyväskylä	4123	33786	1	0.5833	2.3333
1	4	4	Keuruu	760	5919	1	0.5833	2.3333
1	5	4	Saarijärvi	721	4930	1	0.5833	2.3333
1	15	4	Konginkangas	142	675	2	1.2500	5.0000
1	18	4	Kuhmoinen	187	1448	2	1.2500	5.0000
1	26	4	Pihtipudas	331	2543	2	1.2500	5.0000
1	30	4	Toivakka	127	1084	2	1.2500	5.0000
1	31	4	Uurainen	219	1330	2	1.2500	5.0000

Sampling rate for calculation of *unconditional variance*: $8/32 = 0.25$
Sampling rates for calculation of *conditional variance*:
Stratum 1 (Urban) $= 3/7 = 0.43$
Stratum 2 (Rural) $= 5/25 = 0.20$

Table 3.14 Poststratified estimates from a simple random sample drawn without replacement from the *Province'91* population.

(1) Poststratified estimates (conditional)

Statistic	Variables	Estimate	s.e	c.v	deff
Total	UE91	18 106	6014	0.33	0.33
Ratio	UE91, LAB91	12.97%	0.45%	0.03	0.59
Median	UE91	194	36	m.a.	1.09

(2) Poststratified estimates (unconditional)

Statistic	Variables	Estimate	s.e	c.v	deff
Total	UE91	18 106	7364	0.41	0.50
Ratio	UE91, LAB91	12.97%	0.49%	0.03	0.70
Median	UE91	194	50	m.a	1.12

(3) Design-based estimates

Statistic	Variables	Estimate	s.e	c.v	deff
Total	UE91	26 440	13 282	0.50	1.00
Ratio	UE91, LAB91	12.78%	0.41%	0.03	1.00
Median	UE91	226	149	m.a.	1.00

The comparison shows how poststratification affects point estimates. The biggest gain is obtained when estimating the population total. The estimate of the number of unemployed is $\hat{t}_{pos} = 18\,106$, which is closer to the true value $T = 15\,098$ than the design-based estimate $\hat{t} = 26\,440$. The ratio estimate changes only slightly. The median behaves somewhat peculiarly, as has been seen previously.

The reason for a more accurate estimate for the total is obvious. Under SRSWOR, one should have drawn urban and rural municipalities approximately by their respective proportions: $(8/32) \times (7) \approx 2$ towns and $(8/32) \times (25) \approx 6$ rural municipalities. The urban municipalities have larger populations and unemployment figures. If by chance they are over-represented in the sample, then the design-based estimator will overestimate the population total. But poststratification can correct (at least partially) skewnesses. Therefore, we could also get a point estimate closer to its true value.

Poststratification can also improve efficiency. Again, this is true especially for the total. The estimated variance of $\hat{t}_{pos}$ under the conditional assumption is reduced to one-third when compared with the pure design-based estimate $\hat{t}$, which is indicated by deff $= 0.33$. If the unconditional variance is used as a basis, then deff $= 0.50$. The unconditional variance estimate is greater than the conditional variance estimate, because the poststratum sample sizes n_g are by definition random variables whose variance contribution increases the total variance.

Ratio Estimation of Population Total

The estimation of the population total T of a study variable y was considered previously under poststratification using the sample data and a discrete auxiliary variable. *Ratio estimation* can also be used to improve the efficiency of the estimation of T, if a continuous auxiliary variable z is available. The population total T_z and the n sample values z_k of z are required for this method. Such information can often be obtained from administrative registers or official statistics. This information can be used to improve the estimation of T by first calculating the sample estimator $\hat{r} = \hat{t}/\hat{t}_z$ of the ratio $R = T/T_z$ and multiplying $\hat{r}$ by the known total T_z. Ratio estimation of the total can be very efficient if the ratio Y_k/Z_k of the values of the study and auxiliary variables is nearly constant across the population.

Ratio estimators are usually effective but slightly biased. Because of bias, the mean squared error (MSE) could be used instead of the variance when examining the sampling error. It has been shown that the proportional bias of a ratio estimator is $1/n$ and so becomes small when the sample size increases. Thus, the variance serves as an approximation to the MSE in large samples. The properties of ratio estimators have been studied widely in classical sampling theory.

Let us consider ratio estimation of the total T of y under simple random sampling without replacement. We are interested in a *ratio-estimated total* given by

$$\hat{t}_{rat} = \hat{r} \times T_z = \sum_{k=1}^{n} w_k^* y_k, \tag{3.24}$$

where $\hat{r} = \hat{t}/\hat{t}_z = N\overline{y}/N\overline{z} = \sum_{k=1}^{n} yk/\sum_{k=1}^{n} z_k$ and T_z is the population total of the auxiliary variable z. The calibrated weights are $w_k^* = g_k w_k = (T_z/\hat{t}_z)w_k$.

In the estimator (3.24), $\hat{r}$ is a random variable and the total T_z is a constant. Thus, the variance of $\hat{t}_{rat}$ can be written simply as $V_{srs}(\hat{t}_{rat}) = T_z^2 \times V_{srs}(\hat{r})$. If the SRSWOR design variance of the estimator $\hat{r}$ of a ratio (equation (2.9)) is introduced here, an approximative variance of the ratio-estimated total is given by

$$V_{srs}(\hat{t}_{rat}) \doteq N^2 \left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)\sum_{k=1}^{N} \frac{(Y_k - R \times Z_k)^2}{N-1}, \tag{3.25}$$

whose estimator is given by

$$\hat{v}_{srs}(\hat{t}_{rat}) = N^2 \left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)\sum_{k=1}^{n} \frac{(y_k - \hat{r}z_k)^2}{n-1}. \tag{3.26}$$

By studying the sum of squares in the variance equation (3.25), it is possible to find the condition under which ratio estimation results in an improved estimate

of a total. The total sum of squares can be decomposed as follows:

$$\begin{aligned}\sum_{k=1}^{N}(Y_k - R \times Z_k)^2/(N-1) &= \sum_{k=1}^{N}[(Y_k - \overline{Y}) - R(Z_k - \overline{Z})]^2/(N-1) \\ &= \sum_{k=1}^{N}[(Y_k - \overline{Y})^2 - R^2(Z_k - \overline{Z})^2 \\ &\quad - 2R(Y_k - \overline{Y})(Z_k - \overline{Z})]/(N-1) \\ &= S_y^2 + R^2 S_z^2 - 2R\rho_{yz}S_y S_z,\end{aligned}$$

where ρ_{yz} is the finite-population correlation coefficient of the variables y and z. Consider the difference

$$V_{srs}(\hat{t}) - V_{srs}(\hat{t}_{rat}) = N^2\left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)\{S_y^2 - [S_y^2 + R^2 S_z^2 - 2R\rho_{yz}S_y S_z]\}.$$

The ratio estimator improves efficiency if $V_{srs}(\hat{t}) > V_{srs}(\hat{t}_{rat})$, which occurs when

$$R^2 S_z^2 < 2R\rho_{yz}S_z S_y$$

is valid or

$$2\rho_{yz} > \frac{RS_z}{S_y}.$$

It should be noted that $R = \overline{Y}/\overline{Z}$, and that the former condition expressed in terms of coefficients of variation (c.v) of the variables z and y is given by

$$\rho_{yz} > \left(\frac{1}{2}\right)\frac{\text{c.v}_y}{\text{c.v}_z},$$

where $\text{c.v}_y = S_y/\overline{Y}$ and $\text{c.v}_z = S_z/\overline{Z}$ are the coefficients of variation of y and z respectively. Therefore, improvement in efficiency depends on the correlation between the study and auxiliary variables y and z and the c.v of each variable.

Example 3.11

Efficiency of a ratio-estimated total in the *Province'91* population. The variable UE91 is the study variable y and HOU85 is chosen as the auxiliary variable z. The correlation coefficient between UE91 and HOU85 is $\rho_{yz} = 0.9967$, and the corresponding coefficients of variation are $\text{c.v}_y = S_y/\overline{Y} = 743/472 = 1.57$ and $\text{c.v}_z = S_z/\overline{Z} = 4772/2867 = 1.66$. Thus, the condition given above is valid since

$$\rho_{yz} = 0.9967 > 0.4729 = \frac{1}{2} \times \frac{1.57}{1.66}.$$

It can be seen that the ratio estimation improves the efficiency. The improvement can also be measured directly as a design effect. In addition to the parameters given, the ratio $R = \overline{Y}/\overline{Z} = 472/2867 = 0.1646$ is required. The value of the design effect of the ratio-estimated total $\hat{t}_{rat}$ in the *Province'91* population is given by

$$\begin{aligned}\mathrm{DEFF}_{srs}(\hat{t}_{rat}) &= \frac{S_y^2 + R^2 S_z^2 - 2R\rho_{yz} S_y S_z}{S_y^2} \\ &= \frac{743^2 + 0.1646^2 \times 4772^2 - 2 \times 0.1646 \times 0.9967 \times 743 \times 4772}{743^2} \\ &= 0.0102\end{aligned}$$

which is close to 0. This substantial improvement in efficiency is due to the favourable relationship between UE91 and HOU85 such that the ratio Y_k/Z_k is nearly constant across the population.

The ratio-estimated total is in practice calculated using the available survey data under the actual sample design. If the design is, say, stratified SRS, the corresponding parameters would be estimated by using appropriate stratum weights. The present example was evaluated under simple random sampling without replacement, which will also be used in the following example. There, the use of g weights will also be illustrated.

Example 3.12

Calculating a ratio-estimated total from a simple random sample drawn without replacement from the *Province'91* population. Again we use UE91 as the study variable and HOU85 as the auxiliary variable. The estimated ratio is $\hat{r} = \overline{y}/\overline{z} = 0.1603$, which is calculated from the sample in Table 3.15. The sample identifiers are STR = 1, ID is the cluster identifier, and the weight is WGHT = 4.

Table 3.15 A simple random sample drawn without replacement from the *Province'91* population prepared for ratio estimation.

Sample design identifiers			Element	Study var.	Aux. var.	g	Adj.
STR	CLU	WGHT	LABEL	UE91	HOU85	WGHT	WGHT
1	1	4	Jyväskylä	4123	26 881	0.5562	2.2248
1	4	4	Keuruu	760	4896	0.5562	2.2248
1	5	4	Saarijärvi	721	3730	0.5562	2.2248
1	15	4	Konginkangas	142	556	0.5562	2.2248
1	18	4	Kuhmoinen	187	1463	0.5562	2.2248
1	26	4	Pihtipudas	331	1946	0.5562	2.2248
1	30	4	Toivakka	127	834	0.5562	2.2248
1	31	4	Uurainen	219	932	0.5562	2.2248

Sampling rate: $8/32 = 0.25$.

To carry out ratio estimation of the total, the calibrated weights w_k^* are first calculated. The sampling weight w_k is a constant $w_k = N/n = 32/8 = 4$ as before. The values of the g weight are $g_k = T_z/\hat{t}_z$. The population total of the auxiliary variable is $T_z = 91\,753$ and its estimate calculated from the sample is $\hat{t}_z = 164\,952$. Thus, the g weight is the constant $g_k = 91\,753/164\,952 = 0.5562$. Multiplying the weight w_k by the g weight gives the value for the calibrated weight $w_k^* = 4 \times 0.5562 = 2.2248$.

The ratio estimate for the total is calculated as

$$\hat{t}_{rat} = \sum_{k=1}^{n} w_k^* y_k = \hat{r} \times T_z = 0.1603 \times 91\,753 = 14\,707,$$

which is much closer to the population total $T = 15\,098$ than the SRSWOR estimate $\hat{t} = 26\,440$ for the total number of unemployed. The variance estimate for the total estimator is

$$\hat{v}_{srs}(\hat{t}_{rat}) = 32^2 \frac{(1 - 0.25)}{8} \times 91^2 = 892^2.$$

The corresponding deff estimate is

$$\text{deff}_{srs}(\hat{t}_{rat}) = \frac{\hat{v}_{srs}(\hat{t}_{rat})}{\hat{v}_{srs}(\hat{t})} = 892^2/13\,282^2 = 0.0045,$$

which also shows that ratio estimation improves the efficiency. The minimal auxiliary information of the population total T_z and the sample values of z yield good results.

It is also possible to calculate the DEFF when using the ratio-estimated total since the variance of $V_{srs}(\hat{t}_{rat})$ is

$$\begin{aligned} V_{srs}(\hat{t}_{rat}) &\doteq N^2 \left(1 - \frac{n}{N}\right) \left(\frac{1}{n}\right) \sum_{k=1}^{N} \frac{(Y_k - R \times \mathrm{Z}_k)^2}{(N-1)} \\ &= 32^2 \frac{(1 - 0.25)}{8} \times 75^2 = 736^2. \end{aligned}$$

Division by the corresponding SRSWOR design variance of $\hat{t}$ gives

$$\text{DEFF}_{srs}(\hat{t}_{rat}) = \frac{V_{srs}(\hat{t}_{rat})}{V_{srs}(N\overline{y})} = 736^2/7283^2 = 0.0102,$$

which is the same figure presented previously in Example 3.11.

For these data, ratio estimation considerably improves efficiency and brings the point estimate for the total close to its population value. The value of the

ratio estimator is based on the fact that across the population, the ratio Y_k/Z_k remains nearly constant. It should be noted that even a high correlation between the variables does not guarantee this, because the ratio estimator assumes that the regression line of y and z goes near the origin. Thus, an intercept term is not included in the corresponding regression equation. The ratio estimator may therefore be unfavourable if the population regression line intercepts the y-axis far from the origin, even if the correlation is not close to zero. For these situations, the method presented next would be more appropriate.

Regression Estimation of Totals

Regression estimation of the population total T of a study variable y is based on the linear regression between y and a continuous auxiliary variable z. The linear regression can, for example, be given by $E_M(y_k) = \alpha + \beta \times z_k$ with a variance $V_M(y_k) = \sigma^2$, where y_k are independent random variables with the population values Y_k as their assumed realizations, α, β and σ^2 are unknown parameters, Z_k are known population values of z, and E_M and V_M refer respectively to the expectation and variance under the model. The finite-population analogues of α and β, denoted respectively by A and B, are estimated from the sample using weighted least squares estimation so that the sampling design is properly taken into account. It is immediately obvious that multiple auxiliary variables can also be incorporated in the model. Note that the model assumption introduces a new type of randomness; in the estimation considered previously, the sample selection was the only source of random variation.

We consider the basic principles of regression estimation for SRS without replacement using the above regression model with a single auxiliary variable. The finite-population quantities A and B are estimated by the ordinary least squares method giving $\hat{b} = \hat{s}_{yz}/\hat{s}_z^2$ as an estimator of the slope B and $\hat{a} = \overline{y} - \hat{b}\overline{z}$ as an estimator of the intercept A. Using the estimator $\hat{b}$, the *regression estimator* of the total T of y is given by

$$\hat{t}_{reg} = N(\overline{y} + \hat{b}(\overline{Z} - \overline{z})) = \hat{t} + \hat{b}(T_z - \hat{t}_z) \tag{3.27}$$

where $\hat{t} = N\overline{y}$ is the SRSWOR estimator of T, $\hat{t}_z = N\overline{z}$ is the SRSWOR estimator of T_z and $\overline{Z} = T_z/N$. Alternatively, if transformed values $z_k^* = \overline{Z} - z_k$ are used in the regression instead of z_k, an estimated intercept for this model is $\hat{a}^* = \hat{a} + \hat{b}\overline{Z}$ giving $\hat{t}_{reg} = N\hat{a}^*$, because (3.27) can be written also as $\hat{t}_{reg} = N\hat{a} + \hat{b}T_z$. Note that the regression estimation of the total T presupposes only knowledge of the population total T_z and the sample values z_k of the auxiliary variable z.

Regression estimators constitute a wide class of estimators. For example, the previous ratio estimator $\hat{t}_{rat} = \hat{r}T_z$ is a special case of (3.27) such that the intercept A is assumed 0 and the slope B is estimated by $\hat{b} = \hat{r} = \hat{t}/\hat{t}_z$.

Alternatively, we can calculate calibrated weights $w_k^* = w_k \times g_k$ where w_k is the sampling weight and the g weight is calculated from

$$g_k = \frac{N}{\hat{N}}\left[1 + \frac{\overline{Z} - \overline{z}}{\frac{n-1}{n}\hat{s}_z^2} \times (z_k - \overline{z})\right],$$

where $\overline{Z}$ is the population mean and $\overline{z}$ is the sample mean of the auxiliary variable z, the sum of the sampling weights is $\sum_{k=1}^{n} w_k = \hat{N}$ and

$$\hat{s}_z^2 = \frac{\sum_{k=1}^{n}(z_k - \overline{z})^2}{n-1}.$$

The weights g_k and calibrated weights w_k^* are presented under the model $E_M(y_k) = \alpha + \beta \times z_k$ in Table 3.16 for an SRSWOR sample from the *Province'91* Population. A regression estimate for the population total thus is the calibrated weight w_k^* multiplied by the observed value y_k and summed-up over all sample elements. The regression estimator given in (3.27) can thus also be expressed as $\hat{t}_{reg} = \sum_{k=1}^{n} w_k^* y_k$.

An approximate design variance of $\hat{t}_{reg}$ under SRSWOR is given by

$$V_{srs}(\hat{t}_{reg}) \doteq N^2\left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)S_E^2, \tag{3.28}$$

where $S_E^2 = \sum_{k=1}^{N}(E_k - \overline{E})^2/(N-1)$, $E_k = Y_k - \hat{Y}_k$ and $\overline{E} = \sum_{k=1}^{N} E_k/N$ is the mean of population residuals. The fitted values $\hat{Y}_k = A + B \times Z_k$ are calculated from the population values. An approximate estimator of the design variance of $\hat{t}_{reg}$ under SRSWOR design is given by substituting S_E^2 by an estimate $\hat{s}_{\hat{e}}^2 = \sum_{k=1}^{n}(\hat{e}_k - \overline{\hat{e}})^2/(n-1)$, where $\hat{e}_k = y_k - \hat{y}_k$ and $\overline{\hat{e}} = \sum_{k=1}^{n}\hat{e}_k/n$. Fitted values $\hat{y}_k = \hat{a} + \hat{b} \times z_k$ are calculated from the sample values. An alternative, more conservative estimator, which uses g-weights is given by

$$\hat{v}_{srs}(\hat{t}_{reg}) = N^2\left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)\left(\frac{n-1}{n-p}\right) \times \hat{s}_{\hat{e}^*}^2, \tag{3.29}$$

where $\hat{s}_{\hat{e}^*}^2 = \sum_{k=1}^{n}(\hat{e}_k^* - \overline{\hat{e}}^*)^2/(n-1)$, $e_k^* = g_k \times e_k^*$, $\overline{\hat{e}}^* = \sum_{k=1}^{n}\hat{e}_k^*/n$ and p is the number of estimated model parameters.

The improvement gained in regression estimation, as compared to the corresponding simple-random-sampling estimators, depends on the value of the finite-population correlation coefficient $\rho_{yz} = S_{yz}/(S_y S_z)$ between the variables y and z. This can be seen by writing the approximate variance (3.28) in the form

$$V_{srs}(\hat{t}_{reg}) \doteq N^2\left(1 - \frac{n}{N}\right)\left(\frac{1}{n}\right)S_y^2(1 - \rho_{yz}^2). \tag{3.30}$$

It will be noted that the value of the correlation coefficient has a decisive influence on the possible improvement of the regression estimation. If ρ_{yz} is zero, the variance of the regression estimator $\hat{t}_{reg}$ equals that of the SRSWOR counterpart $\hat{t}$. But with a nonzero correlation coefficient, the variance obviously decreases.

Under certain conditions, the regression estimator of a total is more efficient than the ratio estimator. This will be demonstrated below by considering the variances of the SRSWOR estimator, the ratio estimator and the regression estimator. Simple random sampling without replacement is assumed, and the constant (c) given in the formulae represents $c = N^2(1 - (n/N))(1/n)$ The variances are

Design-based estimator	$V_{srs}(\hat{t}) = cS_y^2$
Ratio estimator	$V_{srs}(\hat{t}_{rat}) = c(S_y^2 + R^2S_z^2 - 2R\rho_{yz}S_yS_z)$
Regression estimator	$V_{srs}(\hat{t}_{reg}) = cS_y^2(1 - \rho_{yz}^2)$

Studying the relationship between the regression coefficient B and the ratio $R = T/T_z$ will reveal the condition where the regression-estimated total is more efficient than the ratio-estimated total. To find this condition, the difference between the two variances is

$$\begin{aligned} V_{srs}(\hat{t}_{rat}) - V_{srs}(\hat{t}_{reg}) &= c[(S_y^2 + R^2S_z^2 - 2R\rho_{yz}S_yS_z) - S_y^2 + \rho_{yz}^2S_y^2] \\ &= c[(R^2S_z^2 - 2R\rho_{yz}S_yS_z) + \rho_{yz}^2S_y^2]. \end{aligned}$$

Regression estimation is more efficient if the difference is positive:

$$R^2S_z^2 - 2R\rho_{yz}S_yS_z + \rho_{yz}^2S_y^2 > 0.$$

The condition can be rewritten as

$$-\rho_{yz}^2S_{y-}^2 < R^2S_z^2 - 2R\rho_{yz}S_yS_z.$$

By dividing the inequality above by S_z^2 and inserting $\rho_{yz} = S_{yz}/S_yS_z$ and $B = S_{yz}/S_z^2$, gives

$$-B^2 < R^2 - 2RB.$$

Regression estimation, then, is more efficient than ratio estimation if

$$(B - R)^2 > 0.$$

Thus the squared difference between the finite-population regression coefficient and the ratio determines when the regression estimation is more efficient.

Regression estimation can also be applied using a multiple regression model as the assisting model. We postulate a linear regression model between the study variable y and p continuous auxiliary variables $z_1, z_2, \ldots, z_p$, given by

$y_k = \alpha + \beta_1 z_{1k} + \beta_2 z_{2k} + \cdots + \beta_p z_{pk} + \varepsilon_k$, where α refers to the intercept and $\beta_j, j = 1, \ldots, p$, are the slope parameters, and ε_k is the residual. For multiple regression estimation, we assume that the population totals $T_{z_1}, T_{z_2}, \ldots, T_{z_p}$ are known for each auxiliary variable. They can come from some source outside the survey, such as published official statistics. The regression estimator of the population total T of y is now given by

$$\hat{t}_{reg} = \hat{t} + \hat{b}_1(T_{z_1} - \hat{t}_{z_1}) + \hat{b}_2(T_{z_2} - \hat{t}_{z_2}) + \cdots + \hat{b}_p(T_{z_p} - \hat{t}_{z_p}), \tag{3.31}$$

where the estimated regression coefficients $\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_p$ are obtained from the sample data set using weighted least squares estimation with $w_k = 1/\pi_k$ as the weights. The estimators $\hat{t}$ and $\hat{t}_{z_j}, j = 1, \ldots, p$, refer to Horvitz–Thompson estimators.

A different form, often referred to as the *generalized regression* (GREG) *estimator* (Särndal *et al.* 1992) is given by

$$\hat{t}_{reg} = \sum_{k=1}^{N} \hat{y}_k + \sum_{k=1}^{n} w_k(y_k - \hat{y}_k), \tag{3.32}$$

where $\hat{y}_k = \hat{a} + \hat{b}_1 z_{1k} + \hat{b}_2 z_{2k} + \cdots + \hat{b}_p z_{pk}$ are fitted values calculated using the estimated regression coefficients and the known values of z-variables. Note the difference between (3.31) and (3.32). In the former we only need to know the population totals of the auxiliary z-variables, but in the latter, the individual values of z-variables are assumed known for every population element (because the first summation is over all N population elements). Thus, (3.32) requires more detailed information on the population than (3.31). Micro-level auxiliary z-data may indeed be available, for example, in a statistical infrastructure where population census registers or similar statistical registers, compiled from various administrative registers, are used as sampling frames. In this case, the frame population often includes the necessary auxiliary z-data at a micro-level (see Chapter 6).

Let us consider the expression (3.32) for a multiple regression estimator in more detail. It is obvious that if the weights are equal for all sample elements, and ordinary least squares estimation had been used for a model that includes an intercept, then the latter part of (3.32) vanishes, and the regression estimate reduces to the sum of the fitted values over the population. This is the case for a self-weighting design such as simple random sampling. But if the weights vary between elements, then the sum of weighted residuals can differ from zero, as can happen for example in stratified SRS with non-proportional allocation. In such cases, the latter part of (3.32) serves as a bias adjustment factor protecting against model misspecification.

Under SRSWOR, an approximate design variance given in (3.28) can be applied by using the fitted values $\hat{Y}_k = A + B_1 Z_{1k} + \cdots + B_p Z_{pk}$. A variance estimator is

obtained by replacing $\hat{Y}_k$ by sample-based fitted values $\hat{y}_k = \hat{a} + \hat{b}_1 z_{1k} + \cdots + \hat{b}_p z_{pk}$. An alternative variance estimator is calculated as

$$\hat{v}_{srs}(\hat{t}_{reg}) = \hat{v}_{srs}(\hat{t})(1 - \hat{R}^2), \tag{3.33}$$

where the multiple correlation coefficient squared $\hat{R}^2$ is calculated for the sample data set. Because this term is always non-negative, the multiple regression estimator is always at least as efficient as simple random sampling without replacement. Efficiency improves when multiple auxiliary z-data that correlates with the study variable y are incorporated in the estimation procedure.

In the next example, we compute a regression-estimated total from a sample data set, first in a single auxiliary variable case and then in the context of multiple regression estimation.

Example 3.13

Single Auxiliary Variable

Regression estimation of the total in the *Province'91* population. The previously selected simple random sample is used. There, the study variable UE91 is regressed with the auxiliary variable HOU85. We conduct regression estimation in two ways, resulting in equal estimates. HOU85 is first used as the predictor and an estimate $\hat{t}_{reg}$ is computed using the estimated slope $\hat{b}$. In Table 3.16, the sample identifiers correspond to the SRSWOR case, and the sampling rate is, as previously, 0.25.

Using UE91 as the dependent variable and HOU85 as the predictor, the slope is estimated as $\hat{b} = 0.152$, giving

$$\hat{t}_{reg} = \hat{t} + \hat{b}(T_z - \hat{t}_z) = 26\,440 + 0.152(91\,753 - 164\,952) = 15\,312.$$

Table 3.16 A simple random sample drawn without replacement from the *Province'91* population prepared for regression estimation.

Sample design identifiers			Element	Study var.	Auxiliary information			
					Variable	Model	WGHT	
STR	CLU	WGHT	LABEL	UE91	HOU85	group	g-weight	w*-weight
1	1	4	Jyväskylä	4123	26 881	1	0.2844	1.1378
1	4	4	Keuruu	760	4896	1	1.0085	4.0341
1	5	4	Saarijärvi	721	3730	1	1.0469	4.1877
1	15	4	Konginkangas	142	556	1	1.1057	4.6058
1	18	4	Kuhmoinen	187	1463	1	1.1216	4.4863
1	26	4	Pihtipudas	331	1946	1	1.1391	4.4227
1	30	4	Toivakka	127	834	1	1.1423	4.5691
1	31	4	Uurainen	219	932	1	1.1515	4.5562

Sampling rate = 8/32 = 0.25.

The same point estimate is obtained using the calibrated weights by calculating $\hat{t}_{reg} = \sum_{k=1}^{8} w_k^* y_k = 15\,312$ (see Table 3.16). For variance estimation, the formula (3.29) or (3.33) can be used. The former gives a conservative estimate especially if the sample size is small as is the case here. Thus, by (3.29) we obtain

$$\hat{v}_{srs}(\hat{t}_{reg}) = N^2 \left(1 - \frac{n}{N}\right) \left(\frac{1}{n}\right) \left(\frac{n-1}{n-p}\right) \times \hat{s}^2_{\hat{e}^*}$$
$$= 32^2 \left(1 - \frac{8}{32}\right) \left(\frac{8-1}{8-2}\right) \left(\frac{1}{8}\right) \times 61.24^2 = 648^2.$$

The corresponding design-based total estimate obtained under SRSWOR was $\hat{t} = 26\,440$, whose standard error was 13 282. Therefore, the deff estimate is deff $= 648^2/13\,282^2 = 0.002$, which is almost zero and is persuasive evidence of the superiority of regression estimation over design-based estimation for the present estimation problem. Improved efficiency is due to the strong linear relationship between UE91 and HOU85.

Multiple Regression Model

Multiple regression estimation of the total in the *Province'91* population. Here, the study variable UE91 is regressed with two auxiliary variables, HOU85 and a variable named URB85 with a value 1 for urban municipalities and zero otherwise (see Table 2.1). We use both the formula (3.31) and the GREG method with equation (3.32). First, the estimated regression coefficients $\hat{b}_1$ and $\hat{b}_2$ are calculated by fitting a two-predictor regression model for the sample data set of $n = 8$ municipalities, as given in Table 3.16. The estimates are $\hat{b}_1 = 0.14956$ and $\hat{b}_2 = +68.107$. The estimated totals of auxiliary variables are $\hat{t}_{z_1} = 164\,952$, as previously, and $\hat{t}_{z_2} = 12$. In addition, we use the known population totals $T_{z_1} = 91\,753$ and $T_{z_2} = 7$. Using (3.31), we obtain

$$\hat{t}_{reg} = \hat{t} + \hat{b}_1(T_{z_1} - \hat{t}_{z_1}) + \hat{b}_2(T_{z_2} - \hat{t}_{z_2}) = 26\,440 + 0.14956\,(91\,753 - 164\,952)$$
$$+ 68.107\,(7 - 12) = 15\,152.$$

Using (3.32), we first calculate the fitted values for all population elements. The sum of the fitted values over the population provides the desired regression estimate. The GREG estimation procedure is summarized in Table 3.17. There also, the estimate 15 152 can be obtained. Note that in the SRSWOR case in which the sampling weights are equal, the sum of the residuals over the sample data set is equal to zero.

• Q2 Calculating the multiple correlation coefficient squared •$R^2 = 0.998$ for the sample data set, we obtain the variance estimate of $\hat{t}_{reg}$ by (3.33), $\hat{v}(\hat{t}_{reg}) = 569^2$, which is smaller than in the previous case where HOU85 was used as the only auxiliary variable. There, an estimate $\hat{v}(\hat{t}_{reg}) = 648^2$ was obtained. Hence, multiple regression estimation appeared to be slightly more efficient in this case. The design effect estimate is now deff $= 569^2/13\,282^2 = 0.0018$.

Table 3.17 Population frame merged with sample data for multiple regression estimation. Simple random sample drawn without replacement from the *Province'91* population.

Population frame					Sample		Model fitting	
ID k	LABEL	URB85 z_{1k}	HOU85 z_{2k}	Sample indicator	WGHT w_k	UE91 y_k	Fitted value $\hat{y}_k$	Residual $\hat{e}_k$
1	**Jyväskylä**	**1**	**26 881**	**1**	**4**	**4123**	**4118.15**	**4.85**
2	Jämsä	1	4663	0	...	...	795.27	...
3	Jämsänkoski	**1**	3019	0	...	...	549.40	...
4	**Keuruu**	1	**4896**	**1**	**4**	**760**	**830.12**	**−70.12**
5	**Saarijärvi**	**1**	**3730**	**1**	**4**	**721**	**655.73**	**65.27**
6	Suolahti	**1**	2389	0	...	...	455.18	...
7	Äänekoski	**1**	4264	0	...	...	735.60	...
8	Hankasalmi	0	2179	0	...	...	355.66	...
9	Joutsa	0	1823	0	...	...	302.42	...
10	J:kylä mlk.	0	9230	0	...	...	1410.20	...
11	Kannonkoski	0	726	0	...	...	138.36	...
12	Karstula	0	1868	0	...	...	309.15	...
13	Kinnula	0	675	0	...	...	130.73	...
14	Kivijärvi	0	634	0	...	...	124.60	...
15	**Konginkangas**	**0**	**556**	**1**	**4**	**142**	**112.93**	**29.07**
16	Konnevesi	0	1215	0	...	...	211.49	...
17	Korpilahti	0	1793	0	...	...	297.93	...
18	**Kuhmoinen**	**0**	**1463**	**1**	**4**	**187**	**248.58**	**−61.58**
19	Kyyjärvi	0	672	0	...	...	130.28	...
20	Laukaa	0	4952	0	...	...	770.39	...
21	Leivonmäki	0	545	0	...	...	111.29	...
22	Luhanka	0	435	0	...	...	94.83	...
23	Multia	0	925	0	...	...	168.12	...
24	Muurame	0	1853	0	...	...	306.91	...
25	Petäjävesi	0	1352	0	...	...	231.98	...
26	**Pihtipudas**	**0**	**1946**	**1**	**4**	**331**	**320.82**	**10.18**
27	Pylkönmäki	0	473	0	...	...	100.52	...
28	Sumiainen	0	485	0	...	...	102.31	...
29	Säynätsalo	0	1226	0	...	...	213.13	...
30	**Toivakka**	0	**834**	**1**	**4**	**127**	**154.51**	**−27.51**
31	**Uurainen**	0	**932**	**1**	**4**	**219**	**169.16**	**49.84**
32	Viitasaari	0	3119	0	...	...	496.25	...
	Sum	**7**	**91 753**	**8**	**32**	**6610**	**15 151.98**	**0.00**

. . .Non-sampled elements.

Regression estimation was illustrated in simple cases where one or two auxiliary variables were used and SRSWOR was assumed. The method can also be applied for more complex designs, and multiple auxiliary variables can be incorporated in the estimation. For this, weighted least squares regression can also be used. Although the use of multivariate regression models for regression estimation is technically straightforward, there are certain complexities when compared

to regression estimation under simple random sampling, such as the possible multicollinearity of the predictor variables. Another generalization is also obvious since discrete covariates can also be incorporated into a linear model. Using this kind of auxiliary variables for regression estimation leads to analysis-of-variance-type models. Further extensions are discussed in Chapter 6 in connection with the estimation for population subgroups.

Comparison of Estimation Strategies

For model-assisted estimation, we created three sets of new weights, denoted w^*. First, we check the calibration property of these weights. For ratio estimation, the calibration equation for the auxiliary variable z is

$$\sum_{k=1}^{n} w_k^* \times z_k = T_z$$

where $T_z = \sum_{k=1}^{N} Z_k = 91\,753$. This holds for the regression estimator as well.

We next compare the model-assisted estimation results obtained previously from a sample drawn with SRSWOR from the *Province'91* population. More specifically, poststratification, ratio estimation and regression estimation results for the population total T of UE91 are compared. The design-based estimate using the standard SRS formula is also included (see Table 3.18). The known population total $T = 15\,098$ of UE91 is the reference figure.

Two obvious conclusions can be drawn. Firstly, point estimates calculated using auxiliary information are closer to the population total than the design-based estimate. Secondly, the model-assisted estimators are much more efficient than SRSWOR.

The poststratified estimator uses, as discrete auxiliary information, the administrative division of municipalities into urban and rural municipalities. Improved

Table 3.18 Estimates for the population total of UE91 under different estimation strategies: an SRSWOR sample of eight elements drawn from the *Province'91* population.

Estimation strategy		Estimator	Estimate	s.e	deff
Desing-based					
SRSWOR		$\hat{t}_{srswor}$	26 440	13 282	1.0000
SRSWR		$\hat{t}_{srswr}$	26 440	15 095	1.2917
Design-based model-assisted					
Poststratified estimator		$\hat{t}_{pos}$	18 106	6021	0.3323
Ratio estimator		$\hat{t}_{rat}$	14 707	892	0.0045
Regression estimator	one z-variable	$\hat{t}_{reg,1}$	15 312	648	0.0020
	two z-variables	$\hat{t}_{reg,2}$	15 152	569	0.0018

estimates result, since this division is in relation to the variation of the study variable in such a way that the variation of unemployment figures is smaller in the poststrata than in the whole population. But the relation is not as strong as that between UE91 and the continuous auxiliary variable HOU85, the number of households. This can be seen from the ratio and regression estimation results. Because ratio estimation assumes that the regression line of UE91 and HOU85 goes through the origin, and this is not the case, regression estimation performs slightly better than ratio estimation.

Summary

Using auxiliary information from the population in the estimation of a finite-population parameter of interest is a powerful tool to get more precise estimates, if the variation of the study variable has some strong relationship with an auxiliary covariate. If so, efficient estimators can be obtained such that they produce estimates close to the true population value and have a small standard error. The auxiliary variable can be a discrete variable, in which case poststratification can be used. If the covariate is a continuous variable, ratio estimation or regression estimation is appropriate.

Model-assisted estimation is often used in descriptive surveys to improve the estimation of the population total of a study variable of interest, whereas in multi-purpose studies, where the number of study variables may be large, it may be difficult to find good auxiliary covariates for this purpose. In such surveys, however, poststratification is often used to adjust for nonresponse.

We have examined here the elementary principles of model-assisted estimation supplemented with computational illustrations. For more details, the reader is encouraged to consult Särndal *et al.* (1992); there, model-assisted survey sampling covering poststratification, ratio estimation and regression estimation is extensively discussed. These methods are considered as special cases of *generalized regression estimation* which is used in many statistical agencies in the production of official statistics (for example Estevao *et al.* 1995). A clear overview of poststratification can be found in Holt and Smith (1979). Further, as a generalization of poststratification, Deville and Särndal (1992) and Deville *et al.* (1993) consider a class of weights calibrated to known marginal totals. Silva and Skinner (1997) address the problem of variable selection in regression estimation.