2003

an introduction to model-based survey sampling

ereduetan oinarritutako laginketari sarrera

Introducción al muestreo basado en modelos

RAY CHAMBERS

42

nazioarteko estatistika mintegia euskadin

seminario internacional de estadística en euskadi

AURKEZPENA

Nazioarteko Estatistika Mintegia antolatzean, hainbat helburu bete nahi ditu EUSTAT- Euskal Estatistika Erakundeak:

- Unibertsitatearekiko eta, batez ere, Estatistika-Sailekiko lankidetza bultzatzea.
- Funtzionarioen, irakasleen, ikasleen eta estatistikaren alorrean interesatuta egon daitezkeen guztien lanbide-hobekuntza erraztea.
- Estatistika alorrean mundu mailan abangoardian dauden irakasle eta ikertzaile ospetsuak Euskadira ekartzea, horrek eragin ona izango baitu, zuzeneko harremanei eta esperientziak ezagutzeari dagokienez.

Jarduera osagarri gisa, eta interesatuta egon litezkeen ahalik eta pertsona eta erakunde gehienetara iristearren, ikastaro horietako txostenak argitaratzea erabaki dugu, beti ere txostengilearen jatorrizko hizkuntza errespetatuz; horrela, gai horri buruzko ezagutza gure herrian zabaltzen laguntzeko.

Vitoria-Gasteiz, 2003ko Iraila

JOSU IRADI ARRIETA
EUSTATeko Zuzendari Nagusia

PRESENTATION

In promoting the International Statistical Seminars, EUSTAT-The Basque Statistics Institute wishes to achieve several aims:

- Encourage the collaboration with the universities, especially with their statistical departments.
- Facilitate the professional recycling of civil servants, university teachers, students and whoever else may be interested in the statistical field.
- Bring to the Basque Country illustrious professors and investigators in the vanguard of statistical subjects, on a worldwide level, with the subsequent positive effect of encouraging direct relationships and sharing knowledge of experiences.

As a complementary activity and in order to reach as many interested people and institutions as possible, it has been decided to publish the papers of these courses, always respecting the original language of the author, to contribute in this way towards the growth of knowledge concerning this subject in our country.

Vitoria-Gasteiz, Septiembre 2003

JOSU IRADI ARRIETA
General Director of EUSTAT

PRESENTACION

Al promover los Seminarios Internacionales de Estadística, el EUSTAT-Euskal Estatistika Erakundea pretende cubrir varios objetivos:

- Fomentar la colaboración con la Universidad y en especial con los Departamentos de Estadística.
- Facilitar el reciclaje profesional de funcionarios, profesores, alumnos y cuantos puedan estar interesados en el campo estadístico.
- Traer a Euskadi a ilustres profesores e investigadores de vanguardia en materia estadística, a nivel mundial, con el consiguiente efecto positivo en cuanto a la relación directa y conocimiento de experiencias.

Como actuación complementaria y para llegar al mayor número posible de personas e Instituciones interesadas, se ha decidido publicar las ponencias de estos cursos, respetando en todo caso la lengua original del ponente, para contribuir así a acrecentar el conocimiento sobre esta materia en nuestro País.

Vitoria-Gasteiz, Septiembre 2003

JOSU IRADI ARRIETA
Director General de EUSTAT

BIOGRAFI OHARRAK

RAY CHAMBERS. Southamptoneko (Erresuma Batua) Unibertsitateko Gizarte-Estatistika Saileko katedraduna eta zuzendaria, lagin-estatistiken diseinu eta analisien ikerketan, estatistika ofizialaren metodologian, inferentzia estatistikoko metodo sendoetan eta talde-egituradun datuen analisian nazioartean entzute handikoa.

BIOGRAPHICAL SKETCH

RAY CHAMBERS. Head of the Department of Social Statistics of the University of Southampton with extensive research interests in the design and analysis of sample surveys, official statistics methodology, robuts methods for statistical inference and analysis of data with group structure.

NOTAS BIOGRAFICAS

RAY CHAMBERS. Catedrático y Director del Departamento de Estadística Social de la Universidad de Southampton (Reino Unido), de reconocido prestigio internacional en la investigación del diseño y análisis de encuestas muestrales, metodología de estadística oficial, métodos robustos de inferencia estadística y análisis de datos con estructura de grupo.

AN INTRODUCTION TO MODEL-BASED SURVEY SAMPLING

Ray Chambers
Southampton Statistical Sciences Research Institute
University of Southampton
Highfield
Southampton SO17 1BJ UK

Email: rc6@soton.ac.uk

CONTENTS

4. METHODS OF VARIANCE ESTIMATION

5. ESTIMATION FOR MULTIPURPOSE SURVEYS

6. ESTIMATION FOR DOMAINS AND SMALL AREAS

1. Fundamental Concepts

This short monograph is about the use of models in survey sampling. In this context we need to first introduce some basic ideas that are fundamental to sampling. To start, we note that it is meaningless to talk about a sample without referring to the population from which it was taken.

The target population of a survey is the population at which the survey is aimed, i.e. the population of interest. In contrast, the actual population from which the survey sample is drawn is called the survey population.

The coverage of the survey population is the degree to which target and survey population overlap. Here we assume there is no difference between target and survey population, i.e. we have complete coverage.

1.1 Sample Frames and Auxiliary Information

A standard method of sampling is to select the sample from a list (or lists) that enumerate the units making up the survey population. This is usually referred to as the (sampling) frame for the survey. The information on the frame is crucial for sample design, e.g. stratified sampling requires the frame to contain enough identifying information about each population unit for its stratum membership to be determined. This information is typically referred to as auxiliary information. For economic populations, the frame may also contain values for each unit in the population that characterise its level of economic activity. These so-called measures of size are another example of auxiliary information.

1.2 Non-Informative Sampling

A sampling method is said to be non-informative for a variable Y if the distribution of the sampled values of Y and the distribution of the non-sampled values of this variable are the same. Non-informative sampling methods are extremely important since a non-informative sampling method allows inference about the non-sampled units in a population on the basis of sample information.

Typically the distributions of interest in survey inference are conditioned on the variables on the frame. Consequently any sampling method whose outcome is determined entirely by the data on the frame is non-informative for inference about the parameters of distributions that condition on these frame values.

1.3 Probability Sampling

A probability sampling method is one that uses a randomisation device to decide which units on the frame are in sample. This means that it is not possible to specify in advance precisely which units on the frame make up the sample under this type of sampling. We note that randomised samples are free of the (often hidden) biases that can occur with sampling methods that are not probability-based. Furthermore, randomised samples are non-informative provided inclusion probabilities are determined entirely by frame data.

1.4 Basic Assumptions

Drawing together the ideas set out above, we shall make the following basic assumptions in the subsequent development

(1) A perfect sample frame exists. That is, we have access to a frame that lists every unit in the population once and only once, and there is a known number *N* of such units.

(2) A non-informative sampling method used to draw the sample from the frame.

Typically (2) is ensured by use of some form of probability sampling, where every unit on the frame has a non-zero probability of selection into the sample. We again note that such a sampling method allows sample data to be used to estimate parameters of the distribution of non-sample data. This is extremely important for the model-based prediction approach developed below.

1.5 Population Variables

A basic aim of a sample survey is to allow inference about one or more characteristics of the population. Typically these are defined by the values of one or more population variables.

A population variable is a quantity that is defined for every unit in the population, and is observable when that unit is included in sample. In practice, surveys are concerned with many population variables. However, most of the theory for sample surveys is developed for a small number of variables, typically one or two. These variables will be referred to as study or *Y*-variables in what follows. In addition, we note another class of variables, defined by those variables with values recorded on the sample frame. These frame variables are known for every unit in the population. We refer to them as auxiliary or *X*-variables in what follows.

Example

The quarterly survey of capital expenditure (CAPEX) is a business survey carried out by the U.K. Office for National Statistics (ONS) that has several study (*Y*) variables. These include collected variables like acquisitions and disposals as well as derived variables like net capital expenditure, defined as the difference between acquisitions and disposals. The frame for CAPEX is the ONS Inter-Departmental Business Register (IDBR). This contains auxiliary variables corresponding to the industry classification (Standard Industry Classification) of a business, the number of employees of the business and its total turnover in the preceding year.

1.6 Finite Population Parameters

The population characteristic(s) that are the focus of a sample survey are sometimes referred to as its target(s) of inference. These are typically well-defined function(s) of the population values of the *Y*-variables of interest. We refer to such quantities as finite population parameter(s) or census parameters. Some common examples (all defined with respect to population values of the *Y*-variables of interest) are their totals or averages, the ratios of these averages, their variances, their medians and more generally the quantiles of their finite population distributions.

1.7 Sample Statistics

Once a sample has been selected, and values of *Y*-variables obtained from this sample, we are in a position to calculate the values of various quantities based on these data. These are typically referred to as sample statistics. Sample survey theory is concerned with the behaviour of two types of such sample statistics:

(i) Statistics that estimate the census parameters of interest;

(ii) Statistics that measure the quality of these estimates.

1.8 Sample Error and Sample Error Distribution

The sample error of a survey estimate is the difference between the value of this estimate and the unknown value of the census parameter it estimates. A high quality survey estimator will have a small sample error. But, since the actual value of the census parameter being estimated is unknown, the sample error of its estimator is also unknown.

If we can specify a distribution for the sample error, then we can measure the quality of the survey estimate in terms of the characteristics of this distribution. Thus, we define the bias of an estimator as the central location of its sample error distribution, as defined by its mean or expectation. Similarly, we define the variance of an estimator as the spread of this distribution around this mean. The mean squared error (MSE) of an estimator is then its variance plus the square of its bias. Consequently a high quality survey estimator will have a sample error distribution that has bias close to or equal to zero and a low variance, i.e. a small mean squared error.

1.9 Which Distribution? The Repeated Sampling Distribution

The repeated sampling distribution of the sample error is the distribution of all possible values that this sample error can take under repetition of the sampling method. This corresponds to repeating the sampling process, selecting sample after sample from the population, calculating the value of the estimate for each sample, generating a (potentially) different sample error each time and hence a distribution for these errors.

We note that the repeated sampling distribution treats the population values as fixed. Consequently, the variability underlying this distribution arises from the different samples selected under the sample selection method. This means that sample selection methods that are not probability based are not suited to evaluation under this distribution.

1.10 Another Distribution: The Concept of a Population Model

A population of Y-values can be thought of as the outcome of many chance occurrences that can be characterised in terms of a statistical model. This model describes the range of possible Y-values that can occur and imposes a probability measure on the chance of occurrence of any particular range of values. Such models are usually based on past exposure to data from other populations very much like the one of interest as well as subject matter knowledge about how the population values ought to be distributed.

Statistical models for populations are often referred to as superpopulation models. Such models therefore constitute a statistical "description" of the distribution of the population *Y*-values, in the sense that the *N* population values are assumed to be realisations of *N* random variables whose joint distribution is described by the model. Since the sample statistics and the census parameters they estimate are all functions of these random variables, it is clear that the assumption of a superpopulation distribution immediately induces a distribution for a sample error.

Such a superpopulation model will typically depend on unknown parameters.

1.11 The Repeated Sampling Distribution vs. the Superpopulation Distribution

The superpopulation distribution for a sample error is not the same as the repeated sampling distribution of this error. The only source of variability for the repeated sampling distribution is the sample selection method, which in turn is characterised by the distribution of the values of the sample inclusion variables. It also treats the population *Y*-values as fixed.
In contrast, the underlying variability for the superpopulation distribution arises because of the distribution of values of population variables. This variability has nothing to do with the sampling process. Here values of sample inclusion variables as treated as fixed and variability arises because of the variability of population *Y*-values.

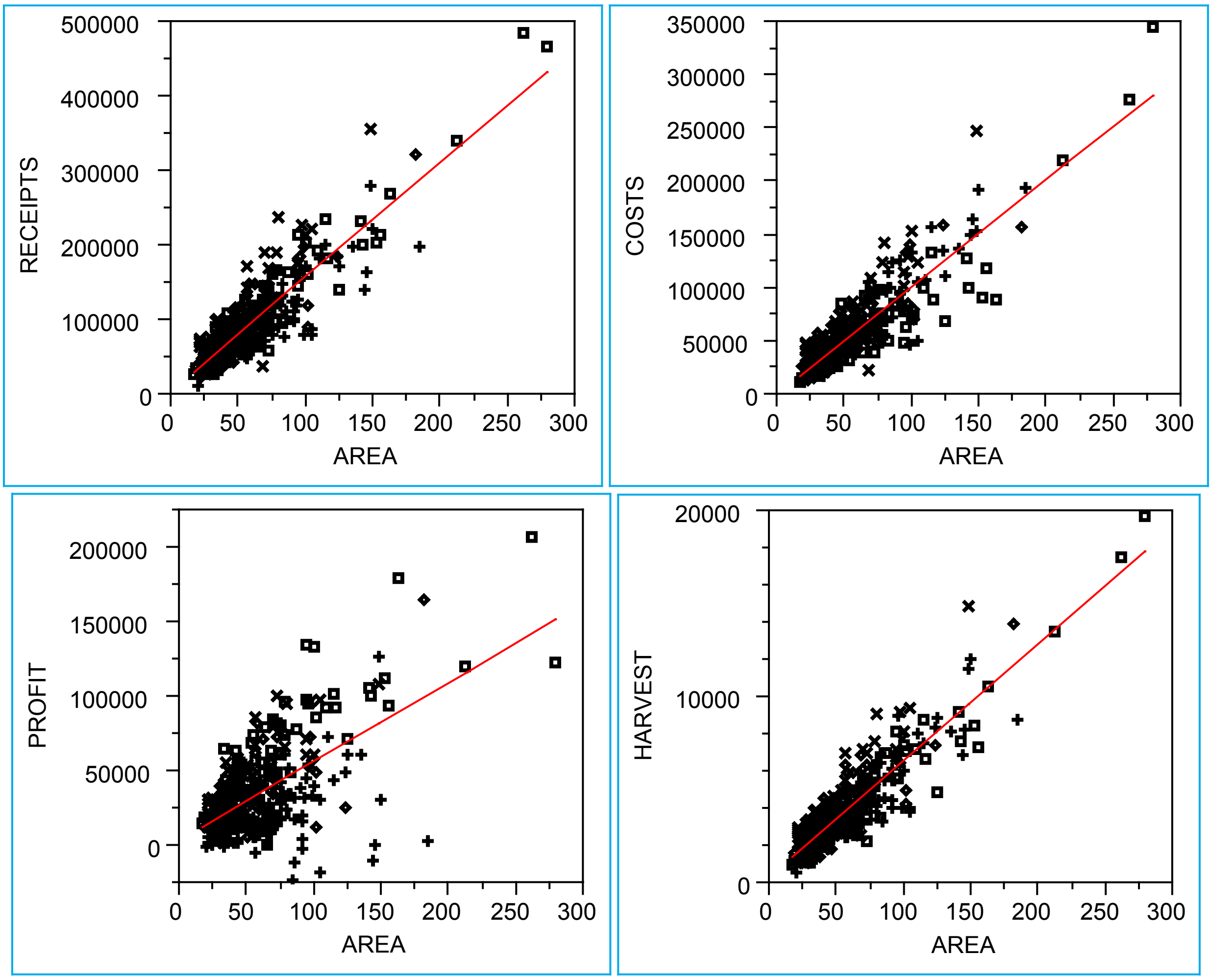
Note that implicit in the use of the superpopulation distribution as the source of variability for the sample error is the assumption that the sampling method is non-informative - i.e. superpopulation distribution of population variables in population and sample is the same.

In what follows we focus on the use of the superpopulation distribution for inference about census parameters. In doing so, we will typically use a subscript of ξ to denote moments with respect to this superpopulation distribution.

1.12 How to specify the superpopulation distribution?

Consider the following four plots that show the relationship between the values of four *Y*-variables (Receipts, Costs, Profit and Harvest) and an *X*-variable (Area for growing sugarcane) measured in a survey of sugarcane farms in Australia in the mid 1980s (n = 338). These show a strong linear relationship. The least squares fit to each plot is also shown, with the coefficients (and associated t-statistics) of this fit set out in the table below.

	RECEIPTS	**COSTS**	**PROFIT**	**HARVEST**
Intercept	3439.82	142.65	3297.17	281.99
t ratio	1.13	0.07	1.27	2.45
AREA	1535.73	1005.34	530.40	62.85
t ratio	35.33	35.17	14.29	38.26



Note how in each case the regression line goes "essentially" through the origin, and the variability about this line tends to increase the further one moves away from the origin. This pattern of behaviour is very common for economic variables. In this case there can be no sugar related economic activity if no sugar is grown and the returns and costs associated with this activity vary on a per unit basis as the amount of effort, i.e. area, increases. All four plots are consistent with the widely used population model underpinning the application of the method of ratio estimation in business surveys. This is

$$E_{\xi}(y_i) = \beta x_i,$$
$$Var_{\xi}(y_i) = \sigma^2 x_i,$$
$$Cov_{\xi}(y_i, y_j) = 0.$$

For data collected at household level in social surveys the linearity assumption often remains valid, but there is typically no reason why the regression should go through the origin or why variability should increase with increasing *X*. In this case a model appropriate for standard regression estimation is often appropriate. This is given by

$$E_{\xi}(y_i) = \alpha + \beta x_i,$$
$$Var_{\xi}(y_i) = \sigma^2,$$
$$Cov_{\xi}(y_i, y_j) = 0.$$

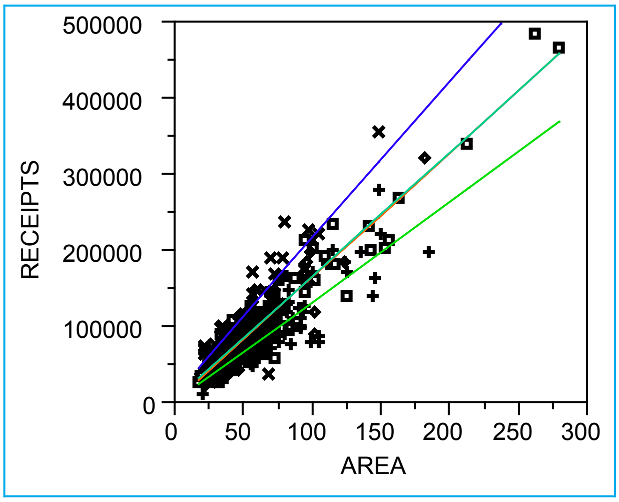
Both the ratio and regression models are special cases of a general regression model for data collected at PSU level (hence the assumption of uncorrelated errors) given by

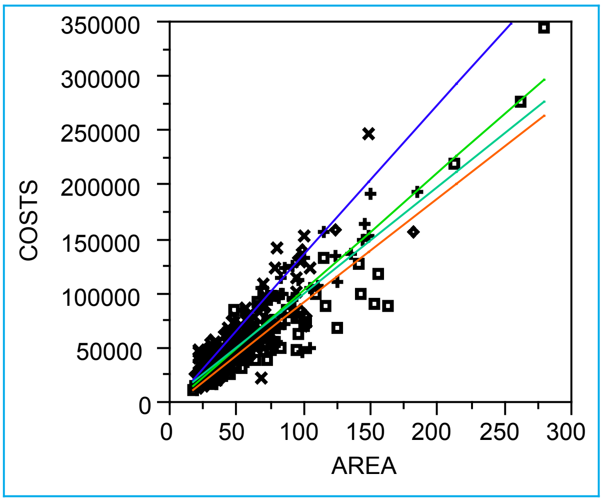
$$\begin{aligned}
&E_{\xi}(y_i) = \mu(x_i;\omega),\\
&Var_{\xi}(y_i) = \sigma^2(x_i;\omega),\\
&Cov_{\xi}(y_i, y_j) = 0,
\end{aligned}$$

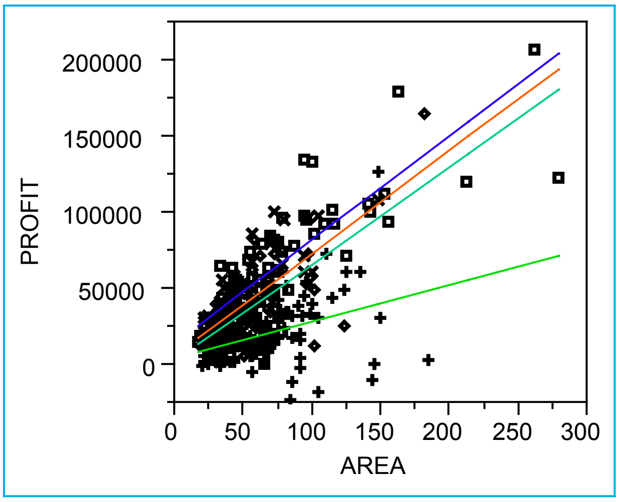
where $\mu(x)$ and $\sigma(x)$ are specified functions of x whose values depend on ω, an unknown (typically vector valued) parameter.

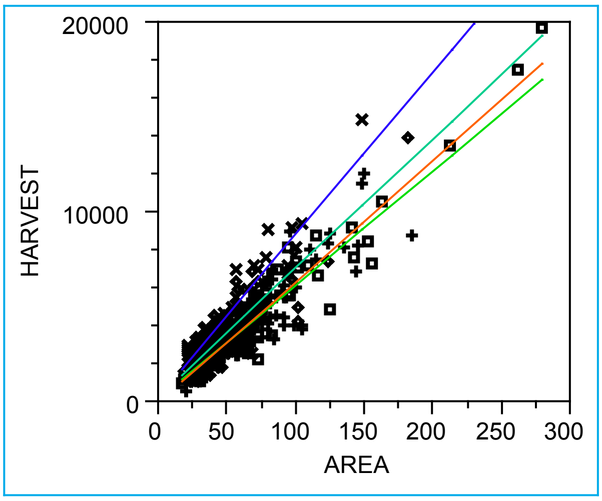
It is highly unlikely that any particular survey population will be adequately modelled via the same relationship between Y and X holding everywhere. In this case it is usual to split the target population into strata and apply a stratified sample design. The simplest model for this type of population is where population units are mutually uncorrelated, with means and variances of the population Y-variables the same for all units within a stratum, but different across strata. This is a special case of the general PSU level model above, and we refer to it as the Homogeneous Strata Model in what follows. It is widely used (typically implicitly) in surveys. Here X is a stratum indicator, with strata indexed by h = 1, 2, .., H, and for a population unit i in stratum h: $\mu(x_i;\omega) = \mu_h$ and $\sigma(x_i;\omega) = \sigma_h$. Note this model does not assume any relationship between the stratum means and variances.

Regression models are often combined with stratum effects. To illustrate, we return to the sugarcane farm data and note that these farms actually drawn from four separate regions. In the four plots below the region specific regression lines are displayed, showing clear regional differences in slope. Furthermore, as the values in the accompanying table demonstrate, it is only in one region and for one variable that the intercept coefficients associated with these lines are significantly different from zero. A better model for this population is therefore one where a different version of the ratio model defined earlier held in each region. We call this a Stratified Ratio Model.









RECEIPTS	REGION=1	REGION=2	REGION=3	REGION=4
Intercept	-13.49	10065.20	820.42	2928.41
t ratio	-0.00	1.01	0.19	0.63
AREA	1319.66	2061.99	1635.00	1627.43
t ratio	21.67	12.87	31.30	18.88

COSTS	REGION=1	REGION=2	REGION=3	REGION=4
Intercept	-4629.67	-2770.53	-4709.07	1266.65
t ratio	-1.51	-0.41	-1.22	0.39
AREA	1078.56	1379.30	961.43	987.74
t ratio	24.94	12.58	21.02	16.20

PROFIT	REGION=1	REGION=2	REGION=3	REGION=4
Intercept	4616.18	12835.73	5529.49	1661.75
t ratio	1.28	2.15	1.21	0.46
AREA	241.105	682.68	673.57	639.69
t ratio	4.73	7.08	12.44	9.51

HARVEST	REGION=1	REGION=2	REGION=3	REGION=4
Intercept	136.43	242.99	-72.33	233.24
t ratio	0.72	0.67	-0.42	1.12
AREA	60.06	85.60	63.83	68.05
t ratio	22.38	14.71	31.04	17.67

Generally we can therefore think of defining a stratified linear regression model for a population. Here the auxiliary information corresponding to X contains a mix of stratum identifiers and size variables, so we have a multivariate auxiliary variable $\mathbf{X}$ and a mean function $\mu(\mathbf{x}_i;\omega) = \mathbf{x}_i'\beta$. Furthermore, we can allow heteroskedasticity (i.e. varying variability) in this model to be defined in terms of a single auxiliary variable Z. This can be one of the auxiliary size variables in $\mathbf{X}$ or some positive valued function of the components of this vector (e.g. a power transformation). In any case we then have $\sigma(\mathbf{x}_i;\omega) = \sigma z_i$.

So far we have developed models for populations where the PSU is also the unit of interest. This is usually not the case in social surveys. In the populations underpinning these surveys it is often the case that the explanatory power of available X-variables is weak and the assumption of lack of correlation between different population units is inappropriate.

For example, many "human" populations are intrinsically hierarchical, with individuals grouped together into small non-overlapping clusters (e.g. households). These clusters are often more or less similar in size, and essentially similar in terms of the range of *Y*-values they contain. This is usually manifested by individuals from the same cluster being more alike than individuals from different clusters.

This type of situation can be modelled by an unobservable "cluster effect" variable γ. In particular we assume that the value of *Y* for unit *j* in cluster *i* satisfies $y_{ij} = \mu + \gamma_i + \varepsilon_{ij}$. Where γ_i and ε_{ij} are independent random variables with zero means and variances $\sigma^2(\gamma)$ and $\sigma^2(\varepsilon)$ respectively. It is easy to see that this model induces a within cluster correlation (ρ) for Y, given by $\rho = \dfrac{\sigma^2(\gamma)}{\sigma^2(\gamma)+\sigma^2(\varepsilon)}$.

Example: 1993 Survey of Psychiatric Morbidity (UK Office for National Statistics, n = 9608)

This was a social survey with PSU defined as postal sector (200 in sample). This therefore defines a cluster. There were two further levels of selection - SSU was defined as address (postal delivery point), and FSU was defined as a randomly selected individual living at the selected address with age in the range 16-64.

Key variables measure in the survey were morbidity score, defined as a score on a clinical interview schedule (revised) with values between 0 and 57, job, defined as working status [0-1] for a slected individual, live alone, defined as an indicator of lifestyle [0-1] for individual, owner/occupier, defined as an indicator of tenure [0-1] for the dwelling and one adult hh, which was an indicator of type [0-1] for household. The table below shows the estimated values of the variance parameters in the cluster model defined above when this model was fitted to the survey data. It is clear that there are quite significant clustering effects in this population.

Variable	$\hat{\sigma}^2(\gamma)$	$\hat{\sigma}^2(\varepsilon)$	$\hat{\rho}$
morbidity score	1.639604	50.85795	0.0312
job	0.008944	0.214525	0.0400
live alone	0.004477	0.137527	0.0315
owner/occupier	0.021737	0.189864	0.1027
one adult hh	0.005715	0.200353	0.0277

2. The Model-Based Approach

In this chapter we develop the essentials of the model-based approach to sample survey design and estimation. For a comprehensive description of this approach, see Valliant, Dorfman and Royall (2000). In order to do so we focus on estimation of the census parameter t_y corresponding to the population total of a *Y*-variable. We denote the estimator of this total by $\hat{t}_y$. Under the model-based approach, the properties of this estimator are determined by the distribution of its sample error under the assumed model for the population. See Brewer (1963) and Royall (1970). Consequently, we focus on the superpopulation distribution of $\hat{t}_y - t_y$. Note that this distribution conditions on the actual outcome(s) of the sampling process as well as on the values of auxiliary variable(s). Unless otherwise indicated we assume we have access to the population values of an auxiliary variable *X* that (usually) is related to *Y*. As always, we assume uninformative sampling, so the superpopulation conditional distribution of *Y* given *X* for the sampled units is the same as the corresponding superpopulation distribution of the non-sampled units.

The first step in developing the model-based approach to estimate t_y is to realise that this is really a prediction problem. We can always write $t_y = \sum_s y_i + \sum_r y_i = t_{sy} + t_{ry}$, where *s* denotes the *n* sampled population units and *r* denotes the remaining *N*-*n* non-sampled population units. Here the sample total t_{sy} is known, so the problem is simply one of predicting the value of the random variable t_{ry} defined by the sum of the *Y*-values of the non-sampled population units.

2.1 Bias and Variance under the Model-Based Approach

The model-based statistical properties of $\hat{t}_y$ as an estimator of t_y are defined by the distribution of the sample error $\hat{t}_y - t_y$ under the assumed superpopulation model. Thus, the prediction bias of $\hat{t}_y$ is the mean of this distribution, $E_\xi(\hat{t}_y - t_y)$, while the prediction variance of $\hat{t}_y$ is its variance, $Var_\xi(\hat{t}_y - t_y)$. The prediction mean squared error of $\hat{t}_y$ is then $E_\xi(\hat{t}_y - t_y)^2$ = $Var_\xi(\hat{t}_y - t_y)$ + $(E_\xi(\hat{t}_y - t_y))^2$. Note that both t_y and $\hat{t}_y$ are random variables here!

The estimator $\hat{t}_y$ is then model unbiased under ξ (ξ-unbiased) if its prediction bias $E_\xi(\hat{t}_y - t_y)$ is zero, in which case its prediction mean squared error is just its prediction variance, $Var_\xi(\hat{t}_y - t_y)$.

Our aim is to identify the "best" estimator for t_y under assumed superpopulation model ξ. In order to do so we use the well known result that the minimum mean squared error predictor of a random variable *U* given the value of another random variable *V* is *E*(*U*|*V*). Consequently the minimum mean squared error predictor (MMSEP) of t_y is given by

$$\begin{aligned} t_y^* &= E_\xi(t_y \mid y_i, i \in s; x_i, i = 1, \cdots, N) \\ &= t_{sy} + E_\xi(t_{ry} \mid y_i, i \in s; x_i, i = 1, \cdots, N) \\ &= t_{sy} + E_\xi(t_{ry} \mid x_i, i \notin s), \end{aligned}$$

where the final equality assumes the mutual independence of different population units. Clearly the conditional expectation in this result will depend on parameters of the assumed superpopulation model ξ, so the corresponding empirical version of this predictor is obtained by replacing these unknown parameters by "efficient" sample-based estimators.

2.2 The Homogeneous Population Model (H)

This model represents the basic "building block" for more complex models that can be used to represent real world variability. It is defined by

$$E_\xi(y_i) = \mu$$
$$Var_\xi(y_i) = \sigma^2$$
$$Cov_\xi(y_i, y_j) = \rho\sigma^2.$$

Linear estimates are widely used in survey sampling. These are estimators that can be expressed as linear combinations of the sample *Y*-values. Consequently we now develop the "best" linear estimator of t_y under H. This is the so-called BLUP (Best Linear Unbiased Predictor) of this quantity. Since our estimator is linear it can be expressed as $\hat{t}_y = \sum_s w_i y_i$, where the w_i are weights that have to be determined. From the decomposition

$$\hat{t}_y = \sum_s w_i y_i = \sum_s y_i + \sum_s (w_i - 1) y_i = t_{sy} + \sum_s u_i y_i$$

we see that the sample error can therefore be expressed

$$\hat{t}_y - t_y = \sum_s u_i y_i - \sum_r y_i = \hat{t}_{ry} - t_{ry}.$$

To start, we require unbiasedness. This implies

$$E_\xi(\hat{t}_y - t_y) = \mu\left(\sum_s u_i - (N-n)\right) = 0 \Rightarrow \sum_s u_i = (N-n).$$

Next, we seek to minimise the prediction variance

$$Var_\xi(\hat{t}_y - t_y) = Var_\xi(\hat{t}_{ry}) - 2Cov_\xi(\hat{t}_{ry}, t_{ry}) + Var_\xi(t_{ry})$$

where

$$Var_\xi(\hat{t}_{ry}) = \sigma^2\left[\sum_s u_i^2 + \rho\sum_{i\in s}\sum_{j\neq i\in s} u_i u_j\right]$$

$$Var_\xi(t_{ry}) = (N-n)\sigma^2\left(1 + \rho(N-n-1)\right)$$

$$Cov_\xi(\hat{t}_{ry}, t_{ry}) = \rho\sigma^2\left(\sum_s u_i \sum_r 1\right) = \rho\sigma^2(N-n)^2.$$

Optimal values of u_i (and hence w_i) are then obtained by minimising $Var_\xi(\hat{t}_{ry})$ above subject to the preceding unbiasedness constraint. In order to do so, we form the Lagrangian:

$$L = \sum_s u_i^2 + \rho\sum_{i\in s}\sum_{j\neq i\in s} u_i u_j - 2\lambda\left(\sum_s u_i - (N-n)\right).$$

Differentiating L with respect to u_i and equating to zero we obtain

$$u_i = \lambda - \rho \sum_{j \neq i \in s} u_j = \lambda - \rho\big((N-n) - u_i\big) \Rightarrow u_i = \frac{\lambda - \rho(N-n)}{1-\rho}.$$

Substituting the unbiasedness constraint above then leads to $\lambda = \frac{N-n}{n}\big(1 + (n-1)\rho\big)$ which implies $u_i = (N - n)/n$ and hence $w_i = N/n$. That is, the optimal predictor of t_y under H is the well-known Expansion Estimator $\hat{t}_{Hy} = N\bar{y}_s$.

In order to develop variance estimates and confidence intervals for this situation, we note that the prediction variance of $\hat{t}_{Hy} = N\bar{y}_s$ under H is given by

$$Var_\xi(\hat{t}_{Hy} - t_y) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\sigma^2(1-\rho).$$

Furthermore, a model-unbiased estimator of this variance under H is then

$$\hat{V}_\xi(\hat{t}_{Hy}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\frac{1}{n-1}\sum\nolimits_s (y_i - \bar{y}_s)^2.$$

Proof of these results is left as an exercise. Application of standard central limit theory to this situation then leads to

$$(\hat{t}_{Hy} - t_y)/\sqrt{\hat{V}_\xi(\hat{t}_{Hy} - t_y)} \sim \mathrm{N}(0, 1)$$

for large values of n. Consequently, an approximate 100(1$-\alpha$)% confidence interval for t_y is $\hat{t}_{Hy} \pm z_\alpha \sqrt{\hat{V}_\xi(\hat{t}_{Hy} - t_y)}$, where z_α is the (1$-\alpha/2$)-quantile of an N(0,1) distribution.

2.3 Stratified Homogeneous Population Model (S)

The simple homogeneous population model discussed in the previous section is unlikely to reflect the variability seen in real survey populations. Consequently we now extend it to allow for strata with distinct means and variances. Note, however, that we also assume that distinct population units are uncorrelated. For population unit *i* in stratum *h* this model is defined by

$E_\xi(y_i) = \mu_h$
$Var_\xi(y_i) = \sigma_h^2$
$Cov_\xi(y_i, y_j) = 0$ for all $i \neq j$.

Clearly the assumption of zero correlation between different elements of the population may not hold if the strata are very small. Consequently this model is appropriate provided the stratum population sizes are reasonably large.

2.3.1 Optimal Estimation Under S

We develop a BLUP for this case. In order to do so we observe that each stratum constitutes a special case of the model H, and so the sample mean of *Y* within each of the strata is the BLUP of the corresponding stratum mean. This immediately leads to the fact that the BLUP of the overall population total t_y is the Stratified Expansion Estimator

$$\hat{t}_{Sy} = \sum_h N_h \bar{y}_{sh} = \sum_h \hat{t}_{Hhy} .$$

In order to compute the prediction variance of this BLUP, and hence develop an estimator for it, we observe that since distinct population units are mutually uncorrelated the prediction variance of stratified estimator $\hat{t}_{Sy}$ is the sum of the individual prediction variances of the stratum specific expansion estimators $\hat{t}_{Hhy}$,

$$Var_\xi(\hat{t}_{Sy} - t_y) = \sum_h Var_\xi(\hat{t}_{Hhy} - t_{hy}) = \sum_h (N_h^2 / n_h)(1 - n_h / N_h)\sigma_h^2 .$$

Consequently an unbiased estimator of this prediction variance is the sum of unbiased stratum level prediction variance estimators, given by

$$\hat{V}_\xi(\hat{t}_{Sy} - t_y) = \sum_h \hat{V}_\xi(\hat{t}_{Hhy} - t_{hy}) = \sum_h (N_h^2 / n_h)(1 - n_h / N_h)s_h^2 .$$

Provided strata sample sizes are large enough, the Central Limit Theorem then applies within each stratum, and we can write:

$$(\hat{t}_{Sy} - t_y)/\sqrt{\hat{V}_\xi(\hat{t}_{Sy} - t_y)} \sim \mathrm{N}(0,1).$$

Confidence intervals for t_y follow directly.

2.3.2 Sample Design Under S

A basic assumption of S is that every stratum is sampled. Consequently the most important sample design consideration under S is how to allocate a sample of size n to the strata. One obvious approach is to allocate in proportion to the stratum population sizes. This is usually referred to as Proportional Allocation and is defined by $f_h = n_h/n = F_h = N_h/N$, where f_h is the sample fraction in stratum h, and F_h is the population fraction in stratum h.

Proportional allocation is typically inefficient. A more efficient approach is to choose the allocation that minimises the prediction variance of stratified expansion estimator subject to an overall sample size of n, i.e. $\sum_h n_h = n$. Now

$$Var_\xi(\hat{t}_{Sy} - t_y) = \sum_h N_h^2 \sigma_h^2 / n_h - \sum_h N_h \sigma_h^2$$

so minimising this prediction variance is equivalent to choosing n_h in order to minimise the first summation on the right hand side above. It can be shown (proof is left as an exercise) that this optimal allocation satisfies $n_h \propto N_h \sigma_h$. It is often referred to as Neyman Allocation, after Neyman (1934).

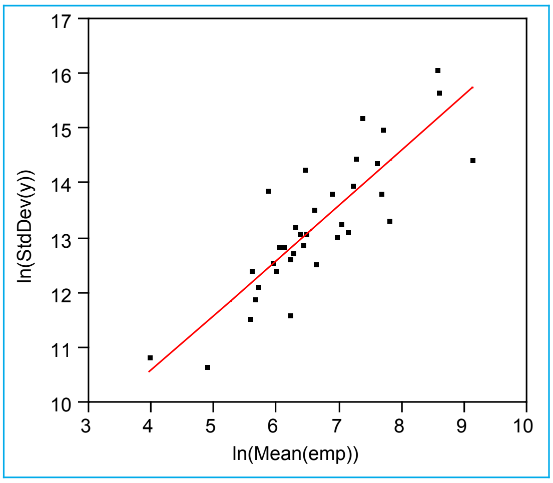
In many cases the strata are defined in terms of the values of an auxiliary variable, e.g. a size variable X. In these situations the question of how to set stratum boundaries arises, given H strata need to be defined. For H strata we require H-1 stratum boundaries: $x_1 < x_2 < ... < x_{H\text{-}1}$.

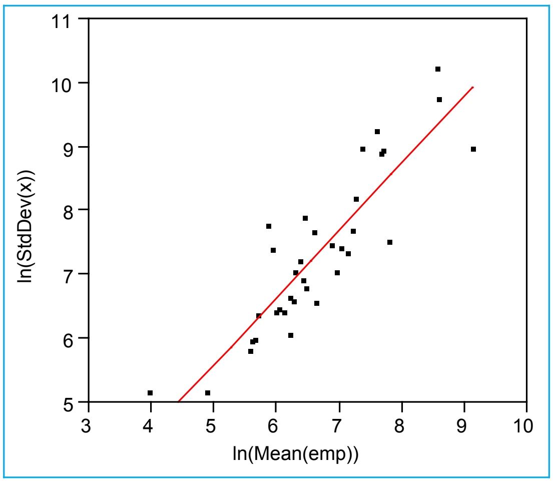
One approach to this problem was suggested by Dalenius and Hodges (1959). This involves partitioning the population into H strata based on the X-values by partitioning the cumulative finite population distribution of the $\sqrt{x_i}$ into H equal size segments. The approach assumes Neyman allocation and many “narrow” strata, allowing an assumption of a uniform distribution for Y within strata.

As an alternative to this approach, we now develop a model-based optimal stratification method that can be useful in very long-tailed populations, as encountered in business surveys, for example. These populations are intrinsically positive and skewed to the right, with “local” variability tending to increase the further out into the tail of the distribution one goes. A model for this behaviour is

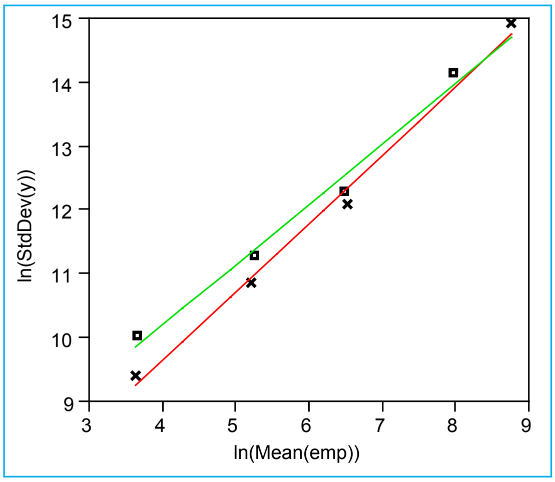
$$\sigma_h^2 = \sigma^2 \bar{x}_h^{2\gamma} \Rightarrow \ln(\sigma_h) = \ln(\sigma) + \gamma \ln(\bar{x}_h)$$

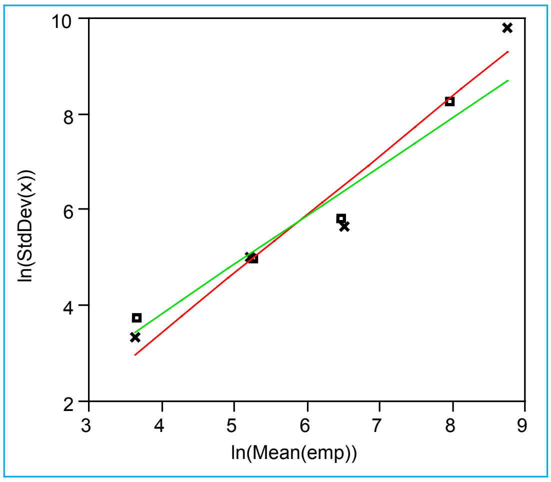
where σ is an unknown scale coefficient and γ is an unknown positive constant. There is considerable empirical evidence that for "long-tailed" economic populations γ typically lies somewhere near 1. Consequently, because our population has a long tail we decide to fix $\gamma = 1$. To justify this decision, consider the following plots, based on data extracted from the UK Monthly Wages and Salaries Survey. For the 35 industry strata used in the design of this survey, the plots show the relationship between the logarithms of the standard deviations of two important survey variables (left = wages, right = employment) and the logarithm of the average value of the size measure (register employment) used in the survey. The slopes of the least squares lines fitted to these plots (also shown) are 1.0065 (wages) and 1.0517 (employment). This provides some evidence that $\gamma = 1$ for these variables and this population.





The preceding plots are not broken down by size. Consequently in the plots below we take two of the industry strata and further break them down by size and type of business (public vs. private), showing the same relationships at industry by size and type stratum level. There are now two lines in each plot, corresponding to the different industry strata. The slope coefficients in this case vary from 0.9464 to 1.2323. This remains consistent with a $\gamma = 1$ assumption in the variance model above.





If we set $\gamma = 1$ in the variance model, and use Neyman allocation then $n_h \propto N_h\sigma_h \propto t_{hx} \Rightarrow n_h = nt_{hx}/t_x$ and so

$$Var_\xi(\hat{t}_{Sy} - t_y) = \sigma^2 t_x^2 / n\,.$$

This expression does not depend on the method used to stratify the population. Consequently any method of stratification leads to the same prediction variance for $\hat{t}_{Sy}$ in this case.

However, equal allocation ($n_h = n/H$) is an alternative and often more convenient allocation procedure and we now investigate how to form efficient strata for this situation. To start, we note that under $\gamma = 1$ the leading term of the prediction variance of the stratified expansion estimator is

$$\sigma^2 \sum\nolimits_h N_h^2 / n_h \bar{x}_h^2 = \sigma^2 \sum\nolimits_h t_{hx}^2 / n_h$$

where t_{hx} is the total of the X-values in stratum h. Under equal allocation this leading term reduces to

$$(\sigma^2 H / n) \sum\nolimits_h t_{hx}^2\,.$$

This expression is minimised by choosing the strata so that the t_{hx} are equal. That is, $t_{hx} = H^{-1}t_x$. We refer to this method as Equal stratification. Given this method of stratification and equal allocation we see that

$$Var_\xi(\hat{t}_{Sy} - t_y) = \sigma^2 t_x^2 / n$$

which is exactly the same as the variance under optimal allocation. That is, there is no efficiency loss from equal allocation – provided we form our strata optimally.

Finally, we consider choice of the number of strata (H). Although this is often not decided on the basis of statistical considerations, when stratifying on the basis of a size variable we usually have the option of choosing between a minimum of 2 strata and a maximum of $[n/k]$ strata, where k denotes a minimum acceptable stratum sample size, and [.] denotes integer part.

Cochran (1977, pp 132-134) recommends that the number of strata be kept to around 6 or 7. We can evaluate this recommendation using a model-based analysis. We again assume the variance model

$$\sigma_h^2 = \sigma^2 \bar{x}_h^{2\gamma}$$

and equal allocation. The leading term of the prediction variance of the stratified expansion estimator is then ($F_h = N_h/N$)

$$(\sigma^2 / n) t_x^{2\gamma} N^{2-2\gamma} \left(\frac{\sum_h F_h^{2-2\gamma}}{H^{2\gamma-1}} \right) = (\sigma^2 / n) t_x^{2\gamma} N^{2-2\gamma} D(H).$$

Note that when $\gamma = {}^1/_2$ and $\gamma = 1$, $D(H) = 1$, and the prediction variance does not depend on H. When ${}^1/_2 < \gamma < 1$, $D(H)$ varies little with H. We illustrate below for $\gamma = {}^3/_4$.

H	**(*F_h*)**									***D*(*H*)**
2	{.75	.25}								.9659
3	{.60	.25	.15}							.9595
4	{.50	.25	.15	.10}						.9553
5	{.45	.25	.15	.10	.05}					.9382
6	{.35	.30	.20	.10	.03	.02}				.9052
7	{.30	.25	.20	.15	.07	.02	.01}			.9027
8	{.28	.24	.18	.14	.08	.05	.02	.01}		.9070
9	{.25	.20	.18	.15	.11	.05	.03	.02	.01}	.9096

Clearly there is nothing to be gained by going beyond Cochran's recommended limited of 6-7 size strata.

2.4 Populations with Linear Regression Structure

As noted earlier, many business survey populations can be modelled by the Simple Ratio Population Model (R). This is defined by:

$$E_\xi(y_i \mid x_i) = \beta x_i$$
$$Var_\xi(y_i \mid x_i) = \sigma^2 x_i$$
$$Cov_\xi(y_i, y_j \mid x_i, x_j) = 0 \text{ for all } i \neq j.$$

It is straightforward to see that the minimum mean squared error predictor (MMSEP) of t_y under this model is

$$t_y^* = t_{sy} + \beta \sum\nolimits_r x_i .$$

It is well known that the BLUE (Best Linear Unbiased Estimator) for β in the model R is

$$b_R = \frac{\sum_s y_i}{\sum_s x_i}.$$

"Plugging" this estimator back into the MMSEP above leads to the Ratio Estimator of t_y

$$\hat{t}_{Ry} = t_{sy} + b_R \sum\nolimits_r x_i = \frac{\sum_s y_i}{\sum_s x_i} \sum\nolimits_U x_i = \frac{\bar{y}_s}{\bar{x}_s} t_x .$$

It can be shown that this estimator is the BLUP for t_y under R. Its use is typically confined to situations where

- the model R itself is plausible - i.e. there is a strong correlation between *Y* and *X* and the size variable *X* is strictly positive;
- values of *X* are measured on the sample;
- the population total of *X* is known.

In effect, the "ratio" model R is a crude but convenient way of describing what tends to be observed in data collected in many business surveys.
Turning now to the Simple Linear Population Model (L), which is useful for describing many social populations, and is defined by

$$E_\xi(y_i \mid x_i) = \alpha + \beta x_i$$

$$Var_\xi(y_i \mid x_i) = \sigma^2$$

$$Cov_\xi(y_i, y_j \mid x_i, x_j) = 0 \text{ for all } i \neq j$$

we see that the MMSEP under L is

$$E_\xi(t_y \mid \text{sample data}) = \sum\nolimits_s y_i + \sum\nolimits_r (\alpha + \beta x_i).$$

Substituting the optimal (BLU) estimators for α and β under L (the Ordinary Least Squares estimators of these parameters) leads to the Regression Estimator

$$\hat{t}_{Ly} = \sum\nolimits_s y_i + \sum\nolimits_r (a_L + b_L x_i) = \sum\nolimits_U (a_L + b_L x_i) = N\left[\bar{y}_s + b_L(\bar{x} - \bar{x}_s)\right]$$

where

$$a_L = \bar{y}_s - b_L \bar{x}_s$$

$$b_L = \frac{\sum_s (y_i - \bar{y}_s)(x_i - \bar{x}_s)}{\sum_s (x_i - \bar{x}_s)^2}.$$

2.4.1 Optimal Sample Design Under R

Under the ratio model R the prediction variance (and hence MSEP) of the ratio estimator is just the prediction variance of the implied predictor of the non-sample total of *Y*. This is

$$\begin{aligned} Var_\xi(\hat{t}_{Ry} - t_y) &= Var_\xi\left(b_R \sum\nolimits_r x_i - \sum\nolimits_r y_i\right) \\ &= \frac{\sigma^2}{\sum_s x_i}\left(\sum\nolimits_r x_i\right)^2 + \sigma^2 \sum\nolimits_r x_i \\ &= \sigma^2 \sum\nolimits_r x_i \left(\frac{\sum_U x_i}{\sum_s x_i} \right). \end{aligned}$$

This result is interesting since it is easy to see that it is minimised when the sample s contains those *n* units in the population with largest values of *X*. Consequently, if we believe that R represents a "right" model for our population, then use of the ratio estimator together with this "extreme" sample represents an optimal sampling strategy for estimating t_y.

2.4.2 Optimal Sample Design under L

A similar approach can be used to show (proof is left as an exercise) that the prediction variance of the regression estimator $\hat{t}_{Ly}$ under the model L is

$$Var_\xi(\hat{t}_{Ly} - t_y) = \frac{N^2}{n}\sigma^2\left[\left(1 - \frac{n}{N}\right) + \frac{(\bar{x} - \bar{x}_s)^2}{(1 - n^{-1})s_{xx}}\right]$$

where

$$s_{xx} = (n-1)^{-1}\sum\nolimits_s (x_i - \bar{x}_s)^2 .$$

In this case we see that this prediction variance is minimised by choosing the sample such that $\bar{x}_s = \bar{x}$. This type of sample is often said to be first-order balanced on *X*.

2.4.3 Combining Regression and Stratification

In practice most populations are more complex than implied by either of the simple linear models underpinning R and L. Again, we can accommodate this complexity by stratifying the population so that separate versions of these models hold in different strata. For example, if different versions of R hold in the different strata, with parameters β_h and σ_h, then we refer to the overall model as the Separate Ratio Population Model (RS). It is straightforward to see that the BLUP of the population total t_y under this model is the Separate Ratio Estimator

$$\hat{t}_{RSy} = \sum\nolimits_h \hat{t}_{Rhy} = \sum\nolimits_h \frac{\bar{y}_{sh}}{\bar{x}_{sh}} t_{xh} .$$

Here a subscript of *h* denotes a stratum *h* specific quantity. If the parameters β_h and σ_h are actually the same for the different strata, then $\hat{t}_{RSy}$ is an inefficient estimator compared with the standard ratio estimator $\hat{t}_{Ry}$. This loss in efficiency is price we must pay for an estimator that is optimal across a much wider, and more likely, range of models for the populations that are observed in survey practice.

Exactly the same argument can be used to define the Separate Regression Population Model (LS), with corresponding BLUP for t_y defined by the Separate Regression Estimator

$$\hat{t}_{LSy} = \sum\nolimits_h N_h [\bar{y}_{sh} + b_h(\bar{x}_h - \bar{x}_{sh})].$$

Optimal sample design for the separate ratio estimator follows directly using the same arguments as those already used with the stratified expansion estimator. Thus, under the RS model

$$V_\xi(\hat{t}_{RSy} - t_y) = \sum\nolimits_h \sigma_h^2 \left(\frac{N_h^2 \bar{x}_h^2}{n_h \bar{x}_{sh}} - N_h \bar{x}_h \right).$$

Optimal sample allocation to the different strata is defined by choosing the n_h to minimise this expression, subject to these values summing to the overall sample size n. However since the stratum sample mean $\bar{x}_{sh}$ depends implicitly on n_h, there is no general solution to this minimisation problem. Instead we adopt a conservative approach and set $\bar{x}_{sh} = \bar{x}_h$ (i.e. we assume stratified balanced sampling). In this case a straightforward application of the same constrained minimisation argument that underpins Neyman allocation for the stratified expansion estimator leads an optimal stratum allocation given by

$$n_h = n \left[\frac{\sigma_h N_h \sqrt{\bar{x}_h}}{\sum_g \sigma_g N_g \sqrt{\bar{x}_g}} \right].$$

Furthermore, under this allocation, and still assuming stratified balanced sampling,

$$Var_\xi(\hat{t}_{RSy} - t_y) = n^{-1} \left[\sum\nolimits_h \sigma_h N_h \sqrt{\bar{x}_h} \right]^2 - \sum\nolimits_h \sigma_h^2 N_h \bar{x}_h .$$

Turning now to sample allocation for the separate regression estimator, we observe that stratified balanced sampling is optimal for this estimator under the model LS. With this type of sampling,

$$V_\xi(\hat{t}_{LSy} - t_y) = \sum\nolimits_h (N_h^2 / n_h)(1 - n_h / N_h)\sigma_h^2 .$$

Optimal stratum sample allocation in this case is equivalent to Neyman allocation and is given by

$$n_h = n \left(\frac{N_h \sigma_h}{\sum_g N_g \sigma_g} \right).$$

Note however that, unlike the case with the stratified expansion estimator, the σ_h above refers to a residual standard deviation under the model LS. With this allocation

$$V_\xi(\hat{t}_{LSy} - t_y) = n^{-1} \left[\sum\nolimits_h N_h \sigma_h \right]^2 - \sum\nolimits_h N_h \sigma_h^2 .$$

2.5 Model-based Methods for Hierarchical Populations

This population model is often used with hierarchical populations. For example, populations of individuals are often grouped into households or families. Similarly, populations of households are often grouped into villages, suburbs, towns etc. This hierarchical grouping structure tends to be reflected by the fact that population units associated with the same group "further up" the hierarchy tend to have values for survey variables that are more alike than those associated with population units from different groups. This clustering effect is especially noticeable when there is little auxiliary information available to explain why some units are more alike than others.

We consider the simplest version of this situation, a two level population, where the population elements defining the census parameters of interest are grouped into clusters. We let g = 1, 2, ..., Q index the clusters in the population, and i = 1, 2, ..., M_g index the elements within the g^{th} cluster. Beyond knowing the cluster of a population element, there is no auxiliary information, and the simple model of a common mean and variance for the population Y-values seems appropriate. Each cluster is taken as defining a "micro-realisation" of this population, and the association between different elements in the same cluster is then accounted for by assuming the cluster can be modelled via the simple homogeneous model H. This leads to the Clustered Population Model (denoted by C in what follows):

$$E_\xi(y_{ig}) = \mu$$
$$Var_\xi(y_{ig}) = \sigma^2$$
$$Cov_\xi(y_{ig},y_{jf}) = \begin{cases} \rho\sigma^2 & \text{if } g = f \\ 0 & \text{otherwise} \end{cases}.$$

Sampling methods for such a two level population typically reflect this structure by sampling clusters first and then sampling elements within sampled clusters. This is typically referred to as Two-Stage Sampling. In order to characterise this situation we now introduce some extra notation. Let Q be the total number of clusters in population, with s denoting the clusters selected into the sample. We assume there are q sampled clusters. We put M_g equal to the number of elements in cluster g, with s_g denoting the set of sampled elements in sampled cluster g. There are m_g sampled elements in sampled cluster g. The overall sample size (of elements) is then $n = \sum_s m_g$, while the corresponding overall population size is $N = \sum_U M_g$. A special case of this situation is cluster sampling, where $m_g = M_g$.

2.5.1 Optimal Prediction under C and Two Stage Sampling

The BLUP of the population total t_y for this situation can be derived via the usual "minimise prediction variance subject to prediction unbiased" type of argument. We do not do this here. Instead we motivate this predictor via a more intuitive argument. To start we note that lack of correlation between clusters means that $E_\xi(\bar{y}_g \mid \text{sample data}) = E_\xi(\bar{y}_g) = \mu$ for any non-sample cluster g. What about the means of sampled clusters? Since we know the sample values in these clusters, we need only to specify the expected value of the non-sampled mean in such a cluster. Clearly, the average value of the non-sampled elements in sampled cluster g will depend on the sampled values, and hence on the average value of the sampled elements in that cluster. If the within cluster correlation ρ is high this mean will be very close to the sample mean in the cluster. Conversely, if ρ is close to zero then this mean will be close to the overall population mean μ. A simple model for this dependence is

$$E_\xi(\bar{y}_{rg} \mid \bar{y}_{sg}) = (1-\alpha_g)\mu + \alpha_g \bar{y}_{sg}$$

where $\bar{y}_{rg}$ is the mean of the non-sampled elements in cluster g**,** $\bar{y}_{sg}$ is the mean of the sampled elements in cluster g and α_g is a weight reflecting knowledge about $\bar{y}_{rg}$ given the average value $\bar{y}_{sg}$ of the sampled data from the cluster.

The MMSEP for t_y under C and this conditional mean model is then

$$\begin{aligned} E_\xi\left(t_y \mid y_{ig}; i \in s_g, g \in s\right) &= \sum_s m_g \bar{y}_{sg} + \sum_s (M_g - m_g) E(\bar{y}_{rg} \mid \bar{y}_{sg}) + \sum_r E(\bar{y}_g) \\ &= \sum_s m_g \bar{y}_{sg} + \sum_s (M_g - m_g)\left[(1-\alpha_g)\mu + \alpha_g \bar{y}_{sg}\right] \\ &\quad + \mu\left(N - \sum_s M_g\right) \end{aligned}$$

An efficient predictor of t_y is obtained by substituting an efficient sample-based estimator $\hat{\mu}$ for μ in this expression, leading to

$$\hat{t}_{Cy} = \sum_s m_g \bar{y}_{sg} + \sum_s (M_g - m_g)\left[(1-\alpha_g)\hat{\mu} + \alpha_g \bar{y}_{sg}\right] + \hat{\mu}\left(N - \sum_s M_g\right).$$

How to define $\hat{\mu}$ here? And how to define the weights α_g? To start, we observe that under C the sample cluster means are uncorrelated, with $E_\xi(\bar{y}_{sg}) = \mu$ and $Var_\xi(\bar{y}_{sg}) = \sigma^2(1-\rho+\rho m_g)/m_g$. Consequently, the BLUE of μ based on these sample means is

$$\hat{\mu} = \frac{\sum_s m_g (1-\rho+\rho m_g)^{-1} \bar{y}_{sg}}{\sum_s m_g (1-\rho+\rho m_g)^{-1}} = \sum_s \theta_g \bar{y}_{sg}.$$

Computation of this estimator assumes we know the intra-cluster correlation ρ. If we do, then substituting this value above and in the value for α_g derived below leads to the BLUP for t_y. However, typically we don't know ρ. Here are some options we might consider in this case:

Option 1: Assume $\rho = 0 \Rightarrow \theta_g = n^{-1} m_g \Rightarrow \hat{\mu} = \sum_s m_g \bar{y}_{sg} / \sum_s m_g = \bar{y}_s$.

Option 2: Assume $\rho = 1 \Rightarrow \theta_g = q^{-1} \Rightarrow \hat{\mu} = q^{-1} \sum_s \bar{y}_{sg} = \bar{\bar{y}}_s$.

Option 3: Estimate ρ directly by fitting a 2-level model to sample data.

We turn now to definition of the weights α_g. In this case we assume that the sample and non-sample means within a cluster are normally distributed (a reasonable assumption provided the sample/non-sample sizes in the cluster are large enough to justify invocation of the Central Limit Theorem). Then standard results for the normal distribution allow us to write

$$\begin{pmatrix} \bar{y}_{sg} \\ \bar{y}_{rg} \end{pmatrix} \sim N\left\{\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \sigma^2 \begin{bmatrix} m_g^{-1}(1-\rho+\rho m_g) & \rho \\ \rho & (M_g - m_g)^{-1}[1-\rho+\rho(M_g - m_g)] \end{bmatrix}\right\}$$

from which we immediately obtain

$$E(\bar{y}_{rg} \mid \bar{y}_{sg}) = \mu + \frac{\rho m_g}{1-\rho+\rho m_g}(\bar{y}_{sg} - \mu)$$

and hence $\alpha_g = \dfrac{\rho m_g}{1-\rho+\rho m_g}$. Here again we see a dependence on ρ and so we have the following estimation options

Option 1: Assume ρ = 0 ⇒ $\alpha_g = 0 \Rightarrow \hat{t}_{Cy} = N\bar{y}_s$

Option 2: Assume ρ = 1 ⇒ $\alpha_g = 1 \Rightarrow \hat{t}_{Cy} = \sum_s M_g \bar{y}_{sg} + \bar{\bar{y}}\left(N - \sum_s M_g\right)$

Option 3: Substitute an estimate of ρ in the formula for the optimal α_g (and in the formula for the BLUE for μ derived earlier). This leads to a so-called EBLUP (Empirical Best Linear Unbiased Predictor) for t_y.

2.5.2 Optimal Sample Design Under C and Two Stage Sampling

It is intuitively obvious that any efficient linear estimator for t_y under C must weight all sample elements in the same cluster equivalently. Such an estimator can therefore be written

$$\hat{t}_{Cy} = \sum_{g\in s} w_g \sum_{i\in s_g} y_i = \sum_s m_g \bar{y}_{sg} + \sum_s u_g m_g \bar{y}_{sg}$$

where

$$w_g = 1 + u_g$$

$$Var_\xi(\hat{t}_{Cy} - t_y) = \sigma^2 \left[\begin{array}{c} \sum_s \left\{ \begin{array}{l} (1-\rho)\left(u_g^2 m_g + (M_g - m_g)\right) \\ + \rho\left(u_g m_g - (M_g - m_g)\right)^2 \end{array} \right\} \\ + \sum_r M_g (1 - \rho + \rho M_g) \end{array} \right].$$

Under cluster sampling ($m_g = M_g$)

$$Var_\xi(\hat{t}_{Cy} - t_y) = \sigma^2 \left[\sum_s u_g^2 M_g (1 - \rho + \rho M_g) + \sum_r M_g (1 - \rho + \rho M_g) \right]$$

The MMSEP (and BLUP) is defined by

$$u_g = m_g^{-1} \left[(M_g - m_g)\alpha_g + \theta_g \left\{ N - \sum_s M_f + \sum_s (M_f - m_f)(1 - \alpha_f) \right\} \right]$$

$$\alpha_g = \frac{\rho m_g}{1 - \rho + \rho m_g}$$

$$\theta_g = \frac{m_g (1 - \rho + \rho m_g)^{-1}}{\sum_s m_f (1 - \rho + \rho m_f)^{-1}}$$

so when $\rho = 0 \Rightarrow \alpha_g = 0$, $\theta_g = n^{-1} m_g \Rightarrow u_g = \dfrac{N - n}{n}$ and when $\rho = 1 \Rightarrow \alpha_g = 1$, $\theta_g = 1/q \Rightarrow$ $u_g = \dfrac{1}{m_g}\left(M_g - m_g + \dfrac{1}{q}(N - \sum_s M_f) \right)$.

An important special case is where all clusters are the same size ($M_g = M$, say), with the same sample size in each sampled cluster, i.e. $m_g = m$. In this situation

$$\alpha_g = \frac{\rho m}{1 - \rho + \rho m} \Rightarrow u_g = (MQ/mq) - 1$$

and the prediction variance of the BLUP is

$$Var_{\xi}(\hat{t}_{Cy} - t_y) = \sigma^2\left[(1-\rho)(MQ - mq)(MQ / mq) + \rho M^2 Q^2 (q^{-1} - Q^{-1})\right]$$

This in minimised by choosing q as large as possible. For a fixed overall sample size $n = mq$ this leads to the conclusion that the optimal design (in terms of minimising the prediction variance) in this important special case has $m = 1$ and $q = n$. The case of varying cluster sizes is more complex.

In most practical two-stage sample design situations, however, the design constraint is cost based, depending on the number of clusters selected, the size of the selected clusters and the second stage allocation in the selected clusters. Cochran (1977, pg 313) gives a simple cost function that incorporates these features:

$$B = c_1 q + c_2 \sum\nolimits_s m_g + c_3 \sum\nolimits_s M_g .$$

The optimal design is then obtained by choosing q, s and m_g to minimise the prediction variance of $\hat{t}_{Cy}$, subject to a fixed value for B above.

Again, we consider the important special case of all clusters the same size (M) with the same sample size (m) in each sampled cluster. Here the cost model is $B = q(c_1 + c_2 m + c_3 M)$.

Substituting $q = B(c_1 + c_2 m + c_3 M)^{-1}$ in $Var_{\xi}(\hat{t}_{Cy} - t_y)$ and simplifying leads to an expression that is proportional to $K_1 + m^{-1}K_2 + mK_3$, where

$$\begin{aligned} K_1 &= N^2 C^{-1}\left(\rho(c_1 + c_3 m) + (1-\rho)c_2\right) - N(1-\rho+\rho M) \\ K_2 &= N^2 C^{-1}(1-\rho)(c_1 + c_3 m) \\ K_3 &= N^2 C^{-1}\rho c_2 . \end{aligned}$$

This expression is minimised by $m = \left(\frac{K_2}{K_3}\right)\left(\frac{(1-\rho)(c_1 + c_3 M)}{\rho c_2}\right)$ when $\rho > 0$.

2.6 Optimal Prediction Under The General Linear Model

Finally, we observe that all the models considered in this Chapter can be considered as special cases of the General Linear Model, defined by

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

Here $\mathbf{Y}$ is the N-vector of the population Y-values, $E_{\xi}(\varepsilon) = 0$ and $Var_{\xi}(\varepsilon) = \sigma^2\mathbf{V}$, $\mathbf{X}$ is a $N \times p$ matrix of auxiliary variables and $\mathbf{V} = \mathbf{V}(\mathbf{X})$ is a known positive definite matrix: $\mathbf{V} = \begin{bmatrix} \mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr} \end{bmatrix}$.

We consider optimal prediction of t_y via a linear combination of sample Y-values. That is, we use a predictor of the form

$$\hat{t}_{wy} = \mathbf{w}'\mathbf{y}_s$$

where $\mathbf{y}_s$ denotes the *n* sample *Y*-values and the weights defining the *n*-vector $\mathbf{w}$ are chosen so that $\hat{t}_{wy}$ has minimum prediction variance in the class of all unbiased predictors of t_y under the general linear model above. These optimal weights were derived by Royall (1976) and are given by

$$\mathbf{w}_{opt} = \mathbf{1}_n + \mathbf{H}'_{opt}\left(\mathbf{X}'\mathbf{1}_N - \mathbf{X}'_s\mathbf{1}_n\right) + \left(\mathbf{I}_n - \mathbf{H}'_{opt}\mathbf{X}'_s\right)\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr}\mathbf{1}_{N-n}$$

where $\mathbf{I}_n$ is the identity matrix of order *n*, $\mathbf{1}_m$ is a *m*-vector of one's and $\mathbf{H}_{opt} = \left(\mathbf{X}'_s\mathbf{V}_{ss}^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}'_s\mathbf{V}_{ss}^{-1}$. With these optimal weights, the BLUP of t_y is then given by

$$\hat{t}_{opt,y} = \mathbf{1}'_n\mathbf{Y}_s + \mathbf{1}'_{N-n}\left[\mathbf{X}_r\hat{\boldsymbol{\beta}}_{opt} + \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}(\mathbf{Y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}_{opt})\right]$$

where $\hat{\boldsymbol{\beta}}_{opt} = \left(\mathbf{X}'_s\mathbf{V}_{ss}^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}'_s\mathbf{V}_{ss}^{-1}\mathbf{Y}_s$ is the BLUE of the parameter $\boldsymbol{\beta}$ in the general linear model.

3. Robust Model-Based Inference

The optimal results developed in the previous Chapter depended on the assumed model ξ actually describing the population of interest. Such a model will be referred to as a "working model" below. However, in practice, working models only approximate reality and so we need to investigate whether our inference is sensitive to misspecification of the working model ξ. In particular, if the inference remains valid, or approximately valid, for reasonable alternative model specifications for the population and sample data, then we say that out inference is robust. If not, then we must be cautious in our interpretation of the inference, since it could contain serious errors.

3.1 Misspecification of the Homogeneous Model (H)

Suppose our working model ξ is the Homogeneous Model H, but an alternative model is better for the population values. Is the BLUP $\hat{t}_{Hy}$ under H (the expansion estimator) still unbiased and optimal under this alternative model? To answer this question we need to specify the alternative model (or models). A simple alternative to H, which we denote by η is

$$E_\eta(y_i) = \mu$$
$$Var_\eta(y_i) = \sigma_i^2$$
$$Cov_\eta(y_i, y_j) = \rho\sigma_i\sigma_j, \text{ for } i \neq j.$$

Since the population Y-values still have a common mean under η, it is easy to see that $\hat{t}_{Hy}$ is unbiased for t_y under ηproof is left as an exercise).

However, it is a different story as far as variance estimation is concerned, since the actual prediction variance of $\hat{t}_{Hy}$ under η is then

$$Var_\eta(\hat{t}_{Hy} - t_y) = (N-n)^2\left[\begin{array}{l}\rho\left\{n^{-1}\sum_s \sigma_i - (N-n)^{-1}\sum_r \sigma_i\right\}^2 \\ +(1-\rho)\left\{n^{-2}\sum_s \sigma_i^2 + (N-n)^{-2}\sum_r \sigma_i^2\right\}\end{array}\right]$$

while the expected value of the unbiased estimator (under the working model) of this variance is now

$$E_\eta\left(\hat{V}_\xi(\hat{t}_{Hy} - t_y)\right) = (N^2/n)(1-n/N)\left[\frac{\rho}{n-1}\sum_s(\sigma_i - \bar{\sigma}_s)^2 + \frac{1-\rho}{n}\sum_s \sigma_i^2\right]$$

where $\bar{\sigma}_s$ is the sample mean of the σ_i.

Suppose now that we wanted to construct a confidence interval for t_y using this working model variance estimator. To do this we need to construct the t-statistic

$$t = (\hat{t}_y - t_y)/\sqrt{\hat{V}_\xi(\hat{t}_y - t_y)}\,.$$

The validity of the resulting confidence interval then requires that the distribution of this statistic, at least in large samples, is N(0,1). However, Cressie (1982) points out that the variance of the

numerator of the t-statistic should match the expected value of its squared denominator if the large sample distribution of this statistic is to be N(0,1). When η holds this is not the case in general.
There are two basic approaches we can take in this situation. The first is to not use the unbiased variance estimator under the working model ξ, but instead develop an alternative estimator of the prediction variance of $\hat{t}_{Hy}$ that is at least approximately unbiased under both η and ξ. We shall consider this approach in the next chapter. The second approach is to select a sample such that the bias in $\hat{V}_{\xi}(\hat{t}_{Hy} - t_y)$ is zero (or approximately zero) under the alternative model η and use the standard t-statistic with this specially selected sample.

In order to do this we observe that the leading term in $\hat{V}_{\xi}(\hat{t}_{Hy} - t_y)$ vanishes if the sample is such that

$$n^{-1}\sum\nolimits_s \sigma_i = (N-n)^{-1}\sum\nolimits_r \sigma_i .$$

If the sampling fraction *n*/*N* is negligible, and ρ is not too far from zero, then both $Var_{\eta}(\hat{t}_{Hy} - t_y)$ and $E_{\eta}\left(\hat{V}_{\xi}(\hat{t}_{Hy} - t_y)\right)$ have the same leading term, $N^2(1-n/N)(1-\rho)n^{-2}\sum\nolimits_s \sigma_i^2$. Consequently this restriction on the sample ensures the t-statistic is “safe”.

Obviously, we cannot choose the sample to satisfy this restriction if we do not know the σ_i. However, we intuitively expect that a large sample selected via simple random sampling to be just as likely to have the right hand side in the above condition less than the left hand side as the other way around. Consequently, provided the sample size is large, SRS represents a safe sampling strategy for the t-statistic under the model η.

Suppose there is an auxiliary variable *X* with values $x_i \propto \sigma_i$, then an alternative approach is to order the population elements according to these *X*-values and then sample by selecting every k^{th} population element on this ordered list, where *k* is the integer part of *N*/*n*. Such an ordered systematic sample also stands a good chance of achieving the above equality.

What about the optimality of the expansion estimator $\hat{t}_{Hy}$ under η? It is true that $\hat{t}_{Hy}$ is the BLUP under the working model ξ, but it is not true that it is generally also the BLUP under η. We denote this η-BLUP by $\hat{t}_{\eta y}$. What is the efficiency loss from using $\hat{t}_{Hy}$ instead of $\hat{t}_{\eta y}$ when η is true? In order to answer this question we need to compare $Var_{\eta}(\hat{t}_{\eta y} - t_y)$ with $Var_{\eta}(\hat{t}_{Hy} - t_y)$.

Suppose $\rho = 0$. Then

$$\hat{t}_{\eta y} = \sum\nolimits_s y_i + (N-n)\left(\sum\nolimits_s y_i\sigma_i^{-2}\right)\left(\sum\nolimits_s \sigma_i^{-2}\right)^{-1};$$

$$Var_{\eta}(\hat{t}_{\eta y} - t_y) = (N-n)^2\left(\sum\nolimits_s \sigma_i^{-2}\right)^{-1} + \sum\nolimits_r \sigma_i^2;$$

$$Var_{\eta}(\hat{t}_{Hy} - t_y) = \left((N-n)^2/n^2\right)\sum\nolimits_s \sigma_i^2 + \sum\nolimits_r \sigma_i^2 .$$

It immediately follows that the gain in precision from using the η-BLUP $\hat{t}_{\eta y}$ instead of the ξ-BLUP $\hat{t}_{Hy}$ is

$$d_{\eta}(\hat{t}_{Hy},\hat{t}_{\eta y}) = \left((N-n)^2/n\right)\left[n^{-1}\sum_{s}\sigma_i^2 - \left(n^{-1}\sum_{s}\sigma_i^{-2}\right)^{-1}\right].$$

This expression is minimised when *s* contains those units with largest values of σ_i. This extreme sample will usually be quite different from the "σ-balanced" sample defined above that guarantees the "safeness" of the t-statistic. Consequently, using a "safe" sample and estimating t_y via $\hat{t}_{Hy}$ does not lead to robustness of optimality for this estimator. This is an example of the insurance "premium" one has to typically pay for using a robust approach.

3.2 Robustness under Stratification

We next turn to the stratified version, S, of the simple homogeneous model H. What if the strata are "wrong"? That is, the population has been incorrectly stratified. In fact, a "correct" stratification exists, but we don't know it.

We use *h* to index the "working" (i.e. incorrect) strata and *g* to index the unknown "correct" strata, with η denoting the correctly stratified model (i.e. the one indexed by *g*) and ξ denoting the incorrectly stratified working model (i.e. the one indexed by *h*). We also assume simple random sampling without replacement within the *h*-strata (i.e. stratified random sampling). The ξ-BLUP, which we denote by $\hat{t}_{S(\xi)y}$ here, is then the stratified expansion estimator defined using the *h*-strata.

Let n_{hg} be the stratum *h* sample "take" of stratum *g* elements. Given the stratified random sampling assumption, it immediately follows that, under η, this quantity is distributed as hypergeometric with parameters N_h, N_{hg} = total number of stratum *g* elements in stratum *h*, and n_h. Hence

$$E_{\eta}(n_{hg}) = n_h\left(\frac{N_{hg}}{N_h}\right),$$

$$Var_{\eta}(n_{hg}) = \left(\frac{N_h - n_h}{N_h - 1}\right)n_h\left(\frac{N_{hg}}{N_h}\right)\left(1 - \frac{N_{hg}}{N_h}\right),$$

$$Cov_{\eta}(n_{hg}, n_{hf}) = -\left(\frac{N_h - n_h}{N_h - 1}\right)n_h\left(\frac{N_{hg}}{N_h}\right)\left(\frac{N_{hf}}{N_h}\right).$$

An immediate consequence is that the η-bias of $\hat{t}_{S(\xi)y}$ is zero:

$$\begin{aligned} E_{\eta}(\hat{t}_{S(\xi)y} - t_y) &= E_{\eta}\left(E_{\eta}(\hat{t}_{S(\xi)y} - t_y \mid n_{hg})\right) \\ &= E_{\eta}\left[\sum_g\left(\sum_h n_{hg}N_h/n_h - N_g\right)\mu_g\right] \\ &= 0\,. \end{aligned}$$

We can also show (but the algebra gets messy) that the η-bias of the usual variance estimator for $\hat{t}_{S(\xi)y}$, i.e. $E_{\eta}\left[\hat{V}_{\xi}(\hat{t}_{S(\xi)y}) - Var_{\eta}(\hat{t}_{S(\xi)y} - t_y)\right]$ is also zero. We conclude that stratified random sampling is a "safe" sampling strategy for the stratified expansion estimator.

The preceding analysis assumes that we don't know the correct stratification for the population, so the unbiasedness properties are with respect to all possible values the (unknown) counts n_{hg} can take. However, in many cases it is possible, from analysis of the sample data, to identify these counts (and hence the "correct" strata) after sampling. In this case it may be possible to replace $\hat{t}_{S(\xi)y}$ by $\hat{t}_{S(\eta)y}$, the stratified expansion estimator based on the g-strata. This is typically referred to as post-stratification and $\hat{t}_{S(\eta)y}$ is called the post-stratified expansion estimator. Inference then proceeds on basis of η being correct (i.e. conditional on the realised values $n_g = \sum_h n_{hg}$).

There are two basic problems with this approach. The first is that we have no control over the value of n_g. In some cases this can even be zero for poststrata that are "rare", in which case no unbiased estimator exists under η. In any case, it is clear that $\hat{t}_{S(\eta)y}$ will then have a higher prediction variance than would have been achieved with correct pre-stratification based on η.

The second problem is that use of $\hat{t}_{S(\eta)y}$ requires knowledge of the population counts N_g in the g-strata. If these are unknown, we can substitute $\hat{N}_g = \sum_h (n_{hg}/n_h)N_h$ in $\hat{t}_{S(\eta)y}$ to get the "estimated" η-BLUP $\tilde{t}_{S(\eta)y} = \sum_g (\hat{N}_g/n_g)\sum_{s_g} y_i$. The prediction variance of this "estimated" BLUP will then be greater than that of the actual η-BLUP (N_g known):

$$Var_\eta(\tilde{t}_{\eta Sy} - t_y) = E_\eta\left[Var_\eta\left(\tilde{t}_{\eta Sy} - t_y \mid \hat{N}_g\right)\right] + Var_\eta\left[E_\eta\left(\tilde{t}_{\eta Sy} - t_y \mid \hat{N}_g\right)\right].$$

A consistent estimator of this prediction variance is obtained by substituting unbiased estimators for unknown values in these expressions.

Often we wish to combine pre- and post-stratification. This typically arises when the expected value of the survey variable Y varies according to a number of factors, but we only have frame information on a subset of these factors. In such cases pre- or "sampling" strata are defined using the "frame" factors and post-stratification is used to account for the remaining factors.

To illustrate, suppose an individual's Y-value varies by X_1 = Region (categorical) and X_2 = Age-Sex category in the sense that $E_\xi(Y)$ = region effect + age-sex effect. The sampling frame contains values for region (sampling strata), while data on age and sex obtained from sampled individuals. We also know (from other sources) the total number of people in each age-sex class in the population. This is a standard scenario for many social surveys.

Let $h = 1, 2, \ldots, H$ index the pre-strata (defined by X_1) and let $g = 1, 2, \ldots, G$ indexes the post-strata (defined by X_2). The BLUP for t_y based on the pre-strata is then defined by sample weights $w_i = N_h/n_h$ for individual $i \in h$. Unfortunately, this estimator is biased under the true two-factor model for the population defined in the previous paragraph.

However, we can recover an unbiased linear predictor of t_y under this two-factor model by modifying the sample weights so that they sum to N_h in pre-stratum h as well as sum to N_g in post-stratum g. This can be achieved by iterative re-scaling (raking) of the w_i as follows:

1. Set $w_i(0) = w_i$ and $k = 0$.
2. Put $k = k + 1$.
3. For each value of g, let $W_g(k-1)$ = sum of the weights $w_i(k-1)$ for all sampled individuals $i \in$ post-stratum g. For each such individual calculate $w_{1i} = w_i(k-1) \times N_g/W_g(k-1)$.

4. For each value of h, let W_{1h} denote the sum of the weights w_{1i} for all sampled individuals $i \in$ pre-stratum h. For each such individual calculate $w_i(k) = w_{1i} \times N_h/W_{1h}$.
5. If there is little or no difference between $w_i(k-1)$ and $w_i(k)$, then go to step 6. Otherwise return to step 2.
6. Set the final weight $w_i = w_i(k)$.

Note that these rescaled weights do not define the BLUP for t_y under the two-factor model (i.e. the one with X_1 and X_2 as covariates). The BLUP for this model requires use of multiple regression, and is typically implemented via calibrated weighting. See later.

3.3 Balanced Sampling and Robust Estimation

Suppose now that we have an auxiliary variable X and the ratio model R seems appropriate for the target population. As shown in the previous Chapter, the optimal sampling strategy for minimising the prediction variance of the BLUP under R is one where the n population units with largest values of X are sampled.

In practice, however, such extreme samples are hardly ever chosen. The reason is easy to find - what if the model is wrong? As it stands, however, this is not really an adequate reason for not adopting this optimal approach. Models are always wrong, since they only approximate reality. The real question is how sensitive is the above optimal strategy to misspecification of the model.

Suppose the target population follows model L rather than R: We use (as usual) ξ to denote our working model R, and we denote this alternative by η. Under η (α is the intercept coefficient):

$$E_\eta(\hat{t}_{Ry} - t_y) = \alpha\left[1/\bar{x}_s - 1/\bar{x}_r\right].$$

For $\alpha > 0$ this bias will be negative (and large in absolute value) when s is the "extreme" sample. Hence adopting the optimal strategy under R leaves us extremely vulnerable to a possible large bias if in fact the model L holds.

But, for arbitrary α, the above bias is zero when the sample s satisfies $\bar{x}_s = \bar{x}_r = \bar{x}$. That is, the sample is first order balanced on the auxiliary variable X. Furthermore, under balanced sampling the ratio estimator (the ξ-BLUP) reduces to the expansion estimator and the prediction variance of the ratio estimator under ξ is

$$Var_\xi(\hat{t}_{Ry} - t_y \mid \text{balance}) = \sigma^2(N-n)(N/n)\bar{x}.$$

In contrast, under "extreme" sampling and ξ,

$$Var_\xi(\hat{t}_{Ry} - t_y \mid \text{extreme}) = \min_s\left\{\sigma^2(N-n)(N/n)\bar{x}(\bar{x}_r/\bar{x}_s)\right\}.$$

The relative efficiency of using a balanced sample instead of the extreme sample when the assumed model ξ is correct is therefore

$$\frac{Var_\xi(\hat{t}_{Ry} - t_y \mid \text{extreme})}{Var_\xi(\hat{t}_{Ry} - t_y \mid \text{balance})} = \min_s(\bar{x}_r/\bar{x}_s) = \left(1-\frac{n}{N}\right)^{-1}\frac{\bar{x}}{\max_s(\bar{x}_s)} - \left(\frac{n}{N-n}\right).$$

This ratio can be very small, so adopting a balanced sampling strategy can lead to a large loss of efficiency. That is, as we found out earlier in this chapter, robustness to model misspecification can have a large "insurance premium" in terms of efficiency loss.

There are two aspects to this efficiency loss. The first is the efficiency loss relative to the optimal sampling and estimation strategy under the working model ξ. The second is more subtle. It is the efficiency loss due to not using the η-BLUP. Remember that we are using the ξ-BLUP which is η-unbiased in the balanced sample but not necessarily efficient under η. Are there situations where this second form of efficiency loss is minimised?

The answer to this is yes. To illustrate, suppose the true model for the population data is L rather than the working model R. As noted above, under balanced sampling the ratio estimator reduces to the simple expansion estimator. However, it is easy to see that under L the BLUP (the regression estimator) also reduces to the expansion estimator. That is, the ratio estimator is equivalent to the BLUP under L in a balanced sample.

Royall and Herson (1973a) prove a theorem that generalises this result to polynomial alternatives to the simple model R. This can be stated as follows: Suppose that expectation under the "real" model η is a polynomial of degree k in X. Also the sample is balanced up to order K (i.e. $n^{-1}\sum_s x_i^k = (N-n)^{-1}\sum_r x_i^k$; k = 1, .., K). If a ratio estimator is used when η is the "true" model then ratio estimator is η-unbiased. Furthermore, if such a sample is selected, and

$$Var_\eta(y_i) = \sigma^2 \sum_0^K a_k x_{ki}$$

then the η-BLUP is $N\bar{y}_s$. Since the ratio estimator is also equal to the expansion estimator on such a sample, it follows immediately that it must be the η-BLUP (as well as η-unbiased) on this sample.

In an effort to try to minimise efficiency loss relative to the optimal strategy under R (ratio estimation and an extreme sample), Royall and Herson (1973b) recommend stratified rather than simple balanced sampling. Their argument for this approach goes along the following lines:

1. When a balanced sample is selected within each stratum

   $$Var_\xi(\hat{t}_{RSy} - t_y) = \sigma^2 \sum_h (N_h^2 / n_h)[1-(n_h/N_h)]\bar{x}_{sh}.$$

   This is minimised subject to a sample size of n when the sample stratum allocation is proportional to $N_h\sqrt{\bar{x}_{sh}}$, in which case

   $$Var_\xi(\hat{t}_{RSy} - t_y \mid \text{optimal allocation}) = \sigma^2\left[n^{-1}\left(\sum_h N_h\sqrt{\bar{x}_{sh}}\right)^2 - t_x \right].$$

   This is always less than or equal to the variance of the ratio estimator under simple balanced sampling.

2. The strategy of stratified balanced sampling and the stratified ratio estimator is qualitatively more robust than the strategy consisting of simple balanced sampling and the ordinary ratio estimator (because the former can accommodate non-linear alternatives).

3. The prediction variance of the stratified ratio estimator $\hat{t}_{RSy}$ under this stratified balance strategy can always be made smaller by using equal allocation and equal stratification.

Finally, we set out a general theorem that allows one to identify when a linear estimator is equal to the BLUP under the general linear model. Let **W** be an arbitrary (N - n) × n matrix of sample weights. We can then always write any linear predictor as a special case of the **W**-weighted linear predictor of t_y defined by $\hat{t}_y(\mathbf{W}) = \sum_s y_i + \mathbf{1}'_r\mathbf{W}\mathbf{y}_s$. It is easy to see that the prediction bias of this estimator under the general linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is then $\mathbf{1}'_r(\mathbf{W}\mathbf{X}_s - \mathbf{X}_r)\boldsymbol{\beta}$. This bias is zero if the sample is **W**-balanced, i.e. $\mathbf{1}'_r\mathbf{W}\mathbf{X}_s = \mathbf{1}'_r\mathbf{X}_r$.

The general theorem (Chambers, 1982; see also Tallis, 1978; Tam, 1986) can then be stated as follows: Provided a **W**-balanced sample is selected, a necessary and sufficient condition for $\hat{t}_y(\mathbf{W})$ to be the BLUP of t_y is when the matrix $\mathbf{V}_{ss}^{1/2}(\mathbf{W} - \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1})'\mathbf{1}_r$ is in the vector space spanned by the columns of $\mathbf{V}_{ss}^{-1/2}\mathbf{X}_s$. Here $\mathbf{V}_{ss}$, $\mathbf{V}_{rr}$ and $\mathbf{V}_{rs}$ denote the sample/non-sample components of the variance matrix of the vector $\boldsymbol{\varepsilon}$.

All the results on balancing developed so far can be obtained as special cases of this theorem.

3.4 Outlier Robust Estimation

So far we have focussed on robustness to model misspecification. However, in practice what we often see are isolated data values in our sample that clearly do not follow the working model, while the main mass of our sample data do conform to it. These outliers (or 'wild' data values) are a common feature of many sample surveys, particularly those of highly skewed economic populations. Ignoring these outlying values and calculating an estimate for t_y on the basis that the working model applies to all the sample data can lead to a completely unrealistic value for this estimate.

3.4.1 Basic Ideas

What can we do if we observe outliers in our sample data? To discuss this further we focus on the common situation of a weighted linear estimate $\hat{t}_{wy} = \sum_s w_i y_i$. Here w_i is the sample weight. An outlier may then be a value y_i completely unlike any other sample Y-value. However, it can also be a 'non-small' y_i that is associated with a 'large' weight w_i so the product $w_i y_i$ is large.

The interplay between the sample weight and the sample value when defining an outlier leads to two basic approaches to dealing with these cases.

1. Modify the sample weights of the outliers by reducing them relative to those of sample units that are not outliers. However, leave their Y-values unchanged (modify weight option).

2. Modify the y-values associated with the sample outliers so as to make them more "acceptable", but leave their weights unchanged (modify value option).

Both the above approaches are reflected in the following practical strategies for dealing with outliers that are in common use:

1. Post-stratify sample outliers by placing them in a special stratum with lower (typically unit) weights.

2. Replace outlier value in current survey replaced by an average of its current and past (non-outlier) values and then put it in a special stratum.

3. Force the sample weights to lie in some "acceptable" range (e.g. ±k% of selection weights).

However, there are alternative approaches one can take, where one replaces the optimal BLUP under the working model by a more robust, but less efficient, estimator. The basic idea here is to apply modern robust statistical concepts in estimation, replacing the non-robust linear BLUP by a robust non-linear estimator that 'accommodates' sample outliers.

In order to define such an estimator, we note that all outlier robust estimation strategies involve a bias-variance tradeoff, where some bias in the estimator is accepted in order to downweight the influence of outliers. This downweighting of outliers decreases the variance of the estimator and (hopefully) the mean squared error as well. However one then runs into problems with confidence interval estimation, as we shall see.

To start, we note that there is now a well developed theory of outlier robust estimation, dating from the seminal paper of Huber (1964), where outlier robust M-estimators for the parameters of statistical models were first introduced. These estimators are typically defined by modifying estimating equations to ensure that no one data value has undue influence on the estimate of a model parameter. They are motivated by the idea that the majority of the population (and hence the sample) follow a 'well-behaved' pattern of behaviour as expected under the working model, with a small number of sample outliers or 'contaminants' that are ideally zero weighted in any inference related to the working model.

Chambers (1986) extended this concept to finite population prediction and introduced the concept of 'representative' outliers. These are legitimate population units (i.e. their outlying values are not mistakes) and we have no guarantee that there are no further outliers in the non-sampled part of the population. Clearly it is inappropriate to zero-weight representative outliers. Non-representative outliers (mistakes or unique values) on the other hand should be either given unit weights or be corrected (if they are due to errors) or be discarded (i.e. zero-weighted).

The basic problem is then how to deal with survey data containing representative outliers. Since there may well be (and usually are) more representative outliers in the non-sampled part of the population, it is inadequate to just isolate these outliers in the survey sample and, as in the post-stratification approach described above, give them unit weights. Sample outliers provide (limited) information about non-sample outliers. However, there are few sample outliers (by definition), so any attempt to use 'standard' weighting methods to extract this information is a recipe for disaster.

3.4.2 Robust Bias Calibration

Robust bias calibration of the "delete outliers" approach is one compromise solution to this problem. The idea here is straightforward. First estimate the finite population total of interest as if the working model applies to all non-sample units. That is, either delete or unit weight all sample outliers. Clearly this 'working model' estimator is generally biased if sample outliers are representative, so we now add a bias calibration term to this estimator that uses the information in the sample outliers to compensate for its bias.

To illustrate how this approach works we consider a simple mixture model. That is, we assume that the actual population values of *Y* are drawn from the two component mixture

$$y_i = \Delta_i(\mu_1 + \sigma_1\varepsilon_{1i}) + (1-\Delta_i)(\mu_2 + \sigma_2\varepsilon_{2i})$$

where Δ_i is a zero-one variable denoting outlier/non-outlier status respectively, with $\theta_i = \mathrm{pr}(\Delta_i = 1)$, $1-\delta \le \theta_i < 1$, $\mu_2 >> \mu_1$ and $\sigma_2 >> \sigma_1$, and ε_{1i}, ε_{2i} are independent 'white noise' variables.

Given a sample drawn from this mixture, we can use modern outlier robust methods to estimate μ_1. In particular, let $\hat{\mu}_1$ be a robust M-estimate of μ_1, defined by the estimating equation

$$\sum\nolimits_s \psi_y(\hat{\sigma}_1^{-1}(y_i - \hat{\mu}_1)) = 0$$

where $\hat{\sigma}_1$ is a robust estimate of σ_1 (e.g. the median absolute deviations from the median or MAD estimate) and ψ_y is the influence function underlying $\hat{\mu}_1$ - a bounded skew-symmetric function that behaves like the identity near the origin and drops away to zero for values far from the origin. An example is the bisquare function (Beaton and Tukey, 1974), defined by

$$\psi_y(t) = t\left(1 - t^2/k_y^2\right)^2 I(-k_y \le t \le k_y)$$

where k_y is a tuning constant (default = 4.5). The smaller k_y, the more outlier robust is $\hat{\mu}_1$.

The initial "no non-sample outliers" estimator is then $\hat{t}_{1y} = \sum\nolimits_s y_i + (N-n)\hat{\mu}_1$ (or the even simpler $\tilde{t}_{1y} = N\hat{\mu}_1$), which is very close (in spirit at least) to the commonly used post-stratification estimator. It clearly assumes that there are no non-sample outliers in the population, and predicts unknown *Y*-values on the assumption that all such values follow a working model consistent with the behaviour of the non-outliers in the sample. See Rivest and Rouillard (1991) and Gwet and Rivest (1992).

This initial estimator is biased if sample outliers are representative, with

$$Bias(\hat{t}_{1y}) = (\mu_1 - \mu_2)\sum\nolimits_r (1-\theta_i).$$

To correct for this bias we must estimate it in some way. Suppose the mixture probabilities are constant, i.e. $\theta_i = \theta$, then

$$Bias(\hat{t}_{1y}) = -(N-n)E(\bar{r}),$$

where $\bar{r} = n^{-1}\sum\nolimits_s (y_i - \hat{\mu}_1)$. An obvious "bias corrected" version of this estimator is then

$$\hat{t}_{adj,y} = \sum\nolimits_s y_i + (N-n)(\hat{\mu}_1 + \bar{r}) = \hat{t}_{1y} + (N-n)\bar{r}.$$

This estimator is unbiased under the assumed mixture model for *Y*. Unfortunately, simple algebra then demonstrates that it is in fact just the simple (and highly non-robust) expansion estimator.

The problem is the bias adjustment $\bar{r} = n^{-1}\sum_s (y_i - \hat{\mu}_1)$. This is computed as the mean of the 'raw' residuals $r_{1i} = y_i - \hat{\mu}_1$. These residuals will be 'small' for the well-behaved units in the sample, but will be 'large' for the sample outliers. A more robust bias adjustment is the modified mean

$$\bar{\varepsilon} = \hat{\sigma}_R\, n^{-1} \sum_s \psi_R (r_{1i} / \hat{\sigma}_R).$$

Here $\hat{\sigma}_R$ is a robust estimate of the scale of the r_{1i} (e.g. the MAD estimate) and ψ_R is a 'prediction' influence function that gives relatively more weight to the sample outliers than the 'estimation' influence function ψ_Y underlying $\hat{\mu}_1$. A natural choice is Huber's Proposal 3 (Huber 1964),

$$\psi_R(t) = \begin{cases} k_R & t > k_R \\ t & |t| \leq k_R \\ -k_R & t < -k_R \end{cases}$$

where the tuning constant k_R is quite large, say k_R = 10. A robustly calibrated estimator of the population total is then

$$\hat{t}_{rob,y} = \sum_s y_i + (N - n)(\hat{\mu}_1 + \bar{\varepsilon}).$$

Note that this estimator is equivalent to using the standard expansion estimator with modified Y-values:

$$y_i^* = \frac{n}{N} y_i + (1 - \frac{n}{N})(\hat{y}_i + \bar{\varepsilon})$$

where

$$\hat{y}_i = \left[\frac{\psi_Y(\hat{\sigma}_Y^{-1}(y_i - \hat{\mu}_1)}{y_i - \hat{\mu}_1}\right]\left[n^{-1}\sum_s \frac{\psi_Y(\hat{\sigma}_Y^{-1}(y_j - \hat{\mu}_1))}{y_j - \hat{\mu}_1}\right]^{-1} y_i.$$

This robust estimator is biased (by construction). Consequently, standard large sample arguments for confidence interval estimation based on a consistent estimator of its variance cannot be claimed to have (even asymptotically) nominal coverage properties.

The extension of the above argument to the case where the general linear model is the "working model" is relatively straightforward (see Chambers, 1986).

3.4.3 Outlier Robust Design

Can we use the sample design to provide protection against outliers, in the same way as it can be used to provide protection against model misspecification? In general this is not possible, because we have no idea a priori where outliers will occur. However, a measure of outlier robustness is achieved by implementing a sample design where the weights w_i do not vary too much from one sample unit to another, since this minimises the opportunity for a sample outlier to 'team up' with a large sample weight and hence destabilise the estimator.

Sample weights are typically functions of one or more auxiliary variables (X), and sample designs where these weights do not vary (or at least vary little) are typically designs that are 'balanced' with respect to these variables. Consequently sample designs that attempt to ensure such 'balance' (e.g.

via restricted randomisation) can therefore be expected to be less outlier sensitive than designs that place no restrictions on the sample weights.

3.4.4 A Numerical Study

This study assumed a working model corresponding to the ratio model R, where:

$$E_{\xi}(y_i) = \beta x_i$$
$$Var_{\xi}(y_i) = \sigma^2 x_i$$

and applied robustness ideas to estimation of t_y in two populations. These were:

SUGAR: This consisted of 338 Queensland sugar cane farms, with *Y* = value of cane produced and *X* = area assigned for growing cane. This population was described in the previous chapter and is reasonably described by ratio model R.
BEEF: This population consisted of 453 beef cattle farms, with *Y* = income from sale of cattle and X = number of cattle. There is clear model misspecification if the ratio model is assumed for this population.

A simulation experiment was then carried out, with 500 independent simple random samples selected from each population. The sample size for SUGAR was *n* = 50, while that for BEEF was *n* = 60. For each sample, four estimation strategies were investigated. These were the ratio estimator with a 'misspecification robust' estimate of the variance of this estimator (see the next Chapter for development of this estimator, due to Royall and Cumberland, 1981), the robust bias calibrated version of the ratio estimator (see Chambers, 1986) together with a bootstrap variance estimator (again, this method is described in the next Chapter). This estimator used the bisquare estimation influence function ψ_Y (with default tuning parameter k_y = 4.5) and the Huber prediction influence function ψ_R. Three versions of this estimator were actually calculated, corresponding to k_R = 6, 10 and 15. Bootstrap simulations were then used to construct bootstrap 95% confidence intervals for the population total of *Y*.

Three estimation performance measures were calculated. The average error over the 500simulations (AVE), the root mean squared error over the 500 simulations (RMSE), and the median absolute deviation error over the 500 simulations (MAE). These measures are set out in the table below. In addition, the average estimated standard error over the 500 simulations (AVSE) is also presented in this table.

Method	AVE	RMSE	MAE	AVSE
	SUGAR			
Ratio	-117	3733	2535	3676
Robust				
k_R = 6	-117	3733	2535	3634
k_R =10	-117	3733	2535	3655
k_R = 15	-117	3733	2535	3636
	BEEF			
Ratio	5768	30802	21269	26823
Robust				
k_R = 6	15058	26647	16974	19658
k_R =10	7856	27772	17647	22444
k_R = 15	5415	29995	20720	25036

We comment on each population in turn.

SUGAR: There is nothing to choose (in terms of AVE, RMSE and MAE) between the ratio estimator and the three robust estimators. This is comforting, since it indicates that the robust estimators lose little efficiency when the working model is in fact reasonably valid. Furthermore, all AVSE values 'track' RMSE in SUGAR.

BEEF: The ratio estimator is substantially outperformed by the three robust estimators. In particular, the robust estimator with $k_R = 6$ has the best RMSE and MAE performance at the cost of a substantial bias (AVE), while the estimator with $k_R = 10$ seems to deliver the best compromise between AVE, RMSE and MAE. However, it is clear that for this population, AVSE underestimates RMSE, caused in no small part by the substantial bias due to the presence of outliers.

We now turn to the coverage performance of the confidence intervals generated by the different estimator/variance estimator combinations investigated in this study. We in fact calculated classical "two sigma" confidence intervals as well as 95% bootstrap confidence intervals using each estimator. Again we comment on these results separately for each population, after first displaying them in the table below.

Method	2σ CI non-coverage	Bootstrap-based 95% confidence intervals		
		OK	HI	LO
	SUGAR			
Ratio	.050	.942	.022	.036
Robust				
$k_R = 6$	.056	.938	.022	.040
$k_R = 10$	.050	.936	.022	.042
$k_R = 15$	.052	.940	.026	.034
	BEEF			
Ratio	.092	.894	.058	.048
Robust				
$k_R = 6$	.176	.808	.190	.002
$k_R = 10$	.110	.880	.120	.000
$k_R = 15$	.100	.892	.106	.002

SUGAR: All methods record actual coverages reasonably close to the nominal 95 per cent level.

BEEF: All methods perform poorly, with substantial undercoverage. The robust estimator with $k_R = 6$ had the worst confidence interval coverage performance. There is also clear skewness (HI > LO) in the coverage performance of the bootstrap confidence intervals generated by the robust estimators, indicating a bias problem.

At this stage, the issue of how to construct valid confidence intervals in the presence of outliers remains an open problem.

3.4.5 Practical Problems

There are substantial practical problems with adoption of the outlier robust estimators described above. One of the most important is caused by the intrinsic non-linearity of these estimators. In particular, population totals estimated ‘robustly’ at a lower level of categorisation may not sum to the value of the corresponding ‘robust’ estimate of total at a higher level of categorisation. Similar problems arise when estimates based on derived variables (e.g. a sum of component variables) is calculated from survey data.

Requirement that such kinds of inconsistencies do not arise therefore limit the effectiveness of the robust procedure at certain levels of categorisation or for certain variables.

4. Methods of Variance Estimation

In this Chapter we explore in more detail issues that arise when we wish to estimate the variability associated with an estimate (or prediction) of a census parameter.

4.1 Robust Variance Estimation for the Ratio Estimator

When the simple ratio model R is the working model ξ, the unbiased estimator of the prediction variance of the ratio estimator is

$$\hat{V}_{\xi}(\hat{t}_{Ry}) = \hat{\sigma}_R^2 (N^2/n)(1-n/N)\left(\bar{x}_r \bar{x}\right)/\bar{x}_s$$

where

$$\hat{\sigma}_R^2 = (n-1)^{-1}\sum\nolimits_s \left(y_i - (\bar{y}_s/\bar{x}_s)x_i\right)^2/x_i$$

is an ξ-unbiased estimator of the parameter σ^2. However, this standard “plug-in” approach to variance estimation is non-robust to misspecification of the variance “model” implied under ξ.

To see this, suppose the true model η for the population has $E_\eta(y_i) = \beta x_i$ and $Var_\eta(y_i) = \sigma^2 v(x_i)$. Since the mean function is unchanged from that assumed under ξ, the ratio estimator remains η-unbiased, but now its prediction variance is

$$Var_\eta(\hat{t}_{Ry} - t_y) = \sigma^2(N^2/n)(1-n/N)\left[(1-n/N)\bar{v}_s(\bar{x}_r/\bar{x}_s)^2 + (n/N)\bar{x}_r\right].$$

Furthermore, under η the standard estimator for the parameter σ^2 is no longer unbiased, since

$$E_\eta\hat{\sigma}_R^2 = \sigma^2\left[\overline{(v/x)}_s + (n-1)^{-1}\{1-\bar{v}_s/\bar{x}_s\}\right].$$

Suppose $\bar{x}_s = \bar{x}_r = \bar{x}$. Then

$$\begin{aligned} E_\eta\left(\hat{V}_R(\hat{t}_{Ry})\right) &= \sigma^2\frac{N^2}{n}\left(1-\frac{n}{N}\right)\left[\overline{(v/x)}_s + \frac{\overline{(v/x)}_s - \bar{v}_s/\bar{x}_s}{n-1}\right]\bar{x}_s \\ &\approx \sigma^2(N^2/n)(1-n/N)\overline{(v/x)}_s\bar{x}_s. \end{aligned}$$

In this case the actual prediction variance of the ratio estimator under η and balanced sampling is

$$\sigma^2(N^2/n)(1-n/N)\left[\bar{v}_s + (n/N)(\bar{v}_r - \bar{v}_s)\right] \approx \sigma^2(N^2/n)(1-n/N)\bar{v}_s$$

so the working model variance estimator $\hat{V}_\xi(\hat{t}_{Ry})$ is biased high when $v(z) < z$ and biased low when $v(z) > z$.

Under balanced sampling, inspection of the expressions above shows that the leading term in actual prediction variance of the ratio estimator depends on $\bar{v}_s \propto E_\eta\left(n^{-1}\sum\nolimits_s (y_i - \beta x_i)^2\right)$. This suggests that we consider an alternative variance estimator for this situation, given by

$$\hat{V}_{SRS}(\hat{t}_{Ry}) = (N^2/n)(1-n/N)(n-1)^{-1}\sum_s \left(y_i - (\bar{y}_s/\bar{x}_s)x_i\right)^2 .$$

This is the "traditional" design-based estimator of the variance of the ratio estimator under simple random sampling. Under η this estimator has expectation

$$E_\eta\left(\hat{V}_{SRS}(\hat{t}_{Ry})\right) = \sigma^2 \frac{N^2}{n}\left(1-\frac{n}{N}\right)\bar{v}_s\left[1+(n-1)^{-1}\left\{1-2\frac{\overline{(vx)}_s}{\bar{v}_s\bar{x}_s}+\frac{\overline{(xx)}_s}{\bar{x}_s^2}\right\}\right]$$
$$\approx \sigma^2(N^2/n)(1-n/N)\bar{v}_s$$

which is the same as the leading term of the actual prediction variance of the ratio estimator under η. That is, $\hat{V}_{SRS}(\hat{t}_{Ry})$ is robust to misspecification (of η by ξ) provided we have balanced sampling.

What if the sample is not balanced? The leading term in prediction variance of the ratio estimator under η is then

$$\sigma^2(N^2/n)(1-n/N)\bar{v}_s(\bar{x}_r/\bar{x}_s)^2 .$$

We can compare this expression with the leading term in the η-expectation of $\hat{V}_{SRS}(\hat{t}_{Ry})$. We see that $\hat{V}_{SRS}(\hat{t}_{Ry})$ will tend to overestimate the actual prediction variance of the ratio estimator when $\bar{x}_r < \bar{x}_s$ (a situation where the ratio estimator in fact has a low variance), and will tend to underestimate this prediction variance when the sample is such that $\bar{x}_r > \bar{x}_s$ (a situation when the ratio estimator in fact has a high variance). This bias can be corrected, leading to a robust variance estimator (in the sense of being approximately unbiased under η irrespective of the "balance" of the sample) of the form

$$\hat{V}_{R/rob}(\hat{t}_{Ry}) = (\bar{x}_r/\bar{x}_s)^2\hat{V}_{SRS}(\hat{t}_{Ry}) .$$

Can we do better? Suppose we require our robust variance estimator that is approximately unbiased under η to also be exactly unbiased under ξ. Here we observe that

$$E_\xi\left(\hat{V}_{R/rob}(\hat{t}_{Ry})\right) = \sigma^2(N^2/n)(1-n/N)\left(\bar{x}_r^2/\bar{x}_s\right)\left[1-n^{-1}\left\{s_{xx}/\bar{x}_s^2\right\}\right]$$

while

$$Var_\xi(\hat{t}_{Ry}-t_y) = \sigma^2(N^2/n)(1-n/N)(\bar{x}_r\bar{x}/\bar{x}_s) .$$

Equating these two expressions leads to a modified version of $\hat{V}_{SRS}(\hat{t}_{Ry})$ first described by Royall and Eberhardt (1975):

$$\hat{V}_{R/rob}(\hat{t}_{Ry}) = \left(\bar{x}_r\bar{x}/\bar{x}_s^2\right)\left[1-n^{-1}\left\{s_{xx}/\bar{x}_s^2\right\}\right]^{1}\hat{V}_{SRS}(\hat{t}_{Ry}) .$$

See also Royall and Cumberland (1981).

4.2 Robust Variance Estimation for Linear Estimators

A widely used class of estimators are linear in the sample Y-values, given by

$$\hat{t}_{wy} = \sum_s w_{is} y_i .$$

In general, the sample weight w_{is} above can depend on the values of one or more auxiliary X-variables and can also be sample dependent, in the sense that it can also depend on the X-values of all the sample units (hence the "s" subscript). However, w_{is} is not a function of the sample Y-values.

Our aim is to use the approach described in the previous section to develop an estimator for the prediction variance of $\hat{t}_{wy}$ that is robust to misspecification. In this context, our working model for the distribution of the population Y-values is the very general specification (denoted by ξ in what follows):

$$E_\xi(y_i) = \mu(x_i;\omega) = \mu_i$$
$$Var_\xi(y_i) = \sigma^2(x_i;\omega) = \sigma_i^2 .$$

The ξ-bias of $\hat{t}_{wy}$ is then

$$E_\xi\left(\hat{t}_{wy} - t_y\right) = \sum_s w_{is}\mu_i - \sum_U \mu_i .$$

Since μ_i is $O(1)$, it follows that w_{is} must be $O(N/n)$ if $\hat{t}_{wy}$ is to be unbiased under ξ. We assume this. The prediction variance of $\hat{t}_{wy}$ under ξ is then

$$Var_\xi\left(\hat{t}_{wy} - t_y\right) = \sum_s (w_{is} - 1)^2 \sigma_i^2 + \sum_r \sigma_i^2 .$$

As always we also assume non-informative sampling, so a consistent estimate $\hat{\omega}$ of ω can be calculated from sample data and a simple "plug-in" estimator of σ_i^2 is then $\hat{\sigma}_i^2 = \sigma^2(x_i;\hat{\omega})$. This immediately leads to a consistent estimator of the prediction variance of $\hat{t}_{wy}$:

$$\hat{V}_\xi\left(\hat{t}_{wy}\right) = \sum_s (w_{is} - 1)^2 \hat{\sigma}_i^2 + \sum_r \hat{\sigma}_i^2 .$$

Recollect that w_{is} is $O(N/n)$ so the leading term in this estimated variance is its first (sample) term.

It is clear that the validity of this estimator depends on specification of $\sigma^2(x_i;\omega)$. We therefore now develop a modified version of this estimator that remains valid even when this variance function is misspecified.

The basic idea is exactly the same as the one used to develop a robust estimator of the prediction variance of the ratio estimator in the previous section. We replace the leading term in the "plug-in" estimator of variance above by a term whose validity only depends on the specification of the model ξ being correct to first, rather than second, order.

Suppose $\hat{\mu}_i = \mu(x_i;\hat{\omega})$ is an unbiased estimate of $E_\eta(y_i)$ under the (unknown) "true" model η for the population. This implies $E_\eta(y_i - \hat{\mu}_i)^2 = Var_\eta(y_i) + O(n^{-1})$ irrespective of the actual specification of $Var_\eta(y_i)$. An alternative variance estimator for $\hat{t}_{wy}$ is then

$$\hat{V}_{\xi,robust}\left(\hat{t}_{wy}\right) = \sum_s (w_{is}-1)^2 (y_i - \hat{\mu}_i)^2 + \sum_r \hat{\sigma}_i^2 .$$

To illustrate this approach, consider the working model ξ where all units in some specified part of the population (e.g. a stratum) are assumed to have the same mean μ and the same variance σ^2. Suppose further that we propose to estimate the total of Y for this (sub)population using the linear estimator $\hat{t}_{wy}$ based on fixed weights w_{is}. Under ξ,

$$E_\xi\left(\hat{t}_{wy} - t_y\right) = \mu\left(\sum_s w_{is} - N\right)$$

so the sample weights have to sum to N for unbiasedness under ξ. We assume this. The corresponding prediction variance of $\hat{t}_{wy}$ is then

$$Var_\xi\left(\hat{t}_{wy} - t_y\right) = \sigma^2\left(\sum_s (w_{is}-1)^2 + (N-n)\right).$$

The standard approach to estimating this prediction variance is to calculate an unbiased estimator of μ (under ξ) using the weighted average

$$\hat{\mu}_w = N^{-1}\sum_s w_{is} y_i$$

and then note that

$$E_\xi (y_i - \hat{\mu}_w)^2 = \left(1 - 2\frac{w_{is}}{N} + \frac{1}{N^2}\sum_s w_{js}^2\right)\sigma^2.$$

Consequently an unbiased estimator of σ^2 under ξ is

$$\hat{\sigma}_w^2 = \frac{1}{n}\sum_s \left(1 - 2\frac{w_{is}}{N} + \frac{1}{N^2}\sum_s w_{js}^2\right)^{-1} (y_i - \hat{\mu}_w)^2$$

implying the following unbiased estimator of the prediction variance of $\hat{t}_{wy}$ under ξ:

$$\hat{V}_\xi\left(\hat{t}_{wy}\right) = \hat{\sigma}_w^2\left(\sum_s (w_{is}-1)^2 + (N-n)\right).$$

However, this estimator will be biased if the assumption of constant variance for the y_i is incorrect.

Suppose now that the true variance of unit i in the population is γ_i^2. To distinguish this case from the constant variance model ξ, we use a subscript of η below. The true prediction variance of $\hat{t}_{wy}$ will then be

$$Var_\eta\left(\hat{t}_{wy} - t_y\right) = \sum_s (w_{is}-1)^2 \gamma_i^2 + \sum_r \gamma_i^2 .$$

A robust variance estimator for $\hat{t}_{wy}$ (in the sense of being consistent under both ξ and η) is then

$$\hat{V}_{\xi,robust}\left(\hat{t}_{wy}\right)=\sum\nolimits_{s}(w_{is}-1)^{2}(y_{i}-\hat{\mu}_{w})^{2}+(N-n)\hat{\sigma}_{w}^{2}.$$

It is not difficult to see that this robust variance estimate will not be exactly unbiased under ξ. However, the slightly modified alternative below is:

$$\hat{V}_{\xi D}\left(\hat{t}_{wy}\right)=\sum\nolimits_{s}\left(\frac{(w_{is}-1)^{2}(y_{i}-\hat{\mu}_{w})^{2}}{1-2\frac{w_{is}}{N}+\frac{1}{N^{2}}\sum\nolimits_{s}w_{js}^{2}}\right)+(N-n)\hat{\sigma}_{w}^{2}.$$

In particular, this alternative variance estimator is unbiased for the prediction variance of $\hat{t}_{wy}$ under the constant variance model ξ and approximately unbiased under the more general model η.

Application of this robust approach to prediction variance estimation for the separate ratio estimator is discussed in Royall and Cumberland (1981). This leads to the variance estimator

$$\begin{aligned}\hat{V}_{\xi D}(\hat{t}_{RSy})&=\sum\nolimits_{h}\left(\frac{N_{h}^{2}}{n_{h}}\right)\left(\frac{\bar{x}_{h}}{\bar{x}_{sh}}-\frac{n_{h}}{N_{h}}\right)\left(\frac{\bar{x}_{h}}{\bar{x}_{sh}}\right)\frac{1}{n_{h}}\sum\nolimits_{s_{h}}\left\{\frac{(y_{i}-\hat{\beta}_{h}x_{i})^{2}}{1-\left(x_{i}/n_{h}\bar{x}_{sh}\right)}\right\}\\&\approx\sum\nolimits_{\mathrm{h}}\left(\frac{N_{h}^{2}}{n_{h}}\right)\left(\frac{\bar{x}_{h}}{\bar{x}_{sh}}\right)^{2}\frac{1}{n_{h}}\sum\nolimits_{s_{h}}\left\{\frac{(y_{i}-\hat{\beta}_{h}x_{i})^{2}}{1-\left(x_{i}/n_{h}\bar{x}_{sh}\right)}\right\}.\end{aligned}$$

4.3 The Ultimate Cluster Variance Estimator

Recollect model C where the population elements all had the same mean and variance and were grouped into clusters, with all elements in the same cluster equi-correlated. This homogeneity within clusters means that sample weights will be the same for all elements in a cluster and so a general linear estimator takes on the form:

$$\hat{t}_{Cy}=\sum\nolimits_{s}w_{g}\sum\nolimits_{s_{g}}y_{i}=\frac{1}{q}\sum\nolimits_{s}\hat{t}_{Cgy}$$

where

$$\hat{t}_{Cgy}=qw_{g}\sum\nolimits_{s_{g}}y_{i}=qw_{g}m_{g}\bar{y}_{sg}$$

can be interpreted as the predictor of the overall population total t_y based just on the sample data from sample cluster g. Under C, unbiasedness of $\hat{t}_{Cy}$ requires that $\sum\nolimits_{s}w_{g}m_{g}=N$, while unbiasedness of $\hat{t}_{Cgy}$ requires $w_g m_g = N/q$. Since this is not generally the case, these cluster specific predictors are typically biased.

However, it is also often the case that the cluster sample sizes m_g, and hence the cluster weights w_g, vary little from cluster to cluster. In such cases the $\hat{t}_{Cgy}$ will all be approximately unbiased for t_y and will have approximately the same variance. This suggests that we can estimate the variance of the overall predictor $\hat{t}_{Cy}$ from the variability of the $\hat{t}_{Cgy}$.

Since the $\hat{t}_{Cgy}$ are uncorrelated with one another, this suggests a variance estimator for $\hat{t}_{Cy}$ of the form

$$\hat{V}_{UC}(\hat{t}_{Cy}) = \frac{1}{q(q-1)}\sum\nolimits_s \left(\hat{t}_{Cgy} - \hat{t}_{Cy}\right)^2 .$$

This estimator is often referred to as the ultimate cluster variance estimator. It is straightforward to calculate and does not require an estimate of the intracluster correlation ρ. We also observe that it is an estimator of the variance of $\hat{t}_{Cy}$, not its prediction variance. Hence it is only appropriate when the sampling fraction is small, as is the case in most social surveys.

Under model C (denoted by ξ below)

$$E_\xi\left(\hat{V}_{UC}(\hat{t}_{Cy})\right) = Var_\xi(\hat{t}_{Cy}) + \frac{\mu^2}{q(q-1)}\sum\nolimits_s \left\{qm_g w_g - N\right\}^2 .$$

That is, the ultimate cluster variance estimator is generally upwardly biased for the actual variance of $\hat{t}_{Cy}$. Furthermore, this bias depends on the bias of the cluster specific predictors $\hat{t}_{Cgy}$, in the sense that it vanishes when these predictors are themselves unbiased. It is easy to see that this occurs if the m_g are the same for all $g \in s$, in which case we must have $w_g = N/n$.

We can "bias correct" the ultimate cluster variance estimator. To see this, define $\hat{\mu}_{Cy} = N^{-1}\hat{t}_{Cy}$. Since $E_\xi(\hat{\mu}_{Cy}) = \mu$, it follows

$$E_\xi\left[\hat{V}_{UC}(\hat{t}_{Cy}) - \frac{\hat{\mu}_{Cy}^2}{q(q-1)}\sum\nolimits_s \left(qm_g w_g - N\right)^2\right] = K_s Var_\xi(\hat{t}_{Cy})$$

where

$$K_s = 1 - \frac{1}{N^2 q(q-1)}\sum\nolimits_s \left(qm_g w_g - N\right)^2 .$$

An unbiased estimator of $Var_\xi(\hat{t}_{Cy})$ is therefore

$$\hat{V}_{UC}^*(\hat{t}_{Cy}) = K_s^{-1}\left[\hat{V}_{UC}(\hat{t}_{Cy}) - \frac{\hat{\mu}_{Cy}^2}{q(q-1)}\sum\nolimits_s \left(qm_g w_g - N\right)^2\right].$$

Like the standard ultimate cluster variance estimator, $\hat{V}_{UC}^*(\hat{t}_{Cy})$ is a conservative estimator of the prediction variance of $\hat{t}_{Cy}$, since

$$Var_{\xi}(\hat{t}_{Cy} - t_y) = Var_{\xi}(\hat{t}_{Cy}) - \sigma^2 \begin{bmatrix} 2\sum_s w_g m_g \left\{1 + (M_g - 1)\rho\right\} \\ -\sum_U M_g \left\{1 + (M_g - 1)\rho\right\} \end{bmatrix}.$$

The term in square brackets is a monotonically increasing function of ρ, ranging from a minimum value of $\sigma^2 N$ when $\rho = 0$ to a maximum value approximately equal to $\sigma^2 N\overline{M}$ at $\rho = 1$. Here $\overline{M}$ is the population average of the M_g.

A version of this bias corrected ultimate cluster variance estimator for the more realistic case where the sample weights vary from element to element within a cluster (e.g. after post-stratification) is given by

$$\hat{V}^*_{UC}(\hat{t}_{Cy}) = \frac{1}{K_s q(q-1)} \sum_s \left\{ \left(\hat{t}_{Cgy} - \hat{t}_{Cy}\right)^2 - \hat{\mu}^2_{Cy}\left(q m_g \overline{w}_g - N\right)^2 \right\}$$

where

$$\overline{w}_g = m_g^{-1} \sum_{s_g} w_i$$

$$\hat{\mu}_{wg} = \sum_{s_g} w_i y_i / \sum_{s_g} w_i$$

$$\hat{t}_{Cgy} = q \sum_{s_g} w_i y_i = q m_g \overline{w}_g \hat{\mu}_{wg}$$

$$K_s = 1 - \frac{1}{N^2 q(q-1)} \sum_s \left(q m_g \overline{w}_g - N\right)^2 .$$

4.4 Variance Estimation for Non-Linear Estimators

So far we have considered the case of variance estimation for linear estimators. We now consider variance estimation for more general non-linear estimators. To start, we consider estimators of differentiable functions of population totals.

Here the target census value is $\theta = f(t_1, t_2, \ldots , t_m)$, which is assumed to be a differentiable function of the population totals $t_1, t_2, \ldots , t_m$ of m Y-variables. The natural estimate of θ is the "plug-in" estimator $\hat{\theta}$ = f($\hat{t}_1$, $\hat{t}_2$, ... , $\hat{t}_m$). We note that if the component estimates $\hat{t}_1$, $\hat{t}_2$, ... , $\hat{t}_m$ are unbiased, then $\hat{\theta}$ will be approximately unbiased in large samples.

A first order approximation to the sample error of $\hat{\theta}$ is

$$\hat{\theta} - \theta = f(\hat{t}_1, \cdots, \hat{t}_m) - f(t_1, \cdots t_m) \approx \sum_{a=1}^{m} \frac{\partial f}{\partial t_a}(\hat{t}_a - t_a)$$

where $\partial f / \partial t_a$ is the partial derivative of f with respect to its a^{th} argument, evaluated at $t_1, t_2, \ldots , t_m$. A first order approximation to the variance of this sample error is

$$Var\left(\hat{\theta} - \theta\right) \approx \sum_{a=1}^{m} \sum_{b=1}^{m} \left(\frac{\partial f}{\partial t_a}\right)\left(\frac{\partial f}{\partial t_b}\right) Cov_{\xi}(\hat{t}_a - t_a, \hat{t}_b - t_b).$$

An estimate of this first order approximation is therefore

$$\hat{V}(\hat{\theta}) \approx \sum_{a=1}^{m}\sum_{b=1}^{m}\left(\frac{\partial f}{\partial \hat{t}_a}\right)\left(\frac{\partial f}{\partial \hat{t}_b}\right)\hat{C}_{\xi}(\hat{t}_a,\hat{t}_b)$$

where $\hat{C}_{\xi}(\hat{t}_a,\hat{t}_b)$ is an estimate of the covariance between the prediction errors of $\hat{t}_a$ and $\hat{t}_b$ and $\partial f/\partial \hat{t}_a$ is the partial derivative of f with respect to its a^{th} argument, evaluated at $\hat{t}_1,\ \hat{t}_2, \ldots ,\ \hat{t}_m$.

Note that the covariance estimate $\hat{C}$ can be calculated using any of the different variance estimation methods described so far.
An important special case is where the estimators $\hat{t}_1,\ \hat{t}_2, \ldots ,\ \hat{t}_m$ are all linear. Then

$$Var_{\xi}\left(\hat{\theta}-\theta\right) \approx Var_{\xi}\left(\hat{t}_z - t_z\right) \approx Var_{\xi}\left(\hat{t}_{\hat{z}} - t_z\right)$$

where t_z is the population total of the linearised variable $z_i = \sum_{a=1}^{m}\left(\frac{\partial f}{\partial t_a}\right)y_{ai}$ and

$$\hat{t}_z = \sum_s w_{is}z_i = \sum_s w_{is}\left(\sum_{a=1}^{m}\left(\frac{\partial f}{\partial t_a}\right)y_{ai}\right)$$
$$\hat{t}_{\hat{z}} = \sum_s w_{is}\hat{z}_i = \sum_s w_{is}\left(\sum_{a=1}^{m}\left(\frac{\partial f}{\partial \hat{t}_a}\right)y_{ai}\right).$$

A first order approximation to the variance of $\hat{\theta}$ can be computed as the estimated variance of the sample error of $\hat{t}_{\hat{z}}$, treating the estimated quantities $\hat{z}_i$ as "observed" quantities (Woodruff, 1971).

Many census parameters are defined as solutions to population level estimating equations (e.g. the finite population median). "Plug-in" methods are also used here to calculate the required estimates, and Taylor series linearisation is used to estimate the variance of these estimators. Thus θ is defined by a population level estimating equation if it is a solution to

$$H(\theta) = \sum_U f(\mathbf{y}_i;\theta) = 0$$

where f is assumed to be a differentiable function of θ. To estimate this quantity, we replace the population parameter $H(\theta)$ by a "linear" estimator $\hat{H}_w(\theta)$ and then estimate θ by $\hat{\theta}$, where

$$\hat{H}_w(\hat{\theta}) = \sum_s w_{is} f(\mathbf{y}_i;\hat{\theta}) = 0.$$

Variance estimation in this case is based on Taylor series linearisation

$$0 = \hat{H}(\hat{\theta}) \approx \hat{H}(\theta) + \left(\frac{\partial \hat{H}}{\partial \theta}\right)(\hat{\theta}-\theta) = \hat{H}(\theta) + (\hat{\theta}-\theta)\sum_s w_{is}\frac{\partial f(\mathbf{y}_i;\theta)}{\partial \theta}$$

from which we obtain the first order approximation

$$Var_{\xi}(\hat{\theta}-\theta) \approx \left(\sum_s w_{si}\frac{\partial f(\mathbf{y}_i;\theta)}{\partial \theta}\right)^{-1} Var_{\xi}(\hat{H}(\theta))\left(\sum_s w_{is}\frac{\partial f(\mathbf{y}_i;\theta)}{\partial \theta}\right)^{-1}.$$

The so-called “sandwich” estimator of variance is then

$$\hat{V}_{\xi}(\hat{\theta}) = \left(\sum\nolimits_{s} w_{si} \frac{\partial f(\mathbf{y}_i;\theta)}{\partial \hat{\theta}}\right)^{-1} \hat{V}_{\xi}(\hat{H}(\hat{\theta})) \left(\sum\nolimits_{s} w_{is} \frac{\partial f(\mathbf{y}_i;\theta)}{\partial \hat{\theta}}\right)^{-1}$$

which corresponds to evaluating the partial derivatives at $\hat{\theta}$, and replacing the variance term by an appropriate “plug-in” estimate defined by replacing θ by $\hat{\theta}$ in

$$\hat{V}_{\xi}(\hat{H}(\theta)) = \hat{V}_{\xi}\left(\sum\nolimits_{s} w_{is} f(\mathbf{y}_i;\theta)\right) = \hat{V}_{\xi}\left(\sum\nolimits_{s} w_{is} z_i(\theta)\right)$$

where $z_i(\theta) = f(\mathbf{y}_\mathrm{i};\ \theta)$ is treated as just another population Y-variable.

4.5 Replication-Based Methods of Variance Estimation

The definition of the census parameters of interest may be so complex that application of Taylor series linearisation methods for variance estimation is difficult, if not impossible (e.g. so-called “chain ratio” indexes that are often estimated using business survey data). In such cases we can use alternative variance estimation methods that are “simple” to implement, but are typically numerically intensive.

The basic idea behind these methods is simple. We “simulate” the variance of a statistic by

(i) making repeated draws from a distribution whose variance is related in a simple (and known) way to the variance of interest;
(ii) empirically estimating the variance of this “secondary” distribution;
(iii) adjusting this variance estimate so that it is an estimate of the variance of interest.

The simplest version of this approach is the Random Groups Variance Estimator. Its origin is in the idea of interpenetrating samples (Mahalonobis,1946; Deming, 1960), where the actual sample selected is made up of G independent replicate or interpenetrating subsamples, each one of which is “representative” of the population, being drawn according to the same design and with the same sample size n/G.

Let $\hat{\theta}_g$ be the estimate of the θ based on the g^{th} replicate sample The overall estimate is then $\hat{\theta} = G^{-1}\sum_g \hat{\theta}_g$. By construction, $\{\hat{\theta}_g, g = 1, \ldots, G\}$ are independent and identically distributed and so we can estimate the variance of their (common) distribution by their empirical variance around their average, the overall estimate $\hat{\theta}$. The variance of $\hat{\theta}$ is then this “replicate variance” divided by the number of replicates, G. We therefore estimate the variance of $\hat{\theta}$ by simply dividing this empirical variance by G, leading to

$$\hat{V}_{rep}(\hat{\theta}) = \frac{1}{G(G-1)} \sum_{g=1}^{G} \left(\hat{\theta}_g - \hat{\theta}\right)^2 .$$

The idea still works even if the replicate estimates are not identically distributed. All that is required is that they are independent of one another, and each is unbiased for θ.

The main problem here is that replicated sample designs are rare. Consequently what is often done is to construct the replicates after the sample is selected, by randomly allocating sample units to *G* groups in such a way that each group is at least approximately independent of the other groups.

This allocation is not as straightforward as it sounds. For stratified designs we can do random grouping within strata, provided there is sufficient sample size within each stratum to carry this out. If not, then random grouping can be applied to the sample as a whole, preserving the strata when splitting the sample between the groups. For multistage designs the allocation is typically carried out at PSU (primary sampling unit) level. In addition, the "average" estimate $\hat{\theta}$ in the variance formula above is often replaced by the "full sample" estimate of this quantity.
As we noted with the ultimate cluster variance estimator, the replication variance estimator is an estimator of the variance of $\hat{\theta}$. It is not an estimator of the prediction variance of this estimator. Consequently the variance estimate does not go to zero as the sample size approaches the population size. This is of no great concern when sample sizes within strata are small compared to stratum population sizes (population-based surveys). However, in many business surveys, sample sizes within strata can be a substantial fraction of the strata populations. In such cases, it is standard to multiply the stratum level replicated groups variance estimates by appropriate finite population correction factors.

A problem with the replication-based approach to variance estimation is the stability of these estimates. The more groups there are, the more stable these variance estimates are. However, the more groups there are, the harder it is to "randomly group" the sample. This leads naturally to the idea of using overlapping (non-independent) groups.

There are essentially two approaches to using overlapping groups. The first is Balanced Repeated Replication, where groups are formed using experimental design methods so that covariances induced by the same unit belonging to different groups "cancel out" in the (non-overlapping) random groups variance formula. This can be quite difficult to accomplish in general. The method is typically restricted to certain types of multistage designs, with $G = 2$ and is rarely used in business surveys. See Wolter (1985) and Shao and Tu (1995).

The other, much more commonly used, approach is Jackknife variance estimation. Here again the sample is divided into *G* groups, but this time the *G* estimates are computed by "dropping out" each of the *G* groups from the sample in turn. The variability between these dependent estimates is then used to estimate the variability of the overall estimate of θ.

Let $\hat{\theta}_{(g)}$ be the estimator of θ based on the sample excluding group *g*. The Jackknife estimator of variance is then given by

$$\hat{V}_{jack}(\hat{\theta}) = \frac{G-1}{G}\sum_{g=1}^{G}\left(\hat{\theta}_{(g)} - \hat{\theta}\right)^2 .$$

There are two types of jackknife. The Type 1 jackknife is defined by $\hat{\theta}$ = average of the $\hat{\theta}_{(g)}$. The Type 2 jackknife is defined by $\hat{\theta}$ = "full sample" estimate of θ. Note that the ANOVA identity implies that the Type 2 jackknife will be more conservative (produce larger estimates of variance) than the Type 1 jackknife.

Some important points about applying the jackknife method in practice are:

1. The jackknife variance estimate is typically computed at PSU level in multistage samples. That is, the G groups are defined as groups of PSUs.

2. The most common type of jackknife is when G is equal to the number of PSU's in the sample, that is one PSU is dropped from sample each time a value of $\hat{\theta}_{(g)}$ is calculated.

3. Like the random groups variance estimate, the jackknife variance estimate does not include a finite population correction. This needs to be applied separately.

There can be a heavy computational burden when $G = n$ in the jackknife so it is sometimes convenient to calculate an approximation that can be computed in one "pass" of the sample data. This is the so-called linearised jackknife and is defined by essentially replacing $\hat{V}_{jack}(\hat{\theta})$ by a first order Taylor series approximation to it.

To start, we make the following assumptions:

(i) Single stage sampling.

(ii) A superpopulation model ξ under which $E_\xi(y_i) = \mu_i$ for $i \in s$.

We first approximate $\hat{\theta}$ by $\hat{\theta} = \hat{\theta}(\boldsymbol{\mu}) + \sum_s \left(\frac{\partial \hat{\theta}}{\partial y_i} \right)_{\mathbf{y}=\boldsymbol{\mu}} (y_i - \mu_i)$, where $\boldsymbol{\mu}$ is the n-vector of expected values for the sample values $\mathbf{y}$ and $\hat{\theta}(\boldsymbol{\mu})$ is the value of $\hat{\theta}$ when these sample Y-values are replaced by $\boldsymbol{\mu}$. Similarly, we approximate $\hat{\theta}_{(i)}$ by $\hat{\theta}_{(i)} = \hat{\theta}_{(i)}(\boldsymbol{\mu}_{(i)}) + \sum_{j \neq i \in s} \left(\frac{\partial \hat{\theta}_{(i)}}{\partial y_j} \right)_{\mathbf{y}_{(i)}=\boldsymbol{\mu}_{(i)}} (y_j - \mu_j)$, where $\boldsymbol{\mu}_{(i)} = \boldsymbol{\mu}$ with the expected value for y_i deleted, $\hat{\theta}_{(i)}$ = estimate of θ based on the sample excluding y_i and $\hat{\theta}_{(i)}(\boldsymbol{\mu}_{(i)}) = \hat{\theta}_{(i)}$ evaluated at $\boldsymbol{\mu}_{(\mathrm{i})}$.

Finally, we need two extra assumptions:

(1) $\hat{\theta}(\boldsymbol{\mu}) = \hat{\theta}_{(i)}(\boldsymbol{\mu}_{(i)}) = \theta_0$.

(2) $\left(\frac{\partial \hat{\theta}}{\partial y_j} \right)_{y=\boldsymbol{\mu}} = \frac{n}{n-1} \left(\frac{\partial \hat{\theta}_{(i)}}{\partial y_j} \right)_{y_{(i)}=\boldsymbol{\mu}_{(i)}}$.

We can then replace the approximation to $\hat{\theta}_{(i)}$ above by

$$\hat{\theta}_{(i)} = \frac{n}{n-1} \left\{ \hat{\theta} - \left(\frac{\partial \hat{\theta}}{\partial y_i} \right)_{y=\boldsymbol{\mu}} (y_i - \mu_i) \right\} - \frac{\theta_0}{n-1}.$$

This expression can be calculated for every unit in sample in one "pass" of the data. Its use in the jackknife variance estimator formula then leads to the linearised Type 1 jackknife variance estimator.

$$\hat{V}^{(1)}_{jack,lin}(\hat{\theta}) = \frac{n}{n-1}\sum\nolimits_s \left\{ \left(\frac{\partial\hat{\theta}}{\partial y_i}\right)_{\mathbf{y}=\hat{\boldsymbol{\mu}}} (y_i - \hat{\mu}_i) - \frac{1}{n}\sum\nolimits_s \left(\frac{\partial\hat{\theta}}{\partial y_j}\right)_{\mathbf{y}=\hat{\boldsymbol{\mu}}} (y_j - \hat{\mu}_j) \right\}^2$$

where $\hat{\boldsymbol{\mu}}$ is the full sample estimate of $\boldsymbol{\mu}$. The linearised Type 2 jackknife variance estimator is obtained similarly after replacing θ_0 by $\hat{\theta}$:

$$\hat{V}^{(2)}_{jack,lin}(\hat{\theta}) = \frac{n}{n-1}\sum\nolimits_s \left\{ \left(\frac{\partial\hat{\theta}}{\partial y_i}\right)_{\mathbf{y}=\hat{\boldsymbol{\mu}}} (y_i - \hat{\mu}_i) - \hat{\theta}\left(\frac{n^2 - 3n + 1}{n(n-1)}\right) \right\}^2 .$$

Finally, we briefly describe application of the bootstrapping idea to variance and confidence interval estimation in surveys. For many sample designs sample sizes are too small for central limit behaviour to be applicable (e.g. fine strata containing relatively few units) and the distribution of the sample error may be quite non-normal. Bootstrapping is then a way of estimating the sampling distribution directly in such cases.

As usual, let θ denote the census parameter of interest. To simplify presentation, we assume that θ is defined in terms of the population values of a single Y-variable with superpopulation distribution specified by a general model where

$$E_\xi(y_i) = \mu(x_i;\omega) = \mu_i$$
$$Var_\xi(y_i) = \sigma^2(x_i;\omega) = \sigma_i^2$$

where $\hat{\omega}$ is a model-unbiased estimator of ω calculated from the sample data. Let $\{r_{std,i};\ i \in s\}$ then be the set of studentised residuals generated by the sample data under this model, and satisfying $E_\xi(r_{std,i}) = 0$ and $Var_\xi(r_{std,i}) = 1$. The steps in the bootstrap procedure are then

1. Generate N bootstrap residuals $\{r_i^*;\ i \in U\}$ by sampling at random and with replacement N times from the n studentised residuals $\{r_{std,i};\ i \in s\}$.
2. Generate a bootstrap realisation of the population Y-values: $y_i^* = \mu(x_k;\hat{\omega}) + \sigma(x_k;\hat{\omega})r_i^*$; $i \in U$.
3. Compute a bootstrap estimate $\hat{\theta}^*$ of θ based on the values $\{y_i^*;\ i \in s\}$, together with the actual value θ^* of θ for the bootstrap population. The bootstrap realisation of the sample error is then $\hat{\theta}^*$ - θ^*.
4. Repeat steps 1 - 3 above a large number of times, thus generating a distribution of bootstrap sample errors. Denote the (known) mean of this bootstrap distribution by $E^*(\hat{\theta}^*$ - $\theta^*)$, and its (known) variance by $Var^*(\hat{\theta}^*$ - $\theta^*)$.

The bootstrap estimate of θ is given by $\hat{\theta}_B = \hat{\theta} + E^*(\hat{\theta}^* - \theta^*)$. The bootstrap estimate of variance (of the bootstrap estimate) can be calculated as $Var^*(\hat{\theta}^*$ - $\theta^*)$. However, this is typically an underestimate since it does not take into account the error in estimating ω. It is better to rescale the bootstrap sample error distribution so that its variance is the larger of $Var^*(\hat{\theta}^*$ - $\theta^*)$ and an estimate of the variance which allows for error in estimation of ω (e.g. a jackknife estimate). Note that if $\hat{\theta}$ represents a "best" estimate of θ, then the bootstrap sample error distribution can be centred at zero prior to this rescaling.

A 100(1 - α)% confidence interval for θ can be "read off" from the final bootstrap sample error distribution as

$$\left(\hat{\theta}_B - Q^*(\frac{\alpha}{2}), \hat{\theta}_B + Q^*(1-\frac{\alpha}{2})\right)$$

where $Q^*(\gamma)$ denotes the γ-th quantile of this distribution.

Note that the bootstrap procedure defined above depends on correct specification of the variance function $\sigma^{\square}(x;\omega)$. A robust model-based bootstrap can be defined by replacing the studentised residuals by "raw" residuals $r_{raw,i} = y_i$ - $\mu(x_i; \hat{\omega})$. The remaining steps in the bootstrap procedure are unchanged (Chambers and Dorfman, 1994).

5. Estimation for Multipurpose Surveys

The theory developed in the previous chapters largely assumed a single scalar auxiliary variable X. The exception to this was stratified estimation where X was implicitly defined as a mix of stratum indicators and (for ratio and regression estimation) a single size variable. In general, however, we may have more than one size variable and many different "types" of stratifiers (e.g. regional and industry indicators). In such situations we need to consider models where X is a vector. This chapter therefore focuses on the realistic situation where X is a mix of stratum identifiers of different types and size variables corresponding to different measures of activity of a population unit and where it is reasonable to link the survey variable Y and this vector X via a working model corresponding to the general linear model defined earlier. In particular, the aim is to explore practical issues associated with employing sampling weights based on this type of multiple regression model.

To start, we define $\mathbf{y}$ to be the N-vector of population values of some characteristic of interest, whose total t_y is to be estimated. Associated with these population units we assume there exists a known matrix $\mathbf{X}$ defined as the $N \times p$ matrix of values of p auxiliary variables. This matrix is assumed to be of full rank. As always we assume uninformative sampling and full response (or uninformative nonresponse).

Our working model is the general linear model, defined by $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $E_\xi(\varepsilon) = 0$ and $Var_\xi(\varepsilon) = \sigma^2\mathbf{V}$, where $\mathbf{V} = \mathbf{V}(\mathbf{X})$ is a known positive definite matrix, partitioned conformably into sample and non-sample submatrices as:

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr} \end{bmatrix}.$$

We consider a general linear estimator for the population total of Y based on sample weights that are fixed for the sample units (i.e. they do not vary for different Y-variables). This is of the form

$$\hat{t}_{wy} = \sum\nolimits_s w_{is}(\mathbf{X})y_i = \mathbf{w}_s'\mathbf{y}_s.$$

5.1 Calibrated Weighting

A common requirement for such general purpose sample weights is that they are calibrated on a set of "benchmark" variables. That is, when sampled values of these variables are "weighted up", they recover the known population totals of these variables. If these benchmark variables are a subset of the variables defining $\mathbf{X}$, then it is easy to see that any set of weights that leads to an unbiased estimator (or predictor) of t_y under this working model is also calibrated. In fact, these weights, by definition, must be calibrated on all the variables defining $\mathbf{X}$, since the unbiasedness condition implies

$$E_\xi\left(\hat{t}_y - t_y\right) = E_\xi\left(\mathbf{w}_s'\mathbf{y}_s - \mathbf{1}'\mathbf{y}\right) = \left(\mathbf{w}_s'\mathbf{x}_s - \mathbf{1}'\mathbf{X}\right)\boldsymbol{\beta} = 0$$

which is satisfied if and only if the calibration condition, $\mathbf{w}_s'\mathbf{x}_s = \mathbf{1}'\mathbf{X}$, is satisfied. That is, a set of weights that are calibrated with respect to $\mathbf{X}$ define an unbiased estimator for t_y under the model $E_\xi(\mathbf{y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$.

Since unbiasedness (i.e. calibration) is a rather weak condition, we consider "efficient" unbiased weights. Recollect that the weights defining the Best Linear Unbiased Predictor (BLUP) of t_y under the general linear model (Royall, 1976) are given by

$$\mathbf{w}_L = \mathbf{1}_s + \mathbf{H}'_L(\mathbf{X}'\mathbf{1} - \mathbf{X}'_s\mathbf{1}_s) + (\mathbf{I}_s - \mathbf{H}'_L\mathbf{X}'_s)\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr}\mathbf{1}_r$$

where $\mathbf{I}_s$ is the identity matrix of order n, $\mathbf{1}$ is a N-vector of one's, $\mathbf{1}_s$ is a n-vector of one's, $\mathbf{1}_r$ is a (N-n)-vector of one's and $\mathbf{H}_L = \left(\mathbf{X}'_s\mathbf{V}_{ss}^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}'_s\mathbf{V}_{ss}^{-1}$.

We use these weights to motivate a family of "linear unbiased" (LU) weights In order to do so we observe that $\mathbf{H}_L\mathbf{X}_s = \mathbf{I}_p$ (identity matrix of order p), so we consider all weights of the form

$$\mathbf{w}_H = \mathbf{1}_s + \mathbf{H}'(\mathbf{X}'\mathbf{1} - \mathbf{X}'_s\mathbf{1}_s) + (\mathbf{I}_s - \mathbf{H}'\mathbf{X}'_s)\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr}\mathbf{1}_r$$

where $\mathbf{H}$ is any matrix that satisfies $\mathbf{H}\mathbf{X}_s = \mathbf{I}_p$. Any set of weights that can be written in this form will be referred to as a set of LU weights. Such weights are calibrated on $\mathbf{X}$ since $\mathbf{X}'_s\mathbf{w}_H = \mathbf{X}'\mathbf{1}$ and hence define an unbiased estimator of t_y under the working linear model.

5.2 Nonparametric Weighting

The linear model assumption in the previous section may be going too far. In such a situation, the traditional nonparametric approach uses weights defined by the n-vector of sample inclusion probabilities. These are weights of the form $\mathbf{w}_\pi = \pi_s^{-1}$ (n-vector of inverse sample inclusion probabilities). The resulting estimator is the well-known Horvitz-Thompson estimator (HTE), $\hat{t}_{\pi y}$.

The HTE lacks a model-based justification. However, model-based nonparametric sample weights (Kuo, 1988) can also be defined. This corresponds to replacing the parametric general linear regression model by nonparametric nonlinear regression model, defined in terms of a single scalar auxiliary variable X, leading to an estimator of t_y of the form

$$\hat{t}_{fy} = \sum\nolimits_s y_i + \sum\nolimits_r \hat{f}(x_i)$$

where $\hat{f}(x_i)$ is a suitable nonparametric estimate of $E(y_i|\mathbf{X}_i)$. A simple choice is the Nadaraya-Watson estimate (a locally weighted average) defined by

$$\hat{f}(x) = \left(\sum\nolimits_s K\left(B^{-1}(x - x_i)\right)\right)^{-1}\left(\sum\nolimits_s K\left(B^{-1}(x - x_i)\right)y_i\right)$$

where K is a "kernel" function (typically a density function) and B is the bandwidth of the estimator. The weights associated with this estimator are given by $\mathbf{w}_f = \mathbf{1}_s + \mathbf{m}_s$, where

$$m_i = \sum\nolimits_{j\in r}\left\lfloor K\left(B^{-1}(x_j - x_i)\right)\left(\sum\nolimits_{k\in s} K\left(B^{-1}(x_j - x_k)\right)\right)^{-1}\right\rfloor.$$

Each m_i above can be interpreted as a measure of how many non-sample units are "close" to the corresponding sample unit in X-space. This also helps one understand when inverse probability weighting works – i.e. when π_i^{-1} is a "count" of the number of population units that are "like" the i^{th} sample unit. In particular, the HTE can be expected to fail when this interpretation is violated.

5.3 Calibrating Nonparametric Weights

Unfortunately, in general neither the HTE nor the nonparametric estimator described above are calibrated on **X**, so these nonparametric estimators are biased under the general linear model. There are two general approaches to remedying this situation.

The first approach (Deville and Särndal, 1992) is to choose sample weights that are "close" to the nonparametric weights but at the same time are unbiased under the general linear model (i.e. calibrated on **X**). In order to develop this approach we require a metric for "closeness". We choose the Euclidean metric Q = $(\mathbf{w}_s - \mathbf{w}_f)'\boldsymbol{\Omega}_s(\mathbf{w}_s - \mathbf{w}_f)$, where $\boldsymbol{\Omega}_s$ is a positive definite diagonal matrix of order n.

Minimising Q subject to calibration leads to sample weights

$$\mathbf{w}_f + \mathbf{H}'_\Omega(\mathbf{X}'\mathbf{1} - \mathbf{X}'_s\mathbf{w}_f)$$

where $\mathbf{H}_\Omega$ is the LU matrix $\mathbf{H}_\Omega = \left(\mathbf{X}'_s\boldsymbol{\Omega}_s^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}'_s\boldsymbol{\Omega}_s^{-1}$. Hence we can generalise by replacing $\mathbf{H}_\Omega$ above by an arbitrary LU matrix **H**, which leads to weights of the form

$$\mathbf{w}_{Hf} = \mathbf{w}_f + \mathbf{H}'(\mathbf{X}'\mathbf{1} - \mathbf{X}'_s\mathbf{w}_f).$$

Observe that such weights are calibrated on **X** for any LU matrix **H**. Also, setting $\mathbf{w}_f = \pi_s^{-1}$ leads to Generalised Regression (GREG) estimator (SSW).

The second approach tackles the problem from the other end. The idea here is that we want to use a set of diagonal LU weights, $\mathbf{w}_H = \mathbf{1}_s + \mathbf{H}'(\mathbf{X}'\mathbf{1} - \mathbf{X}'_s\mathbf{1}_s)$ defined by some LU matrix **H**. Such weights are calibrated on **X** and hence define an unbiased estimator of t_y under assumed linear model. However, suppose this model does not actually fit our data. Can we protect ourselves against bias due to potential model misspecification?

The solution is to nonparametrically "bias calibrate" the LU estimator (Chambers, Dorfman and Wehrly, 1993). We use the sample residuals to compute a nonparametric estimate of the bias, and then subtract this bias estimate from the original LU estimate. Under LU weighting, fitted values are defined by $\hat{\mathbf{y}}_s = \mathbf{X}_s\mathbf{H}\mathbf{y}_s$, with residuals $\mathbf{r}_s = (\mathbf{I}_s - \mathbf{X}'_s\mathbf{H})\mathbf{y}_s$. The nonparametrically bias calibrated weights then satisfy

$$\begin{aligned}\mathbf{w}_{Hm} &= \mathbf{w}_H + (\mathbf{I}_s - \mathbf{H}'\mathbf{x}'_s)\mathbf{m}_s \\ &= \mathbf{1}_s + \mathbf{H}'(\mathbf{x}'\mathbf{1} - \mathbf{x}'_s\mathbf{1}_s) + (\mathbf{I}_s - \mathbf{H}'\mathbf{x}'_s)(\mathbf{w}_f - \mathbf{1}_s) \\ &= \mathbf{w}_f + \mathbf{H}'(\mathbf{x}'\mathbf{1} - \mathbf{x}'_s\mathbf{w}_f) \\ &= \mathbf{w}_{Hf}\ .\end{aligned}$$

That is, we end up using the same calibrated weights as under the first approach. In other words, parametrically bias calibrating a nonparametric estimator is the same as nonparametrically bias calibrating a parametric estimator.

5.4 Problems Associated With Calibrated Weights

Since calibration is equivalent to unbiasedness under a general linear model specified by **X**, it is immediately clear that overspecification of **X** (i.e. introduction of too many calibration constraints) will lead to a loss in efficiency. The evidence for this is an increase in the variability of the sample weights as the number of constraints increases. Thus, suppose **X** contains an intercept term and **V** is the identity matrix. Then adding an extra column **z** to **X** (i.e. changing **X** to [**X z**]) increases the variance of the BLUP weights by $G_1^2(\mathbf{G}_2'\mathbf{G}_2)^{-1}$, where $G_1 = \mathbf{1}_r'(\mathbf{X}_r(\mathbf{X}_s'\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{z}_s - \mathbf{z}_r)$ and $\mathbf{G}_2 = \mathbf{z}_s - \mathbf{X}_s(\mathbf{X}_s'\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{z}_s$.

That is, the greater the "size" *p* of the linear model, the greater the variability of a set of LU weights based on that model. Equivalently, the more calibration constraints one imposes, the higher the variability in the resulting set of sample weights. This increased variability usually leads to "outlying" weights, particularly weights that are substantially negative and hence the possibility of negative estimates for strictly positive quantities, especially in domain analysis. It also results in larger standard errors. (As an aside we note that BLUP type LU weights are less liable to be negative than GREG type calibration weights.)

This problem is well known. Huang and Fuller (1978) describe an algorithm that numerically searches for strictly positive calibrated weights. In contrast, Deville and Särndal (1992) suggest replacing the Euclidean Q metric by alternative metrics that guarantee positive weights. However, we then lose the natural interpretability of Q. Also, there is then no finite sample theory (only recourse to asymptotic arguments). Bankier, Rathwell and Majkowski (1992) on the other hand adopt a more pragmatic approach, reducing *p* (removing calibration constraints), until all (calibrated) weights are strictly positive.

Other approaches have focussed on minimum mean squared error rather than minimum variance. Silva and Skinner (1997) search for lower mean squared error by using the sample data to suggest appropriate variables to include in **X** rather than by including all possible benchmark variables in this matrix (smaller *p* - less likely to get negative weights). However, this has the disadvantage of requiring that each survey variable have its own set of sample weights. Bardsley and Chambers (1984) take a different approach, searching for lower mean squared error by "ridging" **X**-based BLUP weights in order to obtain strictly positive weights. This allows some bias since ridged weights are not exactly calibrated.

Chambers (1996) extends this ridge weighting approach to include nonparametric bias calibration. Thus, one starts with an initial set of (nonparametric) weights $\mathbf{w}_f$ and then seeks a modified set of weights $\mathbf{w}_s$ that minimises the penalised Euclidean metric:

$$(\mathbf{w}_s - \mathbf{w}_f)'\boldsymbol{\Omega}_s(\mathbf{w}_s - \mathbf{w}_f) + \frac{1}{\lambda}(\mathbf{x}'\mathbf{1} - \mathbf{x}_s'\boldsymbol{\Omega}_s)'\mathbf{C}(\mathbf{x}'\mathbf{1} - \mathbf{x}_s'\boldsymbol{\Omega}_s).$$

Here λ ιs a positive scalar "ridge" parameter and **C** is a diagonal matrix of order *p* whose entries reflect

(i) the relative "importance" attached to each of the *p* calibration constraints;
(ii) the different scales of measurement for the benchmark variables in **X**.

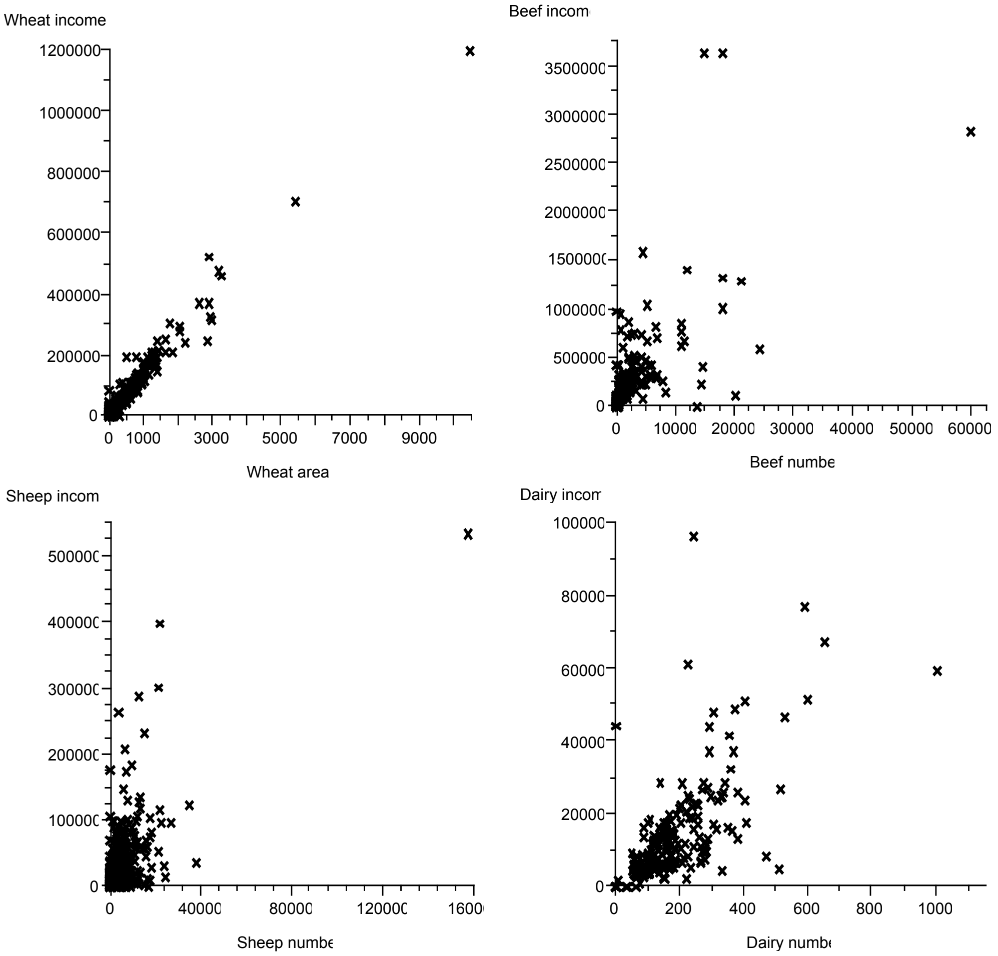
The solution to this optimisation problem is the vector of ridged weights $\mathbf{w}_\lambda = \mathbf{w}_f + \mathbf{G}_\lambda'(\mathbf{X}'\mathbf{1} - \mathbf{X}_s'\boldsymbol{\Omega}_s)$ where $\mathbf{G}_\lambda = \left(\lambda\mathbf{C}^{-1} + \mathbf{X}_s'\boldsymbol{\Omega}_s^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}_s'\boldsymbol{\Omega}_s^{-1}$ is a "ridged" LU matrix. As λ ↓ 0, the ridged weights become standard calibrated weights based on the LU matrix $\left(\mathbf{X}_s'\Omega_s^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}_s'\Omega_s^{-1}$, while as $\lambda \uparrow \infty$, the ridged weights reduce to the uncalibrated weights $\mathbf{w}_f$ (provided all elements of **C** are strictly positive and finite).

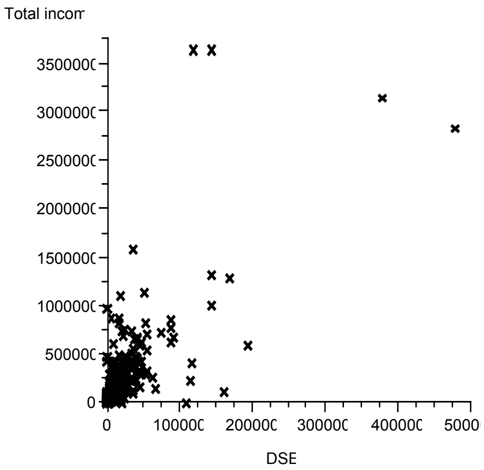
Zeroing some components of **C** allows ridge weights to smoothly interpolate between calibration weights under a “large” model defined by **X** and calibration weights under a “small” model specified by the subset of **X** defined by these “zeroed” components. Thus ridging can be interpreted as a smooth reduction in the dimension of the model.

Finally we note that ridged GREG weights are easily obtained by substituting π_s^{-1} for $\mathbf{w}_f$ and $\Omega_s diag(\pi_s^{-1})$ for Ω_s.

5.5 A Simulation Analysis of Calibrated and Ridged Weighting

In order to illustrate the behaviour of the various calibrated and ridged weighting methods described above, we reproduce results from a simulation study reported in Chambers (1986). The target population here is a group of 904 cropping, livestock and dairy farms that were surveyed in Australia in the 1980s. The scatterplots below show the distribution of the different economic variables that were measured for these farms plotted against relevant “size” variables that were also measured for these farms.





The “framework” variables available for these farms are set out in the following table.

ASIC	Unique industry classification for each farm 181 Wheat growing 182 Wheat growing + Sheep Production 183 Wheat growing + Beef cattle production 184 Sheep + Beef cattle production 185 Sheep production 186 Beef cattle production 187 Dairy farm
State	State/Territory in which farm is located NSW New South Wales VIC Victoria QLD Queensland SA South Australia WA Western Australia TAS Tasmania NT Northern Territory
Region	Identifier for 39 geographically defined regions (nested within State)
DSE	Unique size measure (Dry Sheep Equivalent) for a farm. Defined as a linear combination of the outputs from the farm

In addition we assume a set of benchmark variables, as set out in the following table. These are variables that are measured on the sample and for which population totals are assumed known.

Wheat area	Area (hectares) sown to wheat during the year
Beef number	Number of beef cattle on the farm at the end of the year
Sheep number	Number of sheep on the farm at the end of the year
Dairy number	Number of dairy cattle on the farm at the end of the year

There are five survey variables, as shown on the preceding scatterplots. These are shown on the following table.

Wheat income	Annual income from sale of wheat
Beef income	Annual income from sale of beef cattle
Sheep income	Annual income from sale of wool and sheep
Dairy income	Annual income from sale of milk products
Total income	Annual income from all four activities above

The simulation study involved a number of different sampling methods, designed to mimic actual sampling reality. In all cases 1000 samples were drawn independently according to each design. These are described below.

<table>
<tr><td>Simple Random Sampling</td><td>Random sample of size n = 100 taken without replacement from N = 904. Sample rejected if missing one or more farms from each of the seven ASIC industries, or without production in one of the four farm outputs (wheat, sheep, beef or dairy).</td></tr>
<tr><td>Size Stratification + “Compromise” Allocation</td><td>Independent random samples taken from 4 size strata, defined by values of the size variable DSE. Stratum boundaries defined so that total DSE approximately the same in each stratum. Stratum allocations defined by averaging proportional and Neyman allocation (based on DSE), resulting in the design:
<table>
<tr><th>Stratum</th><th>DSE Range</th><th>N_h</th><th>n_h</th></tr>
<tr><td>1</td><td>200 - 9499</td><td>665</td><td>50</td></tr>
<tr><td>2</td><td>9500 - 24999</td><td>166</td><td>25</td></tr>
<tr><td>3</td><td>25000 - 99999</td><td>52</td><td>18</td></tr>
<tr><td>4</td><td>100000 +</td><td>12</td><td>7</td></tr>
</table>
9 farms with DSE < 200 were excluded from selection. Sample rejected if missing one or more farms from each of the seven ASIC industries, or without wheat, sheep, beef or dairy production.</td></tr>
<tr><td>Size Stratification + “Optimal” Allocation</td><td>Same stratification and sample rejection rule as for size stratification with “compromise” allocation, but with Neyman allocation based on DSE and with the “top” stratum completely enumerated.
<table>
<tr><th>Stratum</th><th>DSE Range</th><th>N_h</th><th>n_h</th></tr>
<tr><td>1</td><td>200 - 9499</td><td>665</td><td>30</td></tr>
<tr><td>2</td><td>9500 - 24999</td><td>166</td><td>29</td></tr>
<tr><td>3</td><td>25000 - 99999</td><td>52</td><td>29</td></tr>
<tr><td>4</td><td>100000 +</td><td>12</td><td>12</td></tr>
</table></td></tr>
</table>

Finally, we show the different estimation methods used with the samples obtained via these three designs. These estimation methods were based on two working models. The first, denoted model “S”, defined **X** purely in terms of the four benchmark variables defined above (plus an intercept). The second, denoted model “L” replaced the intercept by indicators for the seven industry groups.

RATIO	π^{-1}-weighted ratio estimator with estimation benchmarks as follows Y = Wheat income X = Wheat area Y = Beef income X = Beef number Y = Sheep income X = Sheep number Y = Dairy income X = Dairy number Y = Total income X = DSE
S/GREG	GREG case-weights based on model "S"
S/BLUP	BLUP case-weights based on model "S"
S/RIDGE	Ridged BLUP case-weights based on model "S" with C_k = 1000 for each of the four production benchmarks in the model.
S/D3	Nonparametrically calibrated and ridged BLUP weights based on model "S" with local Nadaraya-Watson smoothing against Z = DSE. Same C-values as S/RIDGE.
S/DAR3	Nonparametrically calibrated and ridged BLUP weights based on model "S" with local Nadaraya-Watson smoothing against Z_1 = DSE, Z_2 = ASIC and Z_3 = Region. Same C-values as S/RIDGE.
L/GREG	GREG case-weights based on model "L"
L/BLUP	BLUP case-weights based on model "L"
L/RIDGE	Ridged BLUP case-weights based on model "L" with C_k = 1000 for each of the four production benchmarks in the model, and C_k = 100000 for each of the seven industry benchmarks in the model.
L/D3	Nonparametrically corrected and ridged BLUP weights based on model "L" with local Nadaraya-Watson smoothing against Z = DSE. Same C-values as L/RIDGE.
L/DAR3	Nonparametrically corrected and ridged BLUP weights based on model "L" with local Nadaraya-Watson smoothing against Z_1 = DSE, Z_2 = ASIC and Z_3 = Region. Same C-values as L/RIDGE.

As expected, a number of samples generated negative weights for calibrated weighting methods. The following table shows percentages of samples that generate negative weights under various weighting systems/sample design combinations. Numbers in parentheses are the average number of sample units with a negative weight in samples containing at least one negative weight.

	Simple Random Sampling	**Size Stratification/ “Compromise” Allocation**	**Size Stratification/ “Optimal” Allocation**
S/GREG	44 (4.58)	11 (1.38)	82 (7.59)
L/GREG	77 (4.27)	53 (1.83)	94 (9.73)
S/BLUP	44 (4.58)	6 (1.33)	48 (1.81)
L/BLUP	77 (4.27)	20 (1.83)	93 (5.87)

The efficiencies of the different estimators that were observed in the simulation study are shown in the tables below. These are Root Mean Squared Errors, expressed as a percentage of the population total. The “best” result for each variable is shown in red and the “worst” result in blue.

(a) Simple Random Sampling

	Wheat income	**Beef income**	**Sheep income**	**Dairy income**	**Total income**
RATIO	14.7	**28.9**	**19.1**	**14.4**	16.7
S/GREG	14.0	27.4	17.2	15.6	**17.8**
L/GREG	**13.6**	26.1	17.0	15.0	17.3
S/BLUP	14.0	27.4	17.2	15.6	**17.8**
L/BLUP	**13.6**	26.1	17.0	15.0	17.3
S/RIDGE	**15.8**	24.2	16.3	**20.4**	15.8
L/RIDGE	15.7	23.6	16.0	17.1	15.7
S/D3	15.1	22.2	16.1	18.1	**14.5**
L/D3	15.0	**22.1**	15.9	17.5	14.6
S/DAR3	14.4	22.6	15.9	17.3	14.7
L/DAR3	14.5	22.4	**15.6**	17.0	14.7

(b) Size Stratification with “Compromise” Allocation

	Wheat income	**Beef income**	**Sheep income**	**Dairy income**	**Total income**
RATIO	10.0	11.6	15.5	**19.2**	8.3
S/GREG	10.0	**11.4**	14.7	19.3	**7.9**
L/GREG	**9.9**	11.9	14.8	20.3	8.4
S/BLUP	10.8	**14.5**	14.8	**25.2**	**10.2**
L/BLUP	10.8	12.8	14.3	20.5	8.9
S/RIDGE	11.4	**14.5**	14.8	**25.2**	**10.2**
L/RIDGE	**13.2**	13.1	**15.6**	23.1	9.8
S/D3	10.1	11.8	13.9	19.6	8.1
L/D3	10.5	11.5	14.1	19.8	8.1
S/DAR3	**9.9**	12.1	**13.8**	19.9	8.2
L/DAR3	10.5	11.6	14.1	19.7	8.1

(c) Size Stratification with “Optimal” Allocation

	Wheat income	Beef income	Sheep income	Dairy income	Total income
RATIO	10.1	10.1	15.9	**25.7**	7.9
S/GREG	10.2	10.3	15.6	26.8	7.4
L/GREG	11.6	**11.6**	17.4	32.2	8.4
S/BLUP	**9.1**	11.1	14.8	34.2	8.3
L/BLUP	11.9	11.1	16.4	32.1	8.0
S/RIDGE	12.6	10.7	15.7	37.2	8.7
L/RIDGE	**23.5**	9.6	**21.3**	**47.8**	**11.9**
S/D3	11.5	9.8	**14.3**	29.2	7.4
L/D3	12.5	9.1	15.6	30.7	7.3
S/DAR3	11.5	9.6	14.4	29.7	**7.2**
L/DAR3	12.9	**8.9**	15.7	31.5	7.3

From these results it would appear that combining a ridge weighting strategy and nonparametric bias calibration is a good approach to sample weighting for this population and these variables.

5.6 Interaction Between Sample Weighting and Sample Design

Suppose one has the choice about which sample to select, but calibration is a requirement no matter what sample is selected. Should this influence the way we select the sample? In particular, selection of the sample so that the calibration constraints are automatically satisfied for a fixed set of sample weights is an alternative way of ensuring that the sample weighted estimator remains unbiased under the linear model. Consequently, we can achieve “calibration” by choosing an appropriate sample rather than by modifying sample weights. This idea is an extension of balanced sampling (Royall and Herson, 1973a), where for fixed weights, we define a w-balanced sample as one where $\mathbf{x}_s'\mathbf{w}_s = \mathbf{x}'\mathbf{1}$. Recollect that this is the condition for unbiased prediction using the sample weights.

At the design stage of a survey we therefore have two options:

(i) Select a w-balanced sample, then use $\mathbf{w}_s$ “as is”.
(ii) Select the sample according to other criteria, then use a calibrated version of $\mathbf{w}_s$.

It seems sensible to choose the option that leads to a smaller variance. Since both options lead to an unbiased estimator, this is equivalent to choosing the option that gives smaller mean squared error.

We therefore look at the distribution of values of the Calibration Efficiency Ratio (CER) under different samples:

$$CER = \frac{\operatorname{var}\left(\hat{t}_y - t_y \middle| \text{unbalanced sample, calibrated weights}\right)}{\operatorname{var}\left(\hat{t}_y - t_y \middle| \text{balanced sample, uncalibrated weights}\right)}.$$

If *CER* is generally greater than one we choose option (i), otherwise we choose option (ii).
Given population data and an appropriate working model, together with sample values of auxiliary variables, we can estimate *CER*. We show this in an empirical investigation of two scenarios.

The first scenario combines simple random sampling without replacement with ratio estimation (SRSWOR/RATIO) based on a scalar benchmark variable *Z*. We note that ratio weights, by definition, are calibrated on *Z*, but not on the population count *N*. If these weights must also be calibrated on *N*, then they take the form:

$$\mathbf{w}_{calR} = \frac{N\bar{z}}{n\bar{z}_s}\mathbf{1}_s + \mathbf{V}_{ss}^{-1}\mathbf{X}_s(\mathbf{X}_s'\mathbf{V}_{ss}^{-1}\mathbf{X}_s)^{-1}\begin{pmatrix} N\left(1-\frac{\bar{z}}{\bar{z}_s}\right) \\ 0 \end{pmatrix}$$

where $\mathbf{X}_s = [\mathbf{1}_s : \mathbf{z}_s]$ and $\mathbf{V}_{ss} = \text{diag}(\mathbf{z}_s)$. The question here is - should we select a sample via SRSWOR and then use the ratio calibrated weights, or should we spend some time selecting a balanced sample and then use original ratio weights (in a balanced sample these weights are just *N*/*n*)?

The second scenario we consider is one where the sample units are selected with probability proportional to Z without replacement and the Horvitz-Thompson estimator is used (PPZWOR/HT). In this situation the HTE is the so-called mean of ratios estimator of total, which is the BLUP under the ratio model R, but with the residual variance proportional to the square of *Z*. We note that the inverse probability weights are calibrated on *Z*, but not on *N*. Again, if we require calibration on *N* as well, then these weights need to be replaced by

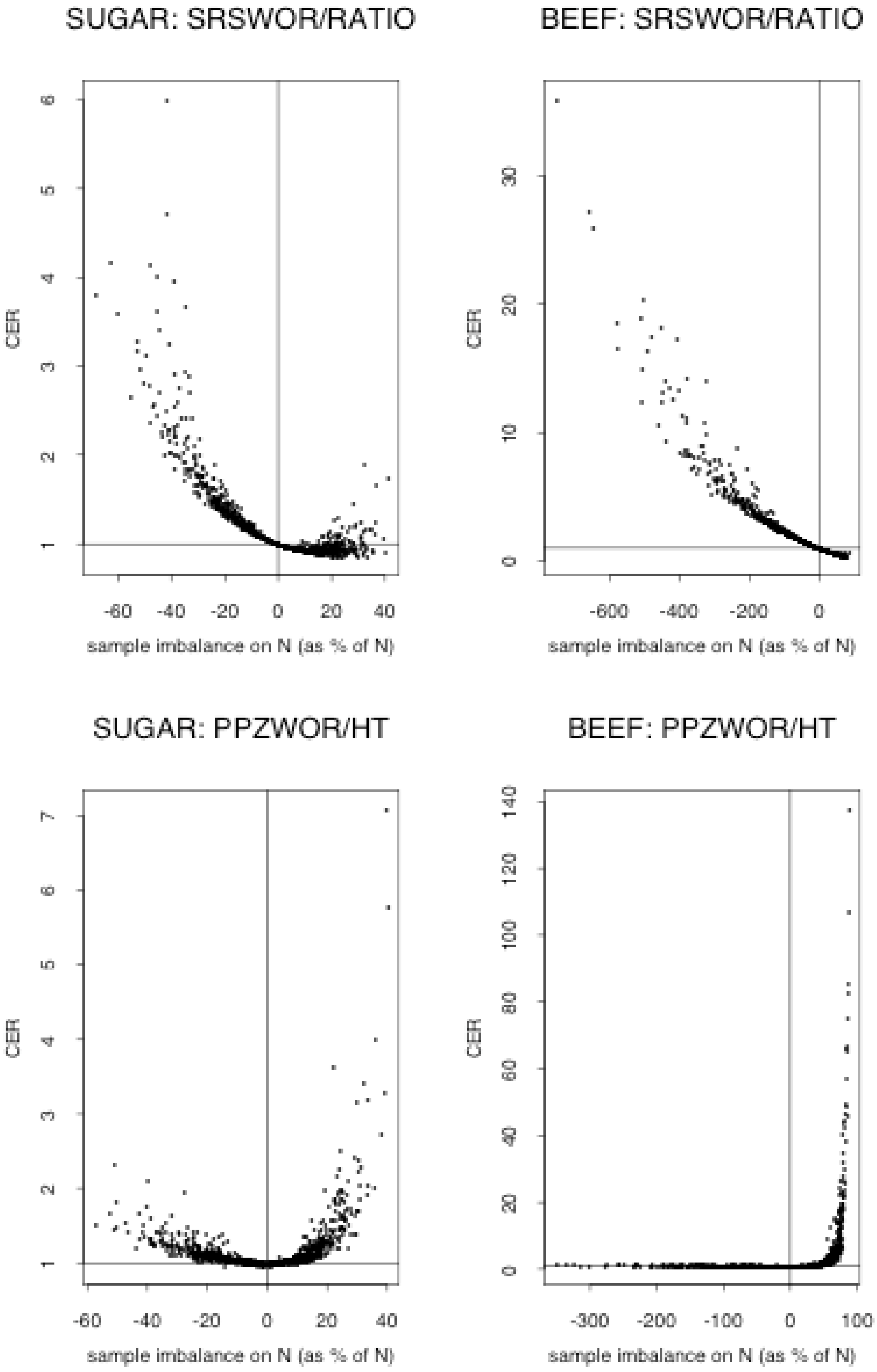
$$\mathbf{w}_{cal\pi} = \boldsymbol{\pi}_s^{-1} + \mathbf{v}_{ss}^{-1}\mathbf{x}_s(\mathbf{x}_s'\mathbf{v}_{ss}^{-1}\mathbf{x}_s)^{-1}\begin{pmatrix} N\left(1-\bar{z}\bar{z}_s^{(-1)}\right) \\ 0 \end{pmatrix}.$$

No calibration is required (sample is *ppz*-balanced) when the sample mean of 1/*Z* is equal to the population mean of *Z*, i.e. $\bar{z}_s^{(-1)} = \bar{z}$. The question then becomes - should we select sample "at random" using PPZWOR and then use calibrated weights, or should we select a *ppz*-balanced sample and then use inverse probability weights?

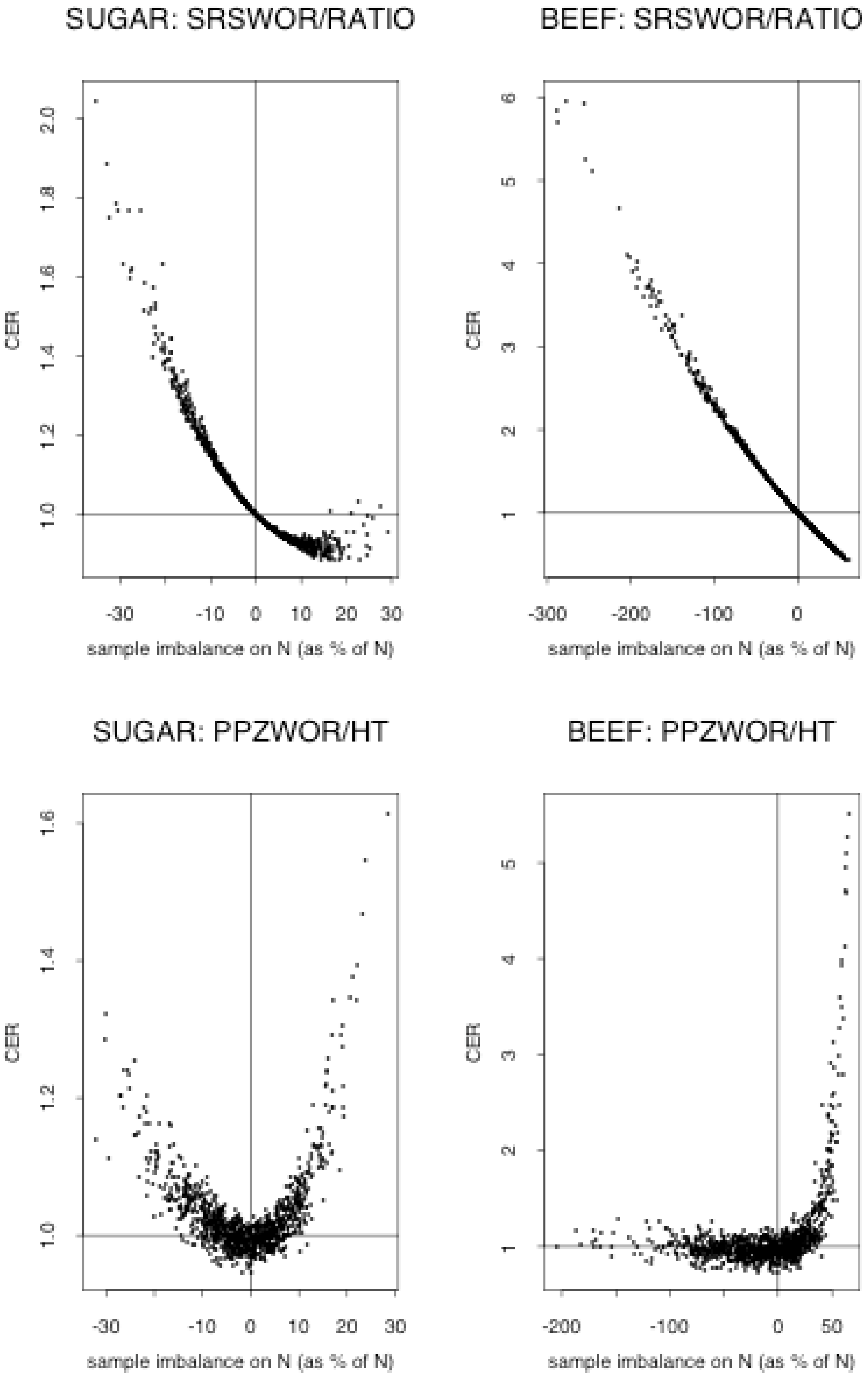
In order to evaluate the tradeoff between these two approaches to calibration (design vs. weighting) we carried out an empirical study with two "real" populations. The first (SUGAR) is the population of 338 sugar growing farms that was described in Chapter 2. Here *Z* = area assigned for cane growing. This population "fits" the ratio model R quite well. The second (BEEF) involves 430 Australian farms involved in beef cattle production, with *Z* = number of beef cattle at the end of the financial year. This population is extremely skewed, with residual variance increasing with at least Z^2, if not a higher power.

In the following three graphs we show *CER* values plotted against sample imbalance for three different sample sizes (*n* = 10, 30 and 100) when the two design/estimation strategies described above are applied to SUGAR and BEEF populations. For the SRSWOR/RATIO strategy, sample imbalance is defined as $\left(1-\frac{\bar{z}}{\bar{z}_s}\right)\times 100$, while for the PPZWOR/HT strategy, sample imbalance is defined as $\left(1-\bar{z}\bar{z}_s^{(-1)}\right)\times 100$.

n = 10


SUGAR: SRSWOR/RATIO
CER
1
2
3
4
5
6
-60
-40
-20
0
20
40
sample imbalance on N (as % of N)
BEEF: SRSWOR/RATIO
CER
0
10
20
30
-600
-400
-200
0
sample imbalance on N (as % of N)
SUGAR: PPZWOR/HT
CER
1
2
3
4
5
6
7
-60
-40
-20
0
20
40
sample imbalance on N (as % of N)
BEEF: PPZWOR/HT
CER
0
20
40
60
80
100
120
140
-300
-200
-100
0
100
sample imbalance on N (as % of N)

n = 30



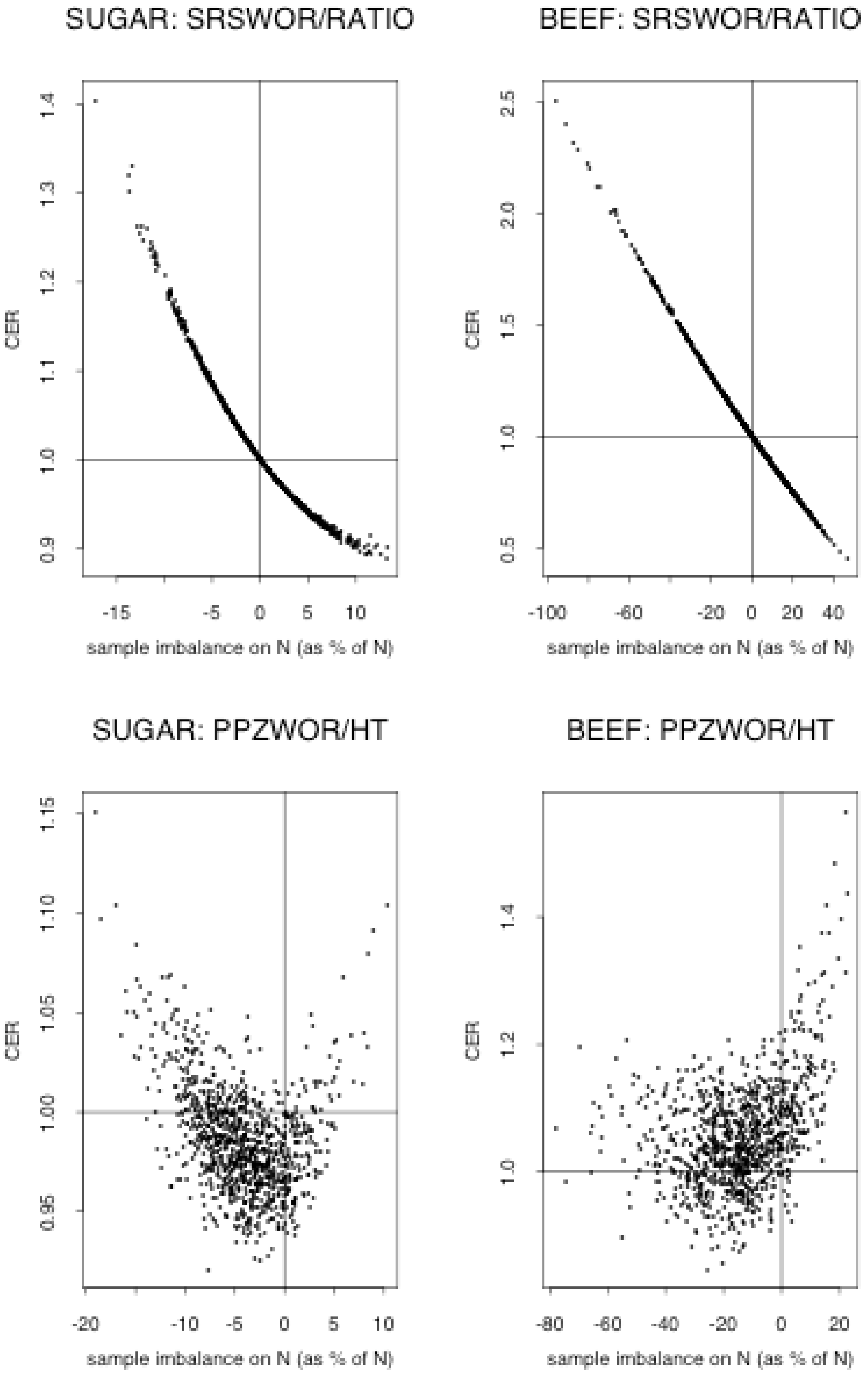
SUGAR: SRSWOR/RATIO

BEEF: SRSWOR/RATIO

SUGAR: PPZWOR/HT

BEEF: PPZWOR/HT

$n = 100$


SUGAR: SRSWOR/RATIO
CER
0.9
1.0
1.1
1.2
1.3
1.4
-15
-5
0
5
10
sample imbalance on N (as % of N)
BEEF: SRSWOR/RATIO
CER
0.5
1.0
1.5
2.0
2.5
-100
-60
-20
0
20
40
sample imbalance on N (as % of N)
SUGAR: PPZWOR/HT
CER
0.95
1.00
1.05
1.10
1.15
-20
-10
-5
0
5
10
sample imbalance on N (as % of N)
BEEF: PPZWOR/HT
CER
1.0
1.2
1.4
-80
-60
-40
-20
0
20
sample imbalance on N (as % of N)

Inspection of these plots leads to the following conclusions:

1. Gains in “robustness” from sample balancing typically outweigh possible efficiency loss.

2. Where calibration leads to more efficient inference than balancing, there appears to be no “preferred” direction of imbalance.

3. Obtaining “exact” w-balance matters little for the ratio estimator, since *CER* values vary little for samples that are “close” to balance.

4. However, the HTE is more sensitive to lack of w-balance, with *CER* values varying considerably for samples that are “close” to balance.

That is, it is typically much better if calibration is achieved by suitable choice of sample than by weight modification.

6. Estimation for Domains and Small Areas

A domain is a subgroup of the sample population for which a separate estimate of the total of *Y* (or mean etc.) is required. For example, in many business surveys the sample frame is out of date, so the industry and size classifications of many units on the frame do not agree with their "current" industry and size classifications. After the survey is carried out, estimates are required for the current industry by size classes. These classes then correspond to domains of interest.

A basic assumption is that domain membership is observable on the sample. Consequently, we can define a domain membership variable *D* with value d_i for population unit *i*, such that $d_i = 1$ if unit *i* is in the domain and is zero otherwise. The number of population units in the domain is then the population sum of *D* and is denoted by N_d. The population total of *Y* for the domain is

$$t_{dy} = \sum\nolimits_U d_i y_i \, .$$

The domain total of interest is therefore just the population total of the derived variable *DY*.

6.1 Model-Based Inference when the Domain Size N_d is Unknown

Consider a working model ξ for the distribution of *Y* values in the domain that is a simple extension of the homogeneous population model H. In particular, we assume that domain membership (*D*) can be modelled as *N* independent and identically distributed realisations of a Bernoulli(θ_d) random variable, and, conditional on *D*, the population values of *Y* are uncorrelated with constant mean and variance, so that we can write

$$\begin{aligned}
&E_\xi(y_i \mid d_i = 1) = \mu_d \\
&Var_\xi(y_i \mid d_i = 1) = \sigma_d^2 \\
&Cov_\xi(y_i, y_j \mid d_i, d_j) = 0 \\
&E_\xi(d_i) = \theta_d \\
&Var_\xi(d_i) = \theta_d(1 - \theta_d) \\
&Cov_\xi(d_i, d_j) = 0.
\end{aligned}$$

As always, we have the implicit assumption that sample inclusion is independent of the values of the variables of interest. Consequently, sample inclusion and domain membership must be independent of one another (so that we can estimate θ_d from the sample data). This will be true if the sample is chosen via simple random sampling.

Under this working model it is straightforward to show that

$$\begin{aligned}
&E_\xi(d_i y_i) = \mu_d \theta_d \\
&Var_\xi(d_i y_i) = \sigma_d^2 \theta_d + \mu_d^2 \theta_d(1 - \theta_d) \\
&Cov_\xi(d_i y_j, d_j y_j) = 0
\end{aligned}$$

which is just a special case of the homogeneous population model H, and so the BLUP for t_{dy} is expansion estimator

$$\hat{t}_{Hdy} = \frac{N}{n}\sum_s d_i y_i = \frac{Nn_d}{n}\bar{y}_{sd}$$

where $\bar{y}_{sd}$ is the mean of the sample *Y* values from domain *d*. Similarly, the prediction variance of this BLUP is

$$Var_\xi\left(\hat{t}_{Hdy} - t_{dy}\right) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\left[\theta_d\sigma_d^2 + \theta_d(1-\theta_d)\mu_d^2\right]$$

with the usual "plug-in" estimate of this variance.

6.2 Model-Based Inference when the Domain Size N_d is Known

This is a more unusual situation. Since we know how many units in the population are in the domain we need to condition on this knowledge. Thus all moments are evaluated conditional on knowing N_d. We denote this conditional version of the working model by ξd. Consequently, $E_{\xi d}(N_d) = N_d$ and $Var_{\xi d}(N_d) = 0$, and symmetry-based arguments can then be used to show

$$E_{\xi d}(d_j) = p_d$$
$$Var_{\xi d}(d_j) = p_d(1 - p_d)$$
$$Cov_{\xi d}(d_j, d_k) = -p_d(1 - p_d)/(N - 1)$$

where $p_d = N_d/N$. If we then further assume that *Y* is independent of N_d conditional on *D* (i.e. knowing N_d tells us nothing extra about y_i than knowing the value of d_i), then

$$E_{\xi d}(d_j y_j) = \mu_d p_d$$
$$Var_{\xi d}(d_j y_j) = \sigma_d^2 \mathrm{p_d} + \mu_d^2 p_d(1 - p_d)$$
$$Cov_{\xi d}(d_j y_j, d_k y_k) = -\mu_d^2 p_d(1 - p_d)/(N - 1)$$
$$Cov_{\xi d}(d_j y_j, d_j) = -\mu_d p_d(1 - p_d)$$
$$Cov_{\xi d}(d_j y_j, d_k) = -\mu_d p_d(1 - p_d)/(N - 1).$$

We see that with respect to this conditional distribution, the "derived" random variable *DY* has a mean and variance that is the same for all population units, and that the covariance between any two population values of *DY* is constant. That is *DY* follows the homogeneous population model (H) as well. Consequently the BLUP of the population total t_{dy} is therefore still the simple expansion estimator $\hat{t}_{Hdy}$, but now

$$Var_{\xi d}\left(\hat{t}_{Hdy} - t_{dy}\right) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\left[Var_{\xi d}(d_i y_i) - Cov_{\xi d}(d_i y_i, d_j y_j)\right]$$
$$= \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\left[\sigma_d^2 p_d + \frac{N}{N-1}\mu_d^2 p_d(1 - p_d)\right].$$

However, in this situation there seems no strong reason why one should restrict attention to estimates that are linear in DY. An obvious alternative is the nonlinear ratio-type estimator that is a "plug-in" version of the MMSEP for this case:

$$\hat{t}_{Rdy} = N_d \bar{y}_{sd} = N_d \frac{\sum_s d_i y_i}{\sum_s d_i}.$$

Observe that $\hat{t}_{Rdy}$ is approximately model-unbiased in large samples, and a first order approximation to its prediction variance is

$$Var_{\xi d}\left(\hat{t}_{Rdy} - t_{dy}\right) \approx \frac{N^2}{n^2} Var_{\xi d}\left[\sum_s d_i y_i - \mu_d \sum_s d_i - \frac{n}{N}\sum_U d_i y_i\right]$$

$$= \frac{N^2}{n}\left(1 - \frac{n}{N}\right)\sigma_d^2 p_d .$$

Comparing this variance with the variance of $\hat{t}_{Hdy}$ above we see that there will typically be large efficiency gains from use of this ratio-type estimate.

Why just condition on N_d? The Conditionality Principle (Cox and Hinkley, 1974) states that one should always condition on ancillary variables in inference. An ancillary variable is one whose distribution depends on parameters that are distinct from those associated with the distribution of the variable of interest. For domain analysis, the parameter(s) associated with the distribution of the domain inclusion variable *D* are distinct from those associated with the distribution of the survey variable Y. Consequently, one should condition on *D* in inference. This is equivalent to conditioning on both the population count N_d of the number of units in the domain, and the corresponding sample count n_d.

If one conditions on both N_d and n_d, then $\hat{t}_{Rdy}$ is in fact the BLUP for t_{dy} with

$$Var_{\xi}\left(\hat{t}_{Rdy} - t_{dy} \mid n_d, N_d\right) = \frac{N_d^2}{n_d}\left(1 - \frac{n_d}{N_d}\right)\sigma_d^2 .$$

This is usually referred to as the variance of the post-stratified estimator for the domain total. It can be seen that this post-stratified variance is zero if $N_d = n_d$, when we know that $\hat{t}_{Rdy}$ has zero error.

Which is the right variance to use with t_{dy}? There is an argument that states that since the distribution of t_{dy} depends on the parameters of *Y* as well as the parameters of *D*, this is a case where the conditionality principle does not apply. A cautious approach may therefore be to use the maximum of the two variance estimates above.

6.3 Model-Based Inference Using Auxiliary Information

Let *X* denote the auxiliary variable. We assume an unknown domain size N_d and a working model ξ for *Y* that satisfies

$E_\xi(y_i \mid d_i = 1) = \mu(x_i;\omega_d)$
$Var_\xi(y_i \mid d_i = 1) = \sigma^2(x_i;\omega_d)$
$Cov_\xi(y_i, y_j \mid d_i, d_j) = 0$ for $i \neq j$.

As above we assume domain membership can be modelled as the outcome of independent and identically distributed Bernoulli trials, independently of the value of *Y*. However, domain membership <u>can</u> depend on *X*, so

$E_\xi(\mathrm{d_i}) = \theta(x_\mathrm{i};\gamma_d)$
$Var_\xi(d_i) = \theta(x_i;\gamma_d)[1 - \theta(x_i;\gamma_d)]$
$Cov_\xi(d_i,d_j) = 0.$

With this set-up we then have

$E_\xi(d_iy_i) = \mu(x_i;\omega_d)\,\theta(x_i;\gamma_d)$
$Var_\xi(d_iy_i) = \sigma^2(x_i;\omega_d)\,\theta(x_i;\gamma_d) + \mu^2(x_i;\omega_d)\,\theta(x_i;\gamma_d)[1 - \theta(x_i;\gamma_d)]$
$Cov_\xi(d_iy_i,d_jy_j) = 0.$

Because sampling is uninformative for both *Y* and *D*, given *X*, we can estimate ω_d and γ_d from the sample data. The "plug-in" model-based estimator of t_{dy} is then

$$\hat{t}_{\xi dy} = \sum\nolimits_s d_i y_i + \sum\nolimits_r \mu(x_i;\hat{\omega}_d)\theta(x_i;\hat{\gamma}_d).$$

This will be a consistent estimator of t_{dy} under our working model. The prediction variance of $\hat{t}_{\xi dy}$ is then

$$Var_\xi\left(\hat{t}_{\xi dy} - t_{dy}\right) = Var_\xi\left(\sum\nolimits_r \mu(x_i;\hat{\omega}_d)\theta(x_i;\hat{\gamma}_d)\right) + \sum\nolimits_r Var_\xi\left(d_i y_i\right) = V_1 + V_2.$$

The leading term in this variance is V_1. The second term V_2 has a simple plug-in estimate:

$$\hat{V}_2 = \sum\nolimits_r \left(\sigma^2(x_i;\hat{\omega}_d)\theta(x_i;\hat{\gamma}_d) + \mu^2(x_i;\hat{\omega}_d)\theta(x_i;\hat{\gamma}_d)[1 - \theta(x_i;\hat{\gamma}_d)]\right).$$

V_1 can be estimated using computer intensive methods like the jackknife or the bootstrap. An alternative is to develop a linearised version of this term

$$V_1 \approx Var_\xi\left(\hat{\gamma}_d \sum\nolimits_r \mu(x_i;\hat{\omega}_d)\frac{\partial\theta(x_i;\gamma_d)}{\partial\gamma_d} + \hat{\omega}_d \sum\nolimits_r \theta(x_i;\gamma_d)\frac{\mu(x_i;\omega_d)}{\partial\omega_d}\right).$$

Estimates of the variances of $\hat{\omega}_d$ and $\hat{\gamma}_d$ and their covariance can calculated from the sample data. The Taylor series estimator of V_1 is then

$$\begin{aligned}\hat{V}_1 = {} & \hat{V}_\xi(\hat{\gamma}_d)\left(\sum\nolimits_r \mu(x_i;\hat{\omega}_d)\frac{\partial\theta(x_i;\hat{\gamma}_d)}{\partial\hat{\gamma}_d}\right)^2 + \hat{V}_\xi(\hat{\omega}_d)\left(\sum\nolimits_r \theta(x_i;\hat{\gamma}_d)\frac{\mu(x_i;\hat{\omega}_d)}{\partial\hat{\omega}_d}\right)^2 \\ & + 2\hat{C}_\xi(\hat{\gamma}_d,\hat{\omega}_d)\left(\sum\nolimits_r \mu(x_i;\hat{\omega}_d)\frac{\partial\theta(x_i;\hat{\gamma}_d)}{\partial\hat{\gamma}_d}\right)\left(\sum\nolimits_r \theta(x_i;\hat{\gamma}_d)\frac{\mu(x_i;\hat{\omega}_d)}{\partial\hat{\omega}_d}\right).\end{aligned}$$

To illustrate, consider the situation where the population is stratified and the regression of *Y* on *X* is linear and through the origin for units in the domain, but the slope of this regression line varies from stratum to stratum. In addition, the proportion of the population in the domain varies significantly from stratum to stratum. Put θ_h = probability that a population unit in stratum *h* lies in the domain and β_h = the slope of regression line for domain units in stratum h. The estimator of the domain total of Y for this working model and population is then:

$$\hat{t}_{\xi dy} = \sum_s d_i y_i + \sum_h p_{shd}\left(N_h \bar{x}_h - n_h \bar{x}_{sh}\right)\hat{\beta}_h$$

where p_{shd} is sample proportion of stratum h units in the domain; $\hat{\beta}_h$ is the stratum h estimate for the slope of the regression of Y on X in the domain; $\bar{x}_h$ is the stratum h average for X; and $\bar{x}_{sh}$is the sample average for X in stratum h. The Taylor series estimate of the leading term in the prediction variance of $\hat{t}_{\xi dy}$ is

$$\hat{V}_1 = \sum_h \left(N_h \bar{x}_h - n_h \bar{x}_{sh}\right)^2 \left(\hat{V}_\xi\left(p_{shd}\right)\hat{\beta}_h^2 + \hat{V}_\xi\left(\hat{\beta}_h\right)p_{shd}^2\right)$$

where $\hat{V}_\xi\left(p_{shd}\right)$ is the estimated variance of p_{hd} and $\hat{V}_\xi\left(\hat{\beta}_h\right)$ is the estimated variance of $\hat{\beta}_h$. Note that independence of D and Y within a stratum causes the covariance term in this estimate to disappear, while the independent Bernoulli realisations assumption gives $\hat{V}_\xi\left(p_{shd}\right) = n_h^{-1} p_{shd}\left(1 - p_{shd}\right)$ and, if $Var_\xi(y_i \mid x_i, d_i = 1) \propto \sigma_{hd}^2 \mathrm{x}_\mathrm{i}$, $\hat{V}_\xi\left(\hat{\beta}_h\right) = \left(n_{hd}\bar{x}_{shd}\right)^{-1}\hat{\sigma}_h^2$. The plug-in estimator of V_2 is defined similarly.

6.4 Small Area Estimation

In many cases large national samples are also used to produce estimates for small sub-national groupings of the population. These groupings are then domains of interest. Typically the groups are defined geographically, in which case they are referred to as small areas.

A basic problem with domains defined in this way is that the estimation methods described earlier become impossible to apply, mainly because sample sizes are typically small or even zero in the small areas of interest, so the direct estimates (i.e. area-specific estimates) investigated above tend to be quite unstable.

Availability of suitable auxiliary information is quite important for resolving this problem. This can be values of the variable of interest in other, similar, areas; values of this variable for the same area in the past; or values of other variables that are related to the variable of interest. In all cases we assume that the relationship between the survey variable and the auxiliary variable is the same across all small areas, and so we "borrow strength" from all small areas by using their data to fit this common model, which we then use in estimation in any single small area.

6.4.1 Fixed Effects Models

These models essentially assume that domain to domain variability in Y can be explained entirely in terms of variability in the auxiliary information. A linear specification is commonly used to model the relationship between Y and X, typically via the general linear model

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

where we use the same notation as in the previous chapter. Note that, as is almost always the case, uninformative sampling is assumed, so this model also holds at sample level. Given an estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, we can then estimate the average of Y in small area d using the estimator

$$\hat{\bar{y}}_d = N_d^{-1}\left(\sum_{i=1}^{n_d} y_{di} + \sum_{i=n_d+1}^{N_d} \mathbf{x}'_{di}\hat{\boldsymbol{\beta}}\right)$$

provided we know (or can estimate) the population count N_d in the small area. This is sometimes referred to as the “synthetic” estimator of this mean.

Given a set of survey weights (typically used for national level estimation), an alternative estimator is the weighted direct estimator:

$$\hat{\bar{y}}_d = \left(\sum_{i=1}^{n_d} w_{di}\right)^{-1}\left(\sum_{i=1}^{n_d} w_{di} y_{di}\right) = \bar{y}_{wd}.$$

It is easy to see that using this estimator corresponds to assuming the simple model $y_{di} = \beta_d + \varepsilon_{di}$.

In many cases, good quality direct estimates for a classification of the population (indexed by g below) are available, and these “cut across” the small area of interest. For example, we may have estimates by age and sex for a population but be interested in an overall mean estimate for a small area. In such a situation we can estimate this mean by assuming that the contributions to it from the auxiliary classification are the same as in the overall population. More generally we can define these contributions in terms of the within area proportions of an auxiliary variable X. This leads to the apportionment estimator

$$\hat{\bar{y}}_d = \sum_{g=1}^{G}\left(\frac{t_{dgx}}{t_{dx}}\right)\hat{\bar{y}}_g .$$

This estimator is the original “synthetic” estimator in the small area estimation literature. It is generally biased, but has a small variance. Under the linear model $E_\xi(y_i \mid i \in g) = \alpha_g + \beta_g X$

$$\begin{aligned}
E_\xi(\hat{\bar{y}}_d - \bar{y}_d) &= \sum_{g=1}^{G}\left(\frac{t_{dgx}}{t_{dx}} E_\xi(\hat{\bar{y}}_g) - \frac{N_{dg}}{N_d} E_\xi(\bar{y}_{dg})\right) \\
&= \sum_{g=1}^{G}\left(\frac{t_{dgx}}{t_{dx}}(\alpha_g + \beta_g \hat{\bar{x}}_g) - \frac{N_{dg}}{N_d}(\alpha_g + \beta_g \bar{x}_{dg})\right) \\
&= \sum_{g=1}^{G}\alpha_g\left(\frac{t_{dgx}}{t_{dx}} - \frac{N_{dg}}{N_d}\right) + \beta_g\left(\frac{t_{dgx}}{t_{dx}}\hat{\bar{x}}_g - \frac{N_{dg}}{N_d}\bar{x}_{dg}\right).
\end{aligned}$$

This is unbiased under a classification-based ANOVA model for the population (i.e. $X = 1$).

6.4.2 Random Effects Models

This is the most commonly used class of models in small area estimation. The assumption here is that unexplained area specific variability remains even after accounting for the auxiliary information. Since there are typically a large number of "exchangeable" small areas, the idea is to introduce a random "effect" for each small area that accounts for this unexplained variability. Without loss of generality we assume that the individual values making up all population and sample vectors are ordered by the D small areas that make up the population. Under a linear specification for Y this leads to a mixed model incorporating domain specific random effects:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

where $\mathbf{Z}$ is a matrix of known covariates (typically defined at area level), $\mathbf{u}$ is a vector made up D independent realisations of a vector-valued area specific effect of dimension q with zero mean vector and covariance matrix $\sigma^2\Lambda$, that is uncorrelated between different areas, and $\boldsymbol{\varepsilon}$ is an individual level "noise" vector, with zero mean and diagonal covariance matrix $\sigma^2\mathbf{D}$, where $\mathbf{D}$ is a known matrix, that is uncorrelated with $\mathbf{u}$. The matrix Λ is often referred to as the matrix of variance components of the model. Typically this matrix is itself parameterised in terms of a lower dimensional parameter, so we write it in the form $\Lambda(\boldsymbol{\varphi})$.

The BLUP of any linear combination of the population Y-values can be obtained using the result of Royall (1976) described at the end of Chapter 2. Let $\mathbf{a}$ be an arbitrary N-vector of known constants. Using the same notation as in that Chapter, the BLUP of $\theta = \mathbf{a}'\mathbf{y}$ is

$$\hat{\theta} = \mathbf{a}'_s\mathbf{y}_s + \mathbf{a}'_r[\mathbf{X}_r\hat{\boldsymbol{\beta}} + \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}(\mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}})]$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'_s\mathbf{V}_{ss}^{-1}\mathbf{X}_s)^{-1}(\mathbf{X}'_s\mathbf{V}_{ss}^{-1}\mathbf{y}_s)$ is the best linear unbiased estimator of $\boldsymbol{\beta}$. Note that in this case $Var_\xi(\mathbf{y}_s) = \sigma^2(\mathbf{D}_s + \mathbf{Z}_s\Lambda\mathbf{Z}'_s) = \mathbf{V}_{ss}$ and $Cov_\xi(\mathbf{y}_r, \mathbf{y}_s) = \sigma^2\mathbf{Z}_r\Lambda\mathbf{Z}'_s = \mathbf{V}_{rs}$ and hence

$$\hat{\theta} = \mathbf{a}'_s\mathbf{y}_s + \mathbf{a}'_r[\mathbf{X}_r\hat{\boldsymbol{\beta}} + \mathbf{Z}_r\hat{\mathbf{u}}],$$

where

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'_s(\mathbf{D}_s + \mathbf{Z}_s\Lambda\mathbf{Z}'_s)^{-1}\mathbf{X}_s)^{-1}(\mathbf{X}'_s(\mathbf{D}_s + \mathbf{Z}_s\Lambda\mathbf{Z}'_s)^{-1}\mathbf{y}_s)$$
$$\hat{\mathbf{u}} = \Lambda\mathbf{Z}'_s(\mathbf{D}_s + \mathbf{Z}_s\Lambda\mathbf{Z}'_s)^{-1}(\mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}})].$$

To illustrate, consider the Random Means (RM) model. In this case $y_{di} = \beta + u_d + e_{di}$, so that

$$\hat{\beta} = \left(\sum_d (\lambda + n_d^{-1})^{-1}\right)^{-1} \sum_d (\lambda + n_d^{-1})^{-1}\bar{y}_{sd}$$
$$\mathbf{V}_{rs} = \sigma^2\lambda\, diag(\mathbf{1}_{rd}\mathbf{1}'_{sd})$$
$$\mathbf{V}_{ss} = \sigma^2[diag(\lambda\mathbf{1}_{sd}\mathbf{1}'_{sd}) + \mathbf{I}_{ss}]$$

which leads to $\hat{u}_d = \left(\frac{n_d\lambda}{1 + n_d\lambda}\right)(\bar{y}_{sd} - \hat{\beta})$ and hence

$$\hat{\bar{Y}}_d = N_d^{-1}\left(n\bar{y}_{sd} + (N_d - n_d)\hat{\beta} + (N_d - n_d)\left(\frac{n_d\lambda}{1 + n_d\lambda}\right)(\bar{y}_{sd} - \hat{\beta})\right).$$

We need to know the matrix Λ if we wish to evaluate the BLUP. A variety of estimation methods are typically used for the parameters of models with random effects. The more common include Empirical Best Linear Unbiased Prediction (EBLUP), Empirical Bayes (EB) and Hierarchical Bayes (HB). The EBLUP method usually involves maximum likelihood (ML) or residual (restricted) maximum likelihood (REML) estimation of the variance components, although sometimes the method of moments is used. In virtually all cases the random effects are assumed to be normally distributed.

6.4.3 Generalized Linear Mixed Models (GLMM) in Small Area Estimation

The data underpinning small area estimates are often categorical. Generalized Linear Models (GLMs) are standard models for such data. Their application in the small area estimation context leads to synthetic GLM-based small area estimates. Similarly, the Generalized Linear Mixed Model (GLMM) extension of GLM can be used to introduce random area effects. Such models are typically written $\mathrm{E}_{\xi}(\mathbf{y} \mid \mathbf{u}) = h(\eta)$ for a specified function *h*, with $\eta = \mathbf{X}\beta + \mathbf{Z}\mathbf{u}$. Here **u** is a vector of random area effects, as above, typically assumed to be normally distributed with zero mean vector and variance-covariance matrix $Var_{\xi}(\mathbf{u}) = \Omega(\varphi)$.

To illustrate, suppose we have a Bernoulli response and a random mean model. Then the small area counts $\mathbf{y}_{s1}, \mathbf{y}_{s2},\dots,\mathbf{y}_{sD}$ are independent binomial random variables. The estimate of the proportion of "successes" in area d is then:

$$\hat{\theta}_d = N_d^{-1}[n_d p_{sd} + (N_d - n_d)(\hat{\beta} + \hat{u}_d)]$$

where p_{sd} is the proportion of successes in the sample in small area *d*, and $\hat{\beta}$ and $\hat{u}_d$ are obtained by fitting a mixed logistic model to the sample data in the small areas of interest.

6.4.4 Estimation of Mean Square Error

This can be quite complex because of the need to include uncertainty associated with estimation of variance component parameters. To illustrate we focus on the linear mixed model and proceed in a number of steps.

1. Suppose β and variance components Λ are known. Put $\tau = \mathbf{a}'_r(\mathbf{X}_r\beta + \mathbf{Z}_r\mathbf{u})$. Then the MMSEP of τ is $\hat{\tau}_{MMSEP} = E_{\xi}(\tau \mid \mathbf{y}_s) = \mathbf{a}'_r(\mathbf{X}_r\beta + \mathbf{Z}_r\hat{\mathbf{u}}_{MMSEP})$, where $\hat{\mathbf{u}}_{MMSEP} = \Lambda\mathbf{Z}'_s\Sigma_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\beta)$ is an unbiased predictor of **u**. Here $\mathbf{V}_{ss} = \sigma^2\Sigma_s$. Then

$$\begin{aligned} MSE_{\xi}(\hat{\tau}_{MMSEP}) &= \mathbf{a}'_r\mathbf{Z}_r Var_{\xi}(\square\mathbf{Z}'_s\square_s^{-1}\mathbf{y}_s - \mathbf{u})\mathbf{Z}'_r\mathbf{a}_r \\ &= \sigma^2\mathbf{a}'_r\mathbf{Z}_r(\square - \square\mathbf{Z}'_s\square_s^{-1}\mathbf{Z}_s\square)\mathbf{Z}'_r\mathbf{a}_r \\ &= g_1(\sigma^2, \square). \end{aligned}$$

   Under the random means model this becomes $g_1 = (1 - f_d)^2\sigma^2 n_d^{-1}\gamma_d$, where $\gamma_d = \lambda(n_d^{-1} + \lambda)^{-1}$.

2. Suppose β is replaced by its weighted least squares estimator $\hat{\beta}$, but the variance components Λ are still assumed known. Then the BLUP of τ is $\hat{\tau}_{BLUP} = \mathbf{a}'_r(\mathbf{X}_r\hat{\beta} + \mathbf{Z}_r\hat{\mathbf{u}}_{BLUP})$, where $\hat{\mathbf{u}}_{BLUP} = \Lambda\mathbf{Z}'_s\Sigma_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\hat{\beta})$. Here we can write

$$\begin{aligned} \hat{\tau}_{BLUP} - \tau &= (\hat{\tau}_{BLUP} - \hat{\tau}_{MMSEP}) + (\hat{\tau}_{MMSEP} - \tau) \\ &= \left[\mathbf{a}'_r[\mathbf{X}_r + \mathbf{Z}_r\Lambda\mathbf{Z}'_s\Sigma_s^{-1}\mathbf{X}_s](\hat{\beta} - \beta)\right] + (\hat{\tau}_{MMSEP} - \tau) \end{aligned}$$

   so $\mathrm{MSE}(\hat{\tau}_{BLUP}) = g_2(\sigma^2, \varphi) + g_1(\sigma^2, \varphi)$. Under the random means model

$$g_2 = \sigma^2 n^{-1}(1-f_d)^2(1-\gamma_d)^2\left[1-n^{-1}\sum\nolimits_{a=1}^{D} n_a\gamma_a\right]^{-1}.$$

3. Finally, we have the EBLUP situation, with $\boldsymbol{\beta}$ and variance components estimated. Then

$$\begin{aligned}\mathrm{MSE}(\hat{\tau}_{EBLUP}) &= \mathrm{E}(\hat{\tau}_{EBLUP} - \hat{\tau}_{BLUP} + \hat{\tau}_{BLUP} - \tau)^2 \\ &= \mathrm{MSE}(\hat{\tau}_{BLUP}) + \mathrm{E}(\hat{\tau}_{EBLUP} - \hat{\tau}_{BLUP})^2 \\ &\quad + 2\mathrm{E}(\hat{\tau}_{BLUP} - \tau)(\hat{\tau}_{EBLUP} - \hat{\tau}_{BLUP}).\end{aligned}$$

The first term on the right hand side is given by the preceding MSE. A naive estimator of the MSE of the EBLUP is defined by disregarding last two terms on the right hand side and replacing the unknown variance components by suitable estimators. However, Kacker and Harville (1984) show that although the third component is neglible, the second is not. They derive a "plug in" approximation to this second term. Prasad and Rao (1990) show that the MSE estimator based on this approximation underestimates true MSE, and so they introduce a second order approximation: $\mathrm{MSE}(\hat{\tau}_{EBLUP}) \cong g_1(\sigma^2,\boldsymbol{\varphi}) + g_2(\sigma^2,\boldsymbol{\varphi}) + g_3(\sigma^2,\boldsymbol{\varphi})$. For the random means model, with variance components estimated via ML, we have

$$g_3 = 2n\sigma^2 n_d^{-2}(1-f_d)^2(n_d^{-1}+\lambda)^{-3}[n\sum\nolimits_{a=1}^{D}(n_a^{-1}+\lambda)^2 - (\sum\nolimits_{a=1}^{D} n_a^{-1}+\lambda)^2]^{-1}.$$

If the variance components are estimated via REML, then

$$\begin{aligned}&g_3 = 2n\sigma^2 n_d^{-2}(n_d^{-1}+\lambda)^{-3}[n\alpha_2 - \alpha_1]^{-1} \\ &\alpha_1 = \sum\nolimits_{d=1}^{D}(n_d^{-1}+\lambda)^{-1} + \sum\nolimits_{d=1}^{D}\Delta_d \\ &\alpha_2 = \sum\nolimits_{d=1}^{D}(n_d^{-1}+\lambda)^{-1} + \lambda\left(\varphi^{-1} - \sum\nolimits_{d=1}^{D}\Delta_d\right)^2 + \lambda^{-1} - 2\sum\nolimits_{d=1}^{D}\Delta_d(1+n_d\lambda)^{-1} \\ &\Delta_d = (n_d^{-1}+\lambda)^{-2} / \sum\nolimits_{a=1}^{D}(n_a^{-1}+\lambda)^{-1}.\end{aligned}$$

It is worth noting that the Prasad and Rao MSE formula above is aimed at estimation of

$$Var_\xi(\hat{\tau}_{EBLUP} - \tau) = Var_\xi\left(\mathbf{a}_r'(\mathbf{X}_r\hat{\boldsymbol{\beta}} + \mathbf{Z}_r\hat{\mathbf{u}}) - \mathbf{a}_r'(\mathbf{X}_r\boldsymbol{\beta} + \mathbf{Z}_r\mathbf{u})\right).$$

However, what we are really interested in is

$$Var_\xi(\hat{\theta}_{EBLUP} - \theta) = Var_\xi\left(\mathbf{a}_r'(\mathbf{X}_r\hat{\boldsymbol{\beta}} + \mathbf{Z}_r\hat{\mathbf{u}}) - \mathbf{a}_r'(\mathbf{X}_r\boldsymbol{\beta} + \mathbf{Z}_r\mathbf{u} + \boldsymbol{\varepsilon}_r)\right).$$

This leads to a fourth term in the MSE formula

$$\mathrm{MSE}(\hat{\theta}_{EBLUP}) \cong g_1(\sigma^2,\boldsymbol{\varphi}) + g_2(\sigma^2,\boldsymbol{\varphi}) + g_3(\sigma^2,\boldsymbol{\varphi}) + g_4(\sigma^2).$$

6.4.5 An Example: Comparing GLM and GLMM-Based Predictors for LAD level Estimates of ILO Unemployment from the UK LFS

We now demonstrate the GLM and GLMM approaches to small area estimation by applying them to the problem of estimating the numbers of people who are "ILO Unemployed" within each of the

406 local authority districts (LADs) of Great Britain. The concept of "ILO unemployed" is the standard way unemployment is measured in the UK Labour Force Survey. However, there is another source of information about unemployment. This is the "claimant count", i.e. the number of people who register to claim unemployment benefits. The two measures of unemployment are not the same, although they are closely related. However, the ILO unemployment estimate is the "official" measure of unemployment in Great Britain. Here we treat claimant count as a covariate, since it is available at age by sex level within each LAD. Our direct estimates of ILO unemployment are obtained from the UK LFS, and correspond to weighted sums of the numbers of "ILO unemployed" survey respondents within each age by sex group within each LAD. These direct estimates define the level at which we specify our models (we do not aggregate age and sex groups because we believe there are differences in the relationship between ILO unemployment and claimant count between these groups.

We consider 5 combinations of models and estimation methods for these data.

Model A

This bases estimation on a fixed effects logistic model for the Bernoulli variable corresponding to being "ILO unemployed" at age-sex by LAD level. There are 29 terms in this model, which can be specified as: age-sex effect * logit(claimant count) effect + region effect + socio-economic group effect + logit(total LAD claimant count) effect. Here * denotes all main effects and interaction effects associated with the contributing terms, in this case equivalent to assuming that each age-sex group has a separate regression relationship for the logit of claimant count. The region effects and socio-economic group effects in the model correspond to effects defined by a regional classification of the LADs and a separate socio-economic classification of these areas. Finally, a separate regression effect is included, defined for the total claimant count in the LAD, and reflecting overall employment conditions in the area. The small area estimates obtained by fitting this model to the LFS direct estimates are calibrated (via iterative re-scaling) to agree with direct estimates of unemployment by age-sex, region and socio-economic group. Also, the estimated MSE includes a between area variance term (derived from fitting a separate random effects version of the model to the LFS data).

Model B

Same as (A) except that no between area component is included in the estimated MSE.

Model L1

These estimates are based on a model without the age-sex specific claimant count effects but with LAD level claimant count effect. That is, it assumes that the relationship between ILO unemployment and claimant count is the same for all age-sex groups in a LAD. The model has 22 terms and can be specified: age-sex effect + region effect + socio-economic group effect + logit(LAD claimant count) effect. These estimates obtained by fitting this model are calibrated to agree with national level age-sex, region and socio-economic counts of ILO unemployed. As with model (B) there is no between area component in the MSE.

Model L2

Same approach as with Model L1 but now with age-sex specific claimant count effects and without the LAD level claimant count effect in the model. This model can be specified as: age-sex effect * logit(claimant count) effect + region effect + socio-economic group effect.

Model M

This is the simplest model we considered. It has no claimant count effects at all, and can be defined as: age-sex effect + region effect + socio-economic group effect. As with all models, estimates derived from it are calibrated on age-sex, region and socio-economic group. Also there is no between area component in MSE.

All the above models were used to derive both GLM and GLMM-based predictors. The latter however include an LAD-specific random effect. GLMM model parameters were fitted using PL/REML and MSE estimated using the Prasad-Rao approach. In addition, a synthetic GLM estimator was produced using just the estimates of the fixed parameters in the GLMM model. This is denoted GLMM/Fixed below.

The table below shows selected diagnostics for the "goodness of fit" of the different small area estimates to the direct estimates (for the LADs). The Wald statistic (W) and associated p-value test for closeness of model-based estimates to expected values of the direct estimates (values in parentheses are cross-validation values of this statistic). The further this statistic deviates from the sample size (i.e. the number of LADs) the worse the fit. Similarly, the Non-Overlap statistic measures the proportion of non-overlapping "2 sigma" CIs for the direct and model-based estimates. This should be five percent of 406 (i.e. approximately 20) if the model-based MSEs are valid. Inspection of these results indicate potential 'overfitting" by the GLMM estimates. With the exception of the GLM(M) estimates, however, the estimates based on fixed model specification seem quite reasonable.

Method	W	p-value	Non-Overlap
GLM(A)	355.1483 (391.8445)	0.9672 (0.6841)	12/406
GLM(B)	421.1285 (466.5007)	0.2919 (0.0202)	18/406
GLMM(A)	261.8535	1	6/406
GLMM(A)/Fixed	416.8397	0.3444	17/406
GLM(L1)	425.5356	0.2425	17/406
GLMM(L1)	260.2399	1	6/406
GLM(L2)	421.0064	0.2920	18/406
GLM(M)	584.3274	1.47e-8	38/406
GLMM(M)	184.0444	1	3/406

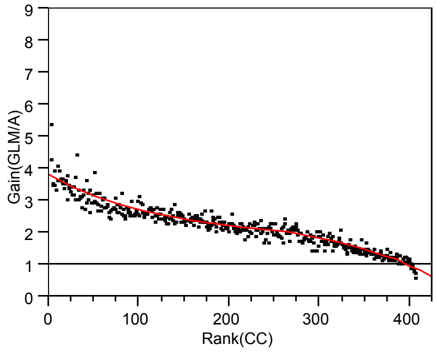
Another set of diagnostics is based on the OLS regression of the square root of the direct estimates on the model-based estimates (estimated parameters with standard errors in parentheses are shown in the following table). The ideal here is a fitted regression line that is not significantly different from a line with unit slope that passes through the origin. Inspection of the entries in the table indicate that the GLM-based estimates tend to be slightly better than the GLMM-based estimates with respect to this diagnostic.

Method	Intercept	Slope	R2
GLM(A)	-0.6278 (1.1048)	0.9995 (0.0169)	0.8960
GLMM(A)	-1.7836 (0.9409)	1.0183 (0.0144)	0.9250
GLMM(A)/Fixed	-0.6213	0.9994	0.8960

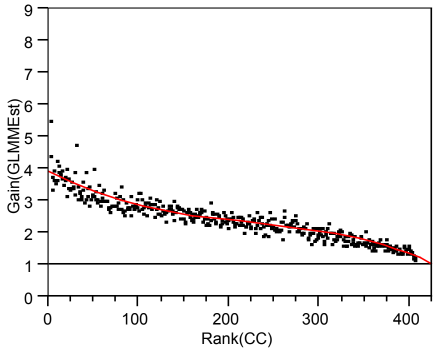
	(1.1049)	(0.0169)	
GLM(L1)	-0.4513 (1.1064)	0.9969 (0.0170)	0.8953
GLMM(L1)	-1.6966 (0.9375)	1.0171 (0.0144)	0.9253
GLM(L2)	-1.20921 (1.1154)	1.0076 (0.0171)	0.8958
GLM(M)	-1.6320 (1.3266)	1.0103 (0.0203)	0.8593
GLMM(M)	-3.7089 (0.8536)	1.0467 (0.0131)	0.9406

Finally, we illustrate the gains from using the model-based small area estimates by showing their gain plots (smooth curves are cubic polynomial fits to gain values). The Gain for a particular LAD is defined as the estimated SE of the Direct Estimate for that LAD divided by the Estimated RMSE of the corresponding model-based estimate. The plots show these gain values for different model/methods when the LADs are ordered by their claimant counts on the *x*-axis. We observe that the gains associated with GLM-based methods are generally larger than those based on the GLMM-based methods. This is what we would expect – the introduction of a random effect tends to ameliorate potential bias at the expense of decreasing precision. However, since the GLM-based methods seem essentially unbiased in this application, the argument for introducing random effects is much weaker. The main impact of introducing a random effect can be seen in the results for model M where the extremely important claimant count covariate was excluded. Here the GLM-based estimates fail the bias test(s) described previously, while the GLMM-based estimates pass. The price for attaining unbiasedness for such a poorly specified model can been seen however in the gains associated with the GLMM in this case. These are rather small.

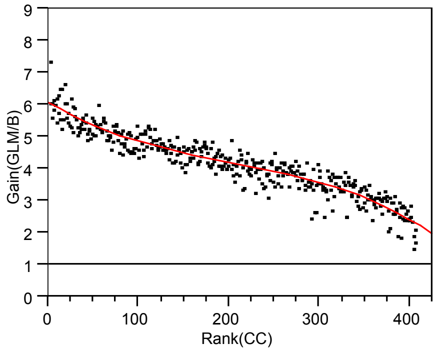
GLM(A)



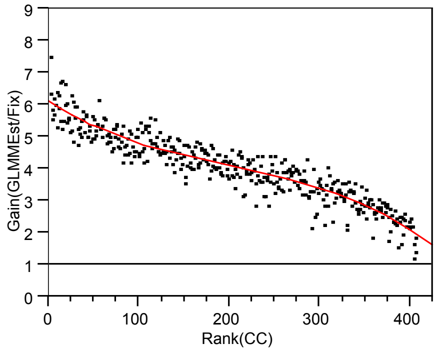
GLMM(A)

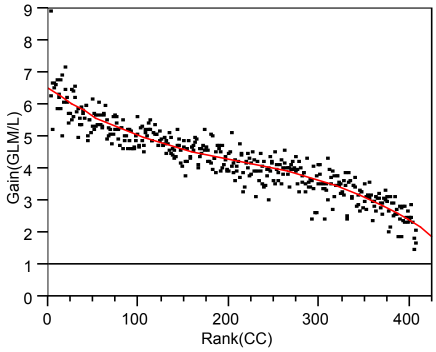


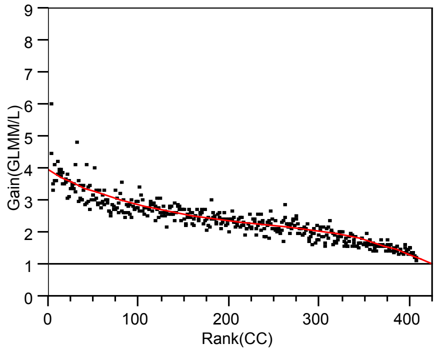
GLM(B)

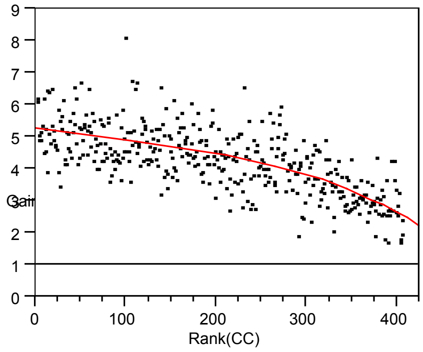


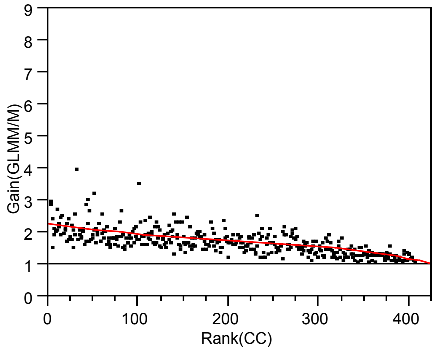
GLMM(A/Fixed)




GLM(L1)


GLMM(L1)


GLM(M)


GLMM(M)

REFERENCES

Bankier, M. D., Rathwell, S. and Majkowski, M. (1992). Two step generalised least squares estimation in the 1991 Canadian census. *Proceedings of the Workshop on Uses of Auxiliary Information in Surveys*, Statistics Sweden, Örebro, Oct 5 - 7 1992.

Bardsley, P. & Chambers, R. L. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33, 290 - 299.

Beaton, A.E. and Tukey, J.W. (1974), The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data, *Technometrics* **16**, 147 - 185.

Brewer, K. R. W. (1963). Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics* **5**, 93 - 105.

Chambers, R.L. (1982), Robust finite population estimation, Unpublished Ph.D thesis, *The Johns Hopkins University*, Baltimore.

Chambers, R. L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063 - 1069.

Chambers, R. L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3 - 32.

Chambers, R. L. and Dorfman, A. H. (1994). Robust sample survey inference via bootstrapping and bias correction: The case of the ratio estimator. *Proceedings of the Joint Statistical Meetings of the ASA, IMS and the Canadian Statistical Society*, Toronto, August 13-18, 1994.

Chambers, R. L., Dorfman, A. H. and Wehrly, T. E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association*, 88, 260-269.

Cochran, W.G. (1977). Sampling Techniques. (Third Edition). New York: Wiley.

Cressie, N. (1982). Playing safe with misweighted means. *Journal of the American Statistical Association* **77**, 754 - 759.

Dalenius, T. and Hodges, J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association* **54,** 88-101.

Deming, W. E. (1960). Sample Design in Business Research. New York: Wiley.

Deville, J. C. and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376 - 382.

Gwet, J.-P. and Rivest, L.-P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association* **87**, 1174 - 1182.

Huang, E. T. and Fuller, W. A. (1978). Nonnegative regression estimation for survey data. *Proceedings of the Social Statistics Section*, American Statistical Association 1978, 300 - 303.

Huber, P.J. (1964), Robust estimation of a location parameter, *The Annals of Mathematical Statistics* **35**, 73 - 101.

Kacker, R.N. and Harville, D.A. (1984). Approximations for standard errors of estimations of fixed and random effects in mixed linear models. *Journal of the American Statistical Association* **79**, 853-862.

Kuo, L. (1988). Classical and prediction approaches to estimating distribution functions from survey data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 280 - 285.

Mahalonobis, P. C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society* **109**, 325-370.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* **97**, 558 - 606.

Prasad, N.G.N and Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association* **85**, 163-171.

Rivest, L.P. and Rouillard, E. (1991), M-estimators and outlier resistant alternatives to the ratio estimator, In *Proceedings of the 1990 Symposium of Statistics Canada*, 271 - 285.

Royall, R. M. (1970). On finite population sampling under certain linear regression models. *Biometrika* **57**, 377 - 387.

Royall, R. M. (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71, 657 - 664.

Royall, R. M. and Cumberland, W. G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association* **73**, 351 - 358.

Royall, R. M. and Cumberland, W. G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association* **76**, 66 - 88.

Royall, R. M. and Eberhardt, K. A. (1975). Variance estimates for the ratio estimator. *Sankhya* **C** **37**, 43 - 52.

Royall, R. M. and Herson, J. (1973a). Robust estimation in finite populations I. *Journal of the American Statistical Association*, 68, 880 - 889.

Royall, R. M. and Herson, J. (1973b). Robust estimation in finite populations II. *Journal of the American Statistical Association* **68**, 890 - 893.

Särndal, C. E., Swensson, B. & Wretman, J. (1992). *Model-assisted survey sampling*. New York: Springer-Verlag.

Silva, P. L. N. and Skinner, C. J. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology* **23**, 23-32.

Shao, J. and Tu, D. (1995). The Jackknife and Bootstrap. New York: Springer-Verlag

Tallis, G. M. (1978). Note on robust estimation in finite populations. *Sankhya* **C 40**, 136 - 138.

Tam, S. M. (1986). Characteristics of best model-based predictors in survey sampling. *Biometrika* **73**, 232 - 235.

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

Wolter, K. M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.