Tilastokeskus
Statistikcentralen
Statistics Finland

Tutkimuksia
Forskningsrapporter
Research Reports 247

Generalised Regression Estimation for Domain Class Frequencies

Mikko Myrskylä

Tilastokeskus
Statistikcentralen
Statistics Finland

Tutkimuksia
Forskningsrapporter
Research Reports 247

Generalised Regression Estimation for Domain Class Frequencies

Mikko Myrskylä

Päätoimittaja – Chefredaktör – Principal editor
Timo Alanko

Toimittaja – Redaktör – Editor
Mikko Myrskylä

Taitto – Ombrytning – Layout
Hilkka Lehtonen

ISSN 0355–2071
= Tutkimuksia
ISBN 978–952–467–712–7

Multiprint Oy

Helsinki – Helsingfors 2007

Acknowledgements

Research is rarely done in a vacuum, and this one is no exception. I am indebted to a large number of people who have helped me start, conduct, and finish this project.

My academic advisors, Risto Lehtonen and Carl-Erik Särndal, have had an enormous influence on this research. Risto introduced me to many of the questions I study in this thesis, and his guidance and support throughout the project have been invaluable, and discussions with both Risto and Carl-Erik have been an endless source of new ideas and encouragement. Pre-examiners Juha Alho and Imbi Traat and my colleague Ari Veijanen read the manuscript of this thesis; their detailed, insightful, and critical suggestions have made this dissertation markedly better.

The people mentioned above have had a direct impact on this thesis. Sometimes the influence has been less direct, but equally important. Before starting this project, I was lucky to attend Imbi's class on survey sampling; the mathematically rigorous approach she advances has saved me from many problems. At that time I was also taking Yrjö Vartia's econometrics classes; I would like to thank him for helping me figure out what is scientific research.

Most of this research was conducted while I was working at Statistics Finland. I am thankful for the facilities and friendship I enjoyed there. Special thanks go to Timo Alanko, Kari Djerf, Janika Konnu, Seppo Laaksonen, Pauli Ollila, Pasi Piela, and all my other colleagues at the Department of Statistical Research and Methodology, and to Mia Kilpiö who improved the language and Hilkka Lehtonen who did the painful job of professionally editing all the formulas and graphs of this thesis. In addition to Statistics Finland, The Finnish School of Statistical Information, Inference, and Data Analysis has financially helped me to do this research.

Even with the best colleagues and research environment, writing a dissertation is a long and lonely job. And during the process of research there are moments when one wants to get mentally as far as possible from the research problems. Luckily I have wonderful friends such as Jari, Juha-Matti, Kalle, and Perttu who could not care less about the topic of this dissertation.

Finally, the greatest gratitude goes to my wife Jaana, who has made all the difference in this project, as in all others.

Philadelphia, July 2007

Mikko Myrskylä

Abstract

Generalised Regression Estimation for Domain Class Frequencies
Doctoral dissertation in Statistics, University of Helsinki, July 1, 2007.
115 pages and 4 appendices.

This study examines the properties of Generalised Regression (GREG) estimators in the estimation of domain class frequencies and proportions.

The family of GREG estimators forms the class of design-based model-assisted estimators. All GREG estimators utilise auxiliary information via modelling. The classic GREG estimator with a linear fixed effects assisting model (GREG-lin) is one example. But when estimating class frequencies, the study variable is binary or polytomous. Then, from the modeller's point of view, logistic-type assisting models (e.g. logistic or probit model) should be preferred over the linear assisting model. However, other GREG estimators than GREG-lin are rarely used, and knowledge about their properties is limited. This study examines the properties of L-GREG estimators, which are GREG estimators with fixed-effects logistic-type models.

First, we study whether and when L-GREG estimators are more accurate than GREG-lin. Both theoretical results and empirical results based on Monte Carlo experiments are given. The experiments cover simple random sampling without replacement (SRSWOR) and fixed size without replacement probability proportional to size (πPS) designs. Several alternative assisting models, including correct, overfitted and weak models, are used. The results show that in standard situations, the difference between L-GREG and GREG-lin is small. But in the case of a strong assisting model, two interesting situations arise: if the domain sample size is reasonably large, L-GREG is more accurate than GREG-lin, and if the domain sample size is very small, estimation of assisting model parameters may be inaccurate, resulting in bias for L-GREG.

Second, we study the goodness of the Standard variance estimator for L-GREG estimators. This variance estimator resembles the Sen-Yates-Grundy variance estimator, but it is a double sum of prediction errors, not of the observed values of the study variable. The Standard variance estimator is widely used for GREG-lin, and in the literature, it has also been suggested for L-GREG estimators. However, Monte Carlo experiments covering both SRSWOR and πPS designs show that the Standard variance estimator underestimates the variance of L-GREG estimators especially if the domain sample size is minor, or if the assisting model is strong. For large domain sample sizes the Standard variance estimator performs well.

Third, we propose a new Augmented variance estimator for L-GREG estimators. The difference between the Standard and Augmented variance estimators is that the latter takes into account the difference between the sample fit model and the census fit model. In Monte Carlo experiments the Augmented variance estimator outperformed the Standard variance estimator in terms of bias, root mean square error and coverage rate. Thus this new estimator provides a good alternative to the Standard variance estimator.

Keywords: generalised regression estimator, class frequencies,
model-assisted domain estimation, logistic model, variance estimation

Tiivistelmä

Alueittaisten luokkafrekvenssien estimointi yleistetyillä regressioestimaattoreilla
Tilastotieteen väitöskirja, Helsingin yliopisto, 1. heinäkuuta 2007
115 sivua ja 4 liitettä.

Tässä työssä tutkitaan yleistettyjen regressioestimaattoreiden (GREG-estimaattoreiden) ominaisuuksia alueittaisten luokkafrekvenssien estimoinnissa.

GREG-estimaattorit muodostavat asetelmaperusteisten malliavusteisten estimaattoreiden joukon. Nämä estimaattorit hyödyntävät lisäinformaatiota tilastollisen mallin avulla. Estimaattoreista tunnetuin, GREG-lin-estimaattori, käyttää lineaarista kiinteiden tekijöiden mallia. Luokkafrekvenssien estimoinnissa kuitenkin mallinnettava vastemuuttuja on kaksi- tai moniluokkainen, jolloin logistis-tyyppiset mallit kuten logistinen tai probit olisivat luontevampia. Silti GREG-estimaattoreissa käytetään vain harvoin muita kuin lineaarisia, kiinteiden tekijöiden malleja, ja kirjallisuudessakin on vain muutamia tutkimuksia muiden GREG-estimaattoreiden kuin GREG-lin-estimaattorin ominaisuuksista. Tässä työssä tarkastellaan L-GREG-estimaattoreiden, eli GREG estimaattoreiden joiden avustava malli on logistis-tyyppinen, ominaisuuksia ja verrataan niitä GREG-lin-estimaattoriin. Tutkimuksessa keskeisessä osassa ovat Monte Carlo -simulointikokeet jotka kattavat palauttamatta-tyyppiset yksinkertaisen satunnaisotannan ja kiinteäkokoisen suhteellisen sisältymistodennäköisyyden otannan.

Ensimmäisessä vaiheessa selvitetään, ovatko L-GREG-estimaattorit GREG-lin-estimaattoreita tarkempia, ja jos kyllä, niin milloin. Osoittautuu, että L-GREG-estimaattori on tarkempi erityisesti jos avustava malli on vahva. Jos malli on heikko, GREG-lin ja L-GREG ovat likimain yhtä tarkkoja. Erittäin pienillä otoskoilla L-GREG-estimaattori voi kuitenkin olla harhainen ja siten epätarkempi kuin GREG-lin.

Toiseksi tutkitaan standardivarianssiestimaattorin (S) tarkkuutta L-GREG-estimaattorien varianssin estimoinnissa. Varianssiestimaattori S muistuttaa muuten tunnettua Sen-Yates-Grundy-estimaattoria, mutta koostuu mallin ennustevirheistä, ei tulosmuuttujan havaituista arvoista. S-estimaattoria käytetään yleisesti GREG-lin-estimaattorin varianssin estimointiin. Kirjallisuudessa sitä on ehdotettu käytettäväksi myös L-GREG-estimaattoreille. Monte Carlo -simulointikokeissa S kuitenkin aliestimoi L-GREG-estimaattorin varianssia erityisesti pienillä otoskoilla sekä tilanteissa, joissa avustava malli oli vahva. Suurilla otoskoilla, tai toisaalta jos malli oli heikko, standardivarianssiestimaattori S oli tarkka.

Kolmanneksi, koska S-estimaattorilla on taipumus aliarvioida varianssia, L-GREG-estimaattorille johdetaan uusi varianssiestimaattori. Tämä estimaattori perustuu samaan approksimaatioon kuin S-estimaattori, mutta se täydentää S-estimaattoria huomioimalla otos- ja perusjoukkosovitteiden eron ennustevirheissä. Simulointikokeissa täydennetty varianssiestimaattori oli selvästi S-estimaattoria tarkempi: sen harha, varianssi ja keskineliövirhe olivat lähes kaikissa tapauksissa pienempiä kuin S-estimaattorin, ja useissa tapauksissa merkittävästi.

Avainsanoja: yleistetty regressioestimaattori, luokkafrekvenssi,
malliavusteinen alue-estimointi, logistinen malli, varianssin estimointi

Contents

Appendices

1 Introduction

The demand for accurate statistics on sub-groups or domains of a population is growing. These statistics, such as regional labour force statistics, or disease prevalence statistics by demographic group, are conventionally produced by survey sampling. But besides sampling, registers may also be used. Register-based data, however, often update and adjust to changing needs relatively slowly, and only occasionally provide information on the study variables that are of interest. Therefore surveys are used in the production of, for example, official monthly unemployment statistics in many European countries.

But registers and surveys need not be mutually exclusive: register-based data may be used both in designing the sample and estimation from the sample. This additional information and in general any information available independently of the sample, is called auxiliary information. Unit level auxiliary information is routinely used in the sampling design, but in the estimation phase, aggregate auxiliary information is commonly used. But as more and more unit level auxiliary data become available, for example, from administrative registers, it is natural to ask how it could be used to make estimation more accurate. The answer depends, among other things, on whether one adopts a design-based or model-based estimation approach.

Design-based and model-based approaches are the two main approaches in survey sampling. In model-based estimation, estimators are generally biased with respect to the sampling design but may have relatively small variance even for domains with a small sample size. From these properties it follows that model-based estimators are often recommended especially for small domain estimation. In design-based estimation, estimators are in most cases approximately unbiased, but may have relatively large variance if the domain sample size is small. Thus, there is a trade-off between bias and variance. In this study, the aim is to study the properties of a certain family of design-based domain class frequency estimators that utilise unit level auxiliary information through explicit modelling. Since the context is *design-based* estimation for domain class frequencies, model-based estimation will be discussed only briefly.

During the past decade, the question of how to use unit level auxiliary information in the design-based framework has been addressed in two ways. In the design-based calibration approach, customarily only aggregate auxiliary information has been used. But recently, calibration methods that utilise unit level auxiliary information have been developed. And in the design-based model-assisted approach where unit level auxiliary data are routinely used, estimators that have some other than the conventional linear fixed effects assisting model have been proposed.

The cornerstone of design-based model-assisted estimators is GREG-lin, the generalised regression estimator with a linear, fixed effects assisting model. By using the structure of GREG-lin as a starting point and setting no restrictions on the choice of the assisting statistical model, we get an estimator family called generalised regression (GREG) estimators. However, GREG estimators with other than a linear, fixed effects assisting model are rarely used, despite the fact

that the linear model formulation is not always the best choice, at least from the modeller's point of view. For class frequencies and proportions, for example, more appropriate assisting models would be models whose predictions can be interpreted as probabilities. In the following, these types of models are called logistic-type models. Examples of logistic-type models are logit, probit, log-log and complementary log-log models. Respectively, GREG estimators with logistic-type assisting models are called logistic-type GREG estimators, or briefly, L-GREG estimators.

The properties such as bias, precision and accuracy of L-GREG estimators are not well known. In this thesis, these properties are studied and compared with the classic GREG-lin. Throughout the study, the accuracy of estimators is considered only with respect to the sampling design. Thus, important sources of non-sampling errors such as nonresponse, frame imperfections and measurement errors are not discussed. This choice has been made since the sampling error often dominates the overall error. It should be stressed that this choice does not mean that non-sampling errors would be unimportant but that they should be a topic of further research.

All GREG estimators, including the classic GREG-lin as well as L-GREG estimators, are design-based model-assisted estimators. This means that the estimators use information about the design by means of sampling weights (thus design-based), and they use models explicitly as an assisting tool to improve accuracy. The GREG estimators are said to be model-assisted, not model-dependent, because in standard situations they are approximately unbiased irrespective of the chosen assisting model.

While the bias of a GREG estimator generally does not depend on the assisting model, the accuracy may depend on the goodness of the assisting model. *Thus, the first set of research questions of this study is whether and when is accuracy gained by changing the assisting model of a GREG estimator from the classic linear fixed effects model to a more natural logistic-type model, and when is accuracy lost?* These questions are studied using both theoretical arguments and empirical results based on the Monte Carlo simulation. The Monte Carlo simulations cover both simple random sampling without replacement (SRSWOR) and fixed size without replacement probability proportional to size (πPS) designs. The experiments also cover a wide range of different L-GREG estimators.

The second set of research questions concerns variance estimation for L-GREG estimators: how well does the Standard variance approximation and the corresponding Standard variance estimator work for L-GREG estimators? These variance approximations and estimators, which resemble the famous Sen-Yates-Grundy formulas but consist of prediction errors, not values of study variables, were originally constructed for GREG-lin. However, in the literature they have also been suggested for a special case of L-GREG estimators, the GREG estimator with a logistic assisting model. Yet, little is known about their accuracy when applied to L-GREG estimators. We study this using Monte Carlo experiments that cover both SRSWOR and πPS designs.

In the course of the study, it turns out that in certain situations the Standard variance estimator severely underestimates the variance of L-GREG estimators. We decompose the total error of the Standard variance estimator into the ap-

proximation error and estimation error, and show that a large part of the total error is due to the estimation error. Another decomposition shows that the source of the estimation error is that the Standard variance estimator does not take into account the difference between the sample fit assisting model and the census fit assisting model.

To improve variance estimation, we propose a new variance estimator, called the Augmented variance estimator. The Augmented variance estimator is based on the Standard variance *approximation*. The difference with respect to the Standard variance *estimator* is that the Augmented variance estimator does take into account the difference between the sample fit and the census fit models.

The third set of research questions concerns the goodness of the Augmented variance estimator: does the Augmented variance estimator provide improvement over the Standard variance estimator, and if yes, when? Under what conditions is the Augmented variance estimator inferior to the Standard variance estimator? The accuracy of the Augmented variance estimator is studied by Monte Carlo experiments that cover both SRSWOR and πPS designs. In addition to the Augmented variance estimator, the performance of two well-known resampling-based variance estimators is studied: the delete-one jackknife and without replacement bootstrap.

The study is organised as follows: In Chapter 2, a brief literature overview is given on design-based model-assisted estimation with focus on GREG estimators. Chapter 3 presents the definitions and notation. The family of GREG estimators, including GREG-lin and L-GREG estimators, is discussed in Chapter 4. In this chapter, we also review the Standard variance approximation and the Standard variance estimator. In Chapter 5, the first set of Monte Carlo experiments is carried out. In these experiments, we study the accuracy of GREG-lin and L-GREG estimators and the goodness of the Standard variance approximation and the Standard Variance estimator. It turns out that the Standard variance estimator fails in certain situations. Thus in Chapter 6, we study the properties of three alternative variance estimators: the jackknife, the bootstrap and the Augmented variance estimator. We conduct a small simulation study in order to compare these estimators with the Standard variance estimator. In terms of bias and coverage rate, the Augmented variance estimator performs best. Thus in the second set of Monte Carlo experiments in Chapter 7, we study the properties of the Augmented variance estimator in more detail. The results of the study and their implications are discussed in Chapter 8.

2 Literature overview

To set the GREG estimators and their variance estimators in context, we discuss briefly different paradigms of survey sampling and present various classifications of estimators in Chapter 2.1. In Chapter 2.2, the literature concerning the use of models under the design-based paradigm is reviewed. In Chapter 2.3, variance estimation methodologies are appraised in finite population estimation in general and especially in the context of model-assisted GREG estimators.

It should be kept in mind that in this study, the focus is on fixed and finite population estimation, where the accuracy of estimators is evaluated with respect to hypothetical repeated sampling. The estimators we consider are all design-based model-assisted. Thus, in Chapter 2.1 we describe more thoroughly the design-based model-assisted approach and only briefly mention the model-based approach. The same holds for Chapters 2.2 and 2.3: modelling in finite population estimation is reviewed only from the point of view of model-assisted estimation, and in the review of variance estimation we focus on variance estimators constructed for design-based model-assisted estimators.

2.1 Classification of estimators

In survey sampling, estimators of finite population parameters are often classified into design-based and model-based. Another classification is direct and indirect estimators. Within these classifications, estimators may then be further classified. An estimator may have properties from several classes and the borders of classes are not always clear. From the practical point of view, this is not a problem: the accuracy of an estimator does not depend on the classification.

2.1.1 Design-based and model-assisted estimators

In the design-based approach, the population is considered *fixed and finite* and its units can be identified and labelled. The study variables are also fixed; therefore the only source of randomness is the randomness of the sample. *Design-based estimators* use information about the sampling design by means of sampling weights.

Within the design-based framework, auxiliary information can be incorporated into the estimation process by means of calibration or modelling. In calibration, the sampling weights are adjusted so that they are consistent with auxiliary population totals. In this approach, models are not necessarily needed. Fundamental papers that discuss the calibration approach are Deville and Särndal (1992), Deville, Särndal and Sautory (1993), and Estevao and Särndal (2006). Recently, Wu and Sitter (2001) and Wu (2003) have studied calibration estimators which utilise micro level auxiliary information.

In design-based model-assisted estimation (or briefly model-assisted estimation), a statistical model is explicitly used as an assisting tool when incorporating auxiliary information into the estimation procedure. This requires that we treat

the fixed and finite population *as if* it had been generated by a statistical model (Särndal *et al.* 1992). This assisting model (or working model) is used to make predictions and both the predictions for the non-sampled units and prediction errors for the sampled units are used in the estimation. Model-assisted estimators use information on the sampling design in the form of sampling weights; therefore they are special cases of design-based estimators.

Which estimators are model-assisted and which are not, is not always clear. In this study, we make the following definition: a design-based estimator is model-assisted only if there is an explicit model that is used to make predictions. It follows that, in the context of this study, all model-assisted estimators are GREG estimators and all GREG estimators are model-assisted. However, all design-based calibration estimators are not model-assisted. Figure 2.1 provides a picture of the design-based calibration and design-based model-assisted estimators. The intersection of C and M is the set of estimators that can be written as both calibration and model-assisted estimators; GREG-lin estimators, for example, belong to this set, but L-GREG estimators generally do not.

Model-assisted estimators, or equivalently GREG estimators in the context of this study, are characterised by design unbiasedness: Even with an ill-fitting model, a model-assisted estimator is in standard situations approximately unbiased. If the model is good, the estimator is also accurate. But for domains with only a few sample observations the variance of a model-assisted estimator can be unacceptably high. Then model-based or composite estimators may be useful.

It should be noted that a GREG estimator is not always approximately unbiased. Musting (2004) and Jurevitš (2005) have derived conditions under which the GREG-lin estimator is biased and also proposed a correction for this bias. The bias correction, however, may increase variance, so accuracy is not necessarily improved. In this study, we study the bias of L-GREG estimators.

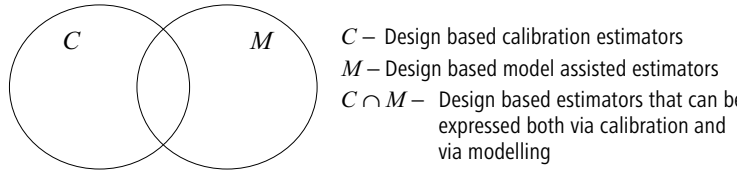


Figure 2.1 Design-based calibration estimators and design-based model-assisted estimators.

2.1.2 Model-based estimators

In the *model-based* framework, the population is considered as a realisation of some hypothetical, infinite superpopulation of populations and the sampling design is ignored. This approach is justified especially when sampling is done with replacement with equal inclusion probabilities (SRSWR design); then the sam-

pled units can be considered as independent realisations of a random variable. Crucial difference with respect to model-assisted estimators is that in model-based estimation, sampling design is often ignored.

Another way to describe the model-based approach is to consider the values of study variables as realisations of random variables. The random variable, or model, that generates the observed values, is then of interest. Inference focuses on the unknown parameters of the distribution of the random variable and the source of uncertainty is the validity of the model assumption.

Model-based estimators rely heavily on models. As in model-assisted estimation, in model-based estimation auxiliary information is also incorporated into the estimation procedure by making use of models. If the model happens to be very good, a model-based estimator is generally more accurate than model-assisted. However, when the model fails, the bias of model-based estimators may be large.

For a comprehensive description of the model-based approach, see Valliant *et al.* (2000) and Chambers (2003). Rao (2003) provides a thorough description of model-based estimation with focus on small areas.

In addition to model-assisted and model-dependent estimators, several estimators which incorporate properties of both the model-assisted and model-dependent approach have been developed. These *composite estimators* are discussed in Valliant *et al.* (2000) and Ghosh (2001).

2.1.3 Direct and indirect estimators

Estimators can also be divided into direct and indirect estimators. This classification is relevant only when domains are studied, since population level estimators are in practice always direct. *Direct estimators* are defined as estimators that use the values of study variables only from the domain under consideration. *Indirect estimators*, on the contrary, utilise information about study variables not directly related to the domain under study (Estevao and Särndal 2004). Table 2.1 shows the situations where direct and indirect estimators are commonly used.

A concept closely related to indirect estimation is *borrowing strength*. Borrowing strength means that one uses information outside the domain under study, and strength can be borrowed cross-sectionally, over time, or both. The information that is borrowed may concern both study and auxiliary variables. For example, in calibration strength is usually borrowed only in terms of auxiliary variables. In model-assisted estimation where models are explicitly used, borrowing strength means that the model is fitted not on the domain level but on a more general level, and strength is borrowed both in terms of study and auxiliary variables.

Table 2.1 The use of direct and indirect estimators by domain type.

Domain \ Estimator	Direct	Indirect
Planned	Often	Rarely
Unplanned	Almost never	Often

Thus, the relationship between direct and indirect estimators and borrowing strength is as follows: Direct estimators do not borrow strength in terms of study variables, but may or may not borrow strength in terms of auxiliary variables. Indirect estimators, by definition, borrow strength in terms of study variables and may or may not borrow strength in terms of auxiliary variables.

The conventional term borrowing *strength* that is used to describe the process where one uses information outside the domain may not be the best possible. This is because borrowing strength may equally well increase or decrease the accuracy of the estimator. Therefore borrowing information might be a more pertinent term. In this study, estimators that borrow strength (or information) always do so in terms of study variables and such estimators will be called indirect estimators. Thus the borrowing terminology is not necessary.

2.2 The role of models in model-assisted estimation

Statistical models have a long history in survey sampling. The earliest known example where a statistical model has been used in the estimation of totals or means seems to be from 1937, when Watson used a regression of leaf area on leaf weight to estimate the average area of the leaves on a plant (see Cochran, 1962 and Knottnerus, 2003). However, the use of models did not really emerge until the 1970's; until that, most of the research in survey methodology focused on using auxiliary information in the sampling phase.

From the 1970's onwards increasing attention has been paid to the use of auxiliary information in the *estimation*. Brewer's, Cochran's and Royall's studies on the ratio estimator and Särndal's work on the general regression (GREG) estimator had a large impact on how survey statisticians started to see the possibilities of auxiliary information: by using auxiliary information in the estimation phase, the sampling design could be kept simple if so desired, without loss in accuracy. The use of auxiliary information also made it possible to reduce the nonresponse bias; this has been especially important in the last quarter of the 20th century when nonresponse rates have been rising (Groves *et al.* 2004, 184–187).

In the design-based context, auxiliary information can be used in the estimation phase either by calibration or by modelling. In the model-assisted approach, statistical models that link the study variable and auxiliary information are utilised to predict values of the study variable. The most widely used model-assisted estimator is the generalised regression estimator with a linear, fixed effects assisting model (GREG-lin). This estimator is comprehensively studied in the textbook by Särndal, Swensson and Wretman (1992) and the paper by Estevao *et al.* (1995).

A wider class of GREG estimators is obtained by letting the assisting model be any statistical model, linear or non-linear, parametric or non-parametric, and so on. In the parametric approach, the parameters of the assisting model may be estimated with or without sampling weights. In the pure design-based approach, which we adopt, sampling weights are used for estimation of model parameters.

Until recently, rarely any other than fixed effects linear models have been used in GREG estimation, irrespective of the nature of the study variable. One reason

for this is that GREG-lin can also be written in the form of a calibration estimator. Consequently, one actually needs only aggregate level auxiliary information.

In the late 1990's, however, other assisting models have also been considered in GREG estimation. Examples are a logistic model (Lehtonen and Veijanen 1998a), polynomial regression models (Breidt and Opsomer 2000), generalised additive models (Opsomer *et al.* 2001), and mixed linear and logistic models (Lehtonen *et al.* 2003). For class frequencies, a logistic model (or some closely related model, such as probit, log-log or complementary log-log) is a natural choice, and it has been shown (Lehtonen and Veijanen 1998a, 1998b, Duchesne 2003, Myrskylä 2004, 2005) that for class frequencies, the accuracy gain can be substantial if a logistic assisting model is used instead of a linear assisting model.

Although more attention has been paid to the selection of the assisting model in the context of GREG estimation, literature on GREG estimators that have some other than a linear model is still quite limited. Simulation-based results of Lehtonen *et al.* (1998a) and Duchesne (2003) indicate that for class frequencies, the L-GREG estimator is more accurate than GREG-lin in some situations, but what exactly these situations are is not clear. Also, the question whether accuracy can be lost by changing the assisting model from linear to logistic-type is unanswered. Moreover, studies that have considered any other than simple random sampling without replacement designs are rare. Lehtonen *et al.* (2006a) study the accuracy of GREG estimators with linear fixed effects and linear mixed models under the unequal probability design, but no studies considering the L-GREG estimator under the unequal probability design have been published. Especially, the accuracy of the L-GREG estimator in the case of the probability proportional to size design has not been studied.

Once the unequal probability sampling design is used, the question about the double use of auxiliary information arises. Specifically, if some auxiliary information is already used in the sampling design, should the same auxiliary information be also used in the estimation phase? Särndal (1996) and Lehtonen *et al.* (2006b) study the effect of double use in the context of continuous study variable and linear models and conclude that double use is profitable. In this study, we try to characterise the conditions under which the L-GREG estimator is more accurate than the classic GREG-lin for domain class frequencies. Both simple random sampling without replacement and probability proportional to size designs are covered, and the double use of auxiliary information is also studied.

2.3 Variance estimation

Figure 2.2, reproduced from Wolter (1985), shows some important dimensions of the estimation strategy (the strategy is a combination of the sampling design and estimator) when it comes to variance estimation.

By simple design we mean sampling designs where the inclusion probability of population units is constant and the sample size fixed. These designs include simple random sampling both with and without replacement. If the sample size is random or inclusion probabilities are unequal, the design is called complex. Examples are stratified sampling and probability proportional to size sampling.

	Simple design	Complex design
Linear estimators	a	b
Nonlinear estimators	c	d

Figure 2.2 Classification of estimation strategies by design and estimator.

By linear estimators we mean estimators that are linear functions of study variables. Examples, such as the Horvitz-Thompson estimator, are easy to use but inefficient. Most GREG estimators are nonlinear; examples are L-GREG estimators and GREG-lin with more detailed auxiliary information than just the population size.

For simple designs and linear estimators (a), variance estimation is straightforward. The Sen-Yates-Grundy variance estimator (Sen 1953, Yates and Grundy 1953), which is sometimes also called the Horvitz-Thompson variance estimator (Horvitz and Thompson 1952), applies directly to strategies where the design is simple and the estimator is linear. The same variance estimator also works in principle in (b), but there is an additional challenge: the calculation of second-order inclusion probabilities (see Chapter 3). In complex designs, second-order inclusion probabilities are often difficult and/or time consuming to calculate. Therefore, several authors have suggested estimators that either totally avoid the calculation of second-order inclusion probabilities or have approximations of these unknown terms. Examples are presented in the papers of Hartley and Rao (1962), Hájek (1964), Särndal (1996) and Berger (2004).

In this study, the cases (c) and (d) are of most interest. This is because the L-GREG estimators we are interested in fall into the category of nonlinear estimators. The literature concerning variance estimation for nonlinear design-based model-assisted estimators, that is, for GREG estimators, may be divided into three parts: linearisation-based variance estimation for the classic GREG-lin, resampling-based techniques which are adaptable to GREG estimators and variance estimation for L-GREG estimators.

The linearisation techniques for GREG-lin are relatively thoroughly studied. Woodruff (1971) and Binder (1983) have presented Taylor series approximations for a general class of estimators for complex surveys. For the GREG-lin estimator, Särndal *et al.* (1989) propose an estimator that is based on the Taylor series approximation and consists of sample weighted residuals. The Taylor series approximation based variance estimation for GREG-lin is also discussed in Särndal *et al.* (1992). Estevao and Särndal (2006) present a method called automated approximation that works for calibration estimators and as a special case, also for the GREG-lin estimator. When the linearisation methods are applied in (d), there is again the problem of second-order inclusion probabilities. In such cases, estimators that either avoid the calculation of second-order inclusion probabilities or have approximations of these unknown terms are often used.

Resampling methods are an alternative to linearisation-based methods. During the last decade, resampling methods have emerged in many areas of statistics

and are becoming popular in survey sampling as well. The predecessor of these modern resampling methods is the random groups technique, where a number of independent samples are drawn from the population (Mahalanobis 1946, Wolter 1985). The obtained sample can also be divided into random groups. However, it is often difficult and expensive to obtain a large enough number of random groups so that the variance estimator would stabilise. Resampling methods, such as the jackknife and bootstrap, try to overcome this problem.

In the jackknife, "random groups" are constructed from the sample, and the jackknife extends the random groups methods so that the random groups may overlap. The method was originally developed in an infinite population context. Quenouille (1949) introduced the method to reduce the bias of an estimator, Tukey (1958) suggested that the method might be used for variance and interval estimation, and Durbin (1959) seems to be the first to consider the jackknife in finite population inference (see Wolter 1985 and Särndal *et al.* 1992). The jackknife is straightforward to apply, but theoretical knowledge of its properties in complex designs is limited. Moreover, it works well when the estimated parameter is a smooth function of population totals (Krewski and Rao 1981), but it does not work for quantiles.

The bootstrap method, like the jackknife, was first introduced in an infinite population context. The originator of this technique was Efron (1979, 1981, 1982). The general idea is to construct a pseudo-population and draw random samples from the population. The distribution of statistics calculated from these pseudo-samples can be used to estimate the distribution of the estimator. The technique works best with independent, identically distributed (iid) observations and in an infinite population context (the standard iid with replacement bootstrap is described in, e.g. Sitter (1992a). Gross (1980) and Bickel and Freedman (1984) have constructed without replacement bootstraps and Sitter (1992a, 1992b) has further developed the method. However, it is still unclear how the technique should be modified if the sampling is not simple random sampling with replacement (Lahiri 2003).

Variance estimation literature concerning the specific estimator family we are studying, the L-GREG estimator, is scarce. Lehtonen and Veijanen (1998a, 1998b) propose a variance estimator that has the form of the famous Sen-Yates-Grundy formula (e.g. Särndal *et al.* 1992, 45) but consists of prediction errors instead of the values of study variables. This estimator is often used for the classic GREG estimator with a linear fixed effects model. We call this variance estimator the Standard variance estimator. Also, Duchesne (2003) and Lehtonen and Pahkinen (2004) use the Standard variance estimator for L-GREG. Lehtonen and Veijanen (1998a, 1998b) and Duchesne (2003) note that the Standard variance estimator underestimates the variance especially when the domain sample size is minor. In this study, we will further study the performance of the Standard variance estimator for L-GREG estimators (Chapter 5) and also examine the performance of several alternative variance estimators (Chapters 6 and 7).

3 Definitions and notation

The notation draws mainly from Traat (2000) and Traat *et al.* (2003). We define a sample as a vector, not a subset of the population; this makes it easy to treat both with and without replacement designs in the same framework.

3.1 Population, auxiliary information and study variables

We study a finite population consisting of units $u_1, u_2, \ldots, u_N$. Let i denote u_i. Now the population is

$$U = \{1, 2, \ldots, N\}. \tag{3.1}$$

Every i is associated with an *identification variable* z, *study variables* y_j, $j = 1, 2, \ldots, J$ and *auxiliary variables* x_m, $m = 1, 2, \ldots, M$. The study variables are unknown prior sampling and the parameters of interest are functions of them. The identification variable and auxiliary variables are known prior sampling for every i. The z and x variables are *auxiliary information;* the z variable on its own is *frame information*.

The population consists of domains $U^{(d)} \subseteq U, d = 1,2,\ldots,D$. Examples of domains could be such as men, women, people belonging to a certain age group, or enterprises whose turnover is below some limit. Domain indicator variables define whether $i \in U$ belongs to a given domain:

$$\delta_i^{(d)} = \begin{cases} 1, \text{ if } i \in U^{(d)} \\ 0 \text{ otherwise} \end{cases} \quad \forall i \in U, d = 1,2,\ldots,D. \tag{3.2}$$

For every $i \in U$, values $\delta_i^{(d)}$, $d = 1,2,\ldots,D$ form a domain indicator column vector $\boldsymbol{\delta}_i = (\delta_i^{(1)}, \delta_i^{(2)}, \ldots, \delta_i^{(D)})$. Correspondingly, let $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iM})$ denote the column vector for auxiliary variables and the vector $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iJ})$ for study variables. All properties of unit i are in the stacked column vector $\mathbf{a}_i = (z_i, \mathbf{x}_i, \boldsymbol{\delta}_i, \mathbf{y}_i)$ of dimension $(1 + M + D + J) \times 1$ and the properties of the population U are

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \vdots \\ \mathbf{a}_N' \end{bmatrix} \begin{bmatrix} z_1 & \mathbf{x}_1' & \boldsymbol{\delta}_1' & \mathbf{y}_1' \\ z_2 & \mathbf{x}_2' & \boldsymbol{\delta}_2' & \mathbf{y}_2' \\ \vdots & \vdots & \vdots & \vdots \\ z_N & \mathbf{x}_N' & \boldsymbol{\delta}_N' & \mathbf{y}_N' \end{bmatrix}. \tag{3.3}$$

The matrix $\mathbf{A}$ of dimension $N \times (1 + M + D + J)$ is called the data matrix. Its row vectors $\mathbf{a}_i$ correspond to units i (for example, persons, households, enterprises) and

column vectors correspond to the properties associated with the units (for example, age of a person, size of a household, revenue of an enterprise). We denote

$$\begin{aligned} \mathbf{z} &= (z_1, z_2, \ldots, z_N), \quad \mathbf{X} = (\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_N') \\ \Delta &= (\delta_1', \delta_2', \ldots, \delta_N'), \quad \mathbf{Y} = (\mathbf{y}_1', \mathbf{y}_2', \ldots, \mathbf{y}_N'). \end{aligned} \tag{3.4}$$

Now the data matrix (3.3) can be written as $\mathbf{A} = [\mathbf{z}\ \mathbf{X}\ \Delta\ \mathbf{Y}]$.

3.2 Parameters of interest

The parameters of interest are frequencies of classes j, $j = 1, 2, \ldots, J$ in domains $U^{(d)} \subseteq U$. Every $i \in U$ may belong to any of the classes and the classes need not be mutually exclusive. Study variables y_j indicate class membership:

$$y_{ij} = \begin{cases} 1, \text{ if } i \text{ belongs to class } j \\ 0 \text{ otherwise.} \end{cases} \qquad \forall\, i \in U, j = 1,2,\ldots,J. \tag{3.5}$$

Domain study variables are defined as products of domain indicator variables and study variables:

$$y_{ij}^{(d)} = \delta_i^{(d)}\, y_{ij}. \tag{3.6}$$

Using (3.6), the number of study variables grows to $J \cdot D$. The parameters of interest, the domain class frequencies, can now be expressed as

$$T_j^{(d)} = \sum_{i \in U^{(d)}} y_{ij} = \sum_{i \in U} y_{ij}^{(d)}. \tag{3.7}$$

Note that the parameter of interest could equally well be the proportion obtained by dividing (3.7) by domain size (we assume that domain indicator variables are known, so also domain size is known).

3.3 Sampling design

To estimate $T_j^{(d)}$, we need information about unknown variables y_j. This information is collected by sampling. We limit ourselves to the design-based framework, where samples are probability samples. This means that every unit has strictly positive probability to be sampled. Sampling can be one-stage, two-stage, and so on, or in general multistage sampling. We consider only one-stage sampling. The *sampling vector*

$$\underline{\mathbf{I}} = (\underline{I}_1, \underline{I}_2, \ldots, \underline{I}_N) \tag{3.8}$$

is a random vector whose elements $\underline{I}_i$ indicate the number selections for i. The realisation $\mathbf{I} = (I_1, I_2, \ldots, I_N)$ of $\underline{\mathbf{I}}$ is called a *sample*. The sampling vector $\underline{\mathbf{I}}$ (and its realisation $\mathbf{I}$) define the *sample set* $\underline{s}$ (and the corresponding s) and the *non-sampled set* $\underline{U}_r$ (and the corresponding U_r) as

$$\underline{s} = \{i : i \in U, \underline{I}_i \geq 1\} \qquad (s = \{i : i \in U, I_i \geq 1\}) \text{ and} \tag{3.9}$$

$$\underline{U}_r = \{i : i \in U, \underline{I}_i = 0\} \qquad (U_r = \{i : i \in U, I_i = 0\}). \tag{3.10}$$

Sampling can be *with replacement* (WR) or *without replacement* (WOR). In WOR sampling, units can be sampled only once, and in WR sampling, more than once. Sampling weights are defined as

$$\underline{w}_i = \frac{\underline{I}_i}{E(\underline{I}_i)} \quad \forall\, i \in U\,, \tag{3.11}$$

these take into account the number of times the unit is sampled. However, the sample set does not contain information about whether the unit is selected more than once. The distinction between sample $\mathbf{I}$ and sample set s should therefore be kept clear: s is a subset of U and its units are determined by $\mathbf{I}$. Sample $\mathbf{I}$, instead, is not a subset of U but a vector in N^N, the N-dimensional space of non-negative integers.

The distribution of $\underline{\mathbf{I}}$, denoted by $p(\cdot)$, is called a *sampling design*. The sampling design assigns a probability $\Pr(\underline{\mathbf{I}} = \mathbf{I}) = p(\mathbf{I})$ for every sample. In survey sampling, the terms *sample design* or *strategy* are also often used; they cover both the sampling design and the estimation plan (Särndal *et al.* 1992, 29). First and second-order inclusion probabilities π_i and π_{ij} are defined as

$$\begin{aligned} \pi_i &= \Pr(\underline{I}_i \geq 1) = \sum_{\mathbf{I}:I_i \geq 1} p(\mathbf{I})\,, \\ \pi_{ij} &= \Pr(\underline{I}_i \geq 1, \underline{I}_j \geq 1) = \sum_{\mathbf{I}:I_i,\, I_j \geq 1} p(\mathbf{I})\,. \end{aligned} \tag{3.12}$$

We consider only designs where both first and second-order inclusion probabilities are strictly positive; this is often expressed by saying that the design is *design measurable*. For designs that are design measurable, expressions for expectation and variance of basic estimators can be obtained.

Covariance of $\underline{I}_i$ and $\underline{I}_j$ is denoted by

$$\operatorname{cov}(\underline{I}_i, \underline{I}_j) = \Delta_{ij}\,. \tag{3.13}$$

For every sampling design, $\pi_i = \pi_{ii}$ and $\underline{w}_i = 0$ if $i \in U_r$. For WOR designs,

$$\pi_i = E(\underline{I}_i),\ \pi_{ij} = E(\underline{I}_i \underline{I}_j),\ \Delta_{ij} = \pi_{ij} - \pi_i \pi_j\,, \text{ and } w_i = \begin{cases} \pi_i^{-1}\,, \text{if } i \in s \\ 0, \text{ if } i \in U_r \end{cases}. \tag{3.14}$$

Depending on the sampling design, the *sample size*

$$n = \sum_{i \in U} I_i \tag{3.15}$$

can be random or pre-determined. If n is non-random, the sampling design is *fixed-size*. Ratio n/N is *sampling fraction*, denoted by f. The sample size and the sample set in domain $U^{(d)}$ are

$$n^{(d)} = \sum_{i \in U^{(d)}} I_i \,, \quad \text{and} \quad s^{(d)} = s \cap U^{(d)} \,. \tag{3.16}$$

The sample size in a domain may be random or fixed. If the domain sample size is fixed, the domain is *planned*; otherwise it is *unplanned*. Usually when a domain is planned, sampling is done with stratification so that from the domain, a sample of certain size is drawn independently of the rest of the sample. In many practical situations, however, fixed-size stratification cannot be done for every domain and domains are often unplanned even when the sample size n is fixed.

We consider two special cases of the general design: simple random sampling without replacement (SRSWOR) and probability proportional to size sampling. SRSWOR is the simplest sampling design that sets a baseline for other designs. It is a fixed-size design under which the first and second-order inclusion probabilities are constants. Under SRSWOR,

$$\pi_i = \frac{n}{N} = f \,, \quad \pi_{ij} = f\frac{n-1}{N-1} \,, \quad \text{and} \quad \Delta_{ij} = \begin{cases} f(1-f), & i = j \\ -f\dfrac{1-f}{N-1}, & i \neq j. \end{cases} \tag{3.17}$$

In SRSWOR sampling the inclusion probabilities are constants. In probability proportional to size sampling the inclusion probabilities satisfy

$$\pi_i \propto x_i \tag{3.18}$$

for some x whose values are known for every unit in the population (as a special case, SRSWOR and SRSWR are obtained from (3.18) when $x_i = x$). The sample size may be fixed or random and sampling may be with or without replacement. The abbreviation PPS is used to denote with replacement probability proportional to size sampling and πPS to denote without replacement probability proportional to size sampling.

We consider only fixed size πPS designs. In πPS designs, both first and second-order inclusion probabilities need to be strictly positive and $\pi_{ik} - \pi_i\pi_k < 0$. This allows the construction of unbiased variance estimators for simple linear estimators. It is not always straightforward to construct sampling algorithms that fulfil these requirements. In the Monte Carlo experiments of this study, we use the Hanurav-Vijayan algorithm (Hanurav 1967, Vijayan 1968).

Once the sample $\mathbf{I}$ is realised, the values of variables y_j, $j = 1, 2, \ldots, J$ are recorded for the units $i \in s$. This information is collected into the matrix $\mathbf{Y}_s$. The unobserved values for U_r are collected correspondingly into the matrix $\mathbf{Y}_r$. Auxiliary

information [**z** **X**], domain indicator matrix Δ and sampling weights $\mathbf{w} = (w_1, w_2, \ldots, w_N)$ are partitioned in a similar way to a sampled and non-sampled part ($w_i = 0$ if $i \notin s$). By re-arranging the rows, the data matrix **A** can be written

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_s \\ \mathbf{A}_r \end{bmatrix} = \begin{bmatrix} \mathbf{z}_s & \mathbf{X}_s & \Delta_s & \mathbf{w}_s & \mathbf{Y}_s \\ \mathbf{z}_r & \mathbf{X}_r & \Delta_r & \mathbf{0} & \mathbf{Y}_r \end{bmatrix}. \tag{3.19}$$

In (3.19), the matrix $\mathbf{Y}_r$ is coloured grey to emphasise the fact that its values are unknown. Under ideal conditions, no nonresponse is present and **A** is fully known apart from $\mathbf{Y}_r$. In practice, some nonresponse always occurs and only in theory it can be harmless. Therefore methods such as re-weighting or imputation are needed to adjust for nonresponse (e.g. Lundström and Särndal 2002). In this study we focus on the sampling error and assume that there is no nonresponse.

3.4 Estimator, estimate and accuracy of an estimator

An estimator is a rule or algorithm that defines how estimates of class frequencies are calculated. It is a random variable whose value depends on the sample and auxiliary information. An estimate, in turn, is the realised value of an estimator. In general notation, an estimator and the corresponding estimate for a population parameter θ are denoted by $\hat{\theta}(\mathbf{I})$ and $\hat{\theta}(\mathbf{I})$, or briefly $\underline{\hat{\theta}}$ and $\hat{\theta}$. For parameters $T_j^{(d)}$, the estimator and estimate are

$$\underline{\hat{T}}_j^{(d)} \text{ and } \hat{T}_j^{(d)} . \tag{3.20}$$

We consider the accuracy of estimators with respect to the sampling design. We do not try to obtain exact sampling distributions, since even if this was possible in principle, it would be computationally impractical. Therefore, measures that summarise important aspects of the sampling distribution, such as bias and variance, are used. These are unknown quantities and have to be estimated.

An estimator is accurate if its bias and variance are small. A *design-unbiased* estimator is one whose expectation with respect to the sampling design equals the true parameter value:

$$E_p(\underline{\hat{\theta}}) = \sum_{\mathbf{I}} p(\mathbf{I})\hat{\theta}(\mathbf{I}) = \theta \tag{3.21}$$

Correspondingly, *design-bias* is defined as the difference between the parameter value and the expectation:

$$B_p(\underline{\hat{\theta}}) = \theta - E_p(\underline{\hat{\theta}}). \tag{3.22}$$

Design-variance is defined as

$$V_p(\underline{\hat{\theta}}) = \sum_{\mathbf{I}} p\left[(\mathbf{I})\hat{\theta}(\mathbf{I}) - E_p(\underline{\hat{\theta}})\right]^2. \tag{3.23}$$

Design-variance is affiliated with precision: the smaller the design-variance, the greater the precision. A measure for accuracy that takes into account both design-bias and design-variance is *design-mean square error* (MSE), defined as

$$MSE_p(\hat{\underline{\theta}}) = E_p(\hat{\underline{\theta}} - \theta)^2 \qquad (3.24)$$

$$= V_p(\hat{\underline{\theta}}) + [B_p(\hat{\underline{\theta}})]^2 .$$

The MSE is, actually, an inverted measure for accuracy: the smaller the MSE, the greater the accuracy.

Since we study the properties only with respect to the sampling design, the prefix design and subscript p are not used in the following. For example, design-bias is called bias and denoted by $B(\hat{\underline{\theta}})$. However, it is important to keep in mind that bias, variance and mean square error generally depend both on the estimator and on the design.

Examples of the distributions of biased, unbiased, precise and imprecise estimators are given in Figure 3.1. The most accurate estimator in the figure is the unbiased, precise estimator (A), and the most inaccurate is the biased, imprecise estimator (B). Both (A) and (B) are atypical in our context: In survey sampling, two main classes of estimators are design-based and model-based estimators. Typical examples of these estimators are given on the right-hand side of Figure 3.1. Design-based estimators are approximately unbiased, but are often relatively imprecise when compared with model-based estimators. Thus, the estimator (C) is a typical design-based estimator. The estimator (D) is a typical model-based estimator, since model-based estimators are often biased but precise. So there is a trade-off between bias and precision. This study focuses on design-based estimators and model-based estimators will not be studied.

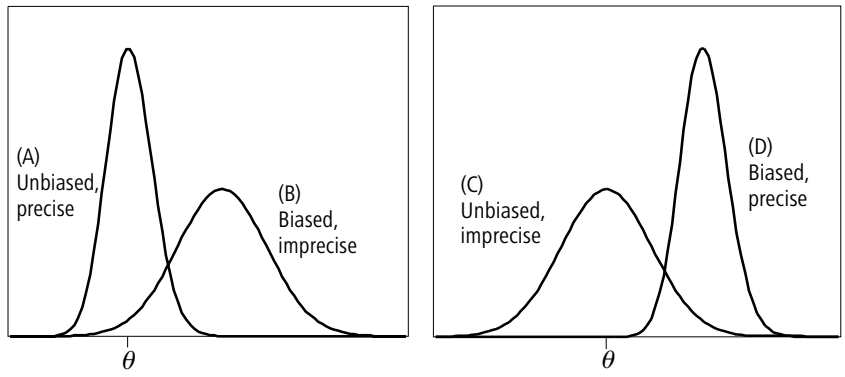


Figure 3.1 Examples of different types of sampling distributions.

In addition to accuracy, one desirable property for total estimators is additivity. Additive estimators are such that for non-overlapping domains $U^{(1)}$, $U^{(2)}$,…, $U^{(k)}$ whose union is $\bigcup_{d=1}^{k} U^{(d)} = U^{(t)}$, it holds that

$$\sum_{d=1}^{k} \hat{\underline{T}}_j^{(d)} = \hat{\underline{T}}_j^{(t)} . \tag{3.25}$$

3.5 The Horvitz-Thompson estimator

The most well-known population total estimator is the Horvitz-Thompson (HT) estimator (Horvitz and Thompson 1952). (It might also be justified to call the estimator the Narain-Horvitz-Thompson, or even the Narain estimator, as it seems that Narain introduced the estimator already in 1951 (see Berger 2003 and Rao 2005)). The HT estimator is linear and unbiased. Since domain class frequencies can be expressed as totals, the HT estimator is appropriate for class frequencies. But HT estimator is inefficient, since it does not use auxiliary information. Thus the HT estimator is introduced here only as a reference estimator and to elaborate certain important concepts.

The parameter of interest is $T_j^{(d)}$, the frequency of class j in domain $U^{(d)}$. The sampling design is fixed-size one-stage design $p(\cdot)$. The weights induced by the design are

$$\underline{w}_i = \begin{cases} \dfrac{\underline{I}_i}{E(\underline{I}_i)}, & i \in \underline{s} \\ 0, & i \in \underline{U}_r . \end{cases} \tag{3.26}$$

For $T_j^{(d)}$, the HT estimator is defined as

$$\hat{\underline{T}}_{j,HT}^{(d)} = \sum_{i \in U^{(d)}} \underline{w}_i \, y_{ij} = \sum_{i \in U} \underline{w}_i \, y_{ij}^{(d)} . \tag{3.27}$$

The corresponding estimate is

$$\hat{T}_{j,HT}^{(d)} = \sum_{i \in s^{(d)}} w_i \, y_{ij} = \sum_{i \in s} w_i \, y_{ij}^{(d)} . \tag{3.28}$$

If the sampling design is with replacement, (3.27) can also be interpreted as the Hansen-Hurwitz estimator. This is possible since the weights w_i carry information about the number of selections of each sample element.

If $U^{(d)} = U$, the estimator (3.27) is a population total estimator:

$$\hat{\underline{T}}_{j,HT} = \sum_{i \in U} \frac{\underline{I}_i}{E(\underline{I}_i)} y_{ij} = \sum_{i \in U} \underline{w}_i \, y_{ij} . \tag{3.29}$$

The HT estimator is always additive. This can be seen from the following. Let domains $U^{(1)}, U^{(2)}, \ldots, U^{(k)}$ be mutually exclusive and $\bigcup_{d=1}^{k} U^{(d)} = U^{(t)}$. Then

$$\begin{aligned}\hat{\underline{T}}_{j,HT}^{(t)} &= \sum_{i \in U^{(t)}} \underline{w}_i y_{ij} \\ &= \sum_{i \in U^{(1)}} \underline{w}_i y_{ij} + \sum_{i \in U^{(2)}} \underline{w}_i y_{ij} + \ldots + \sum_{i \in U^{(k)}} \underline{w}_i y_{ij} \\ &= \sum_{d=1}^{k} \hat{\underline{T}}_{j,HT}^{(d)} .\end{aligned} \tag{3.30}$$

The HT estimator is unbiased since

$$E\left(\hat{\underline{T}}_{j,HT}^{(d)}\right) = \sum_{i \in U^{(d)}} E(\underline{w}_i) y_{ij} = \sum_{i \in U^{(d)}} y_{ij} = T_j^{(d)} . \tag{3.31}$$

The variance of the HT estimator can be obtained by using a known property of variance: if $\underline{x}_i$ are random variables and c_i are constants, then

$$V\left(\sum_{i \in U} \underline{x}_i c_i\right) = \sum_{i \in U} \sum_{k \in U} \operatorname{cov}(\underline{x}_i, \underline{x}_k) c_i c_k . \tag{3.32}$$

In the population level HT estimator (3.29), the constant corresponding to c_i in (3.32) is $y_{ij}/E(\underline{I}_i)$ and the random variable corresponding to $\underline{x}_i$ is $\underline{I}_i$. Therefore the exact variance of (3.29) is

$$\begin{aligned}V\left(\hat{\underline{T}}_{j,HT}\right) &= \sum_{i \in U} \sum_{k \in U} \operatorname{cov}(\underline{w}_i, \underline{w}_k) y_{ij} y_{kj} \\ &= \sum_{i \in U} \sum_{k \in U} \frac{\Delta_{ik}}{E(\underline{I}_i) E(\underline{I}_k)} y_{ij} y_{kj} .\end{aligned} \tag{3.33}$$

The variance (3.33) contains unknown values of y_j and has to be estimated. An unbiased variance estimator is

$$\hat{V}\left(\hat{\underline{T}}_{j,HT}\right) = \sum_{i \in U} \sum_{k \in U} \frac{\Delta_{ik}}{E(\underline{I}_i \underline{I}_k)} \underline{w}_i y_{ij} \underline{w}_k y_{kj} . \tag{3.34}$$

The unbiasedness of (3.34) can be shown by taking expectations:

$$
\begin{aligned}
E\left[\hat{V}\left(\hat{\underline{T}}_{j,HT}\right)\right] &= \sum_{i\in U}\sum_{k\in U}\frac{\Delta_{ik}}{E(\underline{I}_i\underline{I}_k)}y_{ij}y_{kj}E(\underline{w}_i\underline{w}_k) \\
&= \sum_{i\in U}\sum_{k\in U}\frac{\Delta_{ik}}{E(\underline{I}_i\underline{I}_k)}y_{ij}y_{kj}\frac{E(\underline{I}_i\underline{I}_k)}{E(\underline{I}_i)E(\underline{I}_k)} \qquad (3.35) \\
&= \sum_{i\in U}\sum_{k\in U}\frac{\Delta_{ik}}{E(\underline{I}_i)E(\underline{I}_k)}y_{ij}y_{kj} \\
&= V\left(\hat{\underline{T}}_{j,HT}\right).
\end{aligned}
$$

For the domain estimator (3.27), the variance and variance estimator are obtained from (3.33) and (3.34) by replacing y_{ij} by $y_{ij}^{(d)}$ (Särndal *et al.* 1992, 391).

4 Generalised regression (GREG) estimators

Domain class frequencies can always be expressed as totals. Therefore estimators appropriate for totals can be used to estimate domain class frequencies. When totals, or any other population parameters are estimated, auxiliary information can be used in two phases: in the sampling phase and in the estimation phase. We consider the use of auxiliary information in both phases. When using SRSWOR, auxiliary information is used only in estimation, but when using πPS, auxiliary information is used both in sampling and in estimation.

The estimator family we consider is the family of generalised regression (GREG) estimators. The idea underpinning GREG estimators is that there is a statistical connection between auxiliary and study variables. Thus auxiliary information carries information about the study variable and can be utilised in the estimation phase to reduce the variance of an estimator (Särndal *et al.* 1992, 220–221).

GREG estimators constitute a wide class of estimators that utilise auxiliary information by modelling. In the literature, the term GREG estimator sometimes refers to an estimator that by definition has a linear, fixed effects assisting model. In this study, such restrictions are not imposed on the model: the assisting model may be any statistical model. Moreover, the starting point when using GREG estimators is that unit level auxiliary information is available and unit level predictions are made using explicit modelling. For some GREG estimators, however, it can be shown that the estimator can also be constructed using only aggregate auxiliary data.

Models have a key role in GREG estimation. Therefore, after defining the GREG estimator in Chapter 4.1, we discuss the idea of modelling in the model-assisted framework in Chapter 4.2. In this chapter, we also discuss models appropriate for class frequencies. In Chapter 4.3, standard variance estimation for GREG estimators is considered. In Chapter 4.4, the classic GREG estimator with a linear fixed effects model (GREG-lin) is discussed more thoroughly. As is well known, the GREG-lin estimator is also a calibration estimator (e.g. Särndal *et al.* 1992). The classic GREG-lin estimator is also the reference for other GREG estimators, such as logistic-type GREG estimators (L-GREG estimators). L-GREG estimators, which are defined as GREG estimators whose model is appropriate for binary or polytomous variables, are discussed in Chapter 4.5.

The reasons for discussing first the classic GREG-lin and only after that L-GREG estimators are as follows: Historically, GREG-lin precedes the more general family of GREG estimators, GREG-lin is widely used and its properties are reasonably well-known. The family of GREG estimators, on the contrary, includes a great number of more recent estimators (including L-GREG estimators) that have not yet been much used in practice, and knowledge about their properties is still limited.

4.1 The family of GREG estimators

Let us take a look at our data matrix **A** after the sample is observed:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_s \\ \mathbf{A}_r \end{bmatrix} = \begin{bmatrix} \mathbf{z}_s & \mathbf{X}_s & \Delta_s & \mathbf{w}_s & \mathbf{Y}_s \\ \mathbf{z}_r & \mathbf{X}_r & \Delta_r & \mathbf{0} & \mathbf{Y}_r \end{bmatrix}. \tag{4.1}$$

In GREG estimation, we construct a statistical assisting model that connects the study variables $\mathbf{Y}_s$ and auxiliary information $\mathbf{X}_s$. The parameters of the assisting model are estimated from the sample using the sampling weights and then the values of the study variable are predicted using the model. After constructing predictions $\hat{y}_{ij}$ for the whole population and prediction errors e_{ij} for the sampled units, the data matrix can be written as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_s \\ \mathbf{A}_r \end{bmatrix} = \begin{bmatrix} \mathbf{z}_s & \mathbf{X}_s & \Delta_s & \mathbf{w}_s & \mathbf{Y}_s & \hat{\mathbf{Y}}_s & \mathbf{e}_s \\ \mathbf{z}_r & \mathbf{X}_r & \Delta_r & \mathbf{0} & \mathbf{Y}_r & \hat{\mathbf{Y}}_r & .. \end{bmatrix}. \tag{4.2}$$

It is essential that predictions can be made for non-sampled units. This can be seen as a kind of mass imputation where the mechanism that generates the nonresponse is known (the mechanism being the sampling design). But the predictions made for the sampled units are also important: they are the basis for the bias-correction term that ensures approximate unbiasedness of a GREG estimator in a standard situation.

The matrix (4.2) includes all the information needed for the GREG estimator, which is defined as

$$\underline{\hat{T}}_{j,GREG} = \sum_{i \in U} \underline{\hat{y}}_{ij} + \sum_{i \in U} \underline{w}_i \underline{e}_{ij} \ , \ \underline{e}_{ij} = y_{ij} - \underline{\hat{y}}_{ij} \, . \tag{4.3}$$

In (4.3), $\sum_{i \in U} \underline{\hat{y}}_{ij}$ is called the synthetic part and $\sum_{i \in U} \underline{w}_i \underline{e}_{ij}$ the bias-correction part of the estimator.

For domain total $T_j^{(d)}$, the GREG estimator is

$$\underline{\hat{T}}_{j,GREG}^{(d)} = \sum_{i \in U} \underline{\hat{y}}_{ij}^{(d)} + \sum_{i \in U} \underline{w}_i \underline{e}_{ij}^{(d)} \ , \ \underline{e}_{ij}^{(d)} = y_{ij}^{(d)} - \underline{\hat{y}}_{ij}^{(d)} \, . \tag{4.4}$$

Note that by dividing (4.4) by domain size one would get an estimator for proportion (domain size is assumed to be known). The domain estimator (4.4) is almost identical with (4.3) – all that is different is the fact that the study variable is changed to $y_{ij}^{(d)}$. Thus the domain estimator (4.4) also has a synthetic and bias-correction part. Depending on the model, the domain estimator is either direct or indirect. By weighting the bias-correction term $\sum_U \underline{w}_i \underline{e}_{ij}^{(d)}$ by parameter $\underline{\hat{\gamma}}^{(d)} \in [0,1]$, one gets a wider class of estimators. If

$$\begin{cases} \hat{\underline{\gamma}}^{(d)} = 0, \text{ the estimator is pseudo-synthetic,} \\ \hat{\underline{\gamma}}^{(d)} = 1, \text{ the estimator is model-assisted, and} \\ 0 < \hat{\underline{\gamma}}^{(d)} < 1, \text{ the estimator is a composite estimator.} \end{cases} \quad (4.5)$$

We call the estimator with $\hat{\underline{\gamma}}^{(d)} = 0$ pseudo-synthetic, since sampling weights have been used in the estimation of the assisting model. A purely synthetic estimator would not use sampling weights at all, not even in the estimation of the model. If a composite estimator is used, then $\hat{\underline{\gamma}}^{(d)}$ depends on the model's error structure and the sample size in domain (Särndal *et al.* 1992, Lehtonen *et al.* 2003).

In the following, we consider only the case where $\hat{\underline{\gamma}}^{(d)} = 1$, that is, we restrict the study to the model-assisted GREG estimator (4.4). Composite and pseudo-synthetic estimators are not considered since if $\hat{\underline{\gamma}}^{(d)} \neq 1$, the estimator is neither approximately unbiased nor additive.

The additivity property of (4.4) can be seen from the following: Let domains $U^{(d)}, d = 1, 2, \ldots, k \leq D$ be non-overlapping so that $U^{(t)} = \bigcup_{d=1}^{k} U^{(d)}$. Then

$$\begin{aligned} \hat{\underline{T}}^{(t)}_{j,GREG} &= \sum_{i \in U} \hat{\underline{y}}^{(t)}_{ij} + \sum_{i \in U} \underline{w}_i \underline{e}^{(t)}_{ij} \\ &= \left(\sum_{i \in U} \hat{\underline{y}}^{(1)}_{ij} + \sum_{i \in U} \hat{\underline{y}}^{(2)}_{ij} + \ldots + \sum_{i \in U} \hat{\underline{y}}^{(k)}_{ij} \right) + \left(\sum_{i \in U} \underline{w}_i \underline{e}^{(1)}_{ij} + \sum_{i \in U} \underline{w}_i \underline{e}^{(2)}_{ij} + \ldots + \sum_{i \in U} \underline{w}_i \underline{e}^{(k)}_{ij} \right) \quad (4.6) \\ &= \left(\sum_{i \in U} \hat{\underline{y}}^{(1)}_{ij} + \sum_{i \in U} \underline{w}_i \underline{e}^{(1)}_{ij} \right) + \left(\sum_{i \in U} \hat{\underline{y}}^{(2)}_{ij} + \sum_{i \in U} \underline{w}_i \underline{e}^{(2)}_{ij} \right) + \ldots + \left(\sum_{i \in U} \hat{\underline{y}}^{(k)}_{ij} + \sum_{i \in U} \underline{w}_i \underline{e}^{(k)}_{ij} \right) \\ &= \sum_{d=1}^{k} \hat{\underline{T}}^{(t)}_{j,GREG} . \end{aligned}$$

The variance of the GREG estimator depends on the assisting model: the better the assisting model (in terms of small residuals), the smaller the variance (more on variance in Chapter 4.3). Thus, careful modelling has a key role in GREG estimation. In the next chapter, we discuss the role of statistical models in GREG estimation.

4.2 Modelling in the model-assisted framework

In model-assisted estimation, the modelling can be divided into the following two steps: i) the model specification step, which includes the choice of the functional form, parameterisation and decisions regarding fixed and random effects, and ii) the estimation step, which includes decisions concerning the specific estimation method and whether to use weights. In this section we consider some of these aspects and discuss how they relate to modelling in mainstream statistics.

In the model-assisted framework, study variables are considered fixed. Therefore the question sometimes posed in mainstream statistical modelling, namely whether the chosen model is the correct (or true) model that has generated the observations, is irrelevant: in the model-assisted framework, no correct models exist. This does not mean that all models are equally good, nor does it mean that there are no good models.

In mainstream statistical modelling, the correct model may exist, but it cannot be reached – all models are wrong (Box 1979, 202). An exception to this is of course a situation where the data are artificially generated. But in situations of practical interest, a "wrong" model can still be useful (Box 1979, 202). A useful or good model describes the variation in observed values reasonably well with a reasonably simple model.

In the model-assisted framework, the assumption that the study variable is non-random is loosened on pragmatic grounds: to justify the model-fitting procedure, the finite population scatter

$$\left\{(y_{ij}, x_{i1}, x_{i2}, \ldots, x_{iM}) : i \in U\right\} \tag{4.7}$$

is taken *as if* it had been generated by some model ξ (Särndal *et al.* 1992, 226). In this study, instead of using the term model to describe ξ, we use the term *population generating process*. This is done in order to keep clear the distinction between the assisting models and the hypothetical population generating process.

The study variables are considered as if they were realisations of random variables $\underline{y}_{ij}$ and, under the population generating process ξ, expectation of $\underline{y}_{ij}$ is

$$E_{\xi}\left(\underline{y}_{ij}\right) = f(\mathbf{x}_i; \boldsymbol{\beta}) \quad \forall i \in U, \tag{4.8}$$

where f is some function defined by ξ. The expectation of $\underline{y}_{ij}$ depends on what we imagine the population generating process ξ to be.

If the population scatter (4.7) could be observed, the parameters $\boldsymbol{\beta}$ could be estimated using information from the whole population. These *census fit parameters*, which in practice cannot be calculated, are denoted by $\mathbf{B}$. An estimate that can be calculated is based on the observed sample scatter

$$\left\{(y_{ij}, x_{i1}, x_{i2}, \ldots, x_{iM}) : i \in s\right\}. \tag{4.9}$$

The parameters estimated using information (4.9) are denoted by $\hat{\mathbf{B}}$; $\hat{\mathbf{B}}$ estimates the census fit parameter $\mathbf{B}$ which, if it could be calculated, would estimate $\boldsymbol{\beta}$. So there are three parameter levels: on the level of the population generating process ξ, on the (fixed and finite) population level and on the sample level.

It should be emphasised that the main interest here is not in the model or in its parameters, nor in the interpretation of the relationship between variables. The model only assists when estimating the finite population parameters. Therefore the model is called an assisting model. Moreover, the values y_{ij} are taken *only as if* they were realisations of ξ, the assumption that the observations *are* generated by ξ is not necessary.

Despite the philosophical differences between modelling in a model-assisted framework and mainstream statistical modelling, the characteristics of a useful or good model are quite the same. In a model-assisted framework, estimators are approximately unbiased in standard situations; hence the primary aim of modelling is to decrease the variance. This is often achieved when the model fits to the sample scatter well and is parsimonious in the sense that it is as simple as possible, still fitting to the sample scatter. (There is, of course, a trade-off between simplicity and goodness of fit.) Such a model often produces good predictions for non-sampled units and good predictions reduce variance.

When selecting the assisting model, one has to choose the functional form of the model, the effects, whether the effects are random or fixed and finally decide the estimation method. The functional form of the model is often determined by the nature of the response variable. For example, if the response is continuous, a linear model may be adequate, if the response is binary, a logistic model is often recommended and so on. Whether an effect should be treated as fixed or random is discussed in McCulloch and Searle (2001, 16–19). Briefly, if it is reasonable to think that the levels of effects come from a probability distribution, it is reasonable to treat them as random. Also, if the observations are clustered and the number of clusters is large, modelling cluster-specific effects as random terms is reasonable.

The selection of effects also depends on a number of things. In survey sampling, the set of possible effects is often limited. For example, the auxiliary information available at official statistical agencies may contain only a small number of useful variables. But the problem of effect selection still remains.

When selecting effects, one has to look for effects that have the greatest predictive power and guard against overfitting. But the effect, or power, of different effects depends on other effects in the model and also on the functional form of the model; thus it is rarely simple to choose the effects. The strategy of effect selection always depends on the subject matter, and sometimes there is a tome of theory guiding it, sometimes not.

When theoretical knowledge on effects is scarce, one should search for statistically powerful effects. In this process, automatic model selection methods like stepwise selection may be useful especially if the number of possible effects is large. In the stepwise model selection method, effects are included in the model and removed from the model according to certain criteria, often affiliated with an increase or decrease in the goodness of fit (GOF). GOF, in turn, may be measured in numerous ways. In general, observations can be partitioned as

$$\textit{observation} = \textit{model prediction} + \textit{model error}\ , \tag{4.10}$$

and GOF statistics measure the variation of error unexplained by the model. For different model types there are different GOF statistics. The most widely used is R^2, which is appropriate especially for linear models. For logistic models, several other statistics have been proposed; Nagelkerke's maximum likelihood based GOF is recommended for general use (Collett 2002, 90).

But none of the GOF statistics take into account the number of effects in the model. Therefore, if only GOF statistics are used, the modeller will end up with

a full model which explains nothing. When effect selection is done according to Bayesian or Akaike's information criteria (BIC and AIC), the result is different since they have an explicit cost function for a number of effects.

There is still one thing to be considered when using models in GREG estimation: should one use direct or indirect estimation? That is, should the assisting model be fitted on the domain level, or on a more general level? In many cases, in particular when domains are small, it may reasonable to use indirect estimation. In indirect estimation, it is in principle possible to produce estimates for any domain, even for domains whose sample size is zero (then the GREG estimator reduces to the pseudo-synthetic estimator).

4.2.1 Generalised linear models

In this study we consider models from the class of Generalised Linear Models (GLM) as assisting models in GREG estimation. This excludes many models, such as generalised linear mixed models, generalised additive models and nonparametric models, but still the class of GLMs is wide enough to provide appropriate models for most of the situations encountered in survey sampling.

GLM allows both linear and nonlinear model forms. The GLM consists of three parts as follows:

1) The random component: Random variables $\underline{y}_i$ are independent with mean μ_i and distribution

$$f_{\underline{y}_i}(y;\theta_i,\phi) = \exp\left\{\frac{y\theta_i - b(\theta_i)}{\phi} w_i + c(y,\phi,w_i)\right\}, \qquad (4.11)$$

where θ_i and ϕ are unknown parameters, b and c are known functions and w_i is a known weight.

2) The systematic part: Covariates $\mathbf{x}_i$ and a fixed parameter vector $\boldsymbol{\beta}$ form a *linear predictor* $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$.

3) Invertible link function g that links the linear predictor and expectation of $\underline{y}_i : g(\mu_i) = \eta_i$.

The distribution (4.11) is the exponential family of distributions. Special cases of this distribution are normal, gamma, geometric, Poisson and binomial distributions. The expectation and variance of a random variable $\underline{y}_i$ from the exponential family of distributions are

$$E\left(\underline{y}_i\right) = \mu_i = \frac{\partial b(\theta_i)}{\partial \theta_i} \quad \text{and} \quad V\left(\underline{y}_i\right) = \sigma_i^2 = \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2}\frac{\phi}{w_i}. \qquad (4.12)$$

The second component of GLM, linear predictor $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$, determines the effects which affect the distribution. The third part, link function $g(\mu_i) = \eta_i$, is the mechanism through which the linear predictor affects the distribution of $\underline{y}_i$. For example, the standard linear model is obtained by setting $g(\mu_i) = \mu_i$ and the

logistic model by setting $g(\mu_i) = \text{logit}(\mu_i) = \log\left[\mu_i / (1-\mu_i)\right]$. The link function is called a canonical link if

$$\theta_i = g(\mu_i). \tag{4.13}$$

The model equation for a generalised linear model is

$$\underline{y}_i = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta}) + \underline{\varepsilon}_i = \mu_i + \underline{\varepsilon}_i, \tag{4.14}$$

and the model specification is complete when the link function and distribution of $\underline{\varepsilon}_i$ are specified.

The parameters $\boldsymbol{\beta}$ need to be estimated, and the choice of the method is not trivial. Several methods exist; most important of them are least squares and likelihood-based methods. But a thorough treatment of the science of model estimation is beyond the scope of this study. Therefore it will suffice to describe the specific estimation methods used in this study in Chapters 4.4 (for linear models) and 4.5 (for logistic-type models).

4.2.2 Logistic-type models

What is an appropriate model for class frequencies depends on many things, such as the nature of the classification. But when the response is binary or polytomous, it is fair to require that the link function g maps the unit interval $(0,1)$ to the interval $(-\infty,+\infty)$. The reason for this is that the expectation of the response is probability:

$$E\left(\underline{y}_{ij}\right) = \sum_{k=0}^{1} k \cdot P\left(\underline{y}_{ij} = k\right) = P\left(\underline{y}_{ij} = 1\right). \tag{4.15}$$

The restriction that g maps the unit interval to $(-\infty,+\infty)$ thus guarantees that the predictions can be interpreted as probabilities. In this study, models whose link fulfils this condition are called *logistic-type models*.

The number of logistic-type links is infinite. We derive here the canonical link and present three other links that are commonly used. Let $\underline{y} \sim Bin(n, p)$, that is, $\underline{y}$ is binomially distributed with parameters n and p. Consider the distribution of the proportion $\frac{1}{n}\underline{y} \equiv \underline{\bar{y}}$. The distribution of the proportion is essentially the same as the distribution of $\underline{y}$, since $\underline{\bar{y}}$ is $\underline{y}$ divided by a constant. Thus the distribution of $\underline{\bar{y}}$ (the binomial distribution) can be written as

$$\begin{aligned} f_{\underline{\bar{y}}}(\bar{y}, p) &= P(\underline{\bar{y}} = \bar{y}) = \binom{n}{n\bar{y}} p^{n\bar{y}} (1-p)^{n-n\bar{y}} \\ &= \exp\left\{\log\left[\binom{n}{n\bar{y}} p^{n\bar{y}} (1-p)^{n-n\bar{y}}\right]\right\} \end{aligned}$$

$$
\begin{aligned}
&= \exp\left\{ n\bar{y} \underbrace{\log\left(\frac{p}{1-p}\right)}_{\equiv\theta} + n\log(1-p) + \underbrace{\log\binom{n}{n\bar{y}}}_{\equiv c(\bar{y},\phi,w)} \right\} \\
&= \exp\left\{ n\bar{y}\theta - n\underbrace{\log[1+\exp(\theta)]}_{\equiv b(\theta} + c(\bar{y},\phi,w) \right\} \\
&= \exp\left\{ [\bar{y}\theta - b(\theta)]n + c(\bar{y},\phi,w) \right\}.
\end{aligned}
\tag{4.16}
$$

The distribution (4.16) belongs to the family of exponential distributions. From the general form, this distribution is obtained by choosing

$$
\begin{aligned}
&\phi = 1, \quad w_i = n, \quad b(\theta) = \log[1+\exp(\theta)], \\
&\theta = \log\left(\frac{p}{1-p}\right), \text{ and } \quad c(\bar{y},\phi,w) = \log\binom{n}{n\bar{y}}.
\end{aligned}
\tag{4.17}
$$

Previously, it was stated that the canonical link is the link that satisfies $\theta_i = g(\mu_i)$. When binomial distribution is written as a member of the exponential family of distributions, $\theta = \log[p/(1-p)]$. Therefore the link

$$
g(p) = \log\left(\frac{p}{1-p}\right). \tag{4.18}
$$

is the canonical link for binary and binomial random variables. The link (4.18) is called the logit link.

Other three commonly used links are probit, log-log and complementary log-log (cll). The definitions of these three links and the canonical logit link are as follows:

1) logit: $\eta = g(p) = \log\left(\frac{p}{1-p}\right),\ p \in (0,1),$ (4.19)

2) probit: $\eta = g(p) = \Phi^{-1}(p),\ p \in (0,1),$ (4.20)

where Φ is the Normal cumulative distribution function,

3) log-log: $\eta = g(p) = -log[-log(p)],\ p \in (0,1),$ and (4.21)

4) cll: $\eta = g(p) = log[-log(1-p)],\ p \in (0,1).$ (4.22)

Figure 4.1 shows how the inverse of these links transforms the linear predictor η to a probability.

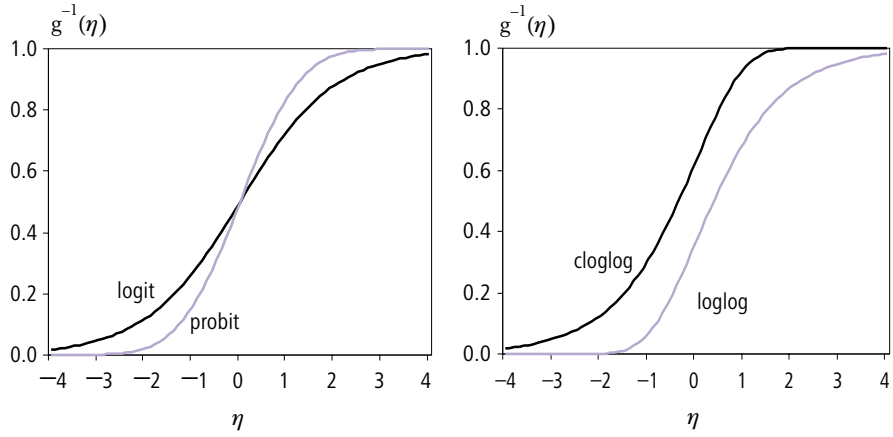


Figure 4.1 Graphs of $p = g^{-1}(\eta), \eta \in [-4, 4]$ where g is logit, probit, loglog and cll.

The choice of a link is rarely straightforward. When a link is chosen among the four common links (logit, probit, log-log, cll), the choice depends mainly on how the probability p is assumed to depend on covariates. Logit and probit links are symmetric about $p = 0.5$ and their predictions are often indistinguishable, but from the computational viewpoint, the logistic transformation is more convenient. Log-log and complementary log-log links, in turn, are not symmetric about $p = 0.5$, therefore their use should be limited to situations where it is appropriate to deal with success probabilities in an asymmetric manner (Collett 2002, 57). However, according to Aldrich and Nelson (1985) and Armitage *et al.* (2002), in most of the practical situations, all commonly used logistic-type links yield very similar predictions.

In the context of finite population inference, there have not been many studies comparing different logistic-type link functions. Firth and Bennett (1998) mention that the logit link has a special advantage over other links because the maximum likelihood fit of a logistic regression can be used to yield a design consistent estimator for total. Laaksonen (2006), on the other hand, has found that asymmetric links might be better than logit or probit when weighting for missingness. In Chapter 5 we conduct a Monte Carlo experiment in order to study the properties of GREG estimators with different logistic-type link functions.

4.3 Variance estimation for GREG

In this chapter, standard variance estimation for the model-assisted population GREG (4.3) and the domain GREG (4.4) are considered. In Chapter 4.3.1, the standard variance approximation and an estimator for the approximation are presented in the case of a general one-stage fixed-size design. In Chapters 4.3.2 and 4.3.3, these approximations and estimators are showed in the case of a SRSWOR design and in Chapter 4.3.4, we consider them in the case of a πPS design.

4.3.1 Standard variance approximation and estimator for GREG

The GREG estimator with an assisting generalised linear model (which is based on the hypothetical population generating process ξ), is

$$\underline{\hat{T}}_{j,GREG} = \sum_{i \in U} \underline{\hat{y}}_{ij} + \sum_{i \in U} \underline{w}_i \underline{e}_{ij} \,. \tag{4.23}$$

The model has the following features:

For every $i \in U$

1) Values y_{ij} are realisations of random variables $\underline{y}_{ij}$,
2) $E_\xi\left(\underline{y}_{ij}\right) = \underline{\mu}_{ij}$, $g(\underline{\mu}_{ij}) = \underline{\eta}_{ij}$,
3) $\underline{\eta}_{ij} = \mathbf{x}_i'\boldsymbol{\beta}$,
4) $V_\xi\left(\underline{y}_{ij}\right) = \sigma^2$, and
5) $Cov_\xi\left(\underline{y}_{ij}, \underline{y}_{kj}\right) = 0$, $i \neq k$.

If values y_{ij} were available for the whole population, a census fit estimate $\mathbf{B}$ for parameters $\boldsymbol{\beta}$ could be obtained. Census fit residuals are defined in terms of census fit predictions and population values as

$$E_{ij} = y_{ij} - \tilde{y}_{ij}, \text{ where} \tag{4.24}$$

$$\tilde{y}_{ij} = g^{-1}(\mathbf{x}_i'\mathbf{B}) \,. \tag{4.25}$$

A sample is observed and a sample fit estimate $\hat{\mathbf{B}}$ is obtained using an estimator that incorporates sampling weights. This estimator is denoted by $\underline{\hat{\mathbf{B}}}$ and predictions by $\underline{\hat{y}}_{ij}$:

$$\underline{\hat{y}}_{ij} = g^{-1}(\mathbf{x}_i'\underline{\hat{\mathbf{B}}}) \,. \tag{4.26}$$

Next, the estimator (4.23) is linearised in order to obtain approximate variance. Steps (5.27)–(5.31) in the linearisation are based on Knottnerus (2003) and Estevao and Särndal (2006). It should be noted that in the case of a generalised linear model, the derivation of the approximation is not based on the Taylor series expansion but on an ad hoc linearisation which we call automated linearisation (Estevao and Särndal (2006) use the term automated linearisation to derive approximate variances for a wide class of calibration estimators). However, if the assisting model is a linear fixed effects model, the approximation corresponds exactly to the approximation resulting from the first-order Taylor series linearisation. The reason for the use of automated linearisation instead of utilising the well-known Taylor series linearisation is that the automated linearisation provides a general variance approximation for all GREG estimators, irrespective

of the functional form of the assisting model and irrespective of whether one uses fixed or random effects models. If the Taylor series linearisation was used, every functional form (linear, logistic, probit, …) would require its own derivation. Thus automated linearisation simplifies things markedly. How well does the automated linearisation work in practice? This is studied in Chapters 5–7.

To obtain the automated linearisation, first write the GREG estimator as

$$\begin{aligned}\hat{\underline{T}}_{j,GREG} &= \sum_{i\in U}\underline{\hat{y}}_{ij} + \sum_{i\in U}\underline{w}_i\left(y_{ij} - \underline{\hat{y}}_{ij}\right)\\ &= \sum_{i\in U}\underline{w}_i y_{ij} + \underbrace{\left[\sum_{i\in U}\underline{\hat{y}}_{ij} - \sum_{i\in U}\underline{w}_i\underline{\hat{y}}_{ij}\right]}_{=h(\underline{\mathbf{I}})}.\end{aligned} \tag{4.27}$$

By adding and subtracting $\sum_{i\in U}\tilde{y}_{ij}$ and $\sum_{i\in U}\underline{w}_i\tilde{y}_{ij}$ from $h(\underline{\mathbf{I}})$ and rearranging, we get

$$\begin{aligned}h(\underline{\mathbf{I}}) &= \sum_{i\in U}\tilde{y}_{ij} + \sum_{i\in U}\left(\underline{\hat{y}}_{ij} - \tilde{y}_{ij}\right) - \sum_{i\in U}\underline{w}_i\tilde{y}_{ij} - \sum_{i\in U}\underline{w}_i\left(\underline{\hat{y}}_{ij} - \tilde{y}_{ij}\right)\\ &= \left[\sum_{i\in U}\tilde{y}_{ij} - \sum_{i\in U}\underline{w}_i\tilde{y}_{ij}\right] + \underbrace{\left[\sum_{i\in U}\left(\underline{\hat{y}}_{ij} - \tilde{y}_{ij}\right) - \sum_{i\in U}\underline{w}_i\left(\underline{\hat{y}}_{ij} - \tilde{y}_{ij}\right)\right]}_{=t(\underline{\mathbf{I}})}.\end{aligned} \tag{4.28}$$

By combining (4.27) and (4.28) and rearranging, we get a linear, HT-type estimator plus constant and $t(\underline{\mathbf{I}})$:

$$\begin{aligned}\hat{\underline{T}}_{j,GREG} &= \sum_{i\in U}\underline{w}_i y_{ij} + \left[\sum_{i\in U}\tilde{y}_{ij} - \sum_{i\in U}\underline{w}_i\tilde{y}_{ij}\right] + t(\underline{\mathbf{I}})\\ &= \sum_{i\in U}\underline{w}_i\left(y_{ij} - \tilde{y}_{ij}\right) + \sum_{i\in U}\tilde{y}_{ij} + t(\underline{\mathbf{I}})\\ &= \sum_{i\in U}\underline{w}_i E_{ij} + \sum_{i\in U}\tilde{y}_{ij} + t(\underline{\mathbf{I}}).\end{aligned} \tag{4.29}$$

In the last row of (4.29), the first term $\sum_U \underline{w}_i E_{ij}$ is a linear HT-type estimator whose variance can be estimated. The second term $\sum_U \tilde{y}_{ij}$ is constant with respect to sampling. In the case of a linear fixed effects model, the third term $t(\underline{\mathbf{I}})$ contributes only little to variance (for example, Estevao and Särndal 2003).

In the case of other GLM, we do not know what the contribution of $t(\underline{\mathbf{I}})$ to variance is. However, $t(\underline{\mathbf{I}})$ consists of only residuals $\underline{\tilde{e}}_{ij} = \underline{\hat{y}}_{ij} - \tilde{y}_{ij}$ so the magnitude of this term is likely to be small. Second, $t(\underline{\mathbf{I}})$ can be written as

$$t(\underline{\mathbf{I}}) = \sum_{i\in U}\left(\underline{\hat{y}}_{ij} - \underline{w}_i\underline{\hat{y}}_{ij}\right) - \sum_{i\in U}\left(\tilde{y}_{ij} - \underline{w}_i\tilde{y}_{ij}\right), \tag{4.30}$$

where the last term on the right-hand side estimates zero unbiasedly (since $\sum_U \underline{w}_i \tilde{y}_{ij}$ is an HT estimator for the fixed quantity $\sum_U \tilde{y}_{ij}$) and the first term on the right-hand side estimates zero unbiasedly if weights $\underline{w}_i$ and predictions $\underline{\hat{y}}_{ij}$ are independent. Third, as both the population and sample size grow, sample fit predictions $\underline{\hat{y}}_{ij}$ tend to census fit predictions $\tilde{y}_{ij}$ and residuals $\underline{\tilde{e}}_{ij}$ tend to zero. These three facts justify the assumption that the contribution of $t(\underline{\mathbf{I}})$ to variance is negligible. Under this assumption we can approximate the variance as

$$\begin{aligned} V\left(\hat{\underline{T}}_{j,GREG}\right) &\approx V\left(\sum_{I\in U} \underline{w}_i E_{ij}\right) \\ &\equiv V_A\left(\hat{\underline{T}}_{j,GREG}\right) = \sum_{i\in U}\sum_{k\in U} \frac{\Delta_{ik}}{E(\underline{I}_i)E(\underline{I}_k)} E_{ij}E_{kj} \ . \end{aligned} \tag{4.31}$$

We call the approximation (4.31) the Standard variance approximation for GREG estimators. The only term that contributes to the Standard variance approximation is $\sum_U \underline{w}_i E_{ij}$, an HT-type linear estimator of census fit residuals, and it is easy to see that the better the fit of the model, the smaller the (approximate) variance.

This Standard variance approximation (4.31) resembles the famous Sen-Yates-Grundy variance formula

$$\begin{aligned} V_{SYG}\left(\hat{\underline{T}}_{j,GREG}\right) &= -\frac{1}{2}\sum_{i\in U}\sum_{k\in U}\Delta_{ij}\left(\frac{y_{ij}}{E(\underline{I}_i)} - \frac{y_{kj}}{E(\underline{I}_k)}\right)^2 \\ &= \sum_{i\in U}\sum_{k\in U}\frac{\Delta_{ik}}{E(\underline{I}_i)E(\underline{I}_k)} y_{ij}y_{kj} \end{aligned} \tag{4.32}$$

where the last equality holds since the design is fixed-size. The only difference between (4.31) and (4.32) is that in the former, the approximation consists of census fit residuals, not of the values of the study variable. This way of taking the model into account in variance estimation is commonly used; see, for example, Thompson (1992), Lehtonen and Veijanen (1998a, 1998b), Lohr (1999), Axelson (2000), Valliant (2002), and Lehtonen and Pahkinen (2004).

If residuals E_{ij} could be observed, an unbiased estimator for the Standard variance approximation would be

$$\hat{V}_A\left(\hat{\underline{T}}_{j,GREG}\right) = \sum_{i\in U}\sum_{k\in U}\frac{\Delta_{ik}}{E(\underline{I}_i\underline{I}_k)} \underline{w}_i E_{ij} \underline{w}_k E_{kj} \ . \tag{4.33}$$

However, the calculation of residuals E_{ij} requires knowledge about census fit parameters $\mathbf{B}$; thus, these residuals are unobservable. Replacing census fit residuals by sample fit residuals $\underline{e}_{ij} = y_{ij} - \underline{\hat{y}}_{ij}$ we obtain the Standard variance estimator

$$\hat{V}_S\left(\hat{\underline{T}}_{j,GREG}\right) = \sum_{i\in U}\sum_{k\in U}\frac{\Delta_{ik}}{E(\underline{I}_i\underline{I}_k)} \underline{w}_i \underline{e}_{ij} \underline{w}_k \underline{e}_{kj}. \tag{4.34}$$

For model-assisted GREG for domains, the same arguments that were used above can be used to derive the approximate variance and its estimator:

$$V_A\left(\hat{\underline{T}}_{j,GREG}^{(d)}\right) = \sum_{i\in U}\sum_{k\in U}\frac{\Delta_{ik}}{E(\underline{I}_i)E(\underline{I}_k)}E_{ij}^{(d)}E_{kj}^{(d)}, \quad E_{ij}^{(d)} = y_{ij}^{(d)} - \tilde{y}_{ij}^{(d)}, \qquad (4.35)$$

$$\hat{V}_S\left(\hat{\underline{T}}_{j,GREG}^{(d)}\right) = \sum_{i\in U}\sum_{k\in U}\frac{\Delta_{ik}}{E(\underline{I}_i\underline{I}_k)}\underline{w}_i\underline{e}_{ij}^{(d)}\underline{w}_k\underline{e}_{kj}^{(d)}, \quad \underline{e}_{ij}^{(d)} = y_{ij}^{(d)} - \hat{\underline{y}}_{ij}^{(d)}. \qquad (4.36)$$

Since we treat $y_j^{(d)}$ as any other study variable, the double sum in (4.35) and (4.36) is over the whole population, not over the domain (recall that if $i \notin U^{(d)}$ then $\delta_i^{(d)} = 0$ and both $\hat{\underline{y}}_{ij}^{(d)} = \hat{\underline{y}}_{ij}\delta_i^{(d)} = 0$ and $\underline{e}_{ij}^{(d)} = \underline{e}_{ij}\delta_i^{(d)} = 0$). From this it also follows that one does not have to account for the possible random sample size in the domain, as long as the overall sample size is fixed.

In the variance approximations (4.31) and (4.35), the assumption that $t(\underline{\mathbf{I}})$ is negligible with respect to the variance was made. The Standard estimator for the approximation is (4.36). For a linear fixed effects assisting model, the Standard variance estimator is sometimes adjusted for the loss of degrees of freedom that comes within the estimation of the assisting model. In the context of calibration, where *g*-weights are explicitly calculated, it is also often suggested to multiply the residuals by *g*-weights (Särndal *et al.* 1989, 1992, Hidiroglou and Särndal 1998).

These adjustments, correction for the loss of degrees of freedom and *g*-weighting of residuals, are discussed more in Chapter 4.4.2. What kind of adjustment is needed (if any) for logistic-type models is not known. In Monte Carlo experiments in Chapter 5, the approximations and estimators presented above are examined for a logistic-type assisting model.

4.3.2 Variance of population GREG under the SRSWOR design

Next we consider the Standard approximate variance and variance estimator for the population GREG estimator in the case of the SRSWOR design. The approximate variance (4.31) can be written as

$$V_A\left(\hat{\underline{T}}_{j,GREG}\right) = -\frac{1}{2}\sum_{i\in U}\sum_{k\in U}\Delta_{ij}\left(\frac{E_{ij}}{E(\underline{I}_i)} - \frac{E_{kj}}{E(\underline{I}_k)}\right)^2, \text{ where} \qquad (4.37)$$

$$E_{ij} = y_{ij} - \tilde{y}_{ij} = y_{ij} - g^{-1}(\mathbf{x}_i'\mathbf{B}), \qquad (4.38)$$

and the parameter vector $\mathbf{B}$ is obtained from the census fit. Under SRSWOR,

$$E(\underline{I}_i) = f, \; E(\underline{I}_i\underline{I}_k) = f\frac{n-1}{N-1} \text{ and } \Delta_{ij} = \begin{cases} f(1-f), & i = j \\ -f\dfrac{1-f}{N-1}, & i \neq j. \end{cases} \qquad (4.39)$$

Using these properties of SRSWOR, the expression (4.37) simplifies so that no double sums are needed. First, the Standard variance approximation (4.37) can

be written as

$$V_A\left(\hat{\underline{T}}_{j,GREG}\right) = \frac{1}{2}\sum_{i\in U}\sum_{k\in U} f\frac{1-f}{N-1}\left(\frac{E_{ij}}{f}-\frac{E_{kj}}{f}\right)^2 = \frac{1-f}{2f(N-1)}\sum_{i\in U}\sum_{k\in U}\left(E_{ij}-E_{kj}\right)^2. \tag{4.40}$$

In (4.40), the distinction in Δ_{ij} when $i = j$ and $i \neq j$ can be ignored, since if $i = j$, $E_{ij} - E_{kj} = 0$. By adding and subtracting $\overline{E}_j = \frac{1}{N}\sum_{i\in U} E_{ij}$, the double sum in (4.40) can be written as

$$\begin{aligned}
\sum_{i\in U}\sum_{k\in U}\left(E_{ij}-E_{kj}\right)^2 &= \sum_{i\in U}\sum_{k\in U}\left[\left(E_{ij}-\overline{E}_j\right)-\left(E_{kj}-\overline{E}_j\right)\right]^2 \\
&= 2\sum_{i\in U}\sum_{k\in U}\left(E_{ij}-\overline{E}_j\right)^2 - 2\sum_{i\in U}\sum_{k\in U}\left(E_{ij}-\overline{E}_j\right)\left(E_{kj}-\overline{E}_j\right) \\
&= 2N\sum_{i\in U}\left(E_{ij}-\overline{E}_j\right)^2 - 2\left[\left(E_{1j}-\overline{E}_j\right)\underbrace{\sum_{k\in U}\left(E_{kj}-\overline{E}_j\right)}_{=0} + \ldots + \left(E_{Nj}-\overline{E}_j\right)\underbrace{\sum_{k\in U}\left(E_{kj}-\overline{E}_j\right)}_{=0}\right] \\
&= 2N\sum_{i\in U}\left(E_{ij}-\overline{E}_j\right)^2.
\end{aligned} \tag{4.41}$$

Inserting (4.41) into (4.40), the approximate variance takes the form

$$\begin{aligned}
V_A\left(\hat{\underline{T}}_{j,GREG}\right) &= \frac{N(1-f)}{f(N-1)}\sum_{i\in U}\left(E_{ij}-\overline{E}_j\right)^2 \\
&= \frac{N^2(1-f)}{n}\frac{\sum_{i\in U}\left(E_{ij}-\overline{E}_j\right)^2}{(N-1)} \\
&= \frac{N^2(1-f)}{n}S^2_{E_j}.
\end{aligned} \tag{4.42}$$

The Standard variance estimator and a Standard variance estimate for (4.42) are obtained by replacing the census fit variance $S^2_{E_j}$ by its estimator $\underline{S}^2_{e_j}$:

$$\hat{V}_S\left(\hat{\underline{T}}_{j,GREG}\right) = \frac{N^2(1-f)}{n}\underline{S}^2_{e_j},\ \underline{S}^2_{e_j} = \frac{1}{n-1}\sum_{i\in U}\underline{I}_i\left(\underline{e}_{ij}-\overline{\underline{e}}_j\right)^2,\ \overline{\underline{e}}_j = \frac{1}{n}\sum_{i\in U}\underline{I}_i\underline{e}_{ij}, \tag{4.43}$$

$$\hat{V}_S\left(\hat{T}_{j,GREG}\right) = \frac{N^2(1-f)}{n}S^2_{e_j},\ S^2_{e_j} = \frac{1}{n-1}\sum_{i\in s}\left(e_{ij}-\overline{e}_j\right)^2,\ \overline{e}_j = \frac{1}{n}\sum_{i\in s}e_{ij}. \tag{4.44}$$

4.3.3 Variance of domain GREG under the SRSWOR design

Let us next consider the variance of the model-assisted domain GREG estimator. When domains are planned, every domain can be treated as a distinct sub-population. Then the Standard variance approximation for the sub-populations is

$$V_A\left(\hat{\underline{T}}_{j,GREG}^{(d)}\right) = \frac{\left(N^{(d)}\right)^2\left(1-f^{(d)}\right)}{n^{(d)}} S_{E_j^{(d)}}^2, \text{ where}$$
$$S_{E_j^{(d)}}^2 = \frac{1}{N^{(d)}-1}\sum_{i\in U^{(d)}}\left(E_{ij}^{(d)} - \bar{E}_j^{(d)}\right)^2 \text{ and } \bar{E}^{(d)} = \frac{1}{N^{(d)}}\sum_{i\in U^{(d)}} E_{ij}^{(d)} \quad . \tag{4.45}$$

The Standard variance estimator for (4.45) is

$$\hat{V}_S\left(\hat{\underline{T}}_{j,GREG}^{(d)}\right) = \frac{\left(N^{(d)}\right)^2\left(1-f^{(d)}\right)}{n^{(d)}} \underline{S}_{e_j^{(d)}}^2, \text{ where}$$
$$\hat{\underline{S}}_{e_j^{(d)}}^2 = \frac{1}{N^{(d)}-1}\sum_{i\in U^{(d)}} \underline{I}_i\left(\underline{e}_{ij} - \bar{\underline{e}}_j^{(d)}\right)^2 \text{ and } \bar{\underline{e}}_j^{(d)} = \frac{1}{n^{(d)}}\sum_{i\in U^{(d)}} \underline{I}_i \underline{e}_{ij} \, . \tag{4.46}$$

If domains are unplanned, the extra variation that comes with random domain sample size must be taken into account – either by using a different variance estimator or by treating the domain total as the population total of variable $y_j^{(d)}$. We use the latter approach. Making use of properties (4.39), the Standard variance approximation (4.35) of domain GREG reduces to

$$V_S\left(\hat{\underline{T}}_{j,GREG}^{(d)}\right) = \frac{N^2(1-f)}{n}\frac{1}{N-1}\sum_{i\in U}\left(E_{ij}^{(d)} - \bar{E}_j^{(d)}\right)^2, \; E_{ij}^d = y_{ij}^{(d)} - \tilde{y}_{ij}^{(d)}. \tag{4.47}$$

The Standard variance estimator and an estimate for approximation (4.47) are

$$\hat{V}_S\left(\hat{\underline{T}}_{j,GREG}^{(d)}\right) = \frac{N^2(1-f)}{n}\frac{1}{n-1}\sum_{i\in U}\underline{I}_i\left(\underline{e}_{ij}^{(d)} - \bar{\underline{e}}_j^{(d)}\right)^2, \text{ where} \tag{4.48}$$

$$\underline{e}_{ij}^{(d)} = y_{ij}^{(d)} - \underline{\hat{y}}_{ij}^{(d)}, \; \bar{\underline{e}}_j^{(d)} = \frac{1}{n}\sum_{i\in U}\underline{I}_i \underline{e}_{ij}^{(d)}, \text{ and}$$

$$\hat{V}_S\left(\hat{T}_{j,GREG}^{(d)}\right) = \frac{N^2(1-f)}{n}\frac{1}{n-1}\sum_{i\in s}\left(e_{ij}^{(d)} - \bar{e}_j^{(d)}\right)^2. \tag{4.49}$$

Note that Standard approximate variance (4.42) and the variance estimator (4.43) for the population GREG are obtained as special cases of (4.47) and (4.48) by setting $y_{ij}^{(d)} = y_{ij}$.

4.3.4 Variance of population and domain GREG under the πPS design

In this chapter, we consider the Standard variance approximation and the Standard variance estimator in the case of the fixed size without replacement probability proportional to size (πPS) design.

We consider variance estimation for the domain GREG and population GREG simultaneously, since the population GREG can be seen as a special case of the domain GREG (setting $U^{(d)} = U$). The case of planned domains is completely analogous to (4.45) and (4.46) and need not be reproduced here. Thus we focus on the case of unplanned domains.

For fixed size without replacement designs, the Standard approximation for the domain GREG is

$$V_A\left(\underline{\hat{T}}_{j,GREG}\right) = \sum_{i\in U}\sum_{k\in U}\frac{\pi_{ik}-\pi_i\pi_k}{\pi_i\pi_k}E_{ij}^{(d)}E_{kj}^{(d)}, \tag{4.50}$$

and the corresponding Standard variance estimator as

$$\hat{V}_S\left(\underline{\hat{T}}_{j,GREG}\right) = \sum_{i\in U}\sum_{k\in U}\frac{\pi_{ik}-\pi_i\pi_k}{\pi_{ik}}\underline{w}_i\underline{e}_{ij}^{(d)}\underline{w}_k\underline{e}_{kj}^{(d)}. \tag{4.51}$$

Unfortunately, expressions (4.50), (4.51) do not simplify as they do under the SRSWOR design. Thus the double sum has to be calculated in order to estimate the variance. In the Monte Carlo experiments that follow, we do not use directly the approximation (4.50) but an unbiased estimator for the approximation. This is done because calculating the second-order inclusion probabilities for the whole population would be computationally very demanding, and when we use the unbiased estimator, it is enough to calculate the π_{ij} for the sample set only. Thus in the Monte Carlo studies, (4.50) is replaced by a pseudo-approximation

$$\overline{\hat{V}}_A\left(\underline{\hat{T}}_{j,GREG}\right) = \frac{1}{K}\sum_{k=1}^{K}\hat{V}_A^{(k)}\left(\underline{\hat{T}}_{j,GREG}\right), \tag{4.52}$$

where $\hat{V}_A^{(k)}$ is the kth estimate

$$\sum_{i\in U}\sum_{k\in U}\frac{\pi_{ik}-\pi_i\pi_k}{\pi_{ik}}\underline{w}_i E_{ij}^{(d)}\underline{w}_k E_{kj}^{(d)} \tag{4.53}$$

of K Monte Carlo replicates. The estimator (4.53) is unbiased for the true approximation (4.50), and given that K is large, the pseudo-approximation (4.52) is close to the true population value (4.50).

4.4 GREG with linear fixed effects assisting model (GREG-lin)

This chapter presents the GREG-lin estimator whose assisting model is a linear, fixed effects model. The estimator is constructed as follows. A hypothetical *linear fixed effects* population generating process ξ is imposed on the population. The process has the following features:

For every $i \in U$
1) values y_{ij} are independent realisations of the random variable $\underline{y}_j$,

2) $E_\xi\left(\underline{y}_{ij}\right) = \mathbf{x}_i'\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a fixed parameter vector,

3) $V_\xi\left(\underline{y}_{ij}\right) = \sigma_i^2$, and

4) $Cov_\xi\left(\underline{y}_{ij}, \underline{y}_{kj}\right) = 0, \ i \neq k.$

The expectation of the study variable is linear in parameters, as can be seen from feature 2). Covariates may be arbitrary functions of the original auxiliary variables, and often there is a constant as one covariate. If the constant is omitted, the resulting estimator is called a *ratio estimator*. In this study the constant is always included in the model. Thus $\mathbf{x}_i = (1, x_1, x_2, \ldots, x_M)$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_M)$.

If values of $y_j, x_1, x_2, \ldots, x_M$ were known for the whole population, a census estimate of $\boldsymbol{\beta}$ could be calculated using generalised least squares (GLS) as

$$\mathbf{B} = \left(\sum_{i \in U} \frac{\mathbf{x}_i \mathbf{x}_i'}{\sigma_i^2}\right)^{-1} \sum_{i \in U} \frac{\mathbf{x}_i y_{ij}}{\sigma_i^2} . \tag{4.54}$$

In practice, the model must be estimated from the sample. An unbiased sample estimator for the census fit parameter $\mathbf{B}$ is the generalised weighted least squares (GWLS) estimator with weights $\underline{w}_i / \sigma_i^2$:

$$\hat{\underline{\mathbf{B}}} = \left(\sum_{i \in U} \underline{w}_i \frac{\mathbf{x}_i \mathbf{x}_i'}{\sigma_i^2}\right)^{-1} \sum_{i \in U} \underline{w}_i \frac{\mathbf{x}_i y_{ij}}{\sigma_i^2} . \tag{4.55}$$

In order to calculate (4.55), one often assumes that the variance parameters σ_i^2 are proportional to known constants so they cancel out, or that $\sigma_i^2 = \sigma^2$ (that is, the study variable is homoscedastic). Under the latter assumption (4.55) simplifies to the weighted least squares (WLS) estimator

$$\hat{\underline{\mathbf{B}}} = \left(\sum_{i \in U} \underline{w}_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i \in U} \underline{w}_i \mathbf{x}_i y_{ij} . \tag{4.56}$$

In this study the assumption $\sigma_i^2 = \sigma^2$ is used, so (4.56) is the model parameter estimator for GREG-lin. However, for class frequencies the assumption $\sigma_i^2 = \sigma^2$ is false, since the response variable is binary and thus the variance depends on mean. This is an unfortunate drawback which follows from the fact that in GREG-lin one fits a linear model to binary data.

When calculating (4.56), problems would arise if the model was not of full rank, so that a unique inverse of $\sum_U \underline{w}_i \mathbf{x}_i \mathbf{x}_i'$ did not exist. In this study all fitted models are such that the unique inverse does exist, so generalised inverses need not be considered.

After estimating the parameters, the micro level auxiliary information is used to construct predictions $\hat{y}_{ij} = \mathbf{x}_i' \hat{\mathbf{B}}$ for every $i \in U$. GREG-lin for the population total T_j is now defined as

$$\hat{\underline{T}}_{j,GREG-lin} = \sum_{i \in U} \underline{\hat{y}}_{ij} + \sum_{i \in U} \underline{w}_i \underline{e}_{ij}, \quad \underline{e}_{ij} = y_{ij} - \underline{\hat{y}}_{ij}. \tag{4.57}$$

When domains are considered, one may want to construct the assisting model on the domain level and estimate the model parameters from $s^{(d)}$. Alternatively, the model can be constructed on the population level and appropriate dummy variables can be used to allow domain specific effects. In both cases, the GREG-lin estimator for domains is

$$\hat{\underline{T}}^{(d)}_{j,GREG-lin} = \sum_{i \in U} \underline{\hat{y}}^{(d)}_{ij} + \sum_{i \in U} \underline{w}_i \underline{e}^{(d)}_{ij}, \quad \underline{e}^{(d)}_{ij} = y^{(d)}_{ij} - \underline{\hat{y}}^{(d)}_{ij}. \tag{4.58}$$

If the domains are planned, each domain is often treated as a population of its own. This means that the model is specified on the domain level and its parameters are estimated using observations from the given domain only; therefore the resulting estimator is called direct.

If the domains are unplanned, indirect estimation is a possible choice. But if direct estimation is chosen, it has to be assumed that domain sample sizes are large enough to justify the model fitting in each domain separately.

If indirect estimation is used, one uses observations from outside the domains under study. Specifically, let $U^{(d)}$ be the domain whose total is estimated and $U^{(m)}$ the set of units that is used in model fitting ($U^{(d)} \subseteq U^{(m)} \subseteq U$). The model is imposed on $U^{(m)}$ and model parameters $\mathbf{B}^{(m)}$ estimated from the sample set $s^{(m)}$ (often $U^{(m)} = U$). Predictions and residuals are constructed using parameters $\hat{\mathbf{B}}^{(m)}$:

$$\underline{\hat{y}}^{(d)}_{ij} = \delta_i^{(d)} \mathbf{x}_i' \underline{\hat{\mathbf{B}}}^{(m)}, \text{ and } \underline{e}^{(d)}_{ij} = y^{(d)}_{ij} - \underline{\hat{y}}^{(d)}_{ij}, \tag{4.59}$$

and the indirect model-assisted GREG-lin for domains is (4.58) with predictions and residuals (4.59).

4.4.1 Some properties of GREG-lin

The GREG-lin estimator (4.57) can be written as a sum of the HT estimator and a correction term:

$$\underline{\hat{T}}_{j,GREG-lin} = \underline{\hat{T}}_{j,HT} + \left(\mathbf{T}_{\mathbf{x}} - \underline{\hat{\mathbf{T}}}_{\mathbf{x},HT}\right)' \underline{\hat{\mathbf{B}}}, \tag{4.60}$$

where $\underline{\hat{\mathbf{B}}}$ is (4.56) (Särndal *et al.* 1992, 225). It can also be written as a calibration estimator:

$$\underline{\hat{T}}_{j,GREG-lin} = \sum_{i \in U} \underline{w}_i \underline{g}_i y_{ij}, \text{ where} \tag{4.61}$$

$$\underline{g}_i = 1 + \left(\mathbf{T}_{\mathbf{x}} - \underline{\hat{\mathbf{T}}}_{\mathbf{x},HT}\right)' \left(\sum_{i \in U} \underline{w}_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \mathbf{x}_i. \tag{4.62}$$

The equivalence of (4.60) and (4.61) can be seen by working out the *g*-weights:

$$\begin{aligned} \sum_{i \in U} \underline{w}_i \underline{g}_i y_{ij} &= \sum_{i \in U} \underline{w}_i \left[1 + \left(\mathbf{T}_{\mathbf{x}} - \underline{\hat{\mathbf{T}}}_{\mathbf{x},HT}\right)' \left(\sum_{i \in U} \underline{w}_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \mathbf{x}_i\right] y_{ij} \\ &= \sum_{i \in U} \underline{w}_i y_{ij} + \left(\mathbf{T}_{\mathbf{x}} - \underline{\hat{\mathbf{T}}}_{\mathbf{x},HT}\right)' \left(\sum_{i \in U} \underline{w}_i \mathbf{x}_i \mathbf{x}_i'\right)^{-1} \sum_{i \in U} \underline{w}_i \mathbf{x}_i y_{ij} \\ &= \underline{\hat{T}}_{j,HT} + \left(\mathbf{T}_{\mathbf{x}} - \underline{\hat{\mathbf{T}}}_{\mathbf{x},HT}\right)' \underline{\hat{\mathbf{B}}}. \end{aligned} \tag{4.63}$$

The weights $\underline{w}_i \underline{g}_i = w_i^{(c)}$ are calibration weights since for any sample they minimise the distance $\sum_s \left(w_i - w_i^{(c)}\right)^2 w_i^{-1}$ to sampling weights while fulfilling the condition $\sum_s w_i g_i \mathbf{x}_i' = \mathbf{T}_{\mathbf{x}}'$, that is, they reproduce the auxiliary totals (e.g. Rao 2003, 25).

The bias-correction term in GREG-lin (4.57), $\sum_U \underline{w}_i \underline{e}_{ij}$ is zero in many cases. A sufficient condition for $\sum_U \underline{w}_i \underline{e}_{ij} = 0$ is (Särndal *et al.* 1992, 231)

1) the estimator of $\mathbf{B}$ is (4.56), and
2) for all samples, there exists a constant column vector $\boldsymbol{\lambda}$ (of dimension M not depending on i) such that for all $i \in U$, $\sigma_i^2 = \boldsymbol{\lambda}' \mathbf{x}_i$.

For example, constant variances and variances proportional to some x-variable satisfy the condition 2). For the domain estimator, the model correction term vanishes if the weights are constants, the model has a constant term and the estimator is direct with parameter estimator (4.56).

The formulas (4.57), (4.60) and (4.61) show that GREG-lin can be written in three useful forms: as an estimator that uses statistical models, as a calibration estimator that does not need explicit models and as an HT estimator plus correction term. When GREG-lin is written as an HT estimator plus correction term, the correction term can be interpreted as follows: If the observed sample is skewed, this skewness is exposed in the error of HT estimates for the auxiliary totals. Hence, a correction can be made, and correction coefficients are regression coefficients (4.56). The more correlation there is between the study variable and auxiliary variables, the more negatively correlated are the error of the HT estimator for $T_j^{(d)}$ and the correction term, and the smaller is the variance of GREG-lin. In the extreme case, where the response variable is linearly dependent on covariates, the variance of the regression estimator is zero.

The *g*-weighted form of the GREG-lin estimator is often most useful in practice. For the users of the data, it may be of great importance that they can handle a data set of size n (instead of N), and using the calibrated weights, produce results that are consistent with some benchmark totals.

However, the construction that explicitly utilises predictions and prediction errors is the most important one in this study. In GREG-lin, predictions are made via linear, fixed effects modelling. This is quite restricting since only particular types of responses can be considered as realisations of a linear model. Binary or polytomous variables, for example, certainly should not be considered as such. In Chapter 4.5, we consider in more detail L-GREG estimators, whose models are appropriate for class frequencies.

4.4.2 Variance of GREG-lin

The variance of GREG-lin is in typical cases smaller than that of the HT estimator (Knottnerus, 2003, 303–304, provides some counterexamples), and with a well-fitting model, the reduction in variance can be substantial. For non-linear estimators like GREG-lin, exact closed form variance formulas are difficult to obtain. Therefore the estimator is linearised in order to obtain the variance approximation and an estimator for the approximation.

The Taylor linearisation based first-order approximate variance for the estimator (4.60) is (Särndal *et al.* 1992, 235)

$$V_A\left(\hat{\underline{T}}_{j,GREG-lin}^{(d)}\right) = \sum_{i\in U}\sum_{k\in U}\frac{\Delta_{ik}}{E(\underline{I}_i)E(\underline{I}_k)}E_{ij}E_{kj}, \quad E_{ij} = y_{ij} - \mathbf{x}_i'\mathbf{B}. \qquad (4.64)$$

An estimator for (4.64) is obtained by replacing census fit residuals by weighted sample fit residuals:

$$\hat{V}\left(\hat{\underline{T}}_{j,GREG-lin}\right) = \sum_{i\in U}\sum_{k\in U}\frac{\Delta_{ik}}{E(\underline{I}_i\underline{I}_k)}\left(\underline{w}_i\underline{e}_{ij}\right)\left(\underline{w}_k\underline{e}_{kj}\right), \quad \underline{e}_{ij} = y_{ij} - \mathbf{x}_i'\hat{\underline{\mathbf{B}}}. \qquad (4.65)$$

Note that the approximation (4.64) and estimator (4.65) are the same as those derived in Chapter 4.3.

The estimator (4.65) performs well if the sample size is reasonably large, but may have downward bias in small samples. A refined estimator for (4.64), suggested by Särndal *et al.* (1989), is obtained by replacing census fit residuals E_{ij} by g-weighted sample fit residuals $\underline{w}_i \underline{g}_i \underline{e}_{ij}$. But this estimator may also be downward biased in small samples (Lundström and Särndal 2002, 50). At least part of the downward bias comes from the fact that the estimators ignore the uncertainty that comes from the estimation of $\mathbf{B}$. This means that the random variable $\underline{\hat{\mathbf{B}}}$ contributes only little to the variance; these types of random variables are sometimes called relatively fixed numbers (Knottnerus 2003, 124). According to Lundström (1997, 43) and Lundström and Särndal (2002, 50), the bias can be reduced by multiplying sample fit residuals by a term that adjusts for the number of degrees of freedom lost in the estimation of $\mathbf{B}$. The simplest adjustment term is

$$\frac{n-1}{n-(M+1)}, \tag{4.66}$$

where $M + 1$ is the number of model parameters. But the adjustment is negligible, unless the number of covariates is very large with respect to the sample size.

An approximate variance and variance estimator for the direct planned-domain GREG-lin estimator are obtained from (4.64) and (4.65) by replacing U with $U^{(d)}$. In the indirect case where the model is fitted in $U^{(m)}$ ($U^{(d)} \subseteq U^{(m)} \subseteq U$) the approximate variance for the model-assisted domain GREG-lin estimator is

$$V_A\left(\underline{\hat{T}}^{(d)}_{j,GREG-lin}\right) = \sum_{i\in U}\sum_{k\in U} \frac{\Delta_{ik}}{E(\underline{I}_i)E(\underline{I}_k)} E^{(d)}_{ij} E^{(d)}_{kj}, \text{ where} \tag{4.67}$$

$$E^{(d)}_{ij} = y^{(d)}_{ij} - \tilde{y}^{(d)}_{ij}, \quad \tilde{y}^{(d)}_{ij} = \delta^{(d)}_i \mathbf{x}'_i \mathbf{B}^{(m)}, \text{ and} \tag{4.68}$$

$$\mathbf{B}^{(m)} = \left(\sum_{i\in U^{(m)}} \mathbf{x}_i \mathbf{x}'_i\right)^{-1} \sum_{i\in U^{(m)}} \mathbf{x}_i y_{ij}.$$

Note that the census fit predictions in (4.68) are zero outside $U^{(d)}$ and no terms of form $E^{(d)}_{ij} = -\tilde{y}^{(d)}_{ij}$ enter in the variance approximation. In the literature there are also regression estimators closely related to (4.60) that do have nonzero terms of form $E^{(d)}_{ij} = -\tilde{y}^{(d)}_{ij}$ in the variance approximation (e.g. Hidiroglou and Patak 2001; Lundström and Särndal 2002; Lehtonen and Pahkinen 2004).

An estimator for (4.67) is

$$\hat{V}\left(\underline{\hat{T}}^{(d)}_{j,GREG-lin}\right) = \sum_{i\in U}\sum_{k\in U} \frac{\Delta_{ik}}{E(\underline{I}_i \underline{I}_k)} \left(\underline{w}_i \underline{e}^{(d)}_{ij}\right)\left(\underline{w}_k \underline{e}^{(d)}_{kj}\right), \text{ where} \tag{4.69}$$

$$\underline{e}^{(d)}_{ij} = y^{(d)}_{ij} - \underline{\hat{y}}^{(d)}_{ij}, \quad \underline{\hat{y}}^{(d)}_{ij} = \delta^{(d)}_{ij} \mathbf{x}'_i \underline{\hat{\mathbf{B}}}^{(m)} \text{ and} \tag{4.70}$$

$$\underline{\hat{\mathbf{B}}}^{(m)} = \left(\sum_{i\in U^{(m)}} \underline{w}_i \mathbf{x}_i \mathbf{x}'_i\right)^{-1} \sum_{i\in U^{(m)}} \underline{w}_i \mathbf{x}_i y_{ij}. \tag{4.71}$$

Note that the estimator (4.69) includes as special cases the situations where the model is fitted on the domain level (direct estimation; $U^{(m)} = U^{(d)}$) and where the model is fitted on the population level (indirect estimation; $U^{(m)} = U$).

4.5 GREG with logistic-type fixed effects assisting model (L-GREG)

The model-assisted GREG estimator for domain class frequencies was defined as

$$\hat{\underline{T}}^{(d)}_{j,GREG} = \sum_{i \in U} \hat{\underline{y}}^{(d)}_{ij} + \sum_{i \in U} \underline{w}_i \underline{e}^{(d)}_{ij}, \quad \underline{e}^{(d)}_{ij} = y^{(d)}_{ij} - \hat{\underline{y}}^{(d)}_{ij}, \tag{4.72}$$

where the model utilised to construct $\hat{y}_{ij}$ could be any statistical model. In Chapter 4.2, we defined logistic-type models as models whose predictions can be interpreted as probabilities. Logistic-type models and the general formulation of the GREG estimator give rise to L-GREG estimators: these are estimators of form (4.72) which have a logistic-type assisting model.

The three logistic-type models we study are the most common models for binary data. These models are the logistic, probit and complementary log-log-model, described in Chapter 4.2.2. A GREG estimator with the logistic model as an assisting model (link (4.19)) will be called GREG-log. Corresponding models with probit and complementary log-log links (links (4.20) and (4.22)) will also be used. The GREG estimators with these models are called GREG-prob and GREG-cll, respectively.

Although it is clear that a logistic-type model is more appropriate than a linear one when class frequencies are estimated, L-GREG is rarely used in survey sampling practice. Lehtonen and Veijanen (1998a, 1998b) have compared GREG-lin and GREG-log, and their simulations indicate that especially if the domain sample sizes are small, GREG-log is more accurate than GREG-lin. Lehtonen *et al.* (2003, 2005) have also further developed the idea of modelling in the context of design-based estimation of class frequencies for domains by using GREG estimators with random effects logistic models. The random effects feature of these models, however, might be unnecessary, since the population consists of fixed points and we have observations for every domain. Thus, in this study, we consider only fixed effects models.

4.5.1 Estimation of model parameters

To estimate the parameters of a logistic-type model, we use the pseudo-maximum likelihood method (PML, Skinner 1989) with the Newton-Raphson algorithm. The method, which is essentially w_i-weighted sample maximum likelihood, is as follows. Let $\underline{y}_i$ be the binary response variable and let $p_i = P(\underline{y}_{ij} = 1)$. The variables $\underline{y}_i, \underline{y}_k$ are uncorrelated for $i \neq k$. Let $\boldsymbol{\beta}$ be the vector of regression parame-

ters of the generalised linear model (the population generating process) ξ. The link of ξ is either logit, probit or complementary log-log, with $\mathbf{x}'\boldsymbol{\beta}$ and p linked as

$$\mathbf{x}'\boldsymbol{\beta} = \begin{cases} \log\left[p/(1-p)\right] & \text{for the logistic model,} \\ \Phi^{-1}(p) & \text{for the probit model, and} \\ \log\left[-\log(1-p)\right] & \text{for the complementary log-log (cll) model.} \end{cases} \tag{4.73}$$

If we observed the whole population U, we would maximise the log-likelihood

$$l_U(\boldsymbol{\beta};\mathbf{y}) = \sum_{i\in U} y_i \log p_i + (1-y_i)\log(1-p_i) \tag{4.74}$$

to obtain the census fit estimate $\mathbf{B}$. In practice, only a sample is observed and one has to choose between weighted and non-weighted likelihoods. We use the w_i-weighted sample log-likelihood, which is essentially an HT estimator for (4.74). The w_i-weighted sample log-likelihood is

$$l_s(\boldsymbol{\beta};\mathbf{y}) = \sum_{i\in U} w_i\left[y_i \log p_i + (1-y_i)\log(1-p_i)\right]. \tag{4.75}$$

To obtain the maximum likelihood (ML) estimate, we calculate the partial derivatives of (4.75) with respect to $\boldsymbol{\beta}$. The first and second partial derivatives, which depend on the link (4.73), are denoted by $\partial l_s(\boldsymbol{\beta};\mathbf{y})/\partial\boldsymbol{\beta}$ and $\partial^2 l_s(\boldsymbol{\beta};\mathbf{y})/\partial\boldsymbol{\beta}^2$, respectively. The gradient $\mathbf{g}_\beta$ and the Hessian matrix $\mathbf{H}_{\beta^2}$ are

$$\mathbf{g}_\beta = \sum_{i\in s} w_i \frac{\partial l(\boldsymbol{\beta};y_i)}{\partial\boldsymbol{\beta}} \text{ and } \mathbf{H}_{\beta^2} = \sum_{i\in s} -w_i \frac{\partial^2 l(\boldsymbol{\beta};y_i)}{\partial\boldsymbol{\beta}^2}, \tag{4.76}$$

and given a start-up value $\hat{\mathbf{B}}(0)$, the PML estimate $\hat{\mathbf{B}}$ is obtained iteratively as

$$\hat{\mathbf{B}}(h+1) = \hat{\mathbf{B}}(h) + \mathbf{H}^{-1}_{\hat{\mathbf{B}}^2(h)}\mathbf{g}_{\hat{\mathbf{B}}(h)}. \tag{4.77}$$

The Newton-Raphson algorithm (4.77) converges to the ML estimate if the estimate exists as unique and finite. However, if the data set is completely or quasi-completely separable, unique and finite estimates do not exist.

The data is completely separable if there exists a parameter vector $\boldsymbol{\beta}$ so that

$$\forall i \in s: \begin{cases} \mathbf{x}_i'\boldsymbol{\beta} > 0 \text{ and } y_i = 0, \\ \mathbf{x}_i'\boldsymbol{\beta} < 0 \text{ and } y_i = 1. \end{cases}$$

The data is quasi-completely separable if there exists a $\boldsymbol{\beta}$ so that

$$\forall i \in s: \begin{cases} \mathbf{x}_i'\boldsymbol{\beta} \geq 0 \text{ and } y_i = 0, \\ \mathbf{x}_i'\boldsymbol{\beta} \leq 0 \text{ and } y_i = 1 \end{cases}$$

and the equality holds for at least one subject.

Complete (or quasi-complete) separation means that the model correctly (or quasi-correctly) allocates all observations to their response groups. If neither complete nor quasi-complete separation holds, the data set is said to overlap, and the Newton-Raphson algorithm converges to the maximum likelihood estimate.

Complete separation and quasi-complete separation are problems typically encountered with small data sets. Moreover, quasi-complete separation is not likely to occur with truly continuous auxiliary variables. In this study, there were truly continuous auxiliary variables in every model except the model in set 5 (see Chapters 5 and 7). Thus, the problem of complete separation is more relevant than quasi-complete separation.

In the simulations of Chapters 5 and 7, the smallest SRSWOR sample sizes are n=1,000. This cannot be considered small, thus complete separation is not likely to be a problem. In the Monte Carlo experiments with πPS sampling, the smallest samples are n=40. For sample sizes this small, complete separation may be a problem. We will discuss this more in conjunction with the simulations.

There are some cases where complete separation is obvious: if all observations in the sample have either $y_i = 0$ or $y_i = 1$, the data set is completely separable. Moreover, if $y_i = 0$ or $y_i = 1$ for every $i \in U^{(d)}$, the data set in the domain $U^{(d)}$ is completely separable. Each Monte Carlo sample that was obviously completely separable ($y_i = 0$ or $y_i = 1$ for every i in the sample or in some domain) was discarded. For the SRSWOR design, there were no such cases. For the πPS design and n=80, obvious complete separation did not occur, but for n=40, 2% of the samples had to be discarded because of obvious complete separation.

Obvious complete separation was also sometimes present when the parameters for the Augmented variance estimator (see Chapter 6) were estimated from pseudo-samples. For the SRSWOR design and the πPS design with n=80, no obvious complete separation occurred in the pseudo-samples. But for the πPS design and n=40, 4% of the pseudo-samples were discarded because of obvious complete separation.

4.5.2 Some properties of GREG-log

The GREG-log estimator

$$\hat{\underline{T}}^{(d)}_{j,GREG-\log} = \sum_{i \in U} \hat{\underline{y}}^{(d)}_{ij} + \sum_{i \in U} \underline{w}_i \underline{e}^{(d)}_{ij}, \text{ where} \tag{4.78}$$

$$\underline{e}^{(d)}_{ij} = y^{(d)}_{ij} - \hat{\underline{y}}^{(d)}_{ij} \text{ and } \quad \hat{\underline{y}}^{(d)}_{ij} = \frac{\exp(\mathbf{x}_i'\hat{\underline{\mathbf{B}}})}{1+\exp(\mathbf{x}_i'\hat{\underline{\mathbf{B}}})} \delta^{(d)}_i, \tag{4.79}$$

has certain interesting properties, shown in Lehtonen and Veijanen (1998b). First, the bias-correction term is zero in many cases. It can be shown that $\sum_{i \in U} \underline{w}_i \underline{e}^{(d)}_{ij} = 0$ if

1) $\hat{\underline{\mathbf{B}}}$ is the PML estimator, and
2) there exists a constant column vector $\boldsymbol{\lambda}$ such that for all $i \in U$, $\boldsymbol{\lambda}'\mathbf{x}_i = \delta_i^{(d)}$.

The latter condition is met if, for example, the covariate vector $\mathbf{x}_i$ includes the domain indicator variable. Second, if the model parameters are estimated using the PML estimator and $\mathbf{x}_i$ includes indicator variables from a saturated (full-interaction) stratification and nothing else, GREG-log is identical with GREG-lin. However, GREG-prob and GREG-cll do not have similar properties. Finally, GREG-log can be written as a calibration estimator if all the covariates are categorical and the model is estimated using the PML estimator.

4.5.3 Comparison of the accuracy of GREG-lin and GREG-log

Here we make an attempt to identify the conditions under which GREG-log is more accurate than GREG-lin. To measure the relative accuracy of GREG-log with respect to GREG-lin, we use the functional form effect *FE*

$$FE\left(\hat{\underline{T}}_{j,GREG}\right) = \sqrt{\frac{V\left(\hat{\underline{T}}_{j,GREG-lin}\right)}{V\left(\hat{\underline{T}}_{j,GREG}\right)}}. \tag{4.80}$$

In the denominator of (4.80), we have the true variance of the GREG estimator with some statistical model and auxiliary information, and in the numerator, the variance of GREG-lin that uses the same auxiliary information via a linear, fixed effects model.

We consider only the GREG-log estimator and the SRSWOR design. Moreover, approximate variances are used instead of the true ones and both GREG-lin and GREG-log are assumed to be direct estimators. Then the approximate functional form effect for GREG-log is

$$FE_A\left(\hat{\underline{T}}_{j,GREG-\log}\right) = \sqrt{\frac{\frac{N^2}{n}(1-f)S^2_{E_{j,lin}}}{\frac{N^2}{n}(1-f)S^2_{E_{j,\log}}}} \approx \sqrt{\frac{\sum_{i\in U} E^2_{j,\text{lin}}}{\sum_{i\in U} E^2_{j,\log}}}, \tag{4.81}$$

where the subscripts $E_{j,\log}$ and $E_{j,\text{lin}}$ refer to census fit prediction errors obtained from the corresponding logistic and linear models. The approximation in (4.81) is obtained by noting that the mean of prediction errors is approximately zero for both models. The approximation is in line with the intuition that if the prediction errors of the logistic model are smaller than those of the linear model, accuracy is gained by changing the model from linear to logistic. But when are the prediction errors of a logistic model smaller than those of a linear model? Let us first check when they are approximately the same. Figure 4.2 illustrates the prediction curves of a logistic and linear model.

Figure 4.2 shows that the linear line approximates the logistic curve well if $p \in (0.25, 0.75)$. But if the predictions are near the borders of the unit interval, differences arise. The predictions of a linear model are not restricted, therefore

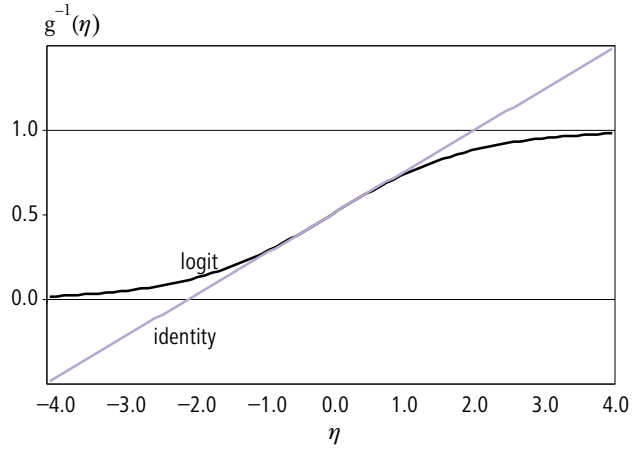


Figure 4.2. Graphs of $g^{-1}(\eta), \eta \in [-4, 4]$ where g is logit and identity.

the prediction errors are also unrestricted. But for a logistic model, predictions are always within the unit interval and the prediction errors within the interval (–1,1). Thus, if the predictions of a logistic model are near the borders of the unit interval, the prediction errors of a linear model are likely to be greater than those of the logistic model.

There are at least two situations where predictions of the logistic model are near the borders of the unit interval. The first is one where the auxiliary information identifies units that very likely either belong or do not belong to the class whose frequency is estimated. For example, in the Finnish Labour Force Survey (LFS), the Ministry of Labour's register of unemployed jobseekers is used as auxiliary information. One of the study variables of the LFS is unemployment, as defined by the International Labour Organisation (ILO). The jobseeker register identifies with great accuracy those who are unemployed according to the ILO criteria.

We say that auxiliary information is strong if it identifies units that very likely either belong or do not belong to the class of interest. If the assisting model uses strong auxiliary information, the model is strong. Respectively, if the predictions are closer to the middle of the unit interval, both the auxiliary information and the model are weak. Note, that correctness of a model and strength of a model are different concepts: A correct model is defined as a model that has the same effects and the same functional form as the population generating process. Then the correct model may be weak, strong, or anything in between. Moreover, a strong model may be correct or incorrect (for example, by adding unnecessary effects to a strong and correct model one often gets an incorrect, strong model). Similarly a weak model may be correct or incorrect.

The result of the reasoning of this chapter is that if the model is strong, GREG-log is expected to be more accurate than GREG-lin. If auxiliary information is weak, GREG-log and GREG-lin are expected to be equally accurate. Situations where GREG-lin would be significantly more accurate than GREG-log, then, seem to be rare at least in the case of the SRSWOR design. In next chapter, we conduct Monte Carlo simulation experiments in order to study empirically the relative accuracy of L-GREG with respect to GREG-lin.

5 Monte Carlo study I: Comparison of GREG-lin and L-GREG

In this chapter, we study the properties of L-GREG estimators by the Monte Carlo simulation. The aim is to study
1) whether there are any accuracy differences between L-GREG estimators,
2) whether and when L-GREG is more accurate than GREG-lin, and
3) the goodness of the Standard variance estimator for L-GREG.

To answer these questions, we conduct Monte Carlo simulation experiments that cover both SRSWOR and πPS designs, several domain sample size categories (minor, medium sized, major), several functional model forms (linear, logistic, probit, cll) and several types of model formulations (weak, moderate, strong, overfitted). First in Chapter 5.1, justification for the use of the Monte Carlo technique is given. From Chapter 5.2 onwards, the experiments are described and results are given. We use the common random numbers technique to reduce Monte Carlo variance. This means that the samples are exactly the same for all GREG estimators, and outlier samples should have only little effect on the estimated difference between the estimators.

5.1 Classical Monte Carlo Approximation

Consider a sum $S = \sum_{x\in\Omega} h(x)f(x)$. By the definition of expectation, this sum can be interpreted as expectation of a function h of discrete random variable $\underline{x}$ with respect to distribution f:$\Omega \to \mathbb{R}$:

$$E_f[h(\underline{x})] = \sum_{x\in\Omega} h(x)f(x)\,. \tag{5.1}$$

For independent $x_1, x_2, \ldots, x_K$ from the distribution f, the mean of $h(x_1)$, $h(x_2), \ldots, h(x_K)$ converges in probability to (5.1) (Robert and Casella 2002, 75):

$$\frac{1}{K}\sum_{k=1}^{K} h(x_k) \xrightarrow[K\to\infty]{P} \sum_{x\in\Omega} h(x)f(x)\,. \tag{5.2}$$

This method of approximation is called the classical Monte Carlo approximation. In the context of survey sampling, the random variable in (5.1) is sampling vector $\underline{\mathbf{I}}$, its distribution is $p(\cdot)$ and function h is the estimator. The result (5.2) enables approximation of bias, variance and other moments (if they are finite) of an estimator $\hat{\underline{\theta}}$. The method is as follows:

1) Draw K independent samples from the population under study
2) For each sample $k = 1, 2, \ldots, K$ calculate the value $\hat{\theta}^{(k)}$ of the estimator $\hat{\underline{\theta}}$.

For expectation and variance we then have

$$\hat{E}(\underline{\hat{\theta}}) = \frac{1}{K}\sum_{k=1}^{K}\hat{\theta}^{(k)} \xrightarrow[K\to\infty]{P} E(\underline{\hat{\theta}}), \tag{5.3}$$

$$\hat{V}(\underline{\hat{\theta}}) = \frac{1}{(K-1)}\sum_{k=1}^{K}\left[\hat{\theta}^{(k)} - \hat{E}(\underline{\hat{\theta}})\right]^2 \xrightarrow[K\to\infty]{P} V(\underline{\hat{\theta}}). \tag{5.4}$$

These results are applied in the simulations in order to study the bias and variance of several estimators. In the simulations, the strategy is the standard design-based strategy where a large number of samples is drawn from a fixed population (in the model-based approach, it is common to generate a large number of populations and draw a small number of samples from each population).

5.2 Experiment I.1: SRSWOR design

In this chapter, we study GREG-lin, L-GREG and the Standard variance estimator under the SRSWOR design. The πPS case is considered in Chapter 5.3.

5.2.1 General setting

The fixed and finite population called Population 1 is generated as follows. First, the sizes of 10 minor, medium sized and major domains are constructed as follows:

$$N^{(d)} = \begin{cases} 250 + \lfloor 50 \cdot \underline{z}^{(d)} \rfloor, & d = 1,2,\ldots,10, \\ 500 + \lfloor 100 \cdot \underline{z}^{(d)} \rfloor, & d = 11,12,\ldots,20, \\ 1250 + \lfloor 150 \cdot \underline{z}^{(d)} \rfloor, & d = 21,22,\ldots,30, \end{cases} \tag{5.5}$$

where $\underline{z}^{(d)} \sim Uni(-1,1)$ and $Uni(a,b)$ is the uniform distribution over the interval $[a,b]$. The resulting population size N = 20,643. Auxiliary variables are generated as follows:

$$\forall i \in U^{(d)}, d = 1,2,\ldots,30: \begin{cases} \underline{x}_{i1} \sim Uni(-0,1) \\ \underline{x}_{i2}^{(d)} \sim Be\left(\underline{t}^{(d)}\right), \quad \underline{t}^{(d)} \sim Uni(0.4, 0.8), \end{cases} \tag{5.6}$$

where $Be(c)$ is the Bernoulli distribution with probability c. Thus, the continuous variable x_1 has the same distribution for every domain, but the binary variable x_2 has its own distribution for every domain. The study variables are generated using the following process: First, a unit-specific linear predictor is defined as

$$\underline{\eta}_i = \beta_0 + \left(\beta_1 + \underline{u}_1^{(d)}\right)\underline{x}_{i1} + \beta_2 \underline{x}_{i2}, \quad \underline{u}_1^{(d)} \sim N(0, 3). \tag{5.7}$$

The intercept β_0 and slope β_2 of x_2 are common over domains, but the coefficient of x_1 has a domain-specific random term. The motivation for the random

term is that is makes it easy to generate variation over domains. Next, three binary study variables are generated from the Bernoulli distribution whose probabilities are constructed using logit, probit and cll links:

$$\forall i \in U: \begin{cases} \underline{y}_{i1}^{(d)} \sim Be\left(\underline{p}_{i1}^{(d)}\right), \quad \underline{p}_{i1}^{(d)} = \exp\left(\underline{\eta}_{i1}\right)\left[1+\exp\left(\underline{\eta}_{i1}\right)\right]^{-1}, \quad (\beta_0, \beta_1, \beta_2) = (1, 10, -10), \\ \underline{y}_{i2}^{(d)} \sim Be\left(\underline{p}_{i2}^{(d)}\right), \quad \underline{p}_{i2}^{(d)} = \Phi\left(\underline{\eta}_{i2}\right), \quad (\beta_0, \beta_1, \beta_2) = (2, -10, 2), \\ \underline{y}_{i3}^{(d)} \sim Be\left(\underline{p}_{i3}^{(d)}\right), \quad \underline{p}_{i3}^{(d)} = 1-\exp\left[-\exp\left(\underline{\eta}_{i3}\right)\right]^{-1}, \quad (\beta_0, \beta_1, \beta_2) = (2, -10, 2). \end{cases} \tag{5.8}$$

In (5.8), $\underline{\eta}_i$ of (5.7) takes three different forms $\underline{\eta}_{i1}$, $\underline{\eta}_{i2}$ and $\underline{\eta}_{i3}$ depending on the parameter values $(\beta_0, \beta_1, \beta_2)$. Table 5.1 shows the totals and proportions of y_1, y_2 and y_3 and expected sample sizes by domain for samples n=5,000, 2,000 and 1,000 for Population 1.

Table 5.1 Domains, totals and proportions of study variables and expected domain sample sizes for Population 1.

Domain type	Domain number	Domain size	Totals of y			Expected sample size		
			y1	y2	y3	n=5000	n=2000	n=1000
Minor	1	289	190	97	109	70	28	14
	2	243	111	105	117	59	24	12
	3	228	109	142	155	55	22	11
	4	201	82	72	84	49	19	10
	5	296	153	125	143	72	29	14
	6	296	167	111	130	72	29	14
	7	274	174	131	156	66	27	13
	8	236	100	108	124	57	23	11
	9	285	117	111	121	69	28	14
	10	252	134	108	126	61	24	12
Medium	11	579	375	193	227	140	56	28
	12	599	208	121	145	145	58	29
	13	569	274	202	237	138	55	28
	14	542	318	355	402	131	53	26
	15	405	251	379	385	98	39	20
	16	599	172	172	206	145	58	29
	17	543	183	140	168	132	53	26
	18	524	303	124	142	127	51	25
	19	511	220	181	210	124	50	25
	20	555	314	115	141	134	54	27
Major	21	1369	878	452	540	332	133	66
	22	1164	519	495	559	282	113	56
	23	1267	504	341	405	307	123	61
	24	1142	495	470	558	277	111	55
	25	1290	801	773	876	312	125	62
	26	1316	543	516	619	319	128	64
	27	1381	922	538	641	334	134	67
	28	1302	786	509	580	315	126	63
	29	1111	725	329	409	269	108	54
	30	1275	729	481	573	309	124	62
Population	31	20643	10857	7996	9288	5000	2000	1000

The parameters of the population generating process are chosen so that complete auxiliary information allows constructing strong assisting models (strong in the sense that predicted p_is are near the borders of the unit interval). Figures 5.1–5.3 show the histograms of p_i for selected domains for all study variables. The domains are representative in the sense that for every domain, most of the p_is were close to either one or zero.

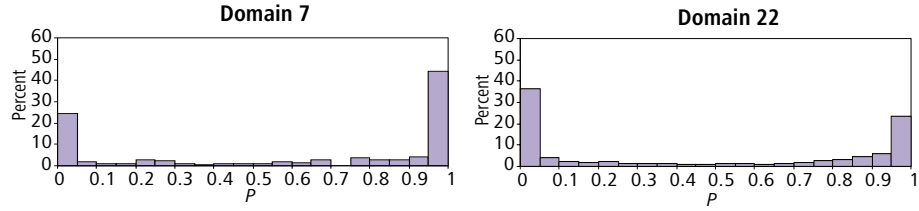


Figure 5.1 Histograms of $p = P(\underline{y}_1 = 1)$ in selected domains of Population 1.

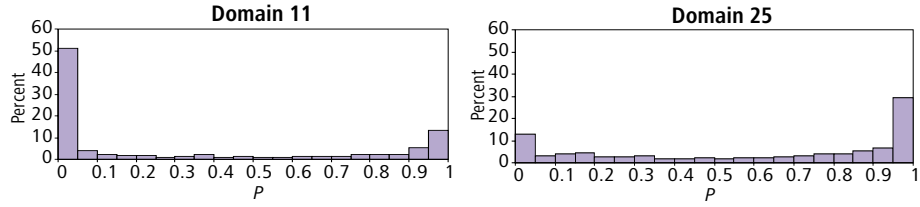


Figure 5.2 Histograms of $p = P(\underline{y}_2 = 1)$ in selected domains of Population 1.

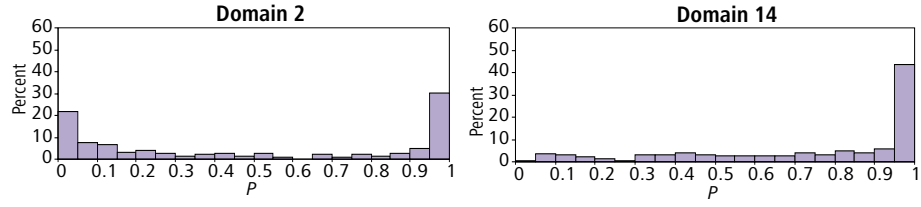


Figure 5.3 Histograms of $p = P(\underline{y}_3 = 1)$ in selected domains of Population 1.

5.2.2 Estimators

From Population 1, K=1,000 independent SRSWOR samples with sample sizes 1,000, 2,000 and 5,000 were drawn. From each sample, totals of y_1, y_2 and y_3 in the population and in the domains were estimated using various GREG estimators. Standard variance estimates and variance approximations (presented in Chapter 4) were also calculated.

By examining GREG estimators with different assisting models, we get an opportunity to study the effect of assisting model strength. Varying the assisting model is also a way to control the amount of auxiliary information that goes into the estimator. The estimators that were used are grouped into sets according to the model strength (Table 5.2). The estimators in sets 1, 3, 5 and 7 have common intercepts. A common intercept was used also when generating the study variables. Estimators in sets 2, 4 and 6 and 8 have domain specific intercepts, so their linear predictors are overspecified compared with (5.7), which, together with (5.8), defines the procedure for creating the y-variables. Estimators in sets 1 and 2 have weak models, since they include only intercepts and a common slope for x_1. The strength of the model in sets 3 and 4 is moderate, since these models also have a domain-specific slope for x_1 (in the population generating process, there was also a domain-specific slope).

Estimators in sets 5–8 have strong models. This is because the auxiliary variable x_2, which had a large role in generating the probabilities $p = P(\underline{y} = 1)$, is included in these models. The estimators in sets 7 and 8 have very strong models, since they use all the auxiliary information that was used in the population generating process.

Only set 7 has estimators with the correct assisting model formulation (to be correct, both the linear predictor and functional form need to be as they are in

Table 5.2 Estimators used in the simulations.

Set	Estimator	Functional form of the model	Linear predictor η*	Description of the model	Over-specified
1	GREG-lin-1	Linear	$\beta_0 + \beta_1 x_1$	Weak	
	GREG-log-1	Logistic	$\beta_0 + \beta_1 x_1$	Weak	
2	GREG-lin-2	Linear	$\beta_0^{(d)} + \beta_1 x_1$	Weak	Yes
	GREG-log-2	Logistic	$\beta_0^{(d)} + \beta_1 x_1$	Weak	Yes
3	GREG-lin-3	Linear	$\beta_0 + \beta_1^{(d)} x_1$	Moderate	
	GREG-log-3	Logistic	$\beta_0 + \beta_1^{(d)} x_1$	Moderate	
4	GREG-lin-4	Linear	$\beta_0^{(d)} + \beta_1^{(d)} x_1$	Moderate	Yes
	GREG-log-4	Logistic	$\beta_0^{(d)} + \beta_1^{(d)} x_1$	Moderate	Yes
5	GREG-lin-5	Linear	$\beta_0 + \beta_2^{(d)} x_2$	Strong	
	GREG-log-5	Logistic	$\beta_0 + \beta_2^{(d)} x_2$	Strong	
6	GREG-lin-6	Linear	$\beta_0^{(d)} + \beta_2^{(d)} x_2$	Strong	Yes
	GREG-log-6	Logistic	$\beta_0^{(d)} + \beta_2^{(d)} x_2$	Strong	Yes
7	GREG-lin-7	Linear	$\beta_0 + \beta_1^{(d)} x_1 + \beta_2 x_2$	Very strong	
	GREG-log-7	Logistic	$\beta_0 + \beta_1^{(d)} x_1 + \beta_2 x_2$	Very strong	
	GREG-prob-7	Probit	$\beta_0 + \beta_1^{(d)} x_1 + \beta_2 x_2$	Very strong	
	GREG-cll-7	Cll	$\beta_0 + \beta_1^{(d)} x_1 + \beta_2 x_2$	Very strong	
8	GREG-lin-8	Linear	$\beta_0^{(d)} + \beta_1^{(d)} x_1 + \beta_2 x_2$	Very strong	Yes
	GREG-log-8	Logistic	$\beta_0^{(d)} + \beta_1^{(d)} x_1 + \beta_2 x_2$	Very strong	Yes

* The linear predictor of the population generating process is $\beta_0 + \beta_1^{(d)} x_1 + \beta_2 x_2$

the population generating process): GREG-log-7 is the correct model for y_1, GREG-prob-7 is the correct model for y_2 and GREG-cll-7 is the correct model for y_3. In sets 5 and 6, the models are missing x_1 and in set 8, the model is overspecified because of the unnecessary domain specific intercepts. The only estimators that are direct are estimators in sets 4 and 6, since all other estimators have model parameters that are common over domains.

The list of assisting models does not cover all the possible models that one could formulate, given the auxiliary information. For instance, interactions, transformations of variables and domain-specific slopes $\beta_2^{(d)}$ are ignored. These, however, would only make the models overspecified. An attempt to strike a balance between completeness of the list and importance of assisting models has been made, and including more overspecified models (for example, models which include $\beta_2^{(d)}$) might not contribute very much to the study.

All the estimators listed in Table 5.2 were not applied to all three study variables. In order to study the differences of various logistic-type GREG estimators under a correct and incorrect specification of the link function, onlyestimators in set 7 were applied to all three study variables. It turned out that differences are small. Therefore, when using the whole list of estimators to study the difference between GREG-lin and GREG-log, only the study variable y_1 (which was generated using the logit link) was used.

Different estimators have different amounts of auxiliary information. We expect that at least in sets 7 and 8, L-GREG estimators should be more accurate than the corresponding GREG-lin estimators, since these estimators have very strong models. The possible accuracy gain should decrease as weaker models are used.

The models were fitted using the SAS v9.1 procedure Reg for the linear fixed effects models (weighted least squares estimation with sweep algorithm to invert matrices) and procedure Logistic for the logistic-type models (PML estimation with Newton-Raphson algorithm). These estimation procedures are described in Chapters 4.4 and 4.5; for details of the computational algorithms, see SAS OnlineDoc (1999).

5.2.3 Accuracy measures

To measure the accuracy of estimators, absolute relative bias (ARB) and standard error (SE) were calculated as

$$ARB_{MC}\left(\underline{\hat{T}}^{(d)}\right) = \frac{\left|\hat{B}_{MC}\left(\underline{\hat{T}}^{(d)}\right)\right|}{T^{(d)}}, \text{ where} \tag{5.9}$$

$$\hat{B}_{MC}\left(\underline{\hat{T}}^{(d)}\right) = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{T}^{(d)(k)} - T^{(d)}\right), \text{ and}$$

$$SE_{MC}\left(\hat{\underline{T}}^{(d)}\right) = \sqrt{\hat{V}_{MC}\left(\hat{\underline{T}}^{(d)}\right)}\text{, where}$$

$$\hat{V}_{MC}\left(\hat{\underline{T}}^{(d)}\right) = \frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{T}^{(d)(k)} - (1/K)\sum_{k=1}^{K}\hat{T}^{(d)(k)}\right)^2, \tag{5.10}$$

where $\hat{T}^{(d)(k)}$ is the estimate calculated from the kth sample. The subscript MC is used in order to make explicit the fact that these statistics are calculated from the Monte Carlo simulation.

Since the aim is to compare L-GREG with GREG-lin, standard error in (5.10) itself is not of direct interest. What is interesting is the change in standard error when changing the linear model to a logistic-type model while keeping the amount of auxiliary information and formulation of the linear predictor fixed. This possible gain in accuracy is measured by the function effect, defined in Chapter 4.5.2 as

$$FE_{MC}\left(\hat{\underline{T}}_{j,GREG}\right) = \sqrt{\frac{V_{MC}\left(\hat{\underline{T}}_{j,GREG-lin}\right)}{V_{MC}\left(\hat{\underline{T}}_{j,GREG}\right)}}. \tag{5.11}$$

We use (5.10) to estimate the function effect (5.11). Monte Carlo estimates of ARB, SE and FE are averaged over domain types. For example, for minor domains,

$$MARB_{MC} = \frac{\sum_{d=1}^{10} ARB_{MC}\left(\hat{\underline{T}}^{(d)}\right)}{10}, \tag{5.12}$$

$$ASE_{MC} = \frac{\sum_{d=1}^{10} SE_{MC}\left(\hat{\underline{T}}^{(d)}\right)}{10}\text{, and} \tag{5.13}$$

$$MFE_{MC} = \frac{\sum_{d=1}^{10} FE_{MC}\left(\hat{\underline{T}}^{(d)}\right)}{10}. \tag{5.14}$$

In (5.13), we denote the mean standard error by *ASE* (=average standard error) to avoid confusion with *MSE* (=mean squared error).

We also study the goodness of the Standard variance approximation and estimator which are

$$V_A\left(\hat{\underline{T}}^{(d)}_{j,GREG}\right) = \frac{N^2(1-f)}{n}\frac{1}{N-1}\sum_{i\in U}\left(E_{ij}^{(d)} - \bar{E}_j^{(d)}\right)^2. \tag{5.15}$$

$$\hat{V}_S\left(\hat{\underline{T}}^{(d)}_{j,GREG}\right) = \frac{N^2(1-f)}{n}\frac{1}{N-1}\sum_{i\in U}\underline{I}_i\left(\underline{e}_{ij}^{(d)} - \bar{\underline{e}}_j^{(d)}\right)^2, \tag{5.16}$$

The emphasis is, of course, on the goodness of the estimator (5.16). This can be measured in numerous ways. For example, the bias and variance could be estimated in the same manner as they are estimated for the GREG estimators. However, the main statistic we use to describe the goodness of (5.16) is the coverage rate. This is because often in practice, variance is used mainly to calculate confidence intervals. The coverage rate is the probability that the true parameter value is inside the estimated confidence interval, given the confidence level. Formally, the coverage rate CR of a variance estimator $\hat{V}\left(\hat{\underline{T}}^{(d)}\right)$ is

$$CR\left[\hat{V}\left(\hat{\underline{T}}^{(d)}\right)\right] = P\left[T^{(d)} \in \left(\hat{\underline{T}}^{(d)} - t_{1-\alpha/2}\sqrt{\hat{V}\left(\hat{\underline{T}}^{(d)}\right)}, \hat{\underline{T}}^{(d)} + t_{1-\alpha/2}\sqrt{\hat{V}\left(\hat{\underline{T}}^{(d)}\right)}\right)\right], \quad (5.17)$$

where $t_{1-\alpha/2}$ is the $(1-\alpha/2)$ th quantile of the standard normal distribution and the probability in (5.17) is evaluated with respect to the sampling distribution. The Monte Carlo estimator for (5.17) is the proportion of estimated confidence intervals that contain the true parameter value. We use the 0.95 confidence level, so the estimator is

$$CR_{MC}\left[\hat{V}_S\left(\hat{\underline{T}}^{(d)}\right)\right] = \frac{\sum_{k=1}^{K} \underline{I}^{(k)}}{K}, \text{ where} \quad (5.18)$$

$$\underline{I}^{(k)} = \begin{cases} 1, \text{if } T^{(d)} \in \left(\hat{T}^{(d)(k)} - 1.96 \cdot \sqrt{\hat{V}_S\left(\hat{T}^{(d)(k)}\right)},\ \hat{T}^{(d)(k)} + 1.96 \cdot \sqrt{\hat{V}_S\left(\hat{T}^{(d)(k)}\right)}\right), \\ 0 \text{ otherwise.} \end{cases}$$

Coverage rates are averaged over domain types. For example for minor domains, the mean coverage rate is

$$MCR_{MC} = \frac{\sum_{d=1}^{10} CR_{MC}\left[\hat{V}_S\left(\hat{\underline{T}}^{(d)}\right)\right]}{10}. \quad (5.19)$$

With K=1,000 samples, acceptable mean coverage rates are in the interval [93.0, 97.0]. It turned out that quite often, the variance estimator failed in terms of the coverage rate: the estimated confidence intervals did not capture the true parameter value as often as would be desirable. To study the reasons for the failure, several statistics were calculated:

First it should be noted that our variance estimator does not estimate variance, it estimates *approximate* variance

$$V_A\left(\hat{\underline{T}}^{(d)}_{j,GREG}\right) = \frac{N^2(1-f)}{n}\frac{1}{N-1}\sum_{i\in U}\left(E_{ij}^{(d)} - \overline{E}_j^{(d)}\right)^2. \quad (5.20)$$

This approximation, in turn, was obtained by linearisation, where certain terms were ignored (Chapter 4.3). So if the estimator fails, it is due to one of the following reasons:

1) The variance estimator is accurate for the approximation, but the approximation is not good.
2) The variance approximation is accurate, but the variance estimator does not estimate it accurately.
3) Both the variance approximation and variance estimator are inaccurate.

In order to see whether the main reason for the variance estimator failing is 1), 2) or 3), we decompose the Standard variance estimator's total relative bias TRB

$$TRB^{(d)} = \frac{1}{SE_{MC}^{(d)}}\left(\overline{SE}_S^{(d)} - SE_{MC}^{(d)}\right) \tag{5.21}$$

into relative approximation error RAE and relative estimation error REE:

$$\begin{aligned} TRB^{(d)} &= \frac{1}{SE_{MC}^{(d)}}\left[\left(\overline{SE}_S^{(d)} - SE_A^{(d)}\right) + \left(SE_A^{(d)} - SE_{MC}^{(d)}\right)\right] \\ &= REE^{(d)} + RAE^{(d)} \ . \end{aligned} \tag{5.22}$$

In (5.21) and (5.22), $SE_{MC}^{(d)}$ is the simulated standard error $\overline{SE}_S^{(d)}$ is the mean of the standard error estimates and $SE_A^{(d)}$ is the approximate standard error (square root of (5.20)). Statistics RAE and REE averaged over domain types are denoted by MRAE (mean RAE) and MREE (mean REE). For instance for minor domains, these are calculated as

$$MRAE = \frac{\sum_{d=1}^{10} RAE^{(d)}}{10}, \text{ where } \quad RAE^{(d)} = \frac{SE_A^{(d)} - SE_{MC}^{(d)}}{SE_{MC}^{(d)}}, \text{ and} \tag{5.23}$$

$$MREE = \frac{\sum_{d=1}^{10} REE^{(d)}}{10}, \text{ where } \quad REE^{(d)} = \frac{\overline{SE}_S^{(d)} - SE_A^{(d)}}{SE_{MC}^{(d)}} \ . \tag{5.24}$$

MRAE describes the mean approximation error and MREE describes the bias of the variance estimator with respect to the approximation it should estimate.

It also turned out that the bias of the variance estimator was not the only problem in variance estimation; the variation of variance estimates was also often quite large. To measure the total accuracy of the variance estimator, we calculated the mean relative root mean square errors for the Standard error estimators:

$$MRRMSE_{MC} = \frac{\sum_{d=1}^{10} RRMSE_{MC}\left[SE_S\left(\hat{\underline{T}}^{(d)}\right)\right]}{10}, \text{ where} \tag{5.25}$$

$$RRMSE_{MC}\left[SE_S\left(\hat{\underline{T}}^{(d)}\right)\right] = \frac{\sqrt{\hat{V}_{MC}\left[SE_S\left(\hat{\underline{T}}^{(d)}\right)\right] + \hat{B}^2_{MC}\left[SE_S\left(\hat{\underline{T}}^{(d)}\right)\right]}}{SE_{MC}\left(\hat{\underline{T}}^{(d)}\right)} \text{ and } \quad (5.26)$$

$$\hat{B}_{MC}\left[SE_S\left(\hat{\underline{T}}^{(d)}\right)\right] = \overline{SE}_S\left(\hat{\underline{T}}^{(d)}\right) - SE_{MC}\left(\hat{\underline{T}}^{(d)}\right). \quad (5.27)$$

The formula (5.25) is for minor domains; medium and major domains were calculated in a similar manner. Note that the accuracy measures Mean Relative Approximation Error MRAE (5.23), Mean Relative Estimation Error MREE (5.24) and Mean Relative Root Mean Square Error MRRMSE (5.25) are all measured in terms of standard error, not in terms of variance.

5.2.4 Results

The results are divided into three sections. First in Chapter 5.2.4.1, we focus on the different L-GREG estimators. It turns out that the differences between them are minor. Therefore in the following, out of the L-GREG family, only GREG-log is used. Second, in Chapter 5.2.4.2, we compare GREG-log and GREG-lin. Third, we study the accuracy of the Standard variance estimator in Chapter 5.2.4.3.

5.2.4.1 GREG-log, GREG-prob and GREG-cll: are there any differences?

We start by considering the differences between three L-GREG estimators: GREG-log, GREG-prob and GREG-cll. The log-log link is not included, because it is the mirror image of cll. As a baseline, we use the GREG-lin estimator. The estimators used in this Chapter are the estimators of Set 7 (see Table 5.2). The research question is whether any of the three L-GREG estimators is more robust than the others to model misspecification. To study this, three study variables were generated using logit, probit and cll links (see Chapter 5.2.1) and all estimators from set 7 were used to estimate totals of these study variables.

Table 5.3 presents the accuracy measures MARB (%), ASE, MFE and MCR (%) for the three L-GREG estimators and GREG-lin for the cases where the population generating link is logit, probit and cll (see (5.8)).

The MARB column shows that all the estimators are approximately unbiased. The largest observed MARB values are less than one per cent, and for all study variables and all estimators, MARB decreases as domain sample size increases.

Since biases are negligible, the ASE and MFE columns summarize the accuracy difference between GREG-log and GREG-lin. Let us first consider the case where the true population generating link is logit. Here it might have been expected that the smallest standard errors had been observed for the GREG-log estimator. However, the results do not indicate any significant differences between the three different L-GREG estimators. It is fair to say that the three L-GREG estimators are equally accurate in terms of bias and variance.

The differences between L-GREG estimators in the cases where the true link was probit and cll are also small: the largest observed differences in MFE (describing

Table 5.3 MARB, ASE, MFE and MCR for GREG estimators in set 7, n=2,000

True link	Domain type	Expected sample size	Estimator	Accuracy of GREG estimators			Accuracy of var. estimator
				MARB %	ASE	MFE	MCR, %
Logit	Population	2000	GREG-lin-7	0.1	129.5		92.3
			GREG-log-7	0.0	107.3	1.21	92.5
			GREG-prob-7	0.0	107.7	1.20	92.1
			GREG-cll-7	0.0	107.9	1.20	92.0
	Major	121	GREG-lin-7	0.2	30.8		93.4
			GREG-log-7	0.2	25.0	1.23	92.9
			GREG-prob-7	0.1	25.1	1.23	92.6
			GREG-cll-7	0.1	25.2	1.22	92.8
	Med	48	GREG-lin-7	0.2	17.2		93.3
			GREG-log-7	0.2	14.5	1.18	86.7
			GREG-prob-7	0.2	14.6	1.18	86.1
			GREG-cll-7	0.2	14.8	1.16	86.6
	Minor	24	GREG-lin-7	0.7	14.7		88.8
			GREG-log-7	0.7	12.5	1.17	79.8
			GREG-prob-7	0.3	12.6	1.16	79.3
			GREG-cll-7	0.3	12.7	1.15	79.9
			Estimator	**MARB**	**ASE**	**MFE**	**MCR**
Probit	Population	2000	GREG-lin-7	0.1	135.6		95.3
			GREG-log-7	0.0	113.1	1.20	94.7
			GREG-prob-7	0.0	112.9	1.20	94.9
			GREG-cll-7	0.0	114.4	1.18	94.6
	Major	121	GREG-lin-7	0.2	34.2		94.2
			GREG-log-7	0.2	28.8	1.19	93.3
			GREG-prob-7	0.2	28.8	1.19	93.3
			GREG-cll-7	0.2	29.1	1.18	93.4
	Med	48	GREG-lin-7	0.6	22.5		92.8
			GREG-log-7	0.3	17.1	1.31	89.2
			GREG-prob-7	0.2	17.1	1.31	89.3
			GREG-cll-7	0.3	17.2	1.31	89.1
	Minor	24	GREG-lin-7	0.5	15.7		92.3
			GREG-log-7	0.6	14.0	1.12	85.8
			GREG-prob-7	0.6	14.0	1.12	86.0
			GREG-cll-7	0.6	14.1	1.11	85.7
			Estimator	**MARB**	**ASE**	**MFE**	**MCR**
Cll	Population	2000	GREG-lin-7	0.0	143.4		94.2
			GREG-log-7	0.1	129.6	1.11	93.7
			GREG-prob-7	0.1	129.8	1.10	94.0
			GREG-cll-7	0.1	128.3	1.12	94.2
	Major	121	GREG-lin-7	0.1	35.0		94.7
			GREG-log-7	0.1	31.7	1.10	93.6
			GREG-prob-7	0.1	31.8	1.10	93.7
			GREG-cll-7	0.1	31.5	1.11	93.7
	Med	48	GREG-lin-7	0.3	22.5		92.8
			GREG-log-7	0.3	18.6	1.21	90.0
			GREG-prob-7	0.3	18.6	1.21	90.4
			GREG-cll-7	0.3	18.5	1.22	90.2
	Minor	24	GREG-lin-7	0.3	16.2		91.8
			GREG-log-7	0.2	14.8	1.09	87.4
			GREG-prob-7	0.2	14.8	1.09	87.7
			GREG-cll-7	0.2	14.8	1.09	87.6

the average accuracy difference with respect to GREG-lin), for example, are less than two per cent. So it is fair to say that the estimators are equally accurate.

If L-GREG estimators are equally accurate, all GREG estimators are not. When comparing the accuracy of L-GREG estimators with GREG-lin, it can be seen that L-GREG estimators are far more accurate. The difference measured by MFE ranges from 9% to 31% in favour of L-GREG estimators. Thus, changing the model from linear to logistic-type does result in accuracy gain. However, it is fair to say that in the setting studied the accuracy gain does not to depend on domain type (minor, medium, major). In Chapter 5.2.4.2, we consider the difference between GREG-lin and L-GREG estimators in more detail.

The MCR (mean coverage rate) column describes the accuracy of the variance estimator. If the variance estimator was accurate, mean coverage rates should be close to 95.0%. The observed MCRs are often far below the expected 95.0% for L-GREG estimators, in some cases even below 80% (true link logit, minor domains). Thus it seems that the Standard variance estimators underestimate the true variance of L-GREG estimators, especially in minor and medium domains. But as the domain sample size increases, the performance of the Standard variance estimator improves. For GREG-lin estimators, the Standard variance estimator performs much better than for L-GREG estimators.

So the three L-GREG estimators clearly outperform GREG-lin, but differences within L-GREG estimators are small. This may be due to the fact that there were continuous variables in the models. From this it follows that response variables are not grouped but are binary. When grouping is done, we observe empirical probabilities in every group and in such cases, the form of the non-linear link may be more important. However, by grouping one always loses information, so it cannot be recommended.

According to these simulations, GREG-log, GREG-prob and GREG-cll yield similar results. Which link to use, then? The choice may be based on practical considerations, such as what link is available in the software that is being used. In the following, we use only the logit link. This is because i) the logit link is the canonical link, ii) use of the logit link makes it easy to construct odds ratios which are often of interest, iii) the probit link yields almost identical results, but the logistic transformation is computationally easier than the probit, and iv) use of asymmetric links, such as cll or log-log, might be difficult to justify in practice.

5.2.4.2 GREG-lin and GREG-log: when are they different?

Next we consider in more detail the differences between GREG-lin and GREG-log. Table 5.4 presents results for estimators in sets 1–4 (weak and moderate models), n=2,000. The differences between GREG-lin and GREG-log seem small; both estimators are approximately unbiased in all cases and average standard errors are almost equivalent (the largest MFE, summarising the accuracy difference between GREG-lin and GREG-log, is 1.01). Also, variance estimation is equally accurate for both GREG-lin and GREG-log. For sample sizes n=5,000 and n=1,000 the results were so similar that there is no need to present them. Thus it is fair to say that there are no differences between GREG-lin and GREG-log when the model is weak or moderate.

Table 5.4 MARB, ASE, MFE and MCR for GREG-lin and GREG-log estimators in sets 1–4 by estimator set, estimator and domain type. Total sample size n=2,000.

n=2000 Estimator		Domain type	Expected sample size	Accuracy of GREG estimators			Accuracy of var. estimator
Set	Link			MARB %	ASE	MFE	MCR, %
1	lin	Population	2000	0.0	224.0		94.8
		Major	121	0.3	50.8		94.8
		Med	48	0.4	34.5		95.4
		Minor	24	0.8	22.2		94.8
	log	Population	2000	0.0	224.2	1.00	94.7
		Major	121	0.3	50.8	1.00	94.8
		Med	48	0.4	34.4	1.00	95.4
		Minor	24	0.8	22.2	1.00	94.8
2	lin	Population	2000	0.0	217.5		93.8
		Major	121	0.3	50.3		94.3
		Med	48	0.5	33.4		93.6
		Minor	24	0.9	22.3		92.1
	log	Population	2000	0.0	217.4	1.00	93.9
		Major	121	0.3	50.4	1.00	94.3
		Med	48	0.5	33.3	1.00	93.7
		Minor	24	0.9	22.3	1.00	92.0
3	lin	Population	2000	0.1	216.9		94.1
		Major	121	0.3	50.0		94.6
		Med	48	0.5	33.1		93.8
		Minor	24	1.0	22.5		92.2
	log	Population	2000	0.0	217.2	1.00	93.8
		Major	121	0.4	50.1	1.00	94.5
		Med	48	0.5	33.1	1.00	93.8
		Minor	24	1.0	22.6	1.00	92.2
4	lin	Population	2000	0.2	216.9		92.9
		Major	121	0.3	50.4		94.1
		Med	48	0.5	33.0		93.1
		Minor	24	1.0	22.8		90.4
	log	Population	2000	0.1	216.5	1.00	93.3
		Major	121	0.3	50.3	1.00	94.2
		Med	48	0.5	32.9	1.00	93.0
		Minor	24	1.1	22.6	1.01	90.2

Table 5.4 also shows that estimators with overfitted models (estimators in sets 2 and 4) and estimators with non-overfitted models (estimators in sets 1 and 3) have similar average standard errors. Thus the accuracy of the GREG estimator is not significantly affected by overfitting, as long as the model is weak or moderate. The performance of the Standard variance estimator, however, does depend on overfitting: Coverage rates drop below 93.0% in medium and minor domains especially if the model is overfitted (sets 2 and 4). Also, as the model gets stronger (from weak, sets 1 and 2, to moderate, sets 3 and 4), MCRs get lower. For example for *set=1*, GREG-log and minor domains, MCR is 94.8%. For *set 4*, GREG-log and minor domains, MCR is 90.2%.

Tables 5.5–5.7 present corresponding statistics for GREG-lin and GREG-log estimators in sets 5–8 (strong and very strong models) for sample sizes *n*=5,000, 2,000 and 1,000. The MARB columns of these tables show that all estimators are approximately unbiased, except in one case: For *n*=1,000, model set (ver strong and overfitted models) and minor domains, the mean absolute relative bias is 4.5% for GREG-log. This is a very extreme situation, since the expected sample size is only 12 per domain and two domain-specific parameters are being estimated. In all other cases, biases are less than two per cent. This bias, which is due to inaccurate estimation of model parameters, is discussed more in Appendix II.

Table 5.5 MARB, ASE, MFE and MCR for GREG-lin and GREG-log estimators in sets 5–8 by estimator set, estimator and domain type. Total sample size n=5,000.

n=5000 Estimator		Domain type	Expected sample size	Accuracy of GREG estimators			Accuracy of var. estimator
Set	Link			MARB %	ASE	MFE	MCR, %
5	lin	Population	5000	0.0	84.7		94.8
		Major	303	0.1	20.4		94.9
		Med	121	0.2	12.3		94.3
		Minor	61	0.3	10.4		93.4
	log	Population	5000	0.0	84.7	1.00	94.8
		Major	303	0.1	20.4	1.00	94.9
		Med	121	0.2	12.3	1.00	94.3
		Minor	61	0.3	10.4	1.00	93.4
6	lin	Population	5000	0.0	82.5		94.9
		Major	303	0.1	20.3		94.7
		Med	121	0.2	11.3		93.7
		Minor	61	0.3	10.3		92.3
	log	Population	5000	0.0	82.2	1.00	94.5
		Major	303	0.1	20.3	1.00	94.6
		Med	121	0.2	10.9	1.03	93.1
		Minor	61	0.3	10.3	1.00	91.8
7	lin	Population	5000	0.0	68.9		95.5
		Major	303	0.1	17.6		94.8
		Med	121	0.1	10.0		94.3
		Minor	61	0.3	8.5		92.6
	log	Population	5000	0.0	56.8	1.21	93.8
		Major	303	0.1	14.3	1.23	94.3
		Med	121	0.1	8.2	1.21	92.4
		Minor	61	0.2	6.9	1.23	90.2
8	lin	Population	5000	0.0	68.7		95.5
		Major	303	0.1	17.6		94.7
		Med	121	0.2	9.9		93.8
		Minor	61	0.2	8.5		92.2
	log	Population	5000	0.0	57.9	1.19	93.7
		Major	303	0.1	14.6	1.21	94.1
		Med	121	0.1	8.3	1.19	91.5
		Minor	61	0.2	7.1	1.20	89.1

Table 5.6 MARB, ASE, MFE and MCR for GREG-lin and GREG-log estimators in sets 5–8 by estimator set, estimator and domain type. Total sample size n=2,000.

n=2000 Estimator		Domain type	Expected sample size	Accuracy of GREG estimators			Accuracy of var. estimator
Set	Link			MARB %	ASE	MFE	MCR, %
5	lin	Population	2000	0.0	152.0		93.8
		Major	121	0.3	36.1		94.2
		Med	48	0.3	20.8		94.3
		Minor	24	0.7	17.2		93.4
	log	Population	2000	0.0	152.0	1.00	93.8
		Major	121	0.3	36.1	1.00	94.2
		Med	48	0.3	20.8	1.00	94.3
		Minor	24	0.7	17.2	1.00	93.4
6	lin	Population	2000	0.0	148.1		93.3
		Major	121	0.3	36.1		93.5
		Med	48	0.3	19.4		92.8
		Minor	24	0.6	17.3		90.1
	log	Population	2000	0.0	147.5	1.00	93.3
		Major	121	0.3	36.2	1.00	93.0
		Med	48	0.3	18.7	1.03	90.5
		Minor	24	0.7	17.2	1.00	88.1
7	lin	Population	2000	0.1	129.5		92.3
		Major	121	0.2	30.8		93.4
		Med	48	0.2	17.2		93.3
		Minor	24	0.7	14.7		88.8
	log	Population	2000	0.0	107.3	1.21	92.5
		Major	121	0.2	25.0	1.23	92.9
		Med	48	0.2	14.5	1.18	86.7
		Minor	24	0.7	12.5	1.17	79.8
8	lin	Population	2000	0.2	130.0		92.0
		Major	121	0.3	31.0		93.2
		Med	48	0.3	17.1		92.5
		Minor	24	0.7	14.8		88.1
	log	Population	2000	0.1	109.2	1.19	91.2
		Major	121	0.2	26.3	1.18	92.3
		Med	48	0.4	14.9	1.15	80.6
		Minor	24	1.0	13.3	1.11	75.4

The MFE columns of Tables 5.5–5.7 show that for estimators in set 5, GREG-lin and GREG-log were practically equally accurate. This is not surprising, since models in set 5 have only categorical covariates and the model parameters are estimated using PML. If interaction terms had been included, the estimators would have been exactly equivalent (see Chapter 4.5). In set 6, GREG-lin and GREG-log were also roughly equally accurate. But MCRs were much lower for GREG-log than for GREG-lin; the difference is largest for n=1,000 (Table 5.7). In set 6, an exception to the general pattern where GREG-lin and GREG-log are equally accurate is the case n=1,000 and minor domains: there GREG-lin is slightly more accurate than GREG-log.

Table 5.7 MARB, ASE, MFE and MCR for GREG-lin and GREG-log estimators in sets 5–8 by estimator set, estimator and domain type. Total sample size n=1,000.

n=1000 Estimator		Domain type	Expected sample size	Accuracy of GREG estimators			Accuracy of var. estimator
Set	Link			MARB %	ASE	MFE	MCR, %
5	lin	Population	1000	0.0	210.6		94.7
		Major	61	0.2	51.5		94.4
		Med	24	0.4	30.6		92.6
		Minor	12	0.4	25.3		89.5
	log	Population	1000	0.0	210.6	1.00	94.7
		Major	61	0.2	51.5	1.00	94.4
		Med	24	0.4	30.6	1.00	92.6
		Minor	12	0.4	25.3	1.00	89.5
6	lin	Population	1000	0.0	211.6		93.8
		Major	61	0.2	51.8		93.1
		Med	24	0.3	28.9		90.3
		Minor	12	0.4	26.3		83.5
	log	Population	1000	0.1	212.9	0.99	92.8
		Major	61	0.2	51.9	1.00	92.5
		Med	24	0.3	28.2	1.02	86.1
		Minor	12	0.5	27.4	0.96	79.1
7	lin	Population	1000	0.1	176.6		93.8
		Major	61	0.2	44.5		92.9
		Med	24	0.3	25.7		90.6
		Minor	12	0.6	22.4		82.6
	log	Population	1000	0.0	155.5	1.14	90.5
		Major	61	0.1	37.2	1.20	89.8
		Med	24	0.3	22.6	1.14	78.6
		Minor	12	0.6	22.6	0.99	62.8
8	lin	Population	1000	0.3	190.5		92.4
		Major	61	0.2	45.1		92.4
		Med	24	0.3	25.9		88.8
		Minor	12	1.6	27.6		79.4
	log	Population	1000	0.6	171.8	1.11	84.9
		Major	61	0.2	41.3	1.09	88.5
		Med	24	0.6	23.7	1.09	60.7
		Minor	12	4.5	25.1	1.10	47.9

In set 7 (very strong models) and set 8 (very strong, overfitted models), GREG-log was far more accurate than GREG-lin when the overall sample size was 5,000 or 2,000. The accuracy gain that is achieved by changing the model from linear to logistic is between 11% and 23%. Also, when n=1,000 GREG-log is better than GREG-lin with one exception: for model set 7 and minor domains, GREG-lin and GREG-log are equally accurate.

The conclusion about the relative accuracies of GREG-lin and GREG-log under SRSWOR are as follows. If the sample size in the domain is not very small and the model is not strong, GREG-lin and GREG-log are equally accurate (Table 5.4). If the sample size in the domain is not very small and the model is

strong, GREG-log is more accurate (Tables 5.5–5.7). And if the sample size in the domain is very small, either one may be more accurate.

Let us now consider the accuracy of the variance estimator, which is measured by the mean coverage rate MCR. For estimators with large MFEs, the Standard variance estimator fails severely. In set 5, where MFEs were close to 1, most MCRs were on an acceptable level. In set 6, where models are overfitted, MCRs start to drop: For n=5,000, only minor domains have too low MCRs, for n=2,000, both minor and medium domains have MCRs below 93.0% and for n=1,000, MCR is below 93.0% for GREG-log even on the population level. And in sets 7 and 8, MCRs are, with a few exceptions, too low in every case.

The conclusion about the Standard variance estimator is that for large domains and weak and non-overfitted models, the Standard variance estimator performs well. But as the sample size in the domain decreases, the model gets stronger and/or the model becomes overfitted, the variance estimator produces too narrow confidence intervals, on average.

In variance estimation, t is important to note the interaction between the model and sample size: if the model is weak (e.g. set 1 in Table 5.4), the Standard variance estimator works well even in minor domains. But if the model is strong, the Standard variance estimator produces catastrophic coverage rates in minor domains. Therefore, the strength and possible overfitting of the model play at least as crucial a role in variance estimation as does the domain sample size. For GREG-lin, the Standard variance estimator is more robust to model choice.

Tables 5.4–5.7 presented results averaged over samples and domain types. To understand better the difference between GREG-lin and GREG-log, we take one sample (sample size 2,000, replicate number 1, medium domain 11) into closer examination. Figures 5.4 and 5.5 show the histograms of sample fit predic-

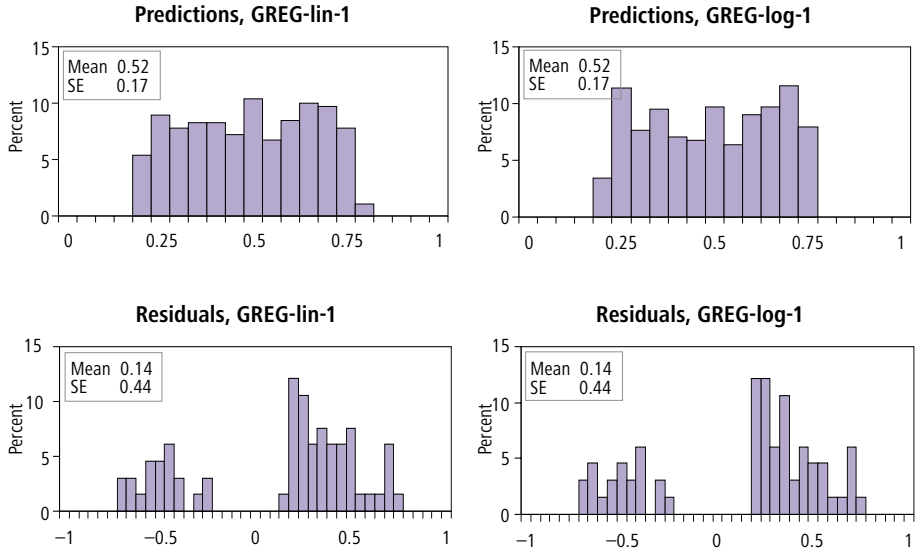


Figure 5.4 Histogram of predictions $\hat{y}_{i1}$ and residuals $e_{i1} = y_{i1} - \hat{y}_{i1}$ for GREG-lin-1 and GREG-log-1. Population 1, domain 11, sample size 2,000, replicate number 1.

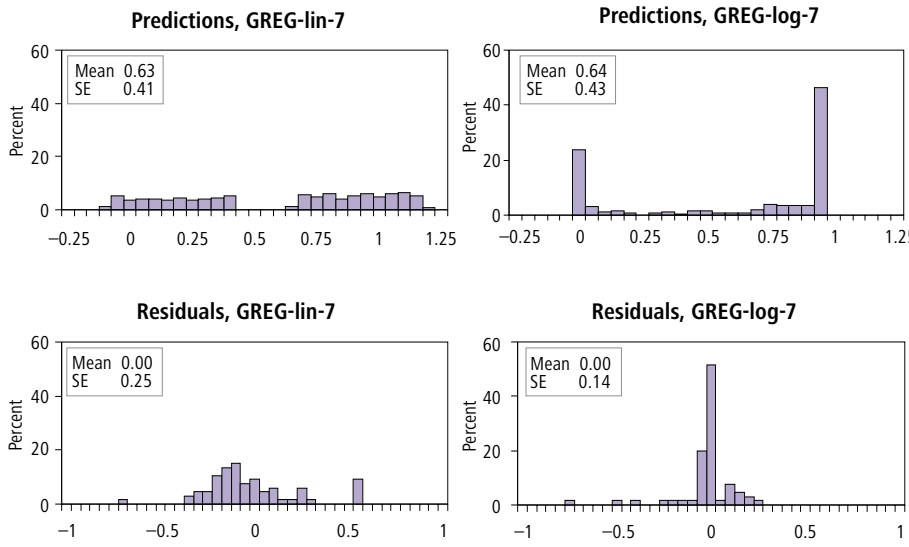


Figure 5.5 Histogram of predictions $\hat{y}_{i1}$ and residuals $e_{i1} = y_{i1} - \hat{y}_{i1}$ for GREG-lin-7 and GREG-log-7. Population 1, domain 11, sample size 2,000, replicate number 1.

tions $\hat{y}_{i1}$ and sample fit residuals $e_{i1} = y_{i1} - \hat{y}_{i1}$ for GREG-lin-1, GREG-log-1, GREG-lin-7 and GREG-log-7.

Figure 5.4 is about estimators whose models are weak (GREG-lin-1 and GREG-log-1). With these models, GREG-lin and GREG-log were equally accurate. This was anticipated, since when the model is not strong, linear and logistic models approximate each other well. The predictions and prediction errors are in line with this reasoning: there are only minor differences between the predictions of linear and logistic models and consequently, also residuals, total estimates and variance estimates are almost equivalent.

Figure 5.5 presents predictions and residuals in the case where the model is very strong (GREG-lin-7 and GREG-log-7). The upper left panel of the figure shows that predictions made with a linear model may well be outside the unit interval if the model is strong. But with a strong logistic model, predictions are within the unit interval (upper right panel). The outlying predictions of the linear model result in large prediction errors. When comparing the residuals of GREG-lin-7 (bottom left panel) and GREG-log-7 (bottom right panel), we notice that the residuals for the linear model are, on average, larger (in absolute value). The standard errors (SEs) of residuals are also very different; for GREG-lin SE is 0.25 and for GREG-log it is 0.14. Since the variance of the GREG estimator grows with the residual variance, GREG-log is more accurate.

5.2.4.3 Performance of the Standard variance estimator

Next we study the reasons for the failure of the Standard variance estimator (4.36). It was previously noted that for GREG-log, the estimator underestimates the variance especially if domains are minor and/or if the model is strong. The same holds when estimating the variance for GREG-lin, but in this case, the Standard variance estimator is more robust.

As in the previous section, the results concerning weak and moderate models (sets 1–4) are given only for the case n=2,000. This is because the results for n=5,000 and n=1,000 are essentially equivalent. Table 5.8 shows accuracy statistics for the Standard variance estimator for estimator sets 1–4 (weak and moderate models) for n=2,000. Tables 5.9–5.11 present the corresponding results for estimator sets 5–8 (strong and very strong models) for n=5,000, n=2,000, and n=1,000, respectively. The population generating link is logit, and the statistics measure the accuracy in terms of standard error.

In Table 5.8, which shows results for weak and moderate models, the average approximation error is small (less than 5% in absolute value in all cases). The average estimation error, MREE, is of similar magnitude, being larger than 5% only for GREG-log-4 in minor domains (–5.1%). The total relative bias (not shown),

Table 5.8 MRAE, MREE, MRRMSE and MCR for the Standard variance estimator for GREG estimators in sets 1–4 by estimator set, link and domain type. Total sample size n=2,000.

n=2000 Estimator		Domain type	Expected sample size	Accuracy of the Standard var. estimators			
Set	Link			MRAE %	MREE %	MRRMSE %	MCR %
1	lin	Population	2000	−1.7	−0.1	2.0	94.8
		Major	121	0.6	−0.1	6.4	94.8
		Med	48	0.5	−0.5	8.6	95.4
		Minor	24	1.4	−0.7	11.4	94.8
	log	Population	2000	−1.6	−0.1	2.0	94.7
		Major	121	0.6	−0.1	6.4	94.8
		Med	48	0.6	−0.5	8.6	95.4
		Minor	24	1.4	−0.7	11.4	94.8
2	lin	Population	2000	−2.2	−1.2	6.8	93.8
		Major	121	0.1	−0.4	6.5	94.3
		Med	48	0.0	−1.3	9.0	93.6
		Minor	24	−0.5	−2.6	12.8	92.1
	log	Population	2000	−2.4	−0.6	6.8	93.9
		Major	121	0.2	−0.4	6.5	94.3
		Med	48	0.1	−1.3	9.1	93.7
		Minor	24	−0.8	−2.6	13.0	92.0
3	lin	Population	2000	−3.0	−0.5	7.3	94.1
		Major	121	0.3	−0.4	6.9	94.6
		Med	48	−0.6	−1.2	8.9	93.8
		Minor	24	−1.5	−2.3	12.7	92.2
	log	Population	2000	−3.2	−0.6	7.3	93.8
		Major	121	0.2	−0.4	6.9	94.5
		Med	48	−0.4	−1.3	8.9	93.8
		Minor	24	−1.7	−2.3	12.9	92.2
4	lin	Population	2000	−3.6	−0.8	8.2	92.9
		Major	121	−0.8	−0.3	6.9	94.1
		Med	48	−1.4	−2.1	9.5	93.1
		Minor	24	−4.4	−4.2	16.0	90.4
	log	Population	2000	−3.1	−1.0	8.2	93.3
		Major	121	−0.6	−0.4	6.8	94.2
		Med	48	−1.2	−2.3	9.8	93.0
		Minor	24	−3.3	−5.1	17.5	90.2

which is a sum of approximation and estimation errors, grows as the model gets stronger and/or the domain size decreases. Large bias, roughly greater than 3% in absolute value, results in MCRs lower than 93.0%.

Large biases are associated with large mean relative root mean square errors (MRRMSE). The MRRMSEs show a similar pattern as both the approximation error and estimation error: as the expected domain sample size gets smaller, the model gets stronger, and/or the model becomes overfitted, MRRMSEs grow. MRRMSEs are mostly of similar magnitude for GREG-lin and GREG-log.

Let us next consider the cases where the model is strong or very strong (sets 5–8, Tables 5.9–5.11). When n=5,000 (Table 5.9), both the approximation error MRAE and estimation error MREE are quite small. Also, the coverage rates are

Table 5.9 MRAE, MREE, RRMSE and MCR for the Standard variance estimator for GREG estimators in sets 5–8 by estimator set, link and domain type. Total sample size 5,000.

n=5000 Estimator		Domain type	Expected sample size	Accuracy of the Standard var. estimators			
Set	Link			MRAE %	MREE %	MRRMSE %	MCR %
5	lin	Population	5000	−1.0	0.0	1.6	94.8
		Major	303	1.1	−0.1	6.2	94.9
		Med	121	−0.9	−0.4	10.2	94.3
		Minor	61	−2.6	−0.7	12.2	93.4
	log	Population	5000	−1.0	0.0	1.6	94.8
		Major	303	1.1	−0.1	6.2	94.9
		Med	121	−0.9	−0.4	10.2	94.3
		Minor	61	−2.6	−0.7	12.2	93.4
6	lin	Population	5000	−0.9	−0.2	1.6	94.9
		Major	303	0.8	−0.2	6.2	94.7
		Med	121	−0.9	−0.8	11.5	93.7
		Minor	61	−3.2	−1.3	12.5	92.3
	log	Population	5000	−1.6	−0.2	2.2	94.5
		Major	303	0.7	−0.2	6.3	94.6
		Med	121	−1.6	−1.3	14.2	93.1
		Minor	61	−3.5	−1.6	13.0	91.8
7	lin	Population	5000	1.1	−0.3	1.3	95.5
		Major	303	0.0	−0.3	5.8	94.8
		Med	121	−0.1	−0.8	10.2	94.3
		Minor	61	−2.8	−1.4	11.2	92.6
	log	Population	5000	−0.6	−0.7	2.3	93.8
		Major	303	0.1	−0.7	8.4	94.3
		Med	121	−1.7	−2.2	17.0	92.4
		Minor	61	−4.4	−3.8	19.2	90.2
8	lin	Population	5000	1.1	−0.4	1.3	95.5
		Major	303	−0.2	−0.2	5.8	94.7
		Med	121	−0.6	−1.1	11.2	93.8
		Minor	61	−3.1	−1.8	11.5	92.2
	log	Population	5000	−2.4	−0.5	3.6	93.7
		Major	303	−1.5	0.0	9.3	94.1
		Med	121	−2.9	−3.5	18.5	91.5
		Minor	61	−6.5	−4.9	20.6	89.1

mostly on an acceptable level; an exception is GREG-log and minor domains, where the coverage rates are around 90% for sets 7 and 8.

When n=2,000 (Table 5.10), both MRAE and MREE grow, but both are still less than 10% (in absolute value) in most cases. The total bias (not shown) and its components MRAE and MREE grow as the domain sample size gets smaller and the model gets more complex, and low MCRs go hand in hand with large bias. When n=1,000 (Table 5.11), both the approximation error and estimation error are already more than 20% in many cases, and the total bias – the sum of approximation and estimation errors – gets close to 50%. As a consequence, MCRs drop below 80%.

The MRRMSE column shows that the relative root mean square error of the standard error estimator is quite the same for GREG-lin and GREG-log in sets 5

Table 5.10 MRAE, MREE, RRMSE and MCR for the Standard variance estimator for GREG estimators in sets 5–8 by estimator set, link and domain type. Total sample size 2,000.

n=2000 Estimator		Domain type	Expected sample size	Accuracy of the Standard var. estimators			
Set	Link			MRAE %	MREE %	MRRMSE %	MCR %
5	lin	Population	2000	−4.9	0.0	5.4	93.8
		Major	121	−0.8	−0.4	10.4	94.2
		Med	48	0.8	−1.2	16.9	94.3
		Minor	24	1.5	−2.3	21.0	93.4
	log	Population	2000	−4.9	0.0	5.4	93.8
		Major	121	−0.8	−0.4	10.4	94.2
		Med	48	0.8	−1.2	16.9	94.3
		Minor	24	1.5	−2.3	21.0	93.4
6	lin	Population	2000	−4.7	−0.6	5.7	93.3
		Major	121	−1.5	−0.6	10.6	93.5
		Med	48	−0.6	−2.4	19.1	92.8
		Minor	24	−0.6	−4.3	21.6	90.1
	log	Population	2000	−5.3	−0.7	6.3	93.3
		Major	121	−2.4	−0.5	11.1	93.0
		Med	48	−1.1	−4.2	24.7	90.5
		Minor	24	−1.0	−5.4	23.2	88.1
7	lin	Population	2000	−4.2	−1.0	5.9	92.3
		Major	121	−0.9	−1.0	8.8	93.4
		Med	48	−1.1	−2.4	17.0	93.3
		Minor	24	−3.0	−4.6	21.7	88.8
	log	Population	2000	−6.0	−2.3	9.8	92.5
		Major	121	−1.6	−2.4	14.9	92.9
		Med	48	−4.4	−7.1	34.9	86.7
		Minor	24	−9.3	−12.2	48.8	79.8
8	lin	Population	2000	−7.8	−1.1	9.1	92.0
		Major	121	−1.7	−0.7	9.3	93.2
		Med	48	−1.3	−3.5	18.2	92.5
		Minor	24	−4.0	−6.0	21.5	88.1
	log	Population	2000	−10.7	−1.2	12.7	91.2
		Major	121	−5.1	−0.1	17.2	92.3
		Med	48	−6.8	−11.9	36.4	80.6
		Minor	24	−13.9	−17.0	44.4	75.4

Table 5.11 MRAE, MREE, RRMSE and MCR for the Standard variance estimator for GREG estimators in sets 5–8 by estimator set, link and domain type. Total sample size 1,000.

n=1000 Estimator		Domain type	Expected sample size	Accuracy of the Standard var. estimators			
Set	Link			MRAE %	MREE %	MRRMSE %	MCR %
5	lin	Population	1000	−0.4	−0.1	5.3	94.7
		Major	61	0.3	−1.1	14.4	94.4
		Med	24	−0.1	−3.2	24.1	92.6
		Minor	12	0.3	−4.9	30.9	89.5
	log	Population	1000	−0.4	−0.1	5.3	94.7
		Major	61	0.3	−1.1	14.4	94.4
		Med	24	−0.1	−3.2	24.1	92.6
		Minor	12	0.3	−4.9	30.9	89.5
6	lin	Population	1000	−3.2	−1.3	5.4	93.8
		Major	61	−0.8	−1.5	14.6	93.1
		Med	24	−2.5	−6.0	27.9	90.3
		Minor	12	−5.0	−9.1	33.5	83.5
	log	Population	1000	−4.8	−1.5	7.0	92.8
		Major	61	−1.5	−1.6	15.4	92.5
		Med	24	−4.1	−10.6	37.1	86.1
		Minor	12	−9.7	−11.7	38.6	79.1
7	lin	Population	1000	−1.2	−2.2	4.3	93.8
		Major	61	−0.9	−2.1	12.8	92.9
		Med	24	−2.9	−5.6	24.7	90.6
		Minor	12	−7.4	−10.2	32.2	82.6
	log	Population	1000	−9.0	−5.0	14.8	90.5
		Major	61	−3.6	−5.3	22.7	89.8
		Med	24	−9.1	−16.4	47.8	78.6
		Minor	12	−26.3	−23.0	61.5	62.8
8	lin	Population	1000	−8.7	−2.5	11.4	92.4
		Major	61	−2.2	−1.7	13.3	92.4
		Med	24	−4.7	−8.0	28.1	88.8
		Minor	12	−17.9	−11.8	40.2	79.4
	log	Population	1000	−17.6	−1.6	20.5	84.9
		Major	61	−11.5	−1.4	27.3	88.5
		Med	24	−13.9	−26.8	58.8	60.7
		Minor	12	−33.7	−31.2	73.5	47.9

and 6 (strong models), but in sets 7 and 8 (very strong models), MRRMSE is roughly twice bigger for GREG-log. These are the cases where GREG-log is significantly more accurate than GREG-lin. For GREG-lin, the approximation error, estimation error and MRRMSE are almost always smaller than for GREG-log. Therefore MCRs are also closer to 95.0%.

The conclusion about the variance estimator is as follows. First, for minor domains and/or for strong models, MCRs are too low. Second, MCR gets closer to the nominal 95.0 as the model gets weaker. Third, overfitting decreases the MCRs. Fourth, for GREG-lin, MCRs are closer to 95.0 than they are for the corresponding GREG-log estimators. Fifth, the larger the expected domain sample size, the better the MCR. Sixth, low MCRs go hand in hand with large bias and approximation and estimation errors contribute to the total bias roughly equally much. All these six features are present in every one of Tables 5.8–5.11.

5.3 Experiment I.2: πPS design

In this Chapter, we study the performance of GREG-lin, GREG-log and the Standard variance estimator under fixed size without replacement probability proportional so size (πPS) design. In Chapter 5.3.1, the general setting of the experiment is described. Chapter 5.3.2 presents the results.

5.3.1 General setting, estimators and accuracy measures

The most important differences in the πPS setting with respect to the SRSWOR setting, is that i) the sampling design is an unequal probability sampling design, and ii) the size of the population (N=400) and sample sizes (n=40, 80) are smaller than in SRSWOR. Smaller population and smaller samples are chosen because the calculation of second-order inclusion probabilities would otherwise be computationally infeasible. Most other features of the population generating process, estimators and accuracy measures are the same as they were in the SRSWOR design.

The population used in the πPS case is called Population 2. The population is generated using the following procedure. First, one minor and one major domain are constructed as follows:

$$N^{(d)} = \begin{cases} 100, d = 1, \\ 300, d = 2. \end{cases} \tag{5.28}$$

The resulting population size is N=400. Auxiliary variables are generated as

$$\begin{cases} \underline{x}_{i1} \sim Uni(0,1) \\ \underline{x}_2^{(d)} \sim Be\left(\underline{t}^{(d)}\right),\ \underline{t}^{(d)} \sim Uni\ (0.4, 0.8) \end{cases} i \in U^{(d)}, d = 1, 2. \tag{5.29}$$

This is the same process that was used in generating Population 1: the continuous variable x_1 has the same distribution for both domains, but the binary variable x_2 has a domain-specific disribution.

One binary study variable is generated from a Bernoulli distribution. The process we use to generate the probabilities $\underline{p}_{i1}$ of the Bernoulli distribution is based on the unit-specific linear predictor $\underline{\eta}_i$ and logistic transformation:

$$\underline{\eta}_i = \beta_0 + \left(\beta_1 + \underline{u}_1^{(d)}\right)\underline{x}_{i1} + \beta_2 \underline{x}_{i2}^{(d)},\ \underline{u}_1^{(d)} \sim N(0, 3), \text{ and} \tag{5.30}$$

$$\underline{p}_{i1} = \frac{\exp\left(\underline{\eta}_i\right)}{1+\exp\left(\underline{\eta}_i\right)},\ (\beta_0, \beta_1, \beta_2) = (1, 10, -10). \tag{5.31}$$

Finally, the study variables are generated from the distribution $\underline{y}_i^{(d)} \sim Be(\underline{p}_i^{(d)})$.

Figure 5.6 shows the histogram of probabilities by domain. The probabilities are intentionally generated so that with full auxiliary information, it is possible to build strong models.

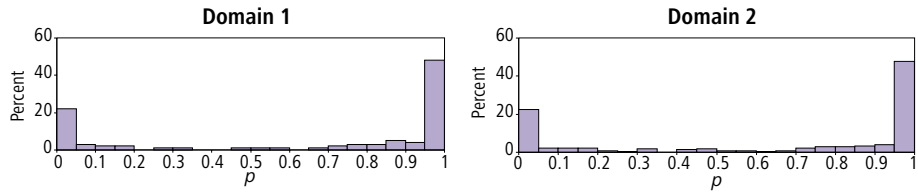


Figure 5.6 Histogram of probabilities $p = P(\underline{y} = 1)$ in domains of Population 2.

The samples are drawn using πPS sampling. Two size variables z are constructed:

$$\begin{cases} \underline{z}_{i1} \sim N(100, 10) \\ \underline{z}_{i2} \sim N(100 + 10p_i, 3) \end{cases} \quad i \in U \,. \tag{5.32}$$

First of the size variables, z_1, is not correlated with the probabilities $p = P(y = 1)$ and the study variable. The second size variable z_2 is strongly correlated with the probabilities and study variable. Figure 5.7 shows the scatter plot of y, p, z_1 and z_2 in Population 2 (the scatter plot looks essentially the same in domains).

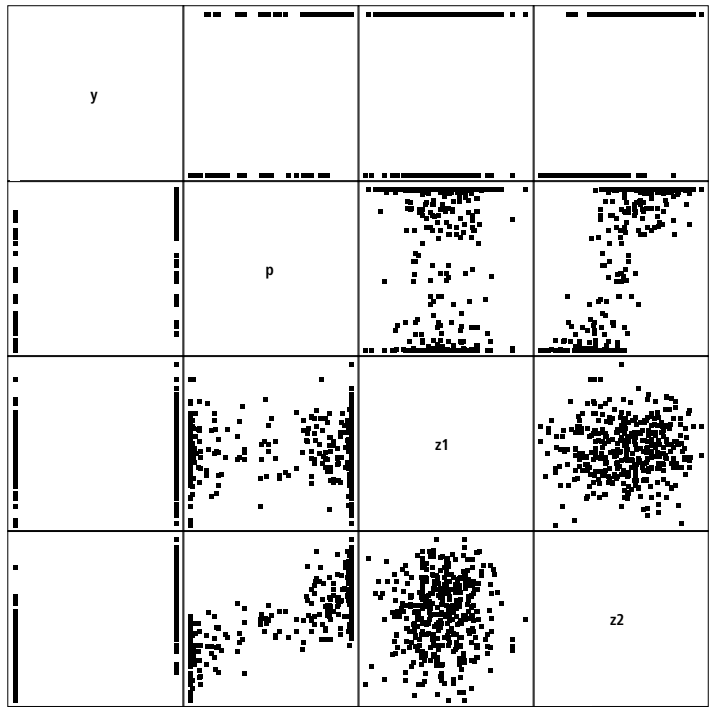


Figure 5.7. Scatter plot of y, p, z_1 and z_2 in Population 2.

For the sample selection, we use the Hanurav-Vijayan algorithm, implemented in the SAS procedure Surveyselect. This algorithm fixed size without replacement πPS samples. The algorithm was introduced by Hanurav (1967) in the context of stratified sampling and generalised by Vijayan (1968) so that more than two units per stratum could be sampled. In the algorithm, the selection probability for unit i is nh_i, where $h_i = z_i \cdot \left(\sum_U z_i\right)^{-1}$ and z is the size measure. The algorithm enables computation of second-order inclusion probabilities, which are necessary for the calculation of the Standard variance estimator. Detailed documentation of the algorithm can be found from SAS OnlineDoc (1999).

Table 5.12 shows the domain sizes and totals and proportions of y in Population 2. Expected sample sizes for the Hanurav-Vijayan πPS samples of size 40 and 80 are also included.

To study empirically the theoretical properties of GREG-lin and GREG-log under πPS sampling, 1,000 independent πPS samples of size 40 and 80 are drawn. From each sample, totals of y in the population and in domains is estimated using the estimators of Table 5.13. Standard variance estimates and variance approximations were calculated in the way they were presented in Chapter 4. Models were estimated as in the SRSWOR study: weighted least squares for the linear models, PML for the logistic models.

We expect that at least in sets 7–10, L-GREG estimators are more accurate than the corresponding GREG-lin estimators, since these estimators have very strong models. The possible accuracy difference between GREG-lin and GREG-log should decrease as weaker models are used. Note that in sets 9–10, double use of auxiliary information is exercised: the size variable is present both in the sampling design and in the assisting model.

All the accuracy measures are the same as they were in the SRSWOR case, with one exception. For the Standard variance approximation, we use the unbiased pseudo-estimator presented in Chapter 4.3.4.

Table 5.12 Domain sizes and totals and proportions of study variables for Population 1.

Domain type	Domain number	Domain size	Total of y	Prop. of y	Exp. sample size with z1		Exp. sample size with z2	
			y1	y1	n=40	n=80	n=40	n=80
Minor	1	100	69	0.69	9.9	19.8	10.0	20.0
Major	2	300	199	0.66	30.1	60.2	30.0	60.0
Pop.	3	400	268	0.67	40	80	40	80

Table 5.13 Estimators used in the simulations.

Set	Estimator	Linear predictor η*	Description of the model	Over-specified
1	GREG-lin-1	$\beta_0 + \beta_1 x_1$	Weak	
	GREG-log-1	$\beta_0 + \beta_1 x_1$	Weak	
2	GREG-lin-2	$\beta_0^{(d)} + \beta_1 x_1$	Weak	Yes
	GREG-log-2	$\beta_0^{(d)} + \beta_1 x_1$	Weak	Yes
3	GREG-lin-3	$\beta_0 + \beta_1^{(d)} x_1$	Moderate	
	GREG-log-3	$\beta_0 + \beta_1^{(d)} x_1$	Moderate	
4	GREG-lin-4	$\beta_0^{(d)} + \beta_1^{(d)} x_1$	Moderate	Yes
	GREG-log-4	$\beta_0^{(d)} + \beta_1^{(d)} x_1$	Moderate	Yes
5	GREG-lin-5	$\beta_0 + \beta_2^{(d)} x_2$	Strong	
	GREG-log-5	$\beta_0 + \beta_2^{(d)} x_2$	Strong	
6	GREG-lin-6	$\beta_0^{(d)} + \beta_2^{(d)} x_2$	Strong	Yes
	GREG-log-6	$\beta_0^{(d)} + \beta_2^{(d)} x_2$	Strong	Yes
7	GREG-lin-7	$\beta_0 + \beta_1^{(d)} x_1 + \beta_2 x_2$	Very strong	
	GREG-log-7	$\beta_0 + \beta_1^{(d)} x_1 + \beta_2 x_2$	Very strong	
8	GREG-lin-8	$\beta_0^{(d)} + \beta_1^{(d)} x_1 + \beta_2 x_2$	Very strong	Yes
	GREG-log-8	$\beta_0^{(d)} + \beta_1^{(d)} x_1 + \beta_2 x_2$	Very strong	Yes
9	GREG-prob-9	$\beta_0 + \beta_1^{(d)} x_1 + \beta_2 x_2 + \beta_3 z$	Very strong	
	GREG-cll-9	$\beta_0 + \beta_1^{(d)} x_1 + \beta_2 x_2 + \beta_3 z$	Very strong	
10	GREG-lin-10	$\beta_0^{(d)} + \beta_1^{(d)} x_1 + \beta_2 x_2 + \beta_3 z$	Very strong	Yes
	GREG-log-10	$\beta_0^{(d)} + \beta_1^{(d)} x_1 + \beta_2 x_2 + \beta_3 z$	Very strong	Yes

* The population generating linear predictor is $\beta_0 + \beta_1^{(d)} x_1 + \beta_2 x_2$

5.3.2 Results

Tables 5.14–5.17 summarize the results from the πPS study. The tables are arranged as they were in the SRSWOR study: first estimators with weak and moderate models (sets 1–4) are presented, then estimators with strong and very strong models (sets 5–10). Since there was only one minor and one major domain, no averaging over domains is done.

Table 5.14 is for weak or moderate models, n=80. The ARB columns show that all estimators are approximately unbiased and the FE columns shows that GREG-log and GREG-lin are equally accurate. The results for the weak size variable z_1 and the strong size variable z_2 are quite similar. Most importantly, whether the size variable is correlated with the study variable or not seems to have little effect on the difference between GREG-lin and GREG-log, and also the effect on the coverage rate is small.

Table 5.15 shows the results for weak and moderate models, n=40. As for the larger sample size (n=80, Table 5.14), the estimators are approximately unbiased, and GREG-lin and GREG-log are essentially equally accurate. Additionally, differences between GREG-lin and GREG-log do not depend on the size

Table 5.14 ARB, SE, FE and CR for GREG-lin and GREG-log estimators in sets 1–4. Total sample size n=80, sampling design πPS, size variables z1 and z2.

n=80 Estimator		Domain type	Exp. sample size	Size variable z1				Size variable z2			
				Accuracy of GREG			Acc. of variance estim.	Accuracy of GREG			Acc. of variance estim.
Set	Link			ARB %	SE	FE	CR, %	ARB %	SE	FE	CR, %
1	lin	Population	80	0.1	18.6		94.0	0.1	20.5		92.4
		Major	60	0.0	15.6		95.2	0.3	17.1		93.5
		Minor	20	0.2	9.3		93.4	0.3	9.2		93.0
	log	Population	80	0.1	18.6	1.00	93.9	0.1	20.5	1.00	92.4
		Major	60	0.0	15.6	1.00	94.7	0.3	17.1	1.00	93.4
		Minor	20	0.2	9.3	1.00	93.3	0.2	9.2	1.00	93.4
2	lin	Population	80	0.1	18.6		94.0	0.1	20.6		91.5
		Major	60	0.1	15.6		94.9	0.3	17.2		93.5
		Minor	20	0.1	9.5		90.7	0.5	9.5		90.4
	log	Population	80	0.1	18.7	1.00	93.9	0.1	20.6	1.00	91.4
		Major	60	0.1	15.7	1.00	94.7	0.3	17.2	1.00	93.3
		Minor	20	0.2	9.5	1.00	90.8	0.4	9.5	1.00	90.2
3	lin	Population	80	0.0	18.7		93.7	0.1	20.5		92.0
		Major	60	0.1	15.6		95.0	0.3	17.2		93.4
		Minor	20	0.0	9.5		91.3	0.4	9.5		91.4
	log	Population	80	0.1	18.7	1.00	93.8	0.1	20.6	1.00	92.1
		Major	60	0.1	15.6	1.00	94.7	0.3	17.2	1.00	93.4
		Minor	20	0.1	9.5	1.01	91.0	0.4	9.5	1.00	91.8
4	lin	Population	80	0.2	18.8		93.6	0.1	20.7		91.2
		Major	60	0.1	15.6		94.8	0.2	17.2		93.1
		Minor	20	0.4	9.8		89.1	0.5	9.7		88.8
	log	Population	80	0.2	18.8	1.00	93.8	0.1	20.7	1.00	91.0
		Major	60	0.1	15.7	1.00	94.9	0.2	17.3	1.00	93.1
		Minor	20	0.5	9.8	1.00	88.6	0.3	9.7	1.00	88.8

variable. This holds also for strong and very strong models; therefore in the following results are presented only for z_1.

When models get stronger (Table 5.16, estimator sets 5–10), differences between GREG-lin and GREG-log arise. In set 5, GREG-log and GREG-lin are still equally accurate. In set 6, where the model is the same as in set 5 but overfitted, GREG-lin is more accurate. With very strong models (sets 7–10) and n=80, GREG-log is more accurate. However, when n=40 (Table 5.17), GREG-lin is more accurate even when the models are very strong. The difference in favor of GREG-lin is large especially in minor domains.

When the model is very strong and domain minor, also bias arises: All estimators in sets 1–4 were approximately unbiased, but in sets 7–10 and for n=40, GREG-log is biased in minor domains. In these cases ARB is between 2.4% and 4.9% for GREG-log, but for GREG-lin bias is always less than 2%.

Table 5.15 ARB, SE, FE and CR for GREG-lin and GREG-log estimators in sets 1–4. Total sample size n=40, sampling design πPS, size variable z1.

n=40 Estimator		Domain type	Exp. sample size	Size variable z1				Size variable z2			
				Accuracy of GREG			Acc. of variance estim.	Accuracy of GREG			Acc. of variance estim.
Set	Link			ARB %	SE	FE	CR, %	ARB %	SE	FE	CR, %
1	lin	Population	40	0.6	27.1		93.3	0.2	27.5		93.5
		Major	30	0.7	22.5		95.3	0.1	23.8		93.3
		Minor	10	0.2	14.0		91.1	0.2	13.7		92.4
	log	Population	40	0.6	27.1	1.00	93.3	0.2	27.4	1.00	93.4
		Major	30	0.7	22.5	1.00	95.1	0.2	23.7	1.00	93.3
		Minor	10	0.2	14.0	1.00	91.0	0.2	13.7	1.00	92.3
2	lin	Population	40	0.5	27.4		92.6	0.1	27.9		92.6
		Major	30	0.8	22.8		94.5	0.2	24.2		92.5
		Minor	10	0.2	14.8		84.4	0.1	14.6		86.2
	log	Population	40	0.6	27.4	1.00	92.2	0.2	28.0	1.00	92.7
		Major	30	0.8	22.8	1.00	93.9	0.3	24.2	1.00	92.6
		Minor	10	0.1	14.9	1.00	82.9	0.1	14.6	1.00	84.6
3	lin	Population	40	0.4	27.5		92.4	0.1	27.8		92.9
		Major	30	0.7	22.8		94.6	0.2	24.1		92.9
		Minor	10	0.5	14.9		86.2	0.1	14.8		87.1
	log	Population	40	0.5	27.4	1.00	92.5	0.2	28.0	1.00	93.0
		Major	30	0.8	22.8	1.00	94.0	0.3	24.1	1.00	92.6
		Minor	10	0.5	14.7	1.01	85.3	0.2	14.8	1.00	86.2
4	lin	Population	40	0.5	28.5		92.0	0.1	28.7		91.7
		Major	30	0.8	22.9		93.9	0.2	24.3		92.2
		Minor	10	0.2	15.9		81.0	0.3	15.9		81.2
	log	Population	40	0.8	27.8	1.03	91.6	0.2	28.4	1.01	91.4
		Major	30	0.8	22.9	1.00	93.5	0.3	24.3	1.00	92.6
		Minor	10	0.7	15.6	1.02	76.5	0.0	15.5	1.02	78.3

Thus, GREG-lin and GREG-log are equally accurate under πPS if the model is not strong (Tables 5.14–5.15). If the model is strong and if the domain sample size is not very small, GREG-log is more accurate (Table 5.16, n=80, set 7–10). However, for small domain samples sizes, GREG-log is inaccurate and may even be biased (Table 5.16, n=40, sets 7–10). This bias is discussed more in Appendix II.

When comparing estimators in sets 7–8 and corresponding estimators in sets 9–10, it can be seen that the latter have larger standard errors. The only difference between sets 7 and 9 and sets 8 and 10 is that in sets 9–10 the models include the size variable. Thus in these simulations, double use of auxiliary information decreases the accuracy.

Let us now consider variance estimation under the πPS design. The CR columns in Tables 5.14–5.16 show that for weak, non-overfitted models the Standard variance estimator performs well. But as the model gets stronger, or overfitted (Table 5.16), and as the domain sample size gets smaller, CRs get lower.

Table 5.16 ARB, SE, FE and CR for GREG-lin and GREG-log estimators in sets 5–10. Sampling design πPS, size variable z1, total sample sizes n=80, and n=40.

Estimator		Total sample size n=80						Total sample size n=80				
		Domain type	Exp. sample size	Accuracy of GREG			Acc. of variance estim.	Exp. sample size	Accuracy of GREG			Acc. of variance estim.
Set	Link			ARB %	SE	FE	CR, %		ARB %	SE	FE	CR, %
5	lin	Population	80	0.1	11.5		93.1	40	0.5	17.4		92.1
		Major	60	0.1	10.0		93.1	30	0.5	15.1		89.0
		Minor	20	0.0	5.7		90.4	10	0.4	8.1		77.1
	log	Population	80	0.1	11.5	1.00	93.1	40	0.5	17.4	1.00	92.1
		Major	60	0.1	10.0	1.00	93.1	30	0.5	15.1	1.00	89.0
		Minor	20	0.0	5.7	1.00	90.4	10	0.4	8.1	1.00	77.1
6	lin	Population	80	0.1	11.6		93.5	40	0.5	17.6		91.5
		Major	60	0.1	10.1		92.8	30	0.5	15.3		88.2
		Minor	20	0.1	5.9		89.0	10	0.3	8.7		71.4
	log	Population	80	0.1	11.6	0.99	93.0	40	0.3	18.3	0.96	89.7
		Major	60	0.1	10.1	1.00	92.6	30	0.6	15.3	1.00	87.2
		Minor	20	0.2	6.0	0.97	87.4	10	0.4	10.2	0.86	65.2
7	lin	Population	80	0.0	10.6		92.7	40	0.6	16.2		90.4
		Major	60	0.0	8.9		92.5	30	0.7	13.6		89.1
		Minor	20	0.1	5.6		87.0	10	0.5	8.3		72.1
	log	Population	80	0.1	9.0	1.17	87.5	40	0.9	16.1	1.01	67.1
		Major	60	0.0	7.5	1.20	88.6	30	0.4	13.1	1.04	67.2
		Minor	20	0.3	5.2	1.08	63.0	10	2.4	9.4	0.88	35.8
8	lin	Population	80	0.1	10.5		92.7	40	0.7	16.3		89.9
		Major	60	0.1	8.9		92.1	30	0.8	13.6		88.9
		Minor	20	0.1	5.6		86.4	10	0.7	9.0		67.9
	log	Population	80	0.1	9.4	1.13	85.3	40	1.4	16.8	0.97	63.1
		Major	60	0.0	7.6	1.17	87.7	30	0.6	13.6	1.01	63.1
		Minor	20	0.4	5.6	1.01	48.4	10	3.8	9.5	0.95	19.8
9	lin	Population	80	0.0	10.8		92.5	40	0.6	16.4		90.4
		Major	60	0.0	9.1		91.5	30	0.7	13.8		89.0
		Minor	20	0.1	5.7		85.7	10	0.5	8.4		71.8
	log	Population	80	0.0	9.4	1.15	83.5	40	1.3	17.3	0.95	51.2
		Major	60	0.1	7.9	1.15	84.9	30	0.7	14.3	0.96	52.7
		Minor	20	0.2	5.2	1.09	60.5	10	3.0	9.6	0.88	26.6
10	lin	Population	80	0.1	10.8		92.3	40	0.8	16.5		90.1
		Major	60	0.1	9.1		91.2	30	0.8	13.8		88.3
		Minor	20	0.0	5.7		85.1	10	0.8	8.8		67.9
	log	Population	80	0.2	9.7	1.11	79.9	40	1.9	18.3	0.90	47.4
		Major	60	0.2	8.1	1.12	83.5	30	0.9	14.7	0.94	49.5
		Minor	20	0.4	5.6	1.03	45.0	10	4.9	10.4	0.84	14.0

The pattern for CRs is the same that was observed in the SRSWOR study in Chapter 5.2. For example, for n=40, size variable z_1, estimator set 1 and major domain, CR is 95.1% for GREG-log (Table 5.15). For the same sample size, the

same domain, the same size variable, but the estimator from set 7, MCR is 67.2% for GREG-log. In minor domains, MCRs are even below 30% for some GREG estimators with very strong models. Thus when the model is very strong, the variance estimator fails completely in minor domains. This holds both for n=80 and for n=40.

It is interesting to notice that both under SRSWOR and πPS, the variance estimator performed best when the model was weak (and the GREG estimator itself was inaccurate). This follows from the fact that the weaker the model, the closer the GREG estimator is to the HT estimator, and for the HT estimator, the Standard variance estimator works well. But when the model is strong, the predictions and prediction errors get a larger role, and variance estimation becomes more challenging.

Next we consider the performance of the Standard variance estimator in more detail. The aim is to see where the variance estimation fails and why the coverage rates are too low. Results are presented only for strong and very strong models (model sets 5–10) since the Standard variance estimator generally works well for weaker models. Only the case of the weak size variable z_1 is considered, since the results were essentially similar for z_2.

Table 5.17 presents accuracy measures for the Standard variance estimator under πPS sampling for strong and very strong models. The table shows that for strong models (sets 5–6) and very strong models (sets 7–10), both the Standard approximation and the Standard estimator fail. This happens both in minor and major domains and when n=80 and when n=40.

When we look at the relative contributions of the approximation and estimation error to total bias, it seems that both contribute, on average, equally much to the bias. When the model is strong or very strong, both approximation and estimation errors are present even in major domains and with n=80. Thus, the strength of the model is at least as important in variance estimation as the sample size.

In the SRSWOR design, we observed that the MRRMSE of the Standard variance estimator was much larger for GREG-log than for GREG-lin. The same holds in the πPS study. For models that are not strong, the differences are not large, but for estimators in sets 5–10, the root mean square error of the variance estimator is in some cases two times larger for GREG-log.

To conclude, the Standard variance estimator is unsatisfactory for GREG-log in the πPS case as well. Especially if the model happens to be good and we gain in accuracy by changing the model from linear to logistic, the variance estimator fails: it has a large downward bias and large variance. And the estimators fail even on the population level with sample size 80, although the failure is much larger in minor domains. Although the errors are large for both GREG-lin and GREG-log, they are much smaller for GREG-lin. Thus there clearly is need for a better variance estimator for GREG-log.

Table 5.17 RAE, REE, RRMSE and CR for the Standard variance estimator for GREG estimators in sets 5–10 by estimator set, link and domain type. Sampling design πPS, size variable z1, total sample size n=80 and n=40.

Estimator		Total sample size n=80						Total sample size n=40				
		Dom. type	Exp. sample size	Standard var. estimator				Exp. sample size	Standard var. estimator			
Set	Link			RAE %	REE %	RRMSE %	CR %		RAE %	REE	RRMSE %	CR %
5	lin	Pop	80	1.3	−0.7	11.4	93.1	40	−1.3	−2.1	19.7	92.1
		Major	60	2.2	−0.8	14.0	93.1	30	−0.6	−2.4	23.7	89.0
		Minor	20	−4.7	−0.9	24.4	90.4	10	−4.1	−3.3	42.0	77.1
	log	Pop	80	1.4	−0.7	11.4	93.1	40	−1.3	−2.1	19.7	92.1
		Major	60	2.2	−0.8	14.0	93.1	30	−0.7	−2.4	23.7	89.0
		Minor	20	−4.7	−0.9	24.4	90.4	10	−4.2	−3.3	42.0	77.1
6	lin	Pop	80	0.9	−1.2	11.4	93.5	40	−2.3	−3.1	19.9	91.5
		Major	60	1.7	−1.0	13.9	92.8	30	−1.8	−2.8	24.1	88.2
		Minor	20	−6.9	−2.6	25.8	89.0	10	−10.3	−7.7	43.6	71.4
	log	Pop	80	0.3	−1.9	11.6	93.0	40	−5.9	−4.5	21.5	89.7
		Major	60	1.4	−1.2	14.0	92.6	30	−2.1	−3.5	25.4	87.2
		Minor	20	−9.6	−6.0	31.6	87.4	10	−23.1	−16.6	58.2	65.2
7	lin	Pop	80	−4.3	0.9	10.1	92.7	40	−3.8	−6.0	19.4	90.4
		Major	60	2.1	−1.9	11.7	92.5	30	−0.6	−5.3	21.9	89.1
		Minor	20	−7.0	−5.2	25.1	87.0	10	−12.6	−11.5	40.9	72.1
	log	Pop	80	−6.7	−7.8	24.9	87.5	40	−22.1	−21.4	55.7	67.1
		Major	60	−1.5	−5.6	24.3	88.6	30	−17.0	−20.4	55.2	67.2
		Minor	20	−25.5	−19.1	59.2	63.0	10	−43.6	−30.5	83.2	35.8
8	lin	Pop	80	−1.0	−2.5	10.1	92.7	40	−5.0	−6.0	19.9	89.9
		Major	60	1.9	−1.8	11.7	92.1	30	−1.3	−5.1	22.1	88.9
		Minor	20	−10.9	−5.1	26.4	86.4	10	−16.8	−12.7	44.6	67.9
	log	Pop	80	−9.7	−9.4	27.6	85.3	40	−25.5	−23.0	58.3	63.1
		Major	60	−3.5	−6.6	26.5	87.7	30	−20.0	−22.7	59.3	63.1
		Minor	20	−30.7	−29.1	72.5	48.4	10	−45.1	−39.1	90.6	19.8
9	lin	Pop	80	−3.3	−2.9	11.2	92.5	40	−5.5	−6.4	20.4	90.4
		Major	60	0.2	−2.3	11.6	91.5	30	−2.1	−5.7	22.1	89.0
		Minor	20	−12.1	−5.0	25.8	85.7	10	−14.1	−11.6	41.0	71.8
	log	Pop	80	−11.8	−9.9	31.1	83.5	40	−27.9	−31.4	69.3	51.2
		Major	60	−7.9	−7.1	28.9	84.9	30	−24.2	−30.2	68.1	52.7
		Minor	20	−28.9	−22.7	63.2	60.5	10	−45.8	−37.5	88.6	26.6
10	lin	Pop	80	−3.2	−3.1	11.3	92.3	40	−6.3	−7.0	21.1	90.1
		Major	60	0.1	−2.5	11.6	91.2	30	−2.3	−6.0	22.3	88.3
		Minor	20	−13.0	−5.8	27.2	85.1	10	−18.1	−13.5	44.7	67.9
	log	Pop	80	−13.7	−12.7	34.6	79.9	40	−32.2	−32.6	72.8	47.4
		Major	60	−8.5	−10.1	31.8	83.5	30	−25.6	−32.9	71.0	49.5
		Minor	20	−34.5	−31.0	75.4	45.0	10	−51.2	−40.3	94.4	14.0

6 Alternative variance estimators

In the Monte Carlo studies of Chapter 5. we observed that the Standard variance estimator ((4.36) for domains, (4.34) for population) often underestimates the variance for GREG estimators. Also, the variance of the estimator may be very large. Large bias and variance are present especially if

i. the assisting model is strong
ii. the sample size in the domain is minor
iii. the link function is logistic-type
iv. the model is overspecified.

The Standard variance estimator performs usually quite well, if

i. the assisting model is weak
ii. the sample size in the domain is major
iii. the link function is linear.

Next, we consider some alternative ways to estimate the variance of a GREG estimator. In Chapter 6.1, we discuss two standard resampling based techniques, the jackknife and the bootstrap and in Chapter 6.2, introduce a new variance estimator called the Augmented variance estimator that is based on the Standard approximation. The performance of these three estimators is studied empirically by a small Monte Carlo simulation in Chapter 6.3. In the simulations the Augmented variance estimator performs best. In Chapter 7, we study the Augmented variance estimator more extensively and conduct all the simulations of Chapter 5 with the Augmented variance estimator.

6.1 Resampling approach

We describe briefly the two basic resampling based variance estimators, the jackknife and bootstrap. Both the jackknife and bootstrap are actually method families. The theoretical literature on these estimators concerns mainly situations where the estimator whose variance is to be estimated is a smooth function of population means. The L-GREG with a continuous auxiliary variable is not such an estimator. Moreover, the dangers related to the jackknife when the estimator is a quantile or order statistics are well known. Thus, we do not plunge deep into the theory, but briefly describe the basic principles of these variance estimation techniques and then study empirically their properties. Estimators are described for the population total estimator, but by setting $y_{ij} = y_{ij}^{(d)}$, these also apply to domain totals. The main references used in this section are Wolter (1985), Särndal *et al.* (1992), Shao (2003) and Lahiri (2003).

6.1.1 Jackknife

The jackknife was originally developed in an infinite population context. Quenouille (1949) introduced the method to reduce bias of an estimator, Tukey (1958) suggested that the method might be used for variance and interval estimation, and Durbin (1959) seems to be the first to consider the jackknife in finite population inference (see Wolter 1985 and Särndal *et al.* 1992).

We construct the jackknife estimator as follows. Let the sampling design be a non-stratified fixed size without replacement design, such as the SRSWOR or πPS design. The population parameter is total of y and we estimate it by $\underline{\hat{T}}$. The sample set is partitioned into k groups of m observations in each group. In practice, it might happen that $km \neq n$. To overcome this problem, we consider only the case where $m = 1$, that is, the delete-one jackknife. This choice is computationally challenging, but with respect to the accuracy of the jackknife, this is the preferred choice (Wolter 1985, p. 164). The pseudo-sample set that is obtained by deleting unit h is denoted by $s_{(-h)}$.

For each of the n sample sets $s_{(-h)}$, we define $\underline{\hat{T}}_{(-h)}$ as the estimator that is calculated from the reduced sample set. The form of the estimator $\underline{\hat{T}}_{(-h)}$ is the same as that of $\underline{\hat{T}}$, but when estimating totals, weights need to be rescaled so that $\sum_{i \in s_{(-h)}} w_i = N$. This is obtained by multiplying the original weights by factor

$$\left(\sum_{i \in s_{(-h)}} w_i \right)^{-1} \sum_{i \in s} w_i \ . \tag{6.1}$$

Next, we define pseudo-values

$$\hat{T}_{-h} = k\hat{T} - (k-1)\hat{T}_{(-h)} = n\hat{T} - (n-1)\hat{T}_{(-h)} \ . \tag{6.2}$$

From these pseudo-values, the jackknife variance estimator is obtained as

$$\hat{V}_J\left(\underline{\hat{T}}\right) = \frac{(1-f)}{n(n-1)} \sum_{h=1}^{n} \left(\hat{T}_{-h} - \overline{\hat{T}} \right)^2 \text{, where } \overline{\hat{T}} = \frac{1}{n} \sum_{h=1}^{n} \hat{T}_{-h} \ . \tag{6.3}$$

Under the SRSWOR sampling and HT estimator, it can be shown that

$$\hat{V}_j\left(\underline{\hat{T}}\right) = \frac{N^2}{n}(1-f)s_y^2 = \hat{V}_S\left(\underline{\hat{T}}_{HT}\right) \ . \tag{6.4}$$

The property (6.4) is called the linear condition. It says that when applied to a simple linear estimator, the resampling based variance estimator is the same as the analytical unblased estimator (the condition may also be defined in terms of expectations). Sometimes the jackknife is defined without the coefficient $1 - f$, which means that the linear condition (6.4) does not hold. We use the linear conditions both in the jackknife and bootstrap. This means that for the SRSWOR sampling and HT estimator, both the bootstrap and jackknife are unbiased estimators for the variance. This, of course, does not necessarily mean that they would be good estimators for the variance of the GREG estimator.

6.1.2 Bootstrap

The bootstrap method, like the jackknife, was first introduced in an infinite population context. The originator of this technique was Efron (1979, 1981, 1982). The technique works best with independent, identically distributed observations and in an infinite population context. Although much research on the bootstrap has been done since the basic technique was introduced (e.g. Bickel and Freedman 1984, Gross 1980, McCarthy and Snowden 1985, Rao and Wu 1984, 1987, Sitter 1992b), it is still unclear how the technique should be modified if the sampling is not SRSWR (Shao 2003, Lahiri 2003). Here, we describe the method following the lines of Särndal *et al.* (1992).

Let the sampling design be a non-stratified fixed size without replacement design. The population parameter is total of y and the estimator is $\underline{\hat{T}}$. In the bootstrap, we proceed as follows:

i. Construct an artificial population U^* from the sample set s
ii. Draw a series of independent bootstrap sample sets $s_1^*, s_2^*, \ldots, s_K^*$ from U^* by a design identical to the one which was used to draw s
iii. For every bootstrap sample k, calculate $\underline{\hat{T}}_k^*$
iv. Use the distribution of $\underline{\hat{T}}_1^*, \underline{\hat{T}}_2^*, \ldots, \underline{\hat{T}}_K^*$ as an estimate of the sampling distribution of $\underline{\hat{T}}$. Specifically, the bootstrap variance estimate is obtained as

$$\hat{V}_B\left(\underline{\hat{T}}\right) = c \cdot \frac{1}{K-1} \sum_{k=1}^{K} \left(\underline{\hat{T}}_k^* - \overline{\underline{\hat{T}}}^*\right)^2, \tag{6.5}$$

where $c = (n - f)/(n - 1)$ is a finite population correction coefficient which ensures that the linear condition with respect to the SRSWOR design and HT estimator holds.

Obviously, there are a number of things that need to be defined more clearly before the method can be implemented in practice. Especially, how does one construct the population U^*? And what does it mean to resample in a way that is identical to the one that was used to draw s? Next, we outline the methods that are used in this study.

When sampling is SRSWOR, every unit in the sample represents N/n units in the population. Thus it makes sense to construct the pseudo-population by replicating every unit N/n times in order to construct the sample. This was proposed by Gross (1980) and McCarthy and Snowden (1985). But there is the problem that N/n is not necessarily an integer. Bickel and Freedman (1984) have proposed a method that tries to overcome this problem by constructing two artificial populations and then drawing randomly samples from these two populations. This method, however, is not feasible when $n^3 < N^2$ (Lahiri 2003). Sitter (1992b) has proposed an alternative to the randomisation between the two populations. We will restrict the study of the bootstrap to a situation where N/n is an integer; thus the method of replicating the sample units N/n times will apply when sampling is SRSWOR. From the bootstrap population, pseudo-samples are drawn using constant inclusion probabilities n/N if the sampling is SRSWOR.

When sampling is πPS, it does not make sense to replicate every sampled unit by a constant, since different sample units represent different numbers of population units. To overcome this problem, we obtain the method described in Särndal *et al.* (1992, 443). Under πPS, every unit in s has a sampling weight w_i. For each $i \in s$ we create w_i artificial elements of U^* if w_i is an integer, and if it is not, we first round the weight to the closest integer and then create the artificial units. This may result in a pseudo-population whose size is not exactly N, but the difference between N^* and N is usually small. Pseudo-samples are drawn from the bootstrap population using the same size variables that were used when drawing the original sample.

6.2 Augmented estimator for the Standard approximation

Like the resampling based variance estimators, the Augmented variance estimator that we propose is also general in the sense that it applies to domain estimators as well. This is obtained by replacing the study variable y_{ij} by $y_{ij}^{(d)}$ (so y_{ij} is zero outside the domain). Also, it applies to all GREG estimators without restrictions to the chosen assisting model. To derive the estimator, let us first collect the variance, Standard approximation and Standard estimator (these were derived in Chapter 4.3).

The variance of the GREG estimator is

$$V\left(\underline{\hat{T}}_{j,GREG}\right). \tag{6.6}$$

The Standard approximation, given previously in (4.31), is

$$V_A\left(\underline{\hat{T}}_{j,GREG}\right) = V\left(\sum_{i\in U} \underline{w}_i E_{ij}\right) = \sum_{i\in U}\sum_{k\in U} \frac{\Delta_{ik}}{E\left(\underline{I}_i\right)E\left(\underline{I}_k\right)} E_{ij}E_{kj}\ , \tag{6.7}$$

where E_{ij} is given by (4.24).

The Standard estimator for the approximation, given previously in (4.34), is

$$\hat{V}_S\left(\underline{\hat{T}}_{j,GREG}\right) = \sum_{i\in U}\sum_{k\in U} \frac{\Delta_{ik}}{E\left(\underline{I}_i\,\underline{I}_k\right)} \underline{w}_i\,\underline{e}_{ij}\,\underline{w}_k\,\underline{e}_{kj}\ , \tag{6.8}$$

where $\underline{e}_{ij}$ is given by (4.3).

Replacing $\underline{e}_{kj}$ by $\underline{e}_{kj}^{(d)}$ in (6.8) and E_{kj} by $E_{kj}^{(d)}$ in (6.7) we get corresponding expressions for the GREG estimator for domains. Previous studies (Myrskylä 2004, 2005) and the simulation in Chapter 5 indicate that the estimation error, that is, the difference between the Standard estimator $\hat{V}_S$ and the Standard approximation V_A, contributes markedly to the total error of the standard variance estimator. Here we focus on this estimation error and aim to decrease it.

First, let us note the difference between the census fit residuals in the approximation (6.7) and the sample fit residuals in the estimator (6.8):

$$\begin{aligned} E_{ij} &= y_{ij} - \tilde{y}_{ij} \\ &= \left(y_{ij} - \underline{\hat{y}}_{ij}\right) + \left(\underline{\hat{y}}_{ij} - \tilde{y}_{ij}\right) \\ &= \underline{e}_{ij} + \underline{\tilde{e}}_{ij} \ . \end{aligned} \tag{6.9}$$

From (6.9) we see that the census fit residual E_{ij} is a sum of two terms: the sample fit residual $\underline{e}_{ij}$ and a residual $\underline{\tilde{e}}_{ij}$, which reflects the difference between the sample fit predictions and census fit predictions. Plugging (6.9) to the approximation, we see that the Standard variance estimator for the Standard variance approximation (6.7) is missing something:

The Standard approximation (6.7) can be written

$$V_A\left(\underline{\hat{T}}_{j,GREG}\right) = V\left(\sum_{i\in U} \underline{w}_i \left(\underline{e}_{ij} + \underline{\tilde{e}}_{ij}\right)\right) = V\left(\sum_{i\in U} \underline{w}_i \underline{e}_{ij} + \sum_{i\in U} \underline{w}_i \underline{\tilde{e}}_{ij}\right). \tag{6.10}$$

If the values E_{ij} (and thus $\underline{\tilde{e}}_{ij}$) were observable, an unbiased estimator for the approximation (6.10) would be

$$\hat{V}_A\left(\underline{\hat{T}}_{j,GREG}\right) = \sum_{i\in U}\sum_{k\in U} \frac{\Delta_{ik}}{E\left(\underline{I}_i \underline{I}_k\right)} \underline{w}_i \left(\underline{e}_{ij} + \underline{\tilde{e}}_{ij}\right) \underline{w}_k \left(\underline{e}_{kj} + \underline{\tilde{e}}_{kj}\right). \tag{6.11}$$

The Standard variance estimator (6.8), however, is

$$\hat{V}_S\left(\underline{\hat{T}}_{j,GREG}\right) = \sum_{i\in U}\sum_{k\in U} \frac{\Delta_{ik}}{E\left(\underline{I}_i \underline{I}_k\right)} \underline{w}_i \underline{e}_{ij} \underline{w}_k \underline{e}_{kj} \ . \tag{6.12}$$

Comparing (6.10), (6.11) and (6.12) it is easy to see that the Standard estimator ignores the terms $\underline{\tilde{e}}_{ij}$ that reflect the difference between the sample fit and population fit. If the values $\underline{\tilde{e}}_{ij}$ could be observed, a better variance estimator might be available. Later, we propose a method to calculate pseudo-estimates for $\underline{\tilde{e}}_{ij}$. Before that, let us look at (6.11) in more detail.

The pseudo-estimator (6.11) has terms that are not observable. To come up with a variance estimator whose values can be calculated, two assumptions are made. The goodness of these assumptions is studied in Monte Carlo simulations in Chapter 6.3. First, we assume that the sampling weights $\underline{w}_i$ and both residuals $\underline{e}_{ij}$ and $\underline{\tilde{e}}_{ij}$ are uncorrelated. Second, we assume that the covariance between $\sum_U \underline{w}_i \underline{e}_{ij}$ and $\sum_U \underline{w}_i \underline{\tilde{e}}_{ij}$ is small compared with the individual variances of these terms and may be ignored. This gives us the estimator

$$\begin{aligned} \hat{V}_{A^*}\left(\underline{\hat{T}}_{j,GREG}\right) &= \sum_{i\in U}\sum_{k\in U} \frac{\Delta_{ik}}{E\left(\underline{I}_i \underline{I}_k\right)} \underline{w}_i \underline{e}_{ij} \underline{w}_k \underline{e}_{kj} + \sum_{i\in U}\sum_{k\in U} \frac{\Delta_{ik}}{E\left(\underline{I}_i \underline{I}_k\right)} \underline{w}_i \underline{\tilde{e}}_{ij} \underline{w}_k \underline{\tilde{e}}_{kj} \\ &= \hat{V}_S + \hat{V}_{M^*} \ . \end{aligned} \tag{6.13}$$

The first term $\hat{V}_S$ in the estimator (6.13) is the familiar Standard variance estimator for GREG, given by (6.8) and derived in Chapter 4.3. This term is easy to calculate for every sample. The second term $\hat{V}_{M^*}$, which reflects the difference between sample fit and census fit models, has the same form as $\hat{V}_S$, but it includes unobservable quantities $\tilde{\underline{e}}_{ij}$. In Chapter 6.3, we study how well the estimator (6.13) estimates the Standard approximation (6.10). However, since (6.13) includes unknown $\tilde{\underline{e}}_{ij}$, it cannot be calculated in practice. Therefore we propose a method for generating pseudo-estimates $\underline{e}^*_{ij}$ for $\tilde{\underline{e}}_{ij}$; by using these pseudo-estimates we can estimate $\hat{V}_{M^*}$. The procedure is as follows.

i. Construct an artificial population U^* from the sample set s. The method to generate the population is the same as in bootstrap estimation. Denote the study variables in U^* by y^*_j.
ii. Draw a series of independent pseudo-sample sets $s^*_1, s^*_2, \ldots, s^*_K$ of size n from U^* by a design identical to the one which was used to draw s.
iii. Using the information $\left\{y^*_j : j \in U^*\right\}$ calculate pseudo-census fit parameters $\mathbf{B}^*$, pseudo-census fit predictions $\hat{y}^*_{ij} = \mathbf{x}_i'\mathbf{B}^*$ and pseudo-census fit residuals $E^*_{ij} = y^*_{ij} - \hat{y}^*_{ij}$. For every pseudo-sample $k = 1,2,\ldots,K$ calculate pseudo-sample fit parameters $\boldsymbol{\beta}^*$, pseudo-sample fit predictions $\hat{y}^{s^*}_{ij} = \mathbf{x}_i'\boldsymbol{\beta}^*$ and pseudo-sample fit residuals $e^0_{ij} = y^*_{ij} - \hat{y}^{s^*}_{ij}$ using information $\left\{y^*_j : j \in s^*\right\}$. Finally, for every $k = 1,2,\ldots,K$, decompose pseudo-census fit residuals E^*_{ij} as

$$\begin{aligned} E^*_{ij} &= \left(y^*_{ij} - \hat{y}^{s^*}_{ij}\right) + \left(\hat{y}^{s^*}_{ij} - \hat{y}^*_{ij}\right) \\ &= e^0_{ij} + e^*_{ij} \end{aligned} \tag{6.14}$$

and calculate the value

$$\hat{V}^{(k)}_M\left(\hat{\underline{T}}_{j,GREG}\right) = \sum_{i \in U^*}\sum_{k \in U^*} \frac{\Delta_{ik}}{E\left(\underline{I}_i \underline{I}_k\right)} w_i e^*_{ij} w_k e^*_{kj} \,. \tag{6.15}$$

iv. From the K values (6.15), calculate an estimate for $\hat{V}_{M^*}$ by

$$\hat{V}_M\left(\hat{\underline{T}}_{j,GREG}\right) = \frac{1}{K}\sum_{k=1}^{K} \hat{V}^{(k)}_M\left(\hat{\underline{T}}_{j,GREG}\right). \tag{6.16}$$

Each of the estimates (6.15) estimate the unobserved term $\hat{V}_{M^*}$. These values vary from a pseudo-sample to a pseudo-sample, so we average them over repeated pseudo-samples (step iv.). After calculating (6.16), the Augmented variance estimator $\hat{V}^*\left(\hat{\underline{T}}_{j,GREG}\right)$ for the Standard approximation can be constructed:

$$\hat{V}^*\left(\hat{\underline{T}}_{j,GREG}\right) = \hat{V}_S\left(\hat{\underline{T}}_{j,GREG}\right) + \hat{V}_M\left(\hat{\underline{T}}_{j,GREG}\right), \tag{6.17}$$

where $\hat{V}_S\left(\hat{\underline{T}}_{j,GREG}\right)$ is the Standard variance estimator, given by (6.8).

6.2.1 Augmented variance estimator under the SRSWOR design

Under SRSWOR, the Augmented variance estimator (6.17) takes the form

$$\hat{V}^*\left(\hat{\underline{T}}_{j,GREG}\right) = \frac{N^2(1-f)}{n}\underline{S}^2_{e_j} + \hat{V}_M\left(\hat{\underline{T}}_{j,GREG}\right), \text{ where} \tag{6.18}$$

$$\begin{aligned} \hat{V}_M\left(\hat{\underline{T}}_{j,GREG}\right) &= \frac{1}{K}\sum_{k=1}^{K}\hat{V}_M^{(k)}\left(\hat{\underline{T}}_{j,GREG}\right), \text{ and} \\ \hat{V}_M^{(v)}\left(\hat{\underline{T}}_{j,GREG}\right) &= \frac{N^2(1-f)}{n}\underline{S}^2_{e_j^*(k)}\,. \end{aligned} \tag{6.19}$$

If domains are fixed-size (i.e. strata) and are treated as populations of their own, the Alternative estimator is analogous to (6.18) with the obvious exception that the pseudo-population and pseudo-sample are of size $N^{(d)}$ and $n^{(d)}$. For unplanned domains, the Augmented variance estimator is

$$\hat{V}^*\left(\hat{\underline{T}}^{(d)}_{j,GREG}\right) = \frac{N^2(1-f)}{n}\underline{S}^2_{e_j^{(d)}} + \hat{V}_M\left(\hat{\underline{T}}^{(d)}_{j,GREG}\right), \text{ where} \tag{6.20}$$

$$\hat{V}_M\left(\hat{\underline{T}}^{(d)}_{j,GREG}\right) = \frac{1}{K}\sum_{k=1}^{K}\frac{N^2(1-f)}{n}\underline{S}^2_{e_j^{(d)*}(k)} \tag{6.21}$$

6.2.2 Augmented variance estimator under the πPS design

Under πPS, the Augmented variance estimator (6.17) is

$$\hat{V}^*\left(\hat{\underline{T}}_{j,GREG}\right) = \hat{V}_S\left(\hat{\underline{T}}_{j,GREG}\right) + \hat{V}_M\left(\hat{\underline{T}}_{j,GREG}\right), \text{ where} \tag{6.22}$$

$$\begin{aligned} \hat{V}_M\left(\hat{\underline{T}}_{j,GREG}\right) &= \frac{1}{K}\sum_{k=1}^{K}\hat{V}_M^{(k)}\left(\hat{\underline{T}}_{j,GREG}\right), \text{ and} \\ \hat{V}_M^{(k)}\left(\hat{\underline{T}}_{j,GREG}\right) &= \sum_{i\in U^*}\sum_{k\in U^*}\frac{\pi_{ik}-\pi_i\pi_k}{\pi_{ik}}\underline{w}_i e^*_{ij}\underline{w}_k e^*_{kj}\,. \end{aligned} \tag{6.23}$$

For unplanned domains, the Augmented variance estimator is

$$\hat{V}^*\left(\hat{\underline{T}}^{(d)}_{j,GREG}\right) = \hat{V}_S\left(\hat{\underline{T}}^{(d)}_{j,GREG}\right) + \hat{V}_M\left(\hat{\underline{T}}^{(d)}_{j,GREG}\right), \text{ where} \tag{6.24}$$

$$\hat{V}_M\left(\hat{\underline{T}}_{j,GREG}^{(d)}\right) = \frac{1}{K}\sum_{k=1}^{K}\hat{V}_M^{(k)}\left(\hat{\underline{T}}_{j,GREG}^{(d)}\right), \text{ and} \tag{6.25}$$

$$\hat{V}_M^{(k)}\left(\hat{\underline{T}}_{j,GREG}^{(d)}\right) = \sum_{i\in U^*}\sum_{k\in U^*}\frac{\pi_{ik}-\pi_i\pi_k}{\pi_{ik}}\underline{w}_i e_{ij}^{(d)^*}\,\underline{w}_k e_{kj}^{(d)^*}.$$

6.3 A small simulation study

In this chapter, we conduct a small simulation study in order to compare the jackknife, the bootstrap and the Augmented variance estimator with the Standard variance estimator. Both SRSWOR and πPS designs are covered. Due to the extensive computing the resampling methods require, both the population and sample size will be relatively small but still realistic.

6.3.1 General setting, estimators and accuracy measures

The population, called Population 3, is of size 150. No domains are considered. Two auxiliary variables and one study variable are generated as follows:

$$\begin{cases}\underline{x}_i \sim Uni(1,2)\\ \underline{z}_i \sim N(100+5p_i,5)\end{cases} \quad i\in U, \tag{6.26}$$

The z-variable is the size variable for πPS sampling. It correlates with the response variable through probabilities $p = P(y = 1)$. These probabilities are generated from x by

$$\underline{p}_i = \frac{\exp\left(\beta_0+\beta_1\underline{x}_i\right)}{1+\exp\left(\beta_0+\beta_1\underline{x}_i\right)},\ \left(\beta_0,\beta_1\right) = (30,-20)$$

The binary study variable is generated as $\underline{y}_i \sim Be\left(\underline{p}_i\right)$. The proportion of y in Population 3 is 72/150 = 0.48. Figure 6.1 shows the histogram of $p = P(y = 1)$ and Figure 6.2 the scatter plot of y, p, z and x.

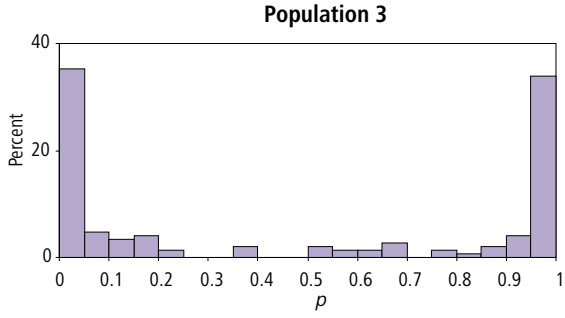


Figure 6.1 Histogram of $p = P(y = 1)$ in Population 3.

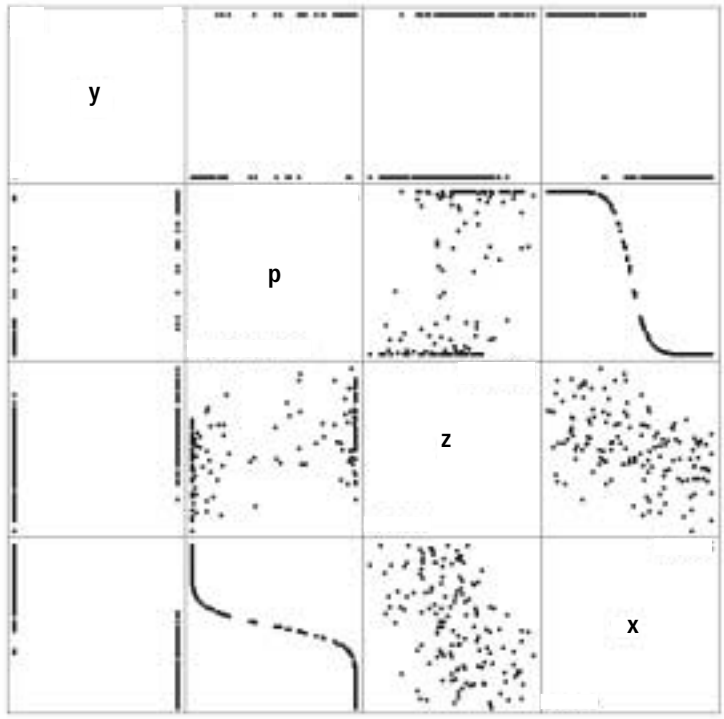


Figure 6.2 Scatter plot of y, p, z and x in Population 3.

K=1,000 independent SRSWOR samples and K=1,000 independent πPS samples (Hanurav-Vijayan algorithm) of n=25 are drawn. From each sample, total of y is estimated with two GREG estimators (Table 6.1).

Both estimators have strong models. The models were estimated using the SAS procedure Reg and weighted least squares for the linear model and SAS procedure Logistic with PML estimation for the logistic model.

For both estimators and for every sample replicate, four variance estimates were calculated. Estimators that were used were the Standard variance estimator, the jackknife, the bootstrap and the Augmented variance estimator. For the bootstrap, 500 pseudo-samples were generated for each sample and for the Augmented variance estimator, 30 pseudo-samples were generated. The number of pseudo-samples was large enough for the Augmented variance estimator to stabilise (see Appendix I).

Table 6.1 Estimators used in the simulations.

Estimator	Functional form of the model	Linear predictor η*	Description of the model
GREG-lin	Linear	$\beta_0 + \beta_1 x_1$	Strong
GREG-log	Logistic	$\beta_0 + \beta_1 x_1$	Strong

* The population generating linear predictor is $\beta_0 + \beta_1 x_1$

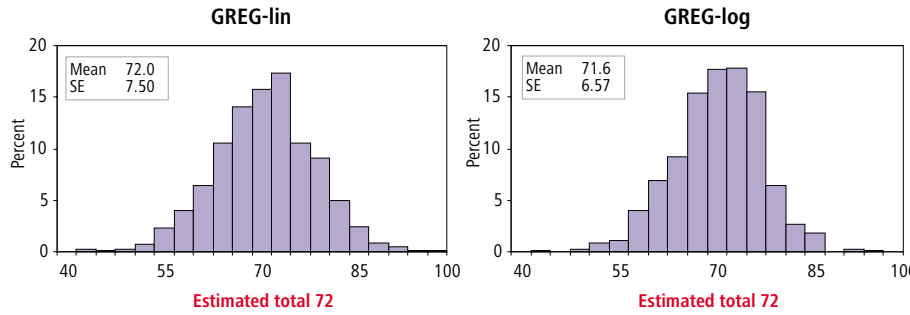


Figure 6.3 Histogram of GREG-lin and GREG-log under the SRSWOR design.

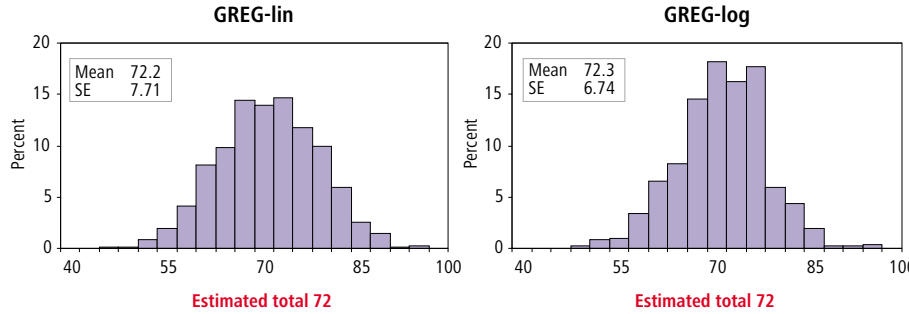


Figure 6.4 Histogram of GREG-lin and GREG-log under the πPS design.

6.3.2 Results

Figures 6.3–6.4 and Tables 6.2–6.5 present the results of the small simulation study. Figures 6.3–6.4 show the distribution of GREG estimates under SRSWOR and πPS designs. The figures show that GREG-log is more accurate than GREG-lin; therefore it is expected that the Standard variance estimator will underestimate the variance of GREG-log.

Table 6.2 shows summary measures for both GREG estimators and the coverage rates of the four variance estimators. Both GREG estimators are approximately unbiased under both designs (ARB close to zero) and GREG-log is significantly better than GREG-lin in terms of standard error SE. The function effect FE summarises the difference: in this simulation, the GREG-log estimator is 14–15% more efficient.

The CR column in Table 6.2 shows that the Standard variance estimator (S) performs relatively well for GREG-lin, but severely underestimates the variance of GREG-log. The Augmented (A), jackknife (J) and bootstrap (B) variance estimators all perform relatively well for GREG-lin, and for GREG-log, each one of these estimators is significantly better than the Standard variance estimator S. In terms of the coverage rate, the Augmented variance estimator A is best, but the differences between A, J and B are small.

Table 6.2 ARB, SE, FE and CR by design and estimator.

Design	Estimator	Accuracy of GREG			CR of variance estimators			
		ARB, %	SE	FE	S	A	J	B
SRSWOR	GREG-lin	0.1	7.5		93.1	93.4	92.3	92.4
	GREG-log	0.5	6.6	1.14	60.1	79.8	78.1	78.1
π**PS**	GREG-lin	0.3	7.7		92.7	92.7	92.5	92.2
	GREG-log	0.4	6.7	1.15	58.7	81.3	81.2	80.1

The first three columns of Table 6.3 show the true (Monte Carlo) standard error, the Standard variance approximation (6.7) and the mean of the pseudo-estimator

$$\hat{V}_{A^*}\left(\hat{\underline{T}}_{j,GREG}\right) = \sum_{i\in U}\sum_{k\in U}\frac{\Delta_{ik}}{E\left(\underline{I}_i\underline{I}_k\right)}\underline{w}_i\underline{e}_{ij}\underline{w}_k\underline{e}_{kj} + \sum_{i\in U}\sum_{k\in U}\frac{\Delta_{ik}}{E\left(\underline{I}_i\underline{I}_k\right)}\underline{w}_i\tilde{\underline{e}}_{ij}\underline{w}_k\tilde{\underline{e}}_{kj}. \quad (6.27)$$

The estimator (6.27), given previously by (6.13), is a pseudo-estimator for the Standard approximation, since it contains terms $\tilde{\underline{e}}_{ij}$ that are not directly observed. From the table, we see that the Standard approximation is quite good when the point estimator is GREG-lin, but for GREG-log, the approximation underestimates the variance. The pseudo-estimator (6.27), however, is accurate for the Standard approximation in both cases.

The last four columns of Table 6.3 present the means of the four standard error estimators. For GREG-log, the bias of S is very large both with respect to the Monte Carlo standard error and with respect to the Standard variance approximation. The Augmented variance estimator A performs much better: its bias with respect to the approximation is very small and the bias with respect to the Monte Carlo standard error comes mainly from the approximation error. In terms of bias, both the bootstrap B and jackknife J are better than S, but not as good as the Augmented variance estimator A.

Table 6.4 presents the total relative biases and relative root mean square errors for the four variance estimators. The bias columns show explicitly that the bias of the Standard variance estimator is huge for GREG-log, smaller for other variance estimators and smallest for the Augmented variance estimator. The RRMSE column of Table 6.4 shows that the standard errors of the variance esti-

Table 6.3 Simulated and approximated standard errors and means of estimates by design and estimator.

Design	Estimator	Standard errors			Mean of estimates			
		MC	Approx.	A* (6.27)	S	A	J	B
SRSWOR	GREG-lin	7.5	7.5	7.5	7.4	7.5	7.3	7.3
	GREG-log	6.6	6.0	5.8	4.2	5.9	5.3	5.4
π**PS**	GREG-lin	7.7	7.5	7.5	7.4	7.5	7.5	7.6
	GREG-log	6.7	5.8	5.8	4.2	5.6	5.4	5.5

Table 6.4 Total relative bias and relative root mean square error of variance estimators by design and estimator.

Design	Estimator	Total relative bias, %				Relative RMSE, %			
		S	A	J	B	S	A	J	B
SRSWOR	GREG-lin	−2.1	−0.7	−3.0	−2.9	12.7	12.9	14.2	20.4
	GREG-log	−36.7	−10.0	−19.1	−18.0	62.0	48.7	51.0	47.3
π**PS**	GREG-lin	−3.7	−2.9	−2.5	−2.0	12.9	12.7	15.7	18.9
	GREG-log	−38.3	−16.4	−19.6	−19.0	62.8	48.0	53.0	47.8

mators vary greatly: although bias was clearly smallest for the Augmented variance estimator, RRMSEs are about the same size for A, J and B.

Table 6.5 shows the empirical lower, upper and total 95 % coverage rates for the four estimators. Differences between A, J and B are small. For GREG-log, the Augmented variance estimator produces best coverage rates (around 80 %). These are too low, but improvement over the Standard variance estimator is significant.

The aim of this small simulation was to find the best alternative for the Standard variance estimator among the three estimators. It turned out that all the three estimators, the jackknife J, the bootstrap B and the Augmented variance estimator A, were clearly better than the Standard estimator, but differences within the triple were small. In terms of bias and coverage rate, the Augmented variance estimator was best by a small margin. In terms of the mean square error, the bootstrap was best, but again, differences between A and B were very small.

Since the coverage rate was previously chosen to be the main measure for the goodness of the variance estimator, the Augmented variance estimator A is chosen for more detailed studies. Another argument for the Augmented estimator is that the number of pseudo-samples it requires is small compared to jackknife and bootstrap. Thus in Chapter 7, we study the properties of the Augmented estimator more extensively and conduct all the simulations of Chapter 5 with this variance estimator.

Table 6.5 Lower, upper and total coverage rates of variance estimators by design and estimator.

Design	Estimator	Lower CR, target 2.5				Upper CR, target 97.5				Total CR, target 95.0			
		S	A	J	B	S	A	J	B	S	A	J	B
SRSWOR	GREG-lin	2.9	2.7	3.4	3.7	96.0	96.1	95.7	96.1	93.1	93.4	92.3	92.4
	GREG-log	22.8	12.6	13.2	13.3	82.9	92.4	91.3	91.4	60.1	79.8	78.1	78.1
π**PS**	GREG-lin	4.2	4.2	4.0	4.3	96.9	96.9	96.5	96.5	92.7	92.7	92.5	92.2
	GREG-log	24.3	11.1	11.3	11.9	83.0	92.4	92.5	92.0	58.7	81.3	81.2	80.1

7 Monte Carlo study II: Comparison of Standard and Augmented variance estimators for L-GREG

In this chapter, we apply the Augmented variance estimator proposed and tentatively tested in Chapters 6.2 and 6.3 to the Monte Carlo simulations of Chapter 5. The aim is to study whether the Augmented variance estimator would provide improvement over the Standard variance estimator when estimating the variance of L-GREG estimators. The settings of the simulations and the GREG estimators are exactly the same as they were in Chapter 5. Chapter 7.1 presents the results for the SRSWOR and Chapter 7.2 the results for the πPS study.

7.1 Experiment II.1: SRSWOR design

Table 7.1 presents accuracy statistics for the Standard variance estimator (S) and the Augmented variance estimator (A) for the case where GREG-log, GREG-prob and GREG-cll were compared under the SRSWOR design (see Chapter 5.2.4.1). The number of pseudo-samples for A was 10. Appendix I provides justification for this choice.

Table 7.1 shows that for every population generating link and every model link, the Augmented variance estimator A performs better than the Standard variance estimator S in terms of bias, root mean square error and coverage rate.

The largest improvement is obtained in terms of estimation error and bias: the mean relative estimation error MREE (the mean relative bias with respect to the Standard approximation) is much smaller for A than S. For example, in the case where the population generating link is logit, MREE for L-GREG estimators is around 11%–12% (in absolute value) for S and around 4% (in absolute value) for A. Consequently, the total bias is also much smaller for A than for S. A similar phenomenon is observed for other population generating links: the Augmented variance estimator A has a much smaller estimation error and total bias than the Standard variance estimator S.

In terms of mean RRMSE, A also outperforms S. In every situation, MRRMSE is smaller for A. The largest difference is obtained on the population level, where MRRMSE for A is often only two thirds of the MRRMSE for S. This holds true for the logit, probit and cll population generating links.

In terms of the coverage rate, the Augmented variance estimator also performs better than the Standard variance estimator. The largest differences are obtained in situations where the CR for S is below 80%. For example, in the case where the population generating link is logit, CRs for L-GREG estimators are around 79%–80% for S and around 84%–85% for A. And when the CRs for S are close to 95%, the CRs for A are also close to 95% (for example, in major domains and on the population level). Similar improvements are observed when the population generating link is probit or cll. Thus the Augmented variance estimator

Table 7.1 Accuracy of the Standard and Augmented variance estimators for GREG-lin and three L-GREG estimators. SRSWOR design, overall sample size 2,000, estimator set 7.

True link	Domain type	Exp. sample size	Link	MRAE	Mean relative estimation error		Mean relative root MSE		Coverage rates	
					S	A	S	A	S	A
Logit	Population	2000	lin	−4.2	−1.0	0.1	5.9	4.7	92.3	92.9
			logit	−6.0	−2.3	0.4	9.8	7.1	92.5	93.5
			probit	−5.9	−2.1	0.4	9.5	7.0	92.1	92.9
			cll	−5.5	−2.1	0.6	9.0	6.4	92.0	93.0
	Major	121	lin	−0.9	−1.0	−0.3	8.8	8.7	93.4	93.9
			logit	−1.6	−2.4	−0.8	14.9	14.5	92.9	93.3
			probit	−1.5	−2.3	−0.7	14.6	14.3	92.6	93.1
			cll	−1.5	−2.0	−0.3	14.7	14.5	92.8	93.2
	Med	48	lin	−1.1	−2.4	−1.1	17.0	16.8	93.3	93.5
			logit	−4.4	−7.1	−3.8	34.9	32.0	86.7	88.0
			probit	−4.2	−6.7	−3.6	33.8	32.0	86.1	88.1
			cll	−3.8	−6.6	−3.3	31.7	31.0	86.6	89.9
	Minor	24	lin	−3.0	−4.6	−1.7	21.7	20.9	88.8	90.0
			logit	−9.3	−12.2	−4.7	48.8	42.2	79.8	84.7
			probit	−9.1	−11.1	−4.2	46.1	42.0	79.3	84.5
			cll	−9.6	−11.6	−4.4	47.3	43.0	79.9	84.6
			Link	**MRAE**	**S**	**A**	**S**	**A**	**S**	**A**
Probit	Population	2000	lin	1.0	−0.6	0.1	1.3	1.2	95.3	95.5
			logit	0.0	−1.9	0.2	3.3	2.7	94.7	95.1
			probit	0.2	−1.8	0.2	3.1	2.7	94.9	95.1
			cll	−0.2	−1.7	0.2	3.3	2.6	94.6	95.2
	Major	121	lin	−1.1	−0.5	−0.2	7.0	7.0	94.2	94.3
			logit	−1.1	−1.7	−0.6	12.6	12.5	93.3	93.7
			probit	−0.9	−1.6	−0.6	12.3	12.1	93.3	93.7
			cll	−1.0	−1.6	−0.5	12.0	11.9	93.4	93.6
	Med	48	lin	−3.1	−1.4	−0.4	12.8	12.5	92.8	93.1
			logit	−4.3	−5.7	−2.1	26.7	25.6	89.2	91.3
			probit	−4.3	−5.3	−2.0	25.9	24.9	89.3	91.5
			cll	−3.9	−5.2	−2.0	25.0	24.1	89.1	91.5
	Minor	24	lin	−1.9	−2.7	−0.9	15.9	15.6	92.3	92.9
			logit	−5.6	−8.5	−3.7	34.7	31.9	85.8	88.5
			probit	−5.5	−7.9	−3.4	33.4	31.8	86.0	88.8
			cll	−5.3	−7.5	−3.1	31.9	30.1	85.7	88.3
			Link	**MRAE**	**S**	**A**	**S**	**A**	**S**	**A**
Cll	Population	2000	lin	−0.7	−0.6	0.1	1.9	1.4	94.2	94.8
			logit	−2.8	−1.4	0.4	5.0	3.5	93.7	94.8
			probit	−2.6	−1.1	0.6	4.6	3.2	94.0	94.8
			cll	−1.6	−1.4	0.3	3.8	2.7	94.2	94.8
	Major	121	lin	0.9	−0.6	−0.2	7.3	7.3	94.7	95.0
			logit	0.2	−1.3	−0.3	11.2	11.2	93.6	94.0
			probit	0.4	−1.2	−0.2	11.0	11.0	93.7	94.1
			cll	0.2	−1.3	−0.4	10.9	10.9	93.7	94.0
	Med	48	lin	−1.7	−1.4	−0.4	13.1	12.9	92.8	93.2
			logit	−2.5	−4.5	−1.5	24.4	24.0	90.0	91.8
			probit	−2.5	−4.1	−1.2	24.0	23.6	90.4	91.8
			cll	−2.3	−4.4	−1.7	23.9	23.3	90.2	92.0
	Minor	24	lin	−2.2	−3.0	−0.9	17.1	16.8	91.8	92.5
			logit	−4.2	−6.9	−2.7	30.0	28.8	87.4	89.8
			probit	−3.9	−6.4	−2.4	28.8	27.7	87.7	89.4
			cll	−3.9	−6.6	−2.6	28.9	27.5	87.6	89.7

improves variance estimation significantly in terms of CR when the CRs of S are too low and when the CR for S is close to 95%, the Augmented variance estimator improves variance estimation in terms of bias and mean square error.

Table 7.2 present corresponding statistics for the Monte Carlo experiments where GREG-log and GREG-lin were compared under the SRSWOR design (see Chapter 5.2.4.2). The number of replicates for the Augmented variance estimator was, as previously, 10. The tables present results only for the GREG-log estimator and model sets 5–8, since for GREG-lin and model sets 1–4, the Standard variance estimator performed relatively well and there was only little room for improvement. (In these cases the A was slightly more accurate than S; selected tables summarizing these results are presented in Appendix III Tables A.III.1 and A.III.2.)

Table 7.2 shows a similar pattern that was observed in Table 7.1: When the Standard variance estimator performs well (CR close to 95%), the Augmented variance estimator also has CR close to 95%, but the bias and mean square error of the Augmented variance estimator are smaller. When the Standard variance estimator S fails (CR lower than 93%), the failure is mainly due to bias. The bias constitutes of the approximation error and estimation error. Especially for minor domains and strong models, the estimation error is large. For example, for n=1,000, minor domains and set 7, the mean relative estimation error for S is –23.0%. For the Augmented variance estimator A, the mean relative estimation errors are in most cases much smaller (in absolute value). For example for n=1,000, minor domains and set 7, the mean relative estimation error for the A is –6.3%. Consequently, total relative biases are also smaller for A than for S.

Thus the general pattern is that the estimation error for A is much smaller than for S. However, the total error is still quite large in some cases. This is due to the approximation error.

There is one exception to the general pattern, where the estimation error for A is smaller than for S: for estimators in set 8, A overestimates the approximation on the population level and in major domains. However, the approximation error and estimation error partly cancel each other, so the total relative bias is still much smaller (in absolute value) for A than for S.

Smaller total bias is associated with smaller relative root mean square error and indeed, A performs better than S in terms of MRRMSE and MCR. The largest differences in terms of MRRMSE are on the population level and the biggest differences in terms of MCR are in minor domains and for strong models. For example for n=1,000, set 6 and population level estimates, MRRMSE is 4.7 for A and 7.0 for S (48 % larger for S). And for n=2,000, set 8 and minor domains, MCR for A is 84.1% and for S, MCR is 75.4%.

Figures 7.1–7.3 visualise the results presented in Tables 7.1–7.2. These figures show the histograms of the Standard (S) and Augmented (A) standard error estimators in selected situations. Appendix IV presents more of these figures for different domains and different estimators.

Figure 7.1 shows the histograms of the standard error estimators for the population level estimator GREG-log-7 for the SRSWOR design and n=5,000. The figure shows that for a large sample, both estimators have a similar, approximately Normal distribution. In this case, both standard error estimators had only a small bias (–0.7% for S, 0.4% for A) and the Standard approximation was also

Table 7.2 Accuracy of the Standard and Augmented variance estimators for GREG–log. Model sets 5–8, SRSWOR design, total sample size 5,000, 2,000 and 1000.

True link	Estimator set	Domain type	Exp. sample size	MRAE	MREE		MRRMSE		MCR	
					S	A	S	A	S	A
n=5000	**5**	Population	5000	−1.0	0.0	0.0	1.6	1.6	94.8	94.8
		Major	303	1.1	−0.1	−0.1	6.2	6.2	94.9	94.9
		Med	121	−0.9	−0.4	−0.4	10.2	10.2	94.3	94.3
		Minor	61	−2.6	−0.7	−0.7	12.2	12.1	93.4	93.4
	6	Population	5000	−1.6	−0.2	0.3	2.2	1.7	94.5	94.6
		Major	303	0.7	−0.2	0.2	6.3	6.4	94.6	94.7
		Med	121	−1.6	−1.3	−0.8	14.2	14.2	93.1	93.3
		Minor	61	−3.5	−1.6	−0.5	13.0	12.6	91.8	92.1
	7	Population	5000	−0.6	−0.7	0.4	2.3	1.9	93.8	94.4
		Major	303	0.1	−0.7	0.0	8.4	8.4	94.3	94.5
		Med	121	−1.7	−2.2	−0.9	17.0	17.0	92.4	92.6
		Minor	61	−4.4	−3.8	−1.0	19.2	18.4	90.2	90.9
	8	Population	5000	−2.4	−0.5	2.1	3.6	2.3	93.7	95.1
		Major	303	−1.5	0.0	2.3	9.3	9.7	94.1	94.5
		Med	121	−2.9	−3.5	−1.1	18.5	18.3	91.5	92.1
		Minor	61	−6.5	−4.9	−0.6	20.6	18.7	89.1	90.6
					S	**A**	**S**	**A**	**S**	**A**
n=2000	**5**	Population	2000	−4.9	0.0	0.0	5.4	5.4	93.8	93.8
		Major	121	−0.8	−0.4	−0.4	10.4	10.4	94.2	94.2
		Med	48	0.9	−1.2	−1.1	16.9	16.8	94.3	94.3
		Minor	24	1.5	−2.3	−2.3	21.0	21.0	93.4	93.4
	6	Population	2000	−5.3	−0.7	0.7	6.3	5.1	93.3	93.6
		Major	121	−2.4	−0.5	0.5	11.1	11.0	93.0	93.3
		Med	48	−1.1	−4.2	−2.7	24.7	24.7	90.5	91.0
		Minor	24	−1.0	−5.4	−2.6	23.2	22.9	88.1	88.8
	7	Population	2000	−6.0	−2.3	0.4	9.8	7.1	92.5	93.5
		Major	121	−1.6	−2.4	−0.8	14.9	14.5	92.9	93.3
		Med	48	−4.4	−7.1	−3.8	34.9	32.0	86.7	88.0
		Minor	24	−9.3	−12.2	−4.7	48.8	42.2	79.8	84.7
	8	Population	2000	−10.7	−1.2	6.7	12.7	6.0	91.2	93.6
		Major	121	−5.1	−0.1	6.0	17.2	17.7	92.3	93.5
		Med	48	−6.8	−11.9	−5.0	36.4	34.5	80.6	83.1
		Minor	24	−13.9	−17.0	−1.6	44.4	34.3	75.4	84.1
					S	**A**	**S**	**A**	**S**	**A**
n=1000	**5**	Population	1000	−0.4	−0.1	0.0	3.1	3.1	94.7	94.7
		Major	61	0.3	−1.1	−1.0	14.4	14.4	94.4	94.4
		Med	24	−0.1	−3.2	−3.1	24.1	24.1	92.6	92.6
		Minor	12	0.3	−4.9	−4.8	30.9	30.9	89.5	89.6
	6	Population	1000	−4.8	−1.5	1.2	7.0	4.7	92.8	93.9
		Major	61	−1.5	−1.6	0.3	15.4	15.4	92.5	92.9
		Med	24	−4.1	−10.6	−7.4	37.1	36.7	86.1	87.4
		Minor	12	−9.7	−11.7	−5.3	38.6	35.2	79.1	82.1
	7	Population	1000	−9.0	−5.0	1.0	14.8	9.3	90.5	92.0
		Major	61	−3.6	−5.3	−2.0	22.7	21.7	89.8	90.7
		Med	24	−9.1	−16.4	−8.5	47.8	44.6	78.6	83.0
		Minor	12	−26.3	−23.0	−6.3	61.5	49.0	62.8	74.8
	8	Population	1000	−17.6	−1.6	14.1	20.5	7.9	84.9	91.4
		Major	61	−11.5	−1.4	9.0	27.3	26.0	88.5	91.5
		Med	24	−13.9	−26.8	−10.1	58.8	50.4	60.7	69.3
		Minor	12	−33.7	−31.2	1.2	73.5	46.7	47.9	73.2

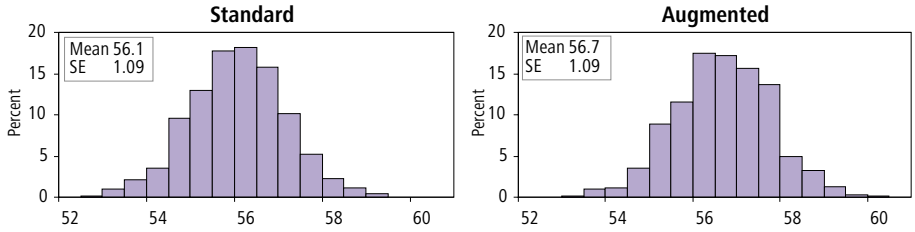


Figure 7.1 Histograms of Standard and Augmented variance estimators. SRSWOR design, n=5,000, population level estimates, estimator set 7, GREG-log. True standard error 56.8.

good: the relative approximation error was only –0.6% (see Table 7.2). Relative RMSE, however, was much smaller for A (1.9 vs. 2.3). Coverage rates were 93.8% for S and 94.4% for A.

Figure 7.2 shows corresponding distributions for the Standard and Augmented estimators for the SRSWOR design, *n*=2,000, domain 5 (minor) and estimator GREG-log-8. Both distributions have a Normal-like shape, but the mean of A is closer to the true standard error (15.5), and the variance of A is smaller than the variance of S. The coverage rate of A was also better in this domain: CR for S was 84.0% and CR for A was 90.2%.

Figure 7.3 shows corresponding distributions for the SRSWOR design, *n*=1,000, domain 15 (medium) and estimator GREG-log-4. In this case, the distributions of both estimators are far from Normal. The distribution of the Standard estimator S has a peak very close to zero (but all estimates are still strictly

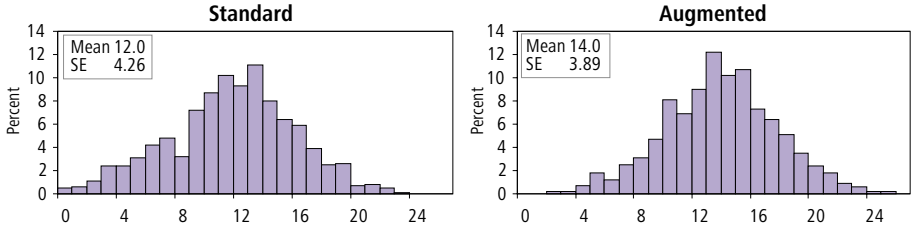


Figure 7.2 Histograms of Standard and Augmented variance estimators. SRSWOR design, n=2,000, domain 5 (minor), estimator set 8, GREG-log. True standard error 15.5.

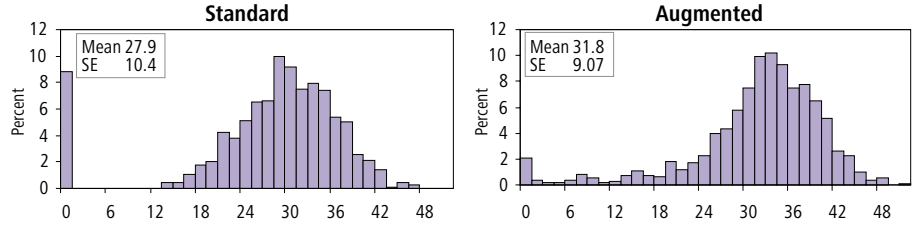


Figure 7.3 Histograms of Standard and Augmented variance estimators. SRSWOR design, n=1,000, domain 15 (medium), estimator set 4, GREG-log. True standard error 33.4.

positive), reflecting the fact that the sample residuals contributing to the variance estimate are all close to zero. These types of variance estimates, which are very close to zero, are clearly undesirable. However, having a variance estimate close to zero is close to the definition of having a strong assisting model: A strong assisting model is such that the outcome of the binary variable is predicted with great precision. This means that residuals are small. The Standard variance estimator takes into account *only* sample fit residuals, which naturally are smaller than (unobserved) out-of-the sample residuals.

The distribution of the Augmented estimator A also has a peak near zero, but compared with the Standard estimator, this peak is very small and the shape of the distribution is much more convenient. The Augmented estimator has fewer very small variance estimates because the term $\hat{V}_M$ takes into account the difference between sample fit and census fit models.

In terms of bias, variance and coverage rate, the Augmented estimator also outperforms the Standard estimator. The mean 32.0 for A is closer to the true standard error 33.4 than the mean 27.9 for S; the variance for A is 81.2 and 109.0 for S, and CR is 88.4 for A, 82.2 for S. These large differences between the Standard and Augmented variance estimator were observed frequently in the experiments; for the interested reader, Appendix IV, Figures A.IV.1–A.IV.2 shows additional histograms similar to Figures 7.1–7.3.

7.2 Experiment II.2: πPS design

Tables 7.3–7.4 present corresponding statistics for Monte Carlo experiments where GREG-log and GREG-lin were compared under the πPS design (see Chapter 5.3). The number of replicates for the Augmented variance estimator A was 10 (Appendix I provides justification for this choice). As in the SRSWOR case, we focus on L-GREG estimators and model sets 5–10. Results for GREG-lin and model sets 1–4 are presented in Appendix III, Tables A.III.3–A.III.6. These tables show that for GREG-lin and the πPS design, the A performs better than S in terms of bias, root mean square error and coverage rate.

Let us first consider cases where the overall sample size was 80 (Table 7.3). The relative estimation error column shows that S underestimates the approximation especially in minor domains. The largest estimation errors (in absolute value) for S are more than 30%, and in many cases, they are close to 20%. The Augmented variance estimator A, in turn, estimates the approximation quite well even in minor domains: in sets 5–8, the largest estimation errors (in absolute value) are only 2.7%, and in sets 9–10, the largest errors are less than 6%.

Smaller estimation errors result in smaller total biases. In Table 7.4, the total bias (the sum of approximation and estimation errors) is smaller for A than for S in almost every case. The largest differences are observed in minor domains and for strong or very strong models, where the total relative bias for A is often less than half of the total relative bias for S.

In terms of root mean square error, the Augmented variance estimator also outperforms the Standard variance estimator. For example, for n=80, size variable z_1, set 7 and major domain, RRMSE for S is 24.3% and for A, RRMSE is

Table 7.3 Accuracy of the Standard and Augmented variance estimators for GREG-log. Model sets 5–10, πPS design, sample size 80, size variables z_1 and z_2.

Size variable	Set	Domain type	Exp. sample size	RAE	REE		RRMSE		CR	
					S	A	S	A	S	A
Z1	**5**	Population	80	1.4	−0.7	0.0	11.4	11.4	93.1	93.2
		Major	60	2.2	−0.8	−0.1	14.0	14.1	93.1	93.3
		Minor	20	−4.7	−0.9	−0.2	24.4	24.1	90.4	90.6
	6	Population	80	0.3	−1.9	0.0	11.6	11.5	93.0	93.6
		Major	60	1.4	−1.2	−0.1	14.0	14.1	92.6	92.9
		Minor	20	−9.6	−6.0	−1.7	31.6	30.4	87.4	88.8
	7	Population	80	−6.7	−7.8	1.5	24.9	17.1	87.5	90.9
		Major	60	−1.5	−5.6	1.5	24.3	19.7	88.6	92.5
		Minor	20	−25.5	−19.1	−0.6	59.2	44.1	63.0	78.0
	8	Population	80	−9.7	−9.4	2.7	27.6	17.3	85.3	91.0
		Major	60	−3.5	−6.6	1.8	26.5	19.4	87.7	92.6
		Minor	20	−30.7	−29.1	1.3	72.5	46.2	48.4	76.6
	9	Population.	80	−11.8	−9.9	4.9	31.1	16.7	83.5	90.8
		Major	60	−7.9	−7.1	5.7	28.9	18.3	84.9	91.8
		Minor	20	−28.9	−22.7	1.6	63.2	42.6	60.5	78.3
	10	Population	80	−13.7	−12.7	5.2	34.6	17.1	79.9	91.2
		Major	60	−8.5	−10.1	4.3	31.8	18.1	83.5	91.3
		Minor	20	−34.5	−31.0	5.5	75.4	43.4	45.0	77.0
					S	**A**	**S**	**A**	**S**	**A**
Z2	**5**	Population	80	−0.5	−0.8	0.0	11.2	11.2	93.6	93.6
		Major	60	0.5	−0.9	−0.1	13.4	13.4	92.9	93.1
		Minor	20	−0.2	−1.2	−0.4	25.5	25.4	91.9	92.1
	6	Population	80	−3.6	−1.8	0.0	12.2	11.6	92.3	92.5
		Major	60	−0.2	−1.2	−0.1	13.6	13.6	92.3	93.1
		Minor	20	−8.4	−6.2	−1.8	32.0	30.8	88.3	89.9
	7	Population	80	−10.4	−7.3	1.0	25.1	17.6	85.5	90.2
		Major	60	−6.1	−5.5	0.7	24.0	18.9	87.9	90.1
		Minor	20	−25.4	−18.1	0.0	58.3	43.6	66.4	77.4
	8	Population	80	−13.7	−9.1	1.6	28.9	19.0	83.6	89.8
		Major	60	−6.9	−6.8	0.7	26.9	19.6	86.4	90.1
		Minor	20	−31.0	−28.4	0.3	72.0	46.6	50.1	75.1
	9	Population	80	−17.1	−9.6	3.5	34.2	20.1	81.5	89.2
		Major	60	−12.4	−8.1	3.3	31.5	19.7	83.8	90.5
		Minor	20	−30.0	−19.3	3.4	62.0	42.5	62.6	77.6
	10	Population	80	−19.2	−12.7	3.4	38.7	21.4	79.4	89.0
		Major	60	−13.6	−10.4	2.8	35.0	20.5	82.2	90.2
		Minor	20	−33.1	−31.1	2.4	74.7	45.4	46.9	75.5

19.7%. When the assisting model is weak or moderate (Table A.III.4), RRMSEs are quite close to each other, but still smaller for A.

In terms of coverage rates, the Augmented variance estimator is also better than the Standard variance estimator. When the Standard variance estimator works well (CR close to 95%, Tables A.III.3–A.III.4), A is also close to 95%. But when the model is strong or very strong (Tables 7.3–7.4), S produces too low coverage rates. In such cases, A is often much closer to the nominal 95% coverage rate. For ex-

Table 7.4 Accuracy of the Standard and Augmented variance estimators for GREG–log. Model sets 5–10, πPS design, sample size 40, size variables z_1 and z_2.

Size variable	Set	Domain type	Exp. sample size	RAE	REE		RRMSE		CR	
					S	A	S	A	S	A
Z1	**5**	Population	40	−1.3	−2.1	−0.8	19.7	19.6	92.1	92.4
		Major	30	−0.7	−2.4	−1.0	23.7	23.6	89.0	89.0
		Minor	10	−4.2	−3.3	−1.7	42.0	41.4	77.1	78.0
	6	Population	40	−5.9	−4.5	−1.2	21.5	20.4	89.7	90.9
		Major	30	−2.1	−3.5	−1.5	25.4	25.3	87.2	87.9
		Minor	10	−23.1	−16.6	−9.1	58.2	54.3	65.2	67.2
	7	Population	40	−22.1	−21.4	0.4	55.7	33.3	67.1	83.1
		Major	30	−17.0	−20.4	−0.8	55.2	36.7	67.2	82.2
		Minor	10	−43.6	−30.5	−0.3	83.2	59.5	35.8	62.7
	8	Population	40	−25.5	−23.0	1.8	58.3	32.9	63.1	81.9
		Major	30	−20.0	−22.7	−0.3	59.3	36.9	63.1	81.5
		Minor	10	−45.1	−39.1	0.9	90.6	56.9	19.8	60.1
	9	Population	40	−27.9	−31.4	2.1	69.3	35.5	51.2	78.8
		Major	30	−24.2	−30.2	1.9	68.1	37.7	52.7	78.1
		Minor	10	−45.8	−37.5	1.0	88.6	54.5	26.6	62.7
	10	Population	40	−32.2	−32.6	3.8	72.8	35.9	47.4	77.8
		Major	30	−25.6	−32.9	1.5	71.0	38.1	49.5	78.8
		Minor	10	−51.2	−40.3	7.3	94.4	50.1	14.0	62.6
					S	**A**	**S**	**A**	**S**	**A**
Z2	**5**	Population	40	−4.0	−1.9	−0.7	18.7	18.4	91.3	91.7
		Major	30	−4.3	−2.1	−0.9	22.3	21.9	87.3	87.5
		Minor	10	−7.4	−3.8	−2.3	41.7	41.0	75.7	76.3
	6	Population	40	−9.2	−4.5	−1.4	21.9	20.5	88.5	89.4
		Major	30	−7.3	−3.2	−1.4	24.5	23.9	85.8	86.6
		Minor	10	−19.9	−18.4	−10.9	58.9	55.4	61.7	65.0
	7	Population	40	−23.5	−22.3	−1.6	57.2	35.8	64.9	79.2
		Major	30	−19.2	−20.9	−3.2	56.3	39.3	65.3	79.9
		Minor	10	−43.3	−32.5	−1.7	84.6	60.1	32.1	61.1
	8	Population	40	−25.4	−23.9	−0.1	59.1	34.8	63.1	79.8
		Major	30	−21.8	−22.7	−2.5	59.7	39.2	63.1	79.2
		Minor	10	−42.8	−41.4	−1.7	91.0	57.9	18.3	57.1
	9	Population	40	−30.3	−30.5	1.1	69.9	37.9	50.5	76.1
		Major	30	−27.2	−29.5	−0.2	68.6	40.5	52.4	75.6
		Minor	10	−46.3	−36.8	2.5	88.9	54.3	25.3	61.8
	10	Population	40	−30.6	−33.8	1.8	72.4	36.5	47.6	78.2
		Major	30	−28.3	−32.0	−0.8	71.2	40.9	49.6	75.2
		Minor	10	−48.1	−43.2	4.9	94.5	50.2	12.6	61.6

ample, for n=80, size variable z_1, set 7 and major domain, CR is 88.6% for S and 92.5% for A. And for n=80, size variable z_2, set 10 and major domains, CR is 82.2% for S and 90.2% for A. Moreover, there are no situations where A would perform worse than S.

Let us next consider the case n=40 (Table 7.4). The general pattern is very much the same that it was with n=80: the relative estimation error for S is large in minor domains. From set 7 onwards, REE for S is always more than 20% (in absolute value). The estimation error for A is, however, much smaller and even in minor domains and for very strong models, REE is less than 10%. Consequently, in most cases, total relative biases for A are only a fraction of the total

relative biases for S. Relative root mean square errors are also smaller for A than for S, and coverage rates for A are much closer to 95% than they are for S.

Let us next look at the distribution of the estimators. Figures 7.4–7.7 present the histograms of the Standard (S) and Augmented (A) standard error estimators for several situations. (Appendix IV, Figures A.IV.3–A.IV.4 shows that the situation observed in Figures 7.6–7.7 is not uncommon.)

Figure 7.4 shows the histograms of the standard error estimators for the population level estimator GREG-log-7 for the πPS design, size variable z_1 and n=80. The true standard error is 9.04 and both the Standard estimator (7.74) and the Augmented estimator (8.58) underestimate this figure. But both bias and variance are smaller for the Augmented estimator. The coverage rate is 90.9% for the Augmented estimator A and 87.5% for the Standard estimator S (see Table 7.3). Thus S is clearly better than A. But in addition to these standard performance statistics, A mostly avoids the problem of very small variance estimates. The distribution of S has a peak near zero, reflecting an almost perfect sample fit. The distribution of A does not have that kind of peak. This is due to the term $\hat{V}_M$ that takes into account the difference between sample fit and census fit residuals.

Figures 7.5–7.7 show corresponding distributions for different sample sizes, size variables, estimators and domain. Figure 7.5 is for πPS design and n=80, size variable z_2, minor domain and GREG-log-4. In this case, the logistic model was not as strong, as it was in Figure 7.4. Thus the Standard estimator S also mostly avoids the problem of very small variance estimates and both estimators have quite Normal distributions. However, in terms of bias, variance and coverage rate, the Augmented estimator A is better than S (see Table 7.4).

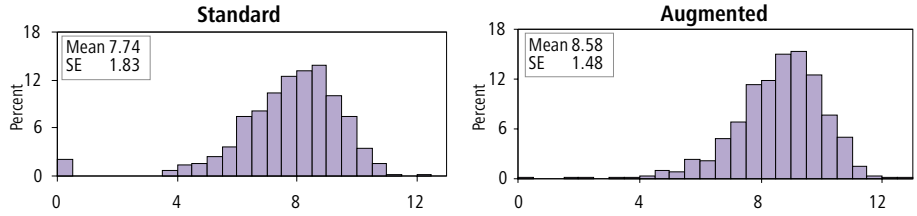


Figure 7.4 Histograms of Standard and Augmented variance estimators. πPS design, size variable z_1, n=80, population level estimates, GREG-log, set 7. True standard error 9.04.

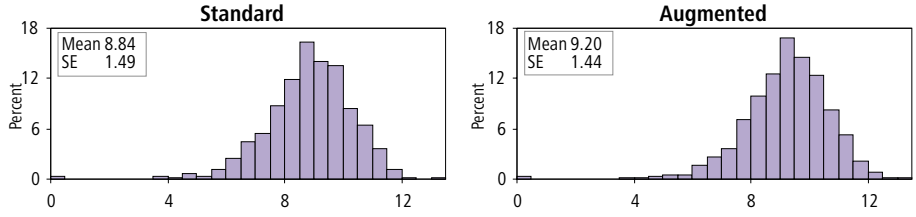


Figure 7.5 Histograms of Standard and Augmented variance estimators. πPS design, size variable z_2, n=80, minor domain, GREG-log, set 4. True standard error 9.67.

Figure 7.6 is for the πPS design and n=40, size variable z_1 and population level GREG-log-9. In this case, the Standard estimator S fails completely in terms of bias and CR (see Table 7.4). Figure 7.6 shows that almost half of the estimates of S are very close to zero. The Augmented estimator A also has quite a large bias, but the bias is less than half of the bias of S. RRMSE and CR are also much better for A than for S. In addition, the problem of very small variance estimates is much smaller for A than for S.

Figure 7.7 is for the πPS design and n=40, size variable z_2 and population level GREG-log-10. Just like in the previous case, S fails completely in terms of bias and CR (see Table 7.4), and almost half of the variance estimates are very close to zero. The bias and variance of A are much smaller, and the coverage rate, although still too low (61.6%) is much better than the coverage rate of S (12.6%). In addition, the problem of very small variance estimates is far more severe for S than for A.

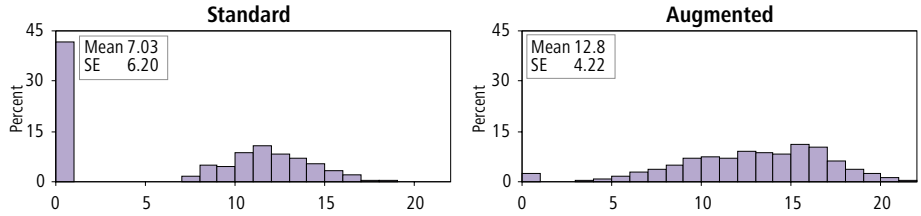


Figure 7.6 Histograms of Standard and Augmented variance estimators. πPS design, size variable z_1, n=40, population level estimates, GREG-log, set 9. True standard error 17.3.

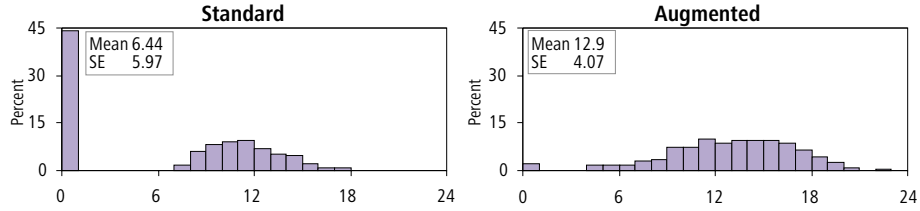


Figure 7.7 Histograms of Standard and Augmented variance estimators. πPS design, size variable z_2, n=40, population level estimates, GREG-log, set 10. True standard error 18.1.

8 Conclusions

In this thesis, we studied the properties of design-based model-assisted domain class frequency GREG estimators that have a logistic-type assisting model (e.g. logit, probit, or complementary log-log model). These estimators are called L-GREG estimators. In the context of this study, the estimators can equally well be applied to proportions, since domain sizes are assumed to be known.

L-GREG estimators belong to the family of GREG estimators, which forms the class of design-based model-assisted estimators. All GREG estimators utilise auxiliary information via modelling. The classic GREG estimator with a linear fixed effects model (GREG-lin) is one example of them.

The linear model formulation is perfectly acceptable when the response variable is continuous. But when estimating class frequencies, the study variable is binary or polytomous. From the modeller's point of view, then, logistic-type models would be preferred over the linear model. However, other GREG estimators than GREG-lin are rarely used, and knowledge about their properties is limited. In this study, some properties of L-GREG estimators for domain class frequencies were examined.

In the Introduction (Chapter 1), three sets of research questions were set. To answer these questions, both simple random sampling without replacement (SRSWOR) and fixed size without replacement probability proportional to size (πPS) designs were covered. The accuracy of estimators was considered only with respect to sampling and the effect of non-sampling errors, such as nonresponse, frame imperfection or measurement errors, was not studied. The research questions were as follows.

The first set of research questions: Are L-GREG estimators more accurate for domain class frequencies than the GREG-lin estimator? If yes, when? And may L-GREG estimators be less accurate than GREG-lin?

Theoretical arguments and Monte Carlo experiments that covered both SRSWOR and πPS designs showed that the accuracy difference between L-GREG and GREG-lin estimators depends especially on the strength of the assisting model. We defined a strong model as a model that predicts the outcome of the binary variable with great precision. A related concept is a correct model, which is a model that has the same effects and the same functional form as the population generating process. It should be noted that a strong model may be correct or incorrect (for example, by adding unnecessary effects to a strong and correct model one often gets an incorrect, strong model), and a weak model may also be correct or incorrect.

If a logistic-type model is strong, the probabilities obtained from the model are close to zero or one. In the Monte Carlo studies we considered a large set of models of different strength, because in practice the amount and quality of auxiliary information set the limits to the strength of the model.

When the model is strong and the sample size in the domain is not very small, the L-GREG estimators outperformed the GREG-lin estimator. For very strong models, L-GREG estimators may have standard error tens of percentages smaller

than GREG-lin, still being approximately unbiased. However, for domains with a very small sample size, L-GREG may be biased if the assisting model is very strong.

When the model is not strong, GREG-lin and L-GREG estimators produced very similar results. Both were approximately unbiased and standard errors were almost identical. These results were obtained both in the SRSWOR and πPS designs. Thus when the model is strong, L-GREG has the potential of being more accurate than GREG-lin. In standard situations, when there are not very strong predictors available, no strong models can be constructed. In such situations, L-GREG and GREG-lin produce approximately equal results.

We also studied the relative accuracies of different L-GREG estimators: GREG with a logistic model (GREG-log), GREG with a probit model (GREG-prob) and GREG with a complementary log-log model (GREG-cll). Differences between these estimators were small. This may be due to the fact that there were continuous variables in the models. From this it follows that response variables are not grouped but are binary. When grouping is done, we observe empirical probabilities in every group, and in such cases, the form of the non-linear link may be more important.

If GREG-log, GREG-prob and GREG-cll yield similar results, which link to use? The choice may be based on practical considerations, such as what link is available in the software that is being used. However, the logit link may be preferred because i) the logit link is the canonical link, ii) use of the logit link makes it easy to construct odds ratios which are often of interest, iii) the probit link yields almost identical results, but the logistic transformation is computationally easier than the probit, and iv) use of asymmetric links, such as complementary log-log, might be difficult to justify in practice.

The second set of research questions: How well does the Standard variance approximation and the corresponding Standard variance estimator work for L-GREG estimators? If they perform well, when? If they fail, when?

The Standard variance estimator resembles the well known Sen-Yates-Grundy variance estimator, but it is a double sum of prediction errors, not a double sum of the observed values of the study variable. The Standard variance estimator is widely used for GREG-lin, and in the literature, the Standard variance estimator has also been suggested for L-GREG estimators. Monte Carlo experiments that covered both SRSWOR and πPS designs showed that the Standard variance estimator may fail for L-GREG estimators. Especially if the domain sample size is small, or if the assisting model is strong and/or overspecified, the Standard variance estimator is biased downwards and produces too narrow confidence intervals. For domains with a large sample size and estimators with weak and non-overfitted models, the estimator performs well. Interestingly, the strength of the model was at least as important in variance estimation as the domain sample size: for weak models, Standard variance estimator performed very well even in smallest domains, but if the model was strong, results were catastrophic.

So, the larger the domain sample size, the better the performance of the Standard variance estimator. The weaker the model, the better the performance of the Standard variance estimator. And for overfitted models, the performance of the Standard variance estimator is worse than for corresponding non-overfitted models.

By decomposing the error of the Standard variance estimator into the approximation error and estimation error, we observed that both error sources contribute to the total error roughly equally much. Moreover, by decomposing the census fit residuals of the Standard variance approximation into sample fit residuals and residuals that reflect the difference between sample fit and census fit, we observed that the source of the estimation error is that the Standard variance estimator does not take into account the difference between sample fit and census fit models. This leads us to the third set of research questions.

The third set of research questions: Does the Augmented variance estimator that takes into account the difference between sample fit and census fit models provide any improvement over the Standard variance estimator? If yes, when? And under what conditions does the Augmented variance estimator perform worse than the Standard variance estimator?

To improve variance estimation, we proposed a new variance estimator called the Augmented variance estimator. The Augmented variance estimator estimates the Standard variance *approximation*, like the Standard variance estimator. The difference with respect to the Standard variance *estimator* is that the Augmented variance estimator does take into account the difference between the sample fit and census fit models.

The goodness of the Augmented variance estimator was examined by Monte Carlo experiments that covered both SRSWOR and πPS designs. In addition to the Augmented variance estimator, the performance of two well-known resampling based variance estimators, the delete-one jackknife and the without replacement bootstrap, were studied.

In the Monte Carlo experiments, the jackknife, the bootstrap and the Augmented variance estimator all performed better than the Standard variance estimator. In terms of mean square error, the bootstrap was most accurate, but in term of bias and coverage rate, the Augmented variance estimator was best. Moreover, the Augmented variance estimator, although being computer intensive, requires less pseudo-samples than the bootstrap: in most cases, only ten replicates were needed to stabilise the estimator. This is much less than what is often recommended for the bootstrap.

The Augmented variance estimator was constructed to estimate the Standard variance approximation. The bias (or estimation error) of the Augmented variance estimator with respect to the Standard variance approximation was generally much smaller than the bias of the Standard variance estimator. But the total bias, which consists of the approximation error of the Standard variance approximation and the estimation error of the variance estimator with respect to the approximation, may be significant even if the latter error source is small. Thus, the Augmented variance estimator provides a promising alternative to the Standard variance estimator: in the Monte Carlo experiments, it almost always performed at least as well as the Standard estimator, and in many cases, significantly better. However, when the assisting model was strong and domain sample size small, also the Augmented variance estimator produced too low confidence intervals; the main source of remaining error was approximation error. Therefore one should be cautious about the estimated confidence intervals if domain sample size is small and the assisting model strong. For large sample sizes the use of an

L-GREG estimator accompanied with the Augmented variance estimator seems safe. But clearly further research is needed in order to address the approximation error of the Standard variance approximation.

To summarise: The L-GREG estimator has the potential of being more accurate than GREG-lin. Especially if the model is very strong and domain sample size is not very small, L-GREG is more accurate. And only in situations where domain sample size was very small and assisting model very strong, GREG-lin was more accurate. In suc situations. L-GREG became biased.

Larger than 2% bias for L-GREG was observed when the expected domain sample size was less than 15 and the assisting logistic-type model was very strong. In such cases, the corresponding GREG-lin estimator retained approximate unbiasedness. The source of the bias was inaccurate estimation of the parameters of the assisting logistic-type model: We compared the distributions of the pseudo-maximum likelihood estimators with census fit parameters, and for domain sample sizes less than 15 and for very strong models, the pseudo-maximum likelihood estimators had large biases and variances. Moreover, the biased L-GREG estimators would have been unbiased if the model parameters had been census fit parameters. Thus the bias emerged when assisting model parameters were estimated from the sample.

We used pseudo-maximum likelihood estimation with Newton-Raphson algorithm to estimate the parameters of the logistic-type assisting model. Sampling weights were always used for model parameter estimation. Whether other existing methods would provide more robust model parameter estimates, and what would be the result if model parameters were estimated without weighting, is a topic of further research.

References

Aldrich, J. H. and Nelson, F. D. (1985). Linear Probability, Logit and Probit Models. Sage Publications, Beverly Hills, California.

Armitage, P., Berry, G. and Matthews, J.N.S (2002). Statistical Methods in Medical Research. Blackwell, Oxford.

Axelson, M. (2000). Variance estimation for the two-phase regression estimator – a calibration approach. Proceedings of the Survey Research Methods Section, American Statistical Association.

Berger, Y.G. (2003). A modified Hajek variance estimator for systematic sampling. Statistics in Transition, June 2003 Vol. 6, No. 1, 5–2.

Berger Y.G. (2004). A simple variance estimator for unequal probability sampling without replacement. Journal of Applied Statistics, Vol. 31, No. 3, 305–315.

Bickel, P. J. and Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. The Annals of Statistics, Vol. 12, No. 2, 470–482.

Binder, D. (1983). On the variance of asymptotically normal estimators from complex surveys. International Statistical Review, Vol. 51, 279–292.

Booth, J.G., Butler, R.W. and Hall, P. (1994). Bootstrap methods for finite populations. Journal of the American Statistical Association, Vol. 89, No. 428, 1282–1289.

Box G.E.P. (1979). Robustness in the strategy of scientific model building. In: Launer R. L. and Wilkinson G. N. (1979). Robustness in Statistics. New York: Academic Press, 202.

Breidt, F.J. and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. Annals of Statistics, Vol. 28, No. 4, 1026–1053.

Brown, H. and Prescott, R. (1999). Applied Mixed Models in Medicine. Wiley, Chichester.

Chambers, R. (2003). An Introduction to Model Based Survey Sampling. Southampton Statistical Sciences Research Institute, University of Southampton, Southampton.

Cochran, W. G. (1962). Sampling Techniques. Wiley, New York.

Collett, D. (2002). Modelling Binary Data. Chapman & Hall/CRC, London.

Cordeiro, G.M. and McCullagh, P. (1991). Bias correction in generalized linear models. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 53, No. 3, 629–643.

Davidson, R. and MacKinnon, J. G. (1993). Estimation and Inference in Econometrics. Oxford University Press, New York.

Deville, J.-C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association, Vol. 87, No. 418, 376–382.

Deville, J.-C., Särndal, C.E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. Journal of the American Statistical Association, Vol. 88, No. 423, 1013–1020.

Djerf, K. (2001). Properties of Some Estimators Under Unit Nonresponse. Statistics Finland, Research Reports 231, Helsinki.

Duchesne, P. (2003). Estimation of a proportion with survey data. Journal of Statistics Education, Vol. 11, No. 3.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. The Annals of Statistics Vol. 7, No. 3,1–26.

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. Biometrika, Vol. 68, 589–599.

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. Society of Industrial and Applied Mathematics CBMS-NSF Monographs, 38.

Estevao, V., Hidiroglou, M.A. and Särndal, C.E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. Journal of Official Statistics, Vol. 11, No. 2, 181–204.

Estevao, V.M. and Särndal, C.E. (2003). A new perspective on calibration estimators. Proceedings of the Survey Research Methods Section, American Statistical Association.

Estevao V.M. and Särndal C.E. (2004). Borrowing strength is not the best technique within a wide class of design-consistent domain estimators. Journal of Official Statistics, Vol. 20, No. 4, 645–669.

Estevao, V.M. and Särndal, C.E. (2006). Survey estimates by calibration on complex auxiliary information. International Statistical Review, Vol. 74, No. 2, 127–147.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. Biometrika 1993, Vol. 80, No. 1, 27–38.

Firth, D. and Bennet, K.E. (1998). Robust models in probability sampling. Journal of the Royal Statistical Society. Series B (Statistical Methodology), Vol. 60, No. 1, 3–21

Ghosh, M. (2001). Lecture notes on estimation for population domains and small areas. Proceedings of the Symposium on Advances in Domain Estimation. Statistics Finland, Reviews 2001/5, Helsinki.

Gross, S.T. (1980). Median estimation in sample surveys. American Statistical Association, Proceedings of Survey Research Methods Section 181–184.

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E. and Tourangeau, R. (2004). Survey Methodology. Hoboken, NJ: Wiley-Interscience.

Hajek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. Annals of Mathematical Statistics, Vol. 35, No. 4, 1491–1523.

Hansen, M.M. and Hurwitz, W.N. (1943). On the theory of sampling from finite populations. Annals of Mathematical Statistics, Vol. 14, No. 4, 333–362.

Hanurav, T.V. (1967). Optimum utilization of auxiliary information: pps sampling of two units from a stratum. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 29, No. 2, 374–391.

Hartley, H.O. and Rao, J.N.K. (1962). Sampling with Unequal Probabilities and without Replacement. Annals of Mathematical Statistics, Vol. 33, 350–374.

Heyde, C. C. and Seneta, E. (editors) (2001). Statisticians of the Centuries. Springer, New York.

Hidiroglou, M. A. and Särndal, C. E. (1998). Use of auxiliary information for two-phase sampling. Survey Methodology, Vol. 24, No. 1, 11–20.

Hidiroglou, M. and Patak, Z. (2001). Domain estimation using linear regression. Proceedings of the Annual Meeting of the American Statistical Association, August 5–9.

Horvitz, D. G., and Thompson D. J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, Vol. 47, No. 260, 663–685.

Jurevitš, N. (2005). The bias-corrected regression estimator. Proceedings of the SAE 2005 Small Area Estimation conference, August 2005, Jyväskylä, Finland.

Knottnerus, P. (2003). Sample Survey Theory. Some Pythagorean Perspectives. Springer, New York.

Krewski, D. and J.N.K. Rao (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. The Annals of Statistics, Vol. 19, 1010–1019.

Laaksonen, S. (2006). Does the choice of link function matter in response propensity modelling? Model Assisted Statistics and Application, Vol. 1, No. 2, 95–100.

Lahiri, P. (2003), On the impact of bootstrap in survey sampling and small-area estimation. Statistical Science, Vol. 18, 199–210.

Lehtonen, R. and Pahkinen, E. (2004). Practical Methods for Design and Analysis of Complex Surveys. Second Edition. Wiley, Chichester.

Lehtonen, R. and Veijanen, A. (1998a). Logistic generalized regression estimators. Survey Methodology, Vol. 24, No.1, 51–55.

Lehtonen, R. and Veijanen, A. (1998b). On multinomial logistic generalized regression estimators. Preprints from the Department of Statistics, University of Jyväskylä, No. 22. Jyväskylä.

Lehtonen, R., Särndal, C.E. and Veijanen, A. (2003). The effect of model choice in estimation for domains, including small domains. Survey Methodology, Vol. 29, No.1, 33–44.

Lehtonen, R., Särndal, C.E. and Veijanen, A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. Statistics in Transition, December 2005, Vol. 7, No. 3, 649–673.

Lehtonen, R., Myrskylä, M., Särndal, C.E. and Veijanen, A. (2006a). Model-assisted and model-dependent estimation for domains and small areas under unequal probability sampling. 9th International Vilnius Conference on Probability Theory and Mathematical Statistics, June 25–30, 2006.

Lehtonen, R., Myrskylä, M., Särndal, C.E. and Veijanen, A. (2006b). The role of models in model-assisted and model-dependent estimation for domains and small areas. Paper presented at the Workshop on Survey Sampling Theory and Methodology, Ventspils, Latvia, August 24–28, 2006.

Lindsey, J.K. (2000). Applying Generalized Linear Models. Springer, New York.

Lohr, S. (1999). Sampling: Design and Analysis. Duxbury Press, Pacific Grove.

Longford, N. T. (1995). Random Coefficient Models. Oxford University Press, Oxford.

Lundström, S. (1997). Calibration as a Standard Method for Treatment of Nonresponse. Department of Statistics, Stockholm University, Edsbruck.

Lundström, S. and Särndal, C.E. (2002). Estimation in the presence of Nonresponse and Frame Imperfections. Statistics Sweden, Örebro.

Mahalanobis P.C. (1946). Recent Experiments in Statistical Sampling in the Indian Statistical Institute. Journal of the Royal Statistical Society, Vol. 109, 325–370.

McCarthy, P.J. and Snowden, C.B. (1985). The bootstrap and finite population sampling. Vital and Health Statistics, Series 2, No. 95, Public Health Service Publication 85–1369.

McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models. Second Edition. Chapman and Hall, London.

McCulloch, C. E. and Searle, S. R. (2001). Generalized, Linear and Mixed Models. Wiley, New York.

Musting, K. (2004). Study of the bias of generalized regression estimator. Workshop on Survey Sampling Theory and Methodology, June 18–22, 2004, Tartu, Estonia, 78–81.

Myrskylä, M. (2004). Estimation of class frequencies with micro level auxiliary information. An application to Finnish Labour Force Survey. Unpublished master's thesis, University of Jyväskylä.

Myrskylä, M. (2005). Logistic-type generalized regression estimators for class frequencies in domains. Unpublished licentiate thesis, University of Jyväskylä.

Opsomer J.D., Moisen, G.G. and Kim, J.Y. (2001). Model-assisted estimation of forest resources with generalized additive models. Proceedings of the Section on Survey Research Methods, American Statistical Association [CD-ROM], Alexandria VA, Article #00369, 6 pp.

Opsomer, J.D., Wang, Y. and Yang Y. (2001). Nonparametric regression with correlated errors. Statistical Science, Vol. 16, No. 2, 134–153.

Pahkinen, E. and Lehtonen, R. (1989). Otanta-asetelmat ja tilastollinen analyysi. (Sampling designs and statistical analysis; in Finnish only). Gaudeamus, Helsinki.

Presnell, P. and Booth, J.G. (1994). Resampling methods for sample surveys. Technical Report 470, Department of Statistics, University of Florida.

Rao, J.N.K. (2005). Interplay between sample survey theory and practice: an appraisal. Survey Methodology, Vol. 31, No. 2, 117–138.

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

Rao, J.N.K. and Wu, C.F.J. (1988). Resampling inference with complex survey data. Journal of the American Statistical Association, Vol. 83, No. 401, 231–241.

Robert, C. P. and Casella, G. (2002). Monte Carlo Statistical Methods. Springer, New York.

Saei, A. and Chambers, R. (2003a). Small Area Estimation Based on a Unit Level Linear Mixed Model with Correlated Time Effect. EURAREA Project, University of Southampton.

Saei, A. and Chambers, R. (2003b). Small Area Estimation Under Linear and Generalized Linear Models With Time and Area Effects. EURAREA Project, University of Southampton.

SAS OnlineDoc (1999). http://v8doc.sas.com/sashtml/ (7.8.2006) SAS Institute Inc., Cary, NC.

Sen, A.R. (1953). On the estimate of the variance in sampling with varying probabilities. Journal of the Indian Society of Agricultural Statistics, Vol. 5, 119–127.

Shao, J. (2003). Impact of the Bootstrap on Sample Surveys. Statistical Science, Vol. 18, 191–198.

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

Sitter, R.R. (1992a). Comparing three bootstrap methods for survey data. Canadian Journal of Statistics, Vol. 20, No. 2, 135–154.

Sitter, R.R. (1992b). A resampling procedure for complex survey data. Journal of the American Statistical Association, Vol. 87, No. 419, 755–765.

Skinner, C.J. (1989). Domain Means, Regression and Multivariate Analysis. In Analysis of Complex Surveys, (C.J. Skinner, D.Holt and T.M.F. Smith, eds.) Wiley, 59–87.

Sova, M. G. (2004). Model based survey inference: a brief introduction. Proceedings of the Workshop on Survey Sampling Theory and Methodology. Tartu, Estonia 2004.

Stigler, S. M. (1986). The History of Statistics. Belknap Press, Cambridge.

Särndal, C.E. (1996). Efficient estimators with simple variance in unequal probability sampling. Journal of the American Statistical Association, Vol. 91, No. 435, 1286–1300.

Särndal, C.E. (2003). A brief history of ideas in survey theory. Presentation held 8.12.2003 at Seminar of survey statistics in the University of Helsinki (unpublished).

Särndal, C.E., Swensson, B., and Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. Biometrika, Vol. 76, No. 3, 527–537.

Särndal, C.E., Swensson, B. and Wretman, J. (1992). Model Assisted Survey Sampling. Springer, New York.

Thompson, S. K. (1992). Sampling. Wiley, New York.

Tille, Y. (1996). Some remarks on unequal probability sampling designs without replacement. Annales d'economie et de statistique, No. 44, (1996).

Traat, I. (2000). Sampling design as a multivariate distribution. New Trends in Probability and Statistics, Vol. 5, 195–207.

Traat, I., Bondesson, L. and Meister, K. (2003). Sampling design and sample selection through distribution theory. Journal of Statistical Planning and Inference, Vol. 123, 395–413.

Valliant, R. (2002). Variance estimation for the general regression estimator. Working paper series 125, Survey Research Center, Institute for Social Research, University of Michigan.

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

Vijayan, K. (1968). An exact pps sampling scheme: generalization of a method of Hanurav. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 30, No. 3, 556–566.

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

Woodruff, R. (1971). A simple method for approximating the variance of a complicated estimate. Journal of the American Statistical Association, Vol. 66, 411–414.

Wright, T. (2001). Selected moments in the development of probability sampling: theory & practice. Survey Research Methods Section Newsletter. American Statistical Association, Alexandria.

Wu, C. (1999). The Effective Use of Complete Auxiliary Information from Survey Data. PhD dissertation, Simon Fraser University, Canada.

Wu, C. (2003). Optimal calibration estimators in survey sampling. Biometrika, Vol. 90, No. 4, 937–951.

Wu, C. and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. Journal of the American Statistical Association, Vol. 96, No. 453, 185–193.

Yates, F., and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 15, No. 2, 253–261.

Appendix I.
How large K is needed for the Augmented variance estimator?

The Augmented variance estimator $\hat{V}^* = \hat{V}_S + \hat{V}_M$ (see Chapter 6.3) consists of two parts: the first term is the Standard variance estimators and the second term corrects for the difference between the sample fit and census fit models. To calculate $\hat{V}_M$, a resampling based procedure was proposed in Chapter 6.3. In this procedure, one constructs a pseudo-population from the sample and draws K pseudo-samples from the pseudo-population. The difference between census fit estimates for the pseudo-population and sample fit estimates for the pseudo-samples allows us to calculate $\hat{V}_M$, which is the mean of values $\hat{V}_M^{(k)}, k = 1,2,\ldots,K$.

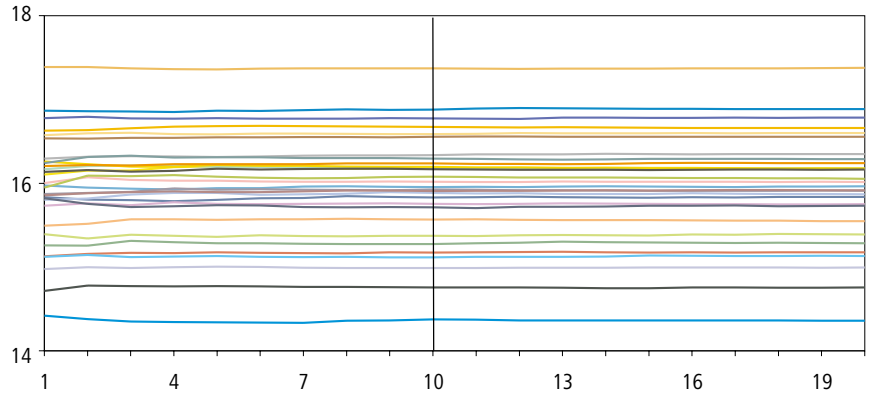
The reason that we draw more than one pseudo-sample from the pseudo-population is that we can decrease the variance of the Augmented estimator by doing this. For small K, the variance of $\hat{V}_M$ may be large, and the variance of $\hat{V}_M$ contributes to the variance of $\hat{V}^*$. As K grows, both the variance of $\hat{V}_M$ and the variance of $\hat{V}^*$ decrease. But how large K do we need?

The number of replicates should be so large that $\hat{V}^*$ (and $\hat{V}_M$) stabilises. Thus, a general rule is to calculate values $\hat{V}_M^{(k)}$ until $\hat{V}^*$ and $\hat{V}_M = K^{-1}\sum_{k=1}^{K}\hat{V}_M^{(k)}$ stabilises. Obviously, the larger the sample size, the faster $\hat{V}^*$ stabilises. Therefore we chose K so that for the smallest sample sizes and the smallest domains, $\hat{V}^*$ stabilises. Then $\hat{V}^*$ also stabilises for larger domains and larger sample sizes.

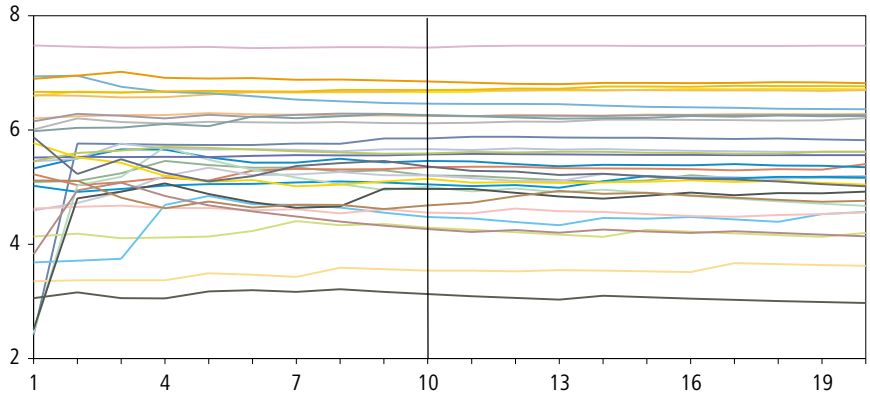
In the Monte Carlo experiments of this study, reasonably small values of K were enough for $\hat{V}^*$ (and $\hat{V}_M$) to stabilise. Figure A.I.1 shows the path of $\sqrt{\hat{V}^*}$ for GREG-log-7 for $K = 1,2,\ldots,20$ in three situations: for n=1,000, SRSWOR design and minor domains, for n=40, πPS design, size variable z_1 and minor domain, and for n=40, πPS design, size variable z_2 and minor domain. In the simulations of Chapter 5 and 7, we used the value $K = 10$. This cut-off line is also given in the figure. In the simulations of Chapter 6, $K = 30$ was used.

From Figure A.I.1, we can see that the Augmented standard error estimator $\sqrt{\hat{V}^*}$ indeed stabilises fast. For the SRSWOR case, K as small as 1 would have given in practice the same results as $K = 10$, which was used. For the πPS case, $\sqrt{\hat{V}^*}$ stabilises slightly slower. However, for $K = 5$, $\sqrt{\hat{V}^*}$ is already quite stable and after $K = 10$, $\sqrt{\hat{V}^*}$ changes only little. Thus the choice $K = 1$ is well justified.

Design SRSWOR, n=1,000, estimator GREG-log-7, minor domains, 30 samples.



Design πPS, size variable z_1, n=40, estimator GREG-log-7, minor domain, 30 samples.



Design πPS, size variable z_2, n=40, estimator GREG-log-7, minor domain, 30 samples.

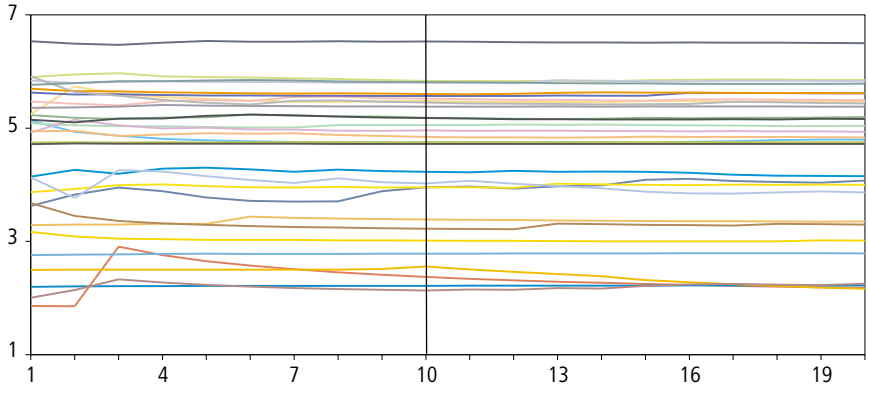


Figure A.I.1. Path of $\sqrt{\hat{V}^*}$ for $K = 1, 2, \ldots, 20$ in three different situations.

Appendix II.

The source of bias in L-GREG estimators

In the Monte Carlo experiments, some bias was observed for GREG estimators with strong models. Table A.II.1 below shows the situations where the mean absolute relative bias was more than 2%.

Table A.II.1 GREG estimators that had MARB larger than 2%.

Design	n	Estimator		Dom. type	Exp. sample size	Accuracy of GREG estimators		
						MARB	MRRMSE	ASE
		Set	Link			%	%	
SRSWOR	**1000**	8	log	Minor	12	4.5	20.4	25.1
πPS, z1	**40**	7	log	Minor	10	2.4	13.9	9.4
		8	log	Minor	10	3.8	14.8	9.9
		9	log	Minor	10	3.0	14.2	9.6
		10	log	Minor	10	4.9	15.9	10.4
πPS, z2	**40**	7	log	Minor	10	2.9	13.9	9.4
		8	log	Minor	10	4.2	14.3	9.5
		9	log	Minor	10	3.6	14.3	9.6
		10	log	Minor	10	4.8	15.7	10.3

The situations where bias occurred have several common characteristics. The bias was present i) only when the expected sample size in domains was small (12 for the SRSWOR design, 10 for the πPS design), ii) only when the assisting model was very strong, and iii) only for GREG-log estimators (GREG-lin estimators did not have any significant bias). We will show that the source of bias is inaccurate and biased estimation of the parameters of the assisting model (it is well known that for small samples, the maximum likelihood estimators for the logistic model parameters are biased; see, for example, McCullagh and Nelder 1989, Cordeiro and McCullagh 1991 and Firth 1993).

The explanation for the bias is as follows.

1. When the assisting model is very strong, sample fit residuals are, by definition, close to zero for GREG-log estimators. However, for GREG-lin estimators, sample fit residuals may be large even with a strong model.
2. When sample fit residuals are small, GREG estimator's bias correction term

$$\underline{\hat{T}}_{j,GREG}^{(d)} = \sum_{i\in U} \underline{\hat{y}}_{ij}^{(d)} + \underbrace{\sum_{i\in U} \underline{w}_i \underline{e}_{ij}^{(d)}}_{\text{Bias-correction}} \quad , \quad \underline{e}_{ij}^{(d)} = y_{ij}^{(d)} - \underline{\hat{y}}_{ij}^{(d)} \qquad \text{(A.1)}$$

is also small. The bias correction term is especially small if the sample size in the domain is small. This follows from the fact that only the units in the sample set have positive weights w_i in the bias correction term of (A.1). Outside the sample set, the weights are zero. For instance, if the number of observations in the domain is ten, the overall number of parameters in the model is 6, two of the parameters are domain-specific and the model is strong (as in the πPS design, estimator set 10), most of the ten sample fit residuals should be close to zero. Figure A.II.1 shows the histograms of the bias-correction term for some GREG-log estimators that suffered from bias. The figures show that the bias correction term indeed is negligible in these cases.

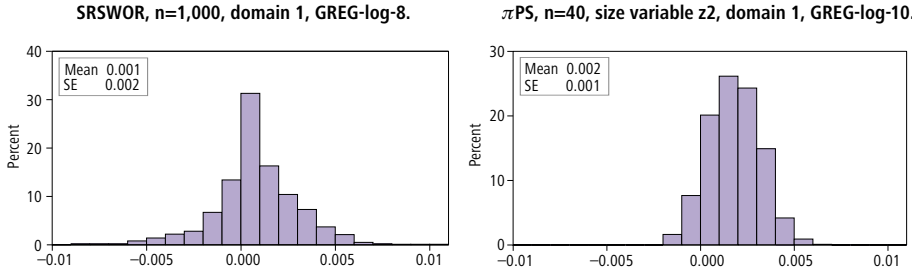


Figure A.II.1 Histograms of the bias-correction term of three biased GREG estimators.

3. When the bias correction term is close to zero, the GREG estimator is close to the pseudo-synthetic estimator:

$$\hat{\underline{T}}^{(d)}_{j,GREG} = \sum_{i \in U} \hat{\underline{y}}^{(d)}_{ij} + \sum_{i \in U} \underline{w}_i \underline{e}^{(d)}_{ij} \qquad \text{(A.2)}$$

$$\approx \sum_{i \in U} \hat{\underline{y}}^{(d)}_{ij} \ .$$

It is well known that the traditional synthetic estimator, which has the same form as (A.2) but does not utilise weights for the estimation of model parameters, may suffer from bias. The bias may be large especially if the model is

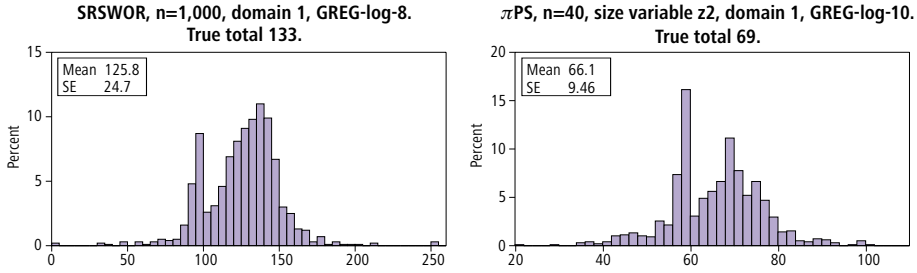


Figure A.II.2 Histograms of two biased GREG estimators.

misspecified (e.g. Lehtonen *et al.* 2005). Here we observed large biases in situations where the model was exactly correct or close to being correct (with respect to the population generating model). Figure A.II.2 shows histograms of some GREG-log estimators that suffered from bias. The histograms resemble a mix of the typical distributions of synthetic and GREG estimators and the synthetic part indeed dominates the distribution.

4. If the model is correct, or close to being correct and the model parameters are close to being correct (with respect to census fit parameters), the bias of the GREG estimator is small. Table A.II.2 shows mean absolute relative biases for GREG estimators

$$\hat{\underline{T}}_{j,GREG}^{(d)} = \sum_{i\in U} \tilde{y}_{ij}^{(d)} + \sum_{i\in U} \underline{w}_i \tilde{e}_{ij}^{(d)} \ , \ \underline{e}_{ij}^{(d)} = y_{ij}^{(d)} - \tilde{y}_{ij}^{(d)} \qquad \text{(A.3)}$$

that utilise census fit predictions instead of the sample fit predictions.

Table A.II.2 Biased GREG estimators with census fit models.

Design	n	Estimator		Dom. type	Exp. sample size	Accuracy of GREG estimators		
						MARB	MRRMSE	ASE
		Set	Link			%	%	
SRSWOR	**1000**	8	log	Minor	12	0.4	19.5	24.6
πPS, z1	**40**	7	log	Minor	10	0.3	8.9	6.1
		8	log	Minor	10	0.3	8.8	6.0
		9	log	Minor	10	0.3	8.7	6.0
		10	log	Minor	10	0.3	8.3	5.8
πPS, z2	**40**	7	log	Minor	10	0.2	9.2	6.4
		8	log	Minor	10	0.3	9.1	6.3
		9	log	Minor	10	0.2	8.9	6.1
		10	log	Minor	10	0.3	9.6	6.3

Table A.II.2 shows that the estimators that were biased with sample fit models become unbiased when the model is the census fit model. Thus the bias is due to errors in the estimation of model parameters.

In the estimator (A.3), only the bias-correction term is random – the synthetic part $\sum_U \tilde{y}_{ij}^{(d)}$ is constant with respect to sampling. When the model is good and model parameters close to census fit parameters, the synthetic part is close to the population total. Figure A.II.3 illustrates this. However, when the model is good and model parameters are estimated from the sample, the synthetic part is not necessarily close to the population total. Figure A.II.4 illustrates this.

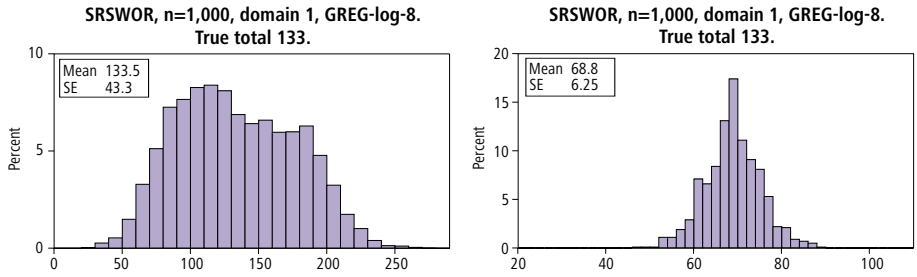


Figure A.II.3 Histograms of two biased GREG estimators with census fit models.

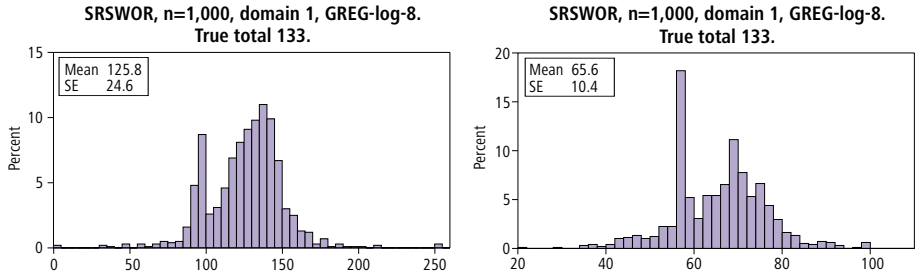


Figure A.II.4 Histograms of the synthetic part of two biased GREG estimators.

5. The sample fit synthetic part $\sum_U \tilde{y}_{ij}^{(d)}$ is not always concentrated around the true population total because PML estimators of the logistic assisting model may be biased when the domain sample size is small. This makes the estimator (A.3) biased. Tables A.II.3 and A.II.4 show that the sample fit maximum likelihood estimators for the logistic assisting models may indeed be severely biased.

Table A.II.3 Mean of census fit parameters and sample fit parameters and absolute mean relative RRMSE for the sample fit estimators.

SRSWOR, n=1000		Mean of census fit parameters	Mean of Estimates
Set	Parameter		
3	b0	–1.1	–1.1
	b1, minor domains	2.5	3.0
	b1, medium domains	2.1	2.2
	b1, major domains	2.6	2.7
8	b0, minor domains	0.2	–36.3
	b0, medium domains	0.3	0.1
	b0, major domains	–0.1	0.2
	b1, minor domains	11.9	95.8
	b1, medium domains	9.6	22.0
	b1, major domains	11.8	14.8
	b2	–10.2	–13.6

Table A.II.3 shows the census fit and sample fit model parameters for the GREG-log-8 estimator that was biased under the SRSWOR design. As a reference, there are also corresponding parameters for an unbiased estimator GREG-log-3. The sample size is n=1,000. We do not consider larger sample sizes, because it is obvious that the sample fit parameters tend to census fit parameters as the sample size grows.

From Table A.II.3 we see that for the unbiased estimator GREG-log-3, sample fit estimators for the logistic assisting model are biased. However, the bias is not huge, as it is for many parameters of the biased estimator GREG-log-8. Especially in minor domains, where the GREG estimator itself was biased, model parameter estimators for the intercept β_0 and slope β_1 for x_1 are severely biased.

Table A.II.4 shows corresponding statistics for selected πPS cases, and as a reference, one unbiased estimator (GREG-log-3) is also included. The table shows that the sample fit estimators are, on average, reasonably close to the census fit parameters for the unbiased estimator GREG-log-3. But for the bi-

Table A.II.4 Mean of census fit parameters and sample fit parameters and absolute mean relative RRMSE for the sample fit estimators. πPS, n=40

Set	Parameter	Census fit parameter	Mean of Estimates	
			Size var z1	Size var z2
3	b0	−0.4	−0.6	−0.5
	b1, minor domain	2.5	2.8	2.7
	b1, major domain	2.3	2.7	2.6
7	b0	2.9	22.1	19.8
	b1, minor domain	9.6	15.1	12.4
	b1, major domain	9.9	9.5	11.8
	b2	−10.6	−35.4	−28.1
8	b0, minor domain	1.8	−15.0	−9.3
	b0, major domain	2.5	29.1	20.1
	b1, minor domain	8.4	35.9	37.9
	b1, major domain	10.3	16.2	13.8
	b2	−10.8	−42.7	−30.3
9	b0	1.2	21.4	24.2
	b1, minor domain	9.6	15.8	15.2
	b1, major domain	9.9	9.5	9.7
	b2	−10.5	−35.7	−34.7
	b3	0.0	0.0	0.0
10	b0, minor domain	0.7	28.8	32.5
	b0, major domain	1.9	−15.0	−13.7
	b1, minor domain	8.2	36.7	33.5
	b1, major domain	10.3	16.3	15.4
	b2	−10.7	−42.5	−41.6
	b3	0.0	0.0	0.0

Table A.II.4 Mean of census fit parameters and sample fit parameters and absolute mean relative RRMSE for the sample fit estimators. πPS, n=40

Set	Parameter	Census fit parameter	Mean of Estimates	
			Size var z1	Size var z2
3	b0	−0.4	−0.6	−0.5
	b1, minor domain	2.5	2.8	2.7
	b1, major domain	2.3	2.7	2.6
7	b0	2.9	22.1	19.8
	b1, minor domain	9.6	15.1	12.4
	b1, major domain	9.9	9.5	11.8
	b2	−10.6	−35.4	−28.1
8	b0, minor domain	1.8	−15.0	−9.3
	b0, major domain	2.5	29.1	20.1
	b1, minor domain	8.4	35.9	37.9
	b1, major domain	10.3	16.2	13.8
	b2	−10.8	−42.7	−30.3
9	b0	1.2	21.4	24.2
	b1, minor domain	9.6	15.8	15.2
	b1, major domain	9.9	9.5	9.7
	b2	−10.5	−35.7	−34.7
	b3	0.0	0.0	0.0
10	b0, minor domain	0.7	28.8	32.5
	b0, major domain	1.9	−15.0	−13.7
	b1, minor domain	8.2	36.7	33.5
	b1, major domain	10.3	16.3	15.4
	b2	−10.7	−42.5	−41.6
	b3	0.0	0.0	0.0

ased estimators, sample fit maximum likelihood estimators for the model parameters are severely biased. The bias is large in major domains, but even larger in minor domains. The only parameter estimates that can be said to be close to census fit parameters are the slope estimates for β_1 in major domains. All other parameter estimators have large biases.

The conclusion is that for domains with a very small sample size, GREG-log may be biased. The bias may occur especially if the estimators for the model parameters are biased. In this study we used maximum likelihood estimators, which performed well when the domain sample size was not very small. For domains with a very small sample size and complex logistic-type models, the model parameter estimators were biased, introducing bias to the GREG estimators. Whether other existing methods would perform better for model parameters and thus decrease bias, might be worth additional study.

Appendix III.
Additional tables from the Monte Carlo stury II

Table A.III.1 Accuracy of the Standard and Augmented variance estimators for GREG-lin. SRSWOR design, overall sample size 5,000.

n=5000 Estimator	Domain type	MRAE	Mean relative estimation error		Mean relative root MSE		Coverage rates	
			S	A	S	A	S	A
1	Population	−1.8	0.0	0.0	1.8	1.8	94.1	94.1
	Major	−0.5	−0.1	−0.1	3.7	3.7	94.7	94.7
	Med	0.3	−0.1	−0.1	5.0	5.0	95.1	95.1
	Minor	−2.4	−0.1	−0.1	7.4	7.4	94.1	94.1
2	Population	−1.0	−0.2	0.1	1.3	1.0	94.4	94.4
	Major	−0.6	−0.1	0.1	3.7	3.7	94.5	94.6
	Med	0.3	−0.4	0.0	5.4	5.5	94.6	94.7
	Minor	−3.5	−0.7	0.1	8.3	8.0	92.8	92.9
3	Population	−3.0	−0.2	0.1	3.2	2.9	93.4	93.7
	Major	−0.9	−0.2	0.0	3.9	3.9	94.4	94.5
	Med	−0.1	−0.4	0.0	5.4	5.4	94.7	94.8
	Minor	−3.5	−0.6	0.1	8.3	8.0	93.0	93.2
4	Population	−2.4	−0.3	0.4	2.7	2.0	93.6	93.6
	Major	−1.1	−0.1	0.4	4.1	4.0	94.5	94.6
	Med	−0.1	−0.7	0.1	5.5	5.5	94.5	94.6
	Minor	−4.3	−1.2	0.2	9.7	9.0	92.2	92.7
5	Population	−1.2	0.0	0.0	1.7	1.7	94.8	94.8
	Major	1.1	−0.1	−0.1	6.2	6.2	94.9	94.9
	Med	−0.9	−0.4	−0.4	10.2	10.2	94.3	94.3
	Minor	−2.7	−0.7	−0.7	12.2	12.1	93.4	93.4
6	Population	−0.9	−0.2	0.2	1.6	1.4	94.9	94.9
	Major	0.8	−0.2	0.0	6.2	6.2	94.7	94.7
	Med	−0.9	−0.8	−0.4	11.5	11.4	93.7	93.8
	Minor	−3.2	−1.3	−0.5	12.5	12.3	92.3	92.5
7	Population	1.1	−0.3	0.2	1.3	1.1	95.5	95.6
	Major	0.0	−0.3	0.0	5.8	5.8	94.8	94.8
	Med	−0.1	−0.8	−0.3	10.2	10.2	94.3	94.4
	Minor	−2.8	−1.4	−0.2	11.2	10.8	92.6	92.9
8	Population	1.1	−0.4	0.4	1.3	1.2	95.5	95.7
	Major	−0.2	−0.2	0.4	5.8	5.8	94.7	94.8
	Med	−0.6	−1.1	−0.3	11.2	11.2	93.8	94.0
	Minor	−3.1	−1.8	0.0	11.5	11.0	92.2	92.8

Table A.III.2 Accuracy of the Standard and Augmented variance estimators for GREG-log. Model sets 1–4, SRSWOR design, total sample size 5,000, 2,000 and 1,000.

Samle size	Estimator set	Domain type	Exp. sample size	MRAE	MREE		MRRMSE		MCR	
					S	A	S	A	S	A
n=5000	**1**	Population	5000	−1.8	0.0	0.0	1.8	1.8	94.1	94.1
		Major	303	−0.5	−0.1	−0.1	3.7	3.7	94.7	94.7
		Med	121	0.3	−0.1	−0.1	5.0	5.0	95.1	95.1
		Minor	61	−2.4	−0.1	−0.1	7.4	7.4	94.1	94.1
	2	Population	5000	−1.2	−0.2	0.1	1.4	1.1	94.4	94.4
		Major	303	−0.6	−0.1	0.1	3.8	3.7	94.4	94.5
		Med	121	0.3	−0.4	0.0	5.5	5.5	94.6	94.7
		Minor	61	−3.6	−0.7	0.1	8.4	8.1	92.7	93.0
	4	Population	5000	−2.8	−0.2	0.1	3.0	2.7	93.7	93.7
		Major	303	−0.8	−0.2	0.0	4.0	4.0	94.3	94.4
		Med	121	0.0	−0.4	0.0	5.4	5.5	94.7	94.8
		Minor	61	−3.6	−0.6	0.1	8.3	8.0	92.9	93.1
	4	Population	5000	−2.6	−0.3	0.4	2.9	2.2	93.7	94.0
		Major	303	−1.0	−0.1	0.4	4.0	3.9	94.4	94.6
		Med	121	0.2	−0.7	0.1	5.7	5.7	94.5	94.6
		Minor	61	−4.2	−1.3	0.3	9.9	9.2	92.2	92.6
					S	**A**	**S**	**A**	**S**	**A**
n=2000	**1**	Population	2000	−1.6	−0.1	0.0	2.0	2.0	94.7	94.7
		Major	121	0.6	−0.1	−0.1	6.4	6.4	94.8	94.8
		Med	48	0.6	−0.5	−0.4	8.6	8.6	95.4	95.4
		Minor	24	1.4	−0.7	−0.7	11.4	11.4	94.8	94.8
	2	Population	2000	−2.4	−0.6	0.2	6.8	6.0	93.9	94.3
		Major	121	0.2	−0.4	0.1	6.5	6.5	94.3	94.4
		Med	48	0.1	−1.3	−0.2	9.1	9.0	93.7	94.0
		Minor	24	−0.8	−2.6	−0.5	13.0	12.6	92.0	92.5
	3	Population	2000	−3.2	−0.6	0.1	7.3	6.6	93.8	94.2
		Major	121	0.2	−0.4	0.0	6.9	6.9	94.5	94.6
		Med	48	−0.4	−1.3	−0.3	8.9	8.8	93.8	94.1
		Minor	24	−1.7	−2.3	−0.4	12.9	12.4	92.2	92.8
	4	Population	2000	−3.1	−1.0	0.8	8.2	6.4	93.3	93.6
		Major	121	−0.6	−0.4	0.8	6.8	6.5	94.2	94.5
		Med	48	−1.2	−2.3	−0.1	9.8	9.1	93.0	93.5
		Minor	24	−3.3	−5.1	−0.3	17.5	15.0	90.2	92.1
					S	**A**	**S**	**A**	**S**	**A**
n=1000	**1**	Population	1000	−1.6	−0.1	0.0	2.0	1.9	94.7	94.7
		Major	61	0.9	−0.4	−0.3	7.5	7.6	95.2	95.2
		Med	24	0.3	−0.8	−0.7	11.5	11.5	94.6	94.6
		Minor	12	0.1	−1.6	−1.5	16.6	16.6	95.2	95.2
	2	Population	1000	−3.9	−1.3	0.4	5.3	3.8	93.9	94.3
		Major	61	−0.5	−0.9	0.1	8.3	8.1	93.9	94.1
		Med	24	−2.6	−2.8	−0.6	14.1	13.5	91.4	92.0
		Minor	12	−6.1	−5.8	−1.8	21.9	19.9	88.3	89.5
	3	Population	1000	−3.9	−1.3	0.2	5.3	3.9	93.8	93.9
		Major	61	−0.1	−1.0	−0.2	8.1	8.0	94.1	94.3
		Med	24	−2.9	−2.6	−0.5	13.8	13.2	91.7	92.2
		Minor	12	−8.2	−5.3	−1.2	22.5	20.2	88.3	89.8
	4	Pop.	1000	−6.5	−2.2	1.5	8.7	5.2	93.3	93.8
		Major	61	−1.6	−0.9	1.3	8.9	8.6	93.6	94.2
		Med	24	−4.7	−5.2	−0.1	17.7	14.9	89.1	91.1
		Minor	12	−10.2	−13.5	−1.9	36.4	26.6	80.0	86.1

Table A.III.3 Accuracy of the Standard and Augmented variance estimators for GREG-lin. πPS design, overall sample size 80, size variable z_1.

Set	Domain type		Relative estimation error		Relative root MSE		Coverage rates	
		MRAE	S	A	S	A	S	A
1	Population	−3.0	−0.5	−0.1	6.3	6.1	94.0	94.4
	Major	0.2	−0.5	0.0	6.6	6.5	95.2	95.2
	Minor	−3.2	−0.7	−0.2	14.3	14.1	93.4	93.5
2	Population	−3.3	−1.1	−0.1	6.9	6.3	94.0	94.3
	Major	−0.2	−0.7	0.0	6.6	6.5	94.9	94.9
	Minor	−5.8	−2.3	−0.5	16.4	15.7	90.7	91.6
3	Population	0.2	−4.6	−3.8	6.9	6.4	93.7	93.8
	Major	0.0	−0.7	−0.1	6.5	6.5	95.0	95.0
	Minor	2.8	−10.9	−9.3	15.9	15.4	91.3	92.0
4	Population	−4.0	−1.5	−0.1	7.7	6.8	93.6	94.1
	Major	−0.4	−0.8	0.0	6.6	6.5	94.8	95.1
	Minor	−8.7	−3.7	−0.3	18.8	16.9	89.1	90.5
5	Population	1.3	−0.7	0.0	11.4	11.4	93.1	93.2
	Major	2.2	−0.8	−0.1	14.0	14.1	93.1	93.3
	Minor	−4.7	−0.9	−0.2	24.4	24.1	90.4	90.6
6	Population	0.9	−1.2	0.0	11.4	11.4	93.5	94.0
	Major	1.7	−1.0	−0.1	13.9	14.0	92.8	93.1
	Minor	−6.9	−2.6	−0.4	25.8	25.2	89.0	89.7
7	Population	−4.3	0.9	2.1	10.1	9.6	92.7	93.6
	Major	2.1	−1.9	−0.2	11.7	11.9	92.5	93.1
	Minor	−2.6	−7.7	−3.1	25.1	23.4	87.0	88.0
8	Population	−1.0	−2.5	0.1	10.1	9.6	92.7	93.7
	Major	1.9	−1.8	0.1	11.7	11.8	92.1	92.9
	Minor	−10.9	−5.1	−0.5	26.4	24.1	86.4	87.5
9	Population	−3.3	−2.9	0.2	11.2	9.9	92.5	93.3
	Major	0.2	−2.3	0.2	11.6	11.4	91.5	92.4
	Minor	−12.1	−5.0	−0.3	25.8	23.1	85.7	87.4
10	Population	−3.2	−3.1	0.3	11.3	9.9	92.3	93.3
	Major	0.1	−2.5	0.3	11.6	11.4	91.2	92.4
	Minor	−13.0	−5.8	−0.1	27.2	23.8	85.1	86.2

Table A.III.4 Accuracy of the Standard and Augmented variance estimators for GREG-lin. πPS design, overall sample size 40, size variable z_1.

Set	Domain type		Relative estimation error		Relative root MSE		Coverage rates	
		MRAE	S	A	S	A	S	A
1	Population	0.0	−1.2	−0.3	7.6	7.6	93.3	93.4
	Major	4.1	−1.2	−0.2	10.4	10.7	95.3	95.6
	Minor	−5.3	−1.5	−0.6	21.7	21.5	91.1	91.6
2	Population	−1.1	−2.4	−0.6	8.4	8.0	92.6	93.1
	Major	2.8	−1.6	−0.3	10.0	10.2	94.5	94.6
	Minor	−10.7	−5.7	−2.1	28.4	26.5	84.4	85.4
3	Population	2.3	−6.0	−4.1	8.2	7.8	92.4	92.9
	Major	2.9	−1.6	−0.4	9.9	10.2	94.6	94.7
	Minor	−7.0	−8.6	−5.0	25.9	24.3	86.2	87.9
4	Population	−5.0	−3.4	0.8	11.3	8.8	92.0	93.1
	Major	2.3	−2.0	−0.3	10.0	10.2	93.9	94.4
	Minor	−16.8	−14.0	−4.4	38.4	32.3	81.0	84.2
5	Population	−1.3	−2.1	−0.8	19.7	19.6	92.1	92.4
	Major	−0.6	−2.4	−1.0	23.7	23.6	89.0	89.0
	Minor	−4.1	−3.3	−1.7	42.0	41.4	77.1	78.1
6	Population	−2.3	−3.1	−0.8	19.9	19.7	91.5	92.1
	Major	−1.8	−2.8	−1.1	24.1	24.0	88.2	88.4
	Minor	−10.3	−7.7	−3.3	43.6	42.4	71.4	72.8
7	Population	−3.8	−6.0	−1.9	19.4	18.0	90.4	92.0
	Major	−0.6	−5.3	−2.1	21.9	21.4	89.1	90.4
	Minor	−12.9	−10.2	−2.9	40.9	38.5	72.1	74.9
8	Population	−5.0	−6.0	0.6	19.9	18.1	89.9	91.8
	Major	−1.3	−5.1	−1.5	22.1	21.5	88.9	90.0
	Minor	−16.8	−12.7	0.7	44.6	38.6	67.9	71.6
9	Population	−5.5	−6.4	−1.3	20.4	18.3	90.4	92.1
	Major	−2.1	−5.7	−1.5	22.1	21.2	89.0	90.1
	Minor	−14.1	−11.6	−3.1	41.0	37.4	71.8	75.3
10	Population	−6.3	−7.0	1.2	21.1	17.3	90.1	92.2
	Major	−2.3	−6.0	−1.4	22.3	21.2	88.3	89.7
	Minor	−18.1	−13.5	2.9	44.7	36.9	67.9	72.4

Table A.III.5 Accuracy of the Standard and Augmented variance estimators for GREG-log. Model sets 1–4, πPS design, sample size 80, size variables z_1 and z_2.

Size variable	Set	Domain type	Exp. sample size	RAE	REE		RRMSE		CR
					S	A	S	A	S
Z1	**1**	Population	80	−3.1	−0.5	−0.1	6.4	6.1	93.9
		Major	60	0.1	−0.5	0.0	6.6	6.6	94.7
		Minor	20	−3.1	−0.7	−0.2	14.4	14.2	93.3
	2	Population	80	−3.3	−1.1	−0.1	7.0	6.4	93.9
		Major	60	−0.5	−0.7	0.0	6.7	6.5	94.7
		Minor	20	−5.2	−2.5	−0.5	16.6	15.8	90.8
	3	Population	80	−3.5	−1.0	−0.1	7.0	6.4	93.8
		Major	60	−0.3	−0.7	0.0	6.6	6.5	94.7
		Minor	20	−5.3	−2.3	−0.4	16.3	15.5	91.0
	4	Population	80	−3.9	−1.6	0.1	7.6	6.6	93.8
		Major	60	−0.6	−0.8	0.0	6.7	6.5	94.9
		Minor	20	−8.1	−3.9	0.1	19.1	16.1	88.6
					S	**A**	**S**	**A**	**S**
Z2	**1**	Population	80	−8.2	−0.7	−0.3	10.4	10.0	92.4
		Major	60	−4.8	−0.8	−0.3	8.8	8.5	93.4
		Minor	20	0.2	−0.9	−0.3	14.1	14.1	93.4
	2	Population	80	−9.0	−1.3	−0.4	11.5	10.7	91.4
		Major	60	−5.2	−1.0	−0.3	9.1	8.7	93.3
		Minor	20	−3.1	−2.6	−0.7	15.7	15.2	90.2
	3	Population	80	−8.7	−1.2	−0.3	11.2	10.5	92.1
		Major	60	−5.2	−1.0	−0.3	9.1	8.7	93.4
		Minor	20	−2.5	−2.2	−0.4	15.1	14.6	91.8
	4	Population	80	−9.4	−1.7	−0.3	12.3	11.0	91.0
		Major	60	−5.6	−1.1	−0.4	9.6	9.0	93.1
		Minor	20	−4.5	−4.2	−0.4	17.6	15.7	88.8

Table A.III.6 Accuracy of the Standard and Augmented variance estimators for GREG-log. Model sets 1–4, πPS design, sample size 40, size variables z_1 and z_2.

Size variable	Set	Domain type	Exp. sample size	RAE	REE		RRMSE		CR
					S	A	S	A	S
Z1	**1**	Population	40	0.0	−1.2	−0.3	7.8	7.6	93.3
		Major	30	4.2	−1.2	−0.2	10.6	10.9	95.1
		Minor	10	−5.2	−1.6	−0.6	22.1	21.8	91.0
	2	Population	40	−1.0	−2.7	−0.6	8.9	8.2	92.2
		Major	30	2.7	−1.7	−0.2	10.3	10.4	93.9
		Minor	10	−11.0	−7.1	−3.3	31.7	30.0	82.9
	3	Population	40	−1.1	−2.5	−0.4	8.8	8.1	92.5
		Major	30	2.7	−1.7	−0.2	10.2	10.5	94.0
		Minor	10	−9.8	−6.9	−2.8	31.5	29.6	85.3
	4	Population	40	−2.4	−4.0	−0.4	10.5	8.8	91.6
		Major	30	2.4	−2.1	−0.2	10.4	10.4	93.5
		Minor	10	−14.9	−13.7	−3.3	42.7	33.8	76.5
					S	**A**	**S**	**A**	**S**
Z2	**1**	Population	40	0.2	−1.3	−0.4	7.5	7.5	93.4
		Major	30	0.2	−1.3	−0.4	10.3	10.3	93.3
		Minor	10	−1.6	−1.6	−0.6	20.6	20.5	92.3
	2	Population	40	−1.9	−2.8	−0.8	9.0	8.2	92.7
		Major	30	−1.9	−1.8	−0.5	10.8	10.5	92.6
		Minor	10	−7.8	−6.8	−3.2	28.5	27.1	84.6
	3	Population	40	−1.8	−2.6	−0.7	8.7	8.0	93.0
		Major	30	−1.4	−1.8	−0.5	10.7	10.4	92.6
		Minor	10	−9.0	−6.5	−2.6	29.0	27.3	86.2
	4	Population	40	−3.3	−4.0	−0.7	10.7	8.9	91.4
		Major	30	−2.2	−2.2	−0.6	11.4	10.5	92.6
		Minor	10	−12.9	−12.9	−3.1	39.6	31.3	78.3

Appendix IV.
Additional graphs from the Monte Carlo study II

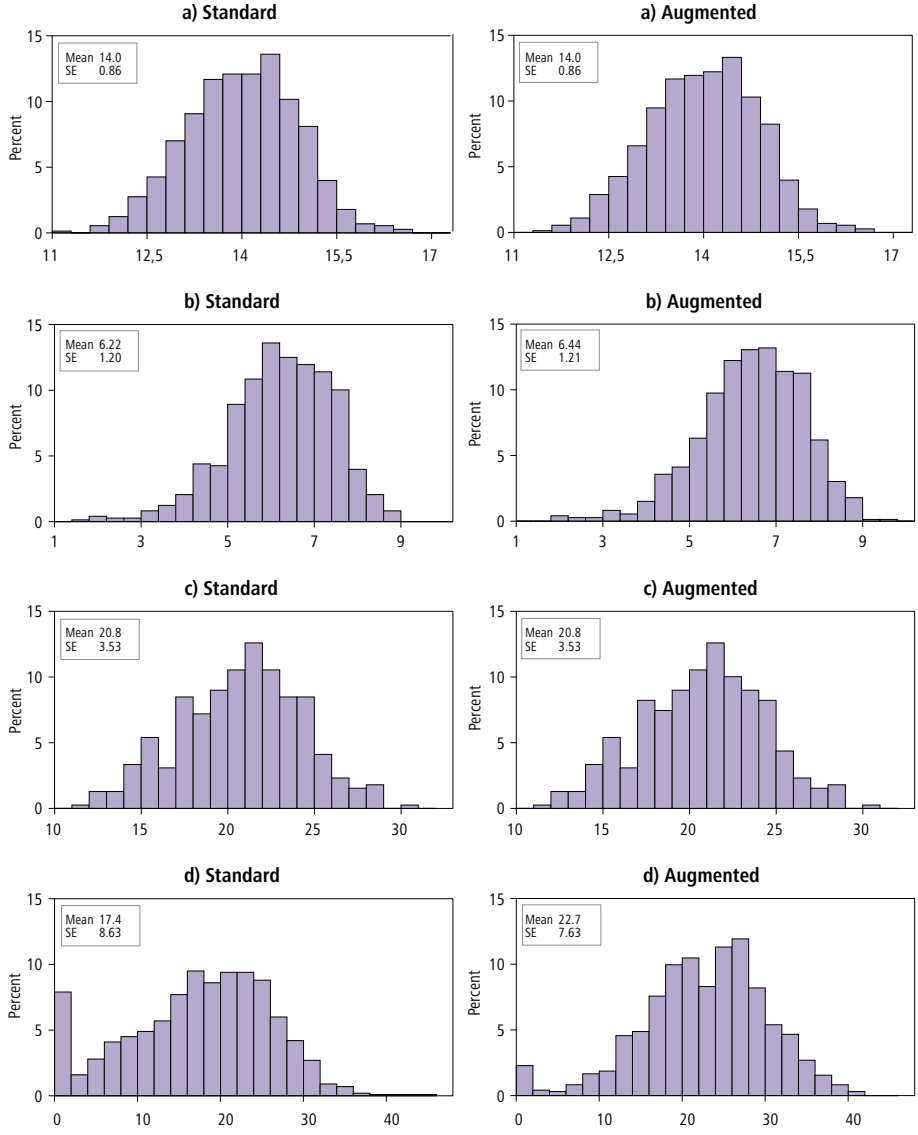


Figure A.IV.1 Histograms of Standard and Augmented variance estimators for GREG-log; SRSWOR design.

- **a)** n=5000, set=1, domain=1 (minor). True SE 15.1.
- **b)** n=5000, set=7, domain=1 (minor). True SE 6.72.
- **c)** n=2000, set=5, domain=11 (medium). True SE 21.3.
- **d)** n=1000, set=8, domain=11 (medium). True SE 23.9.

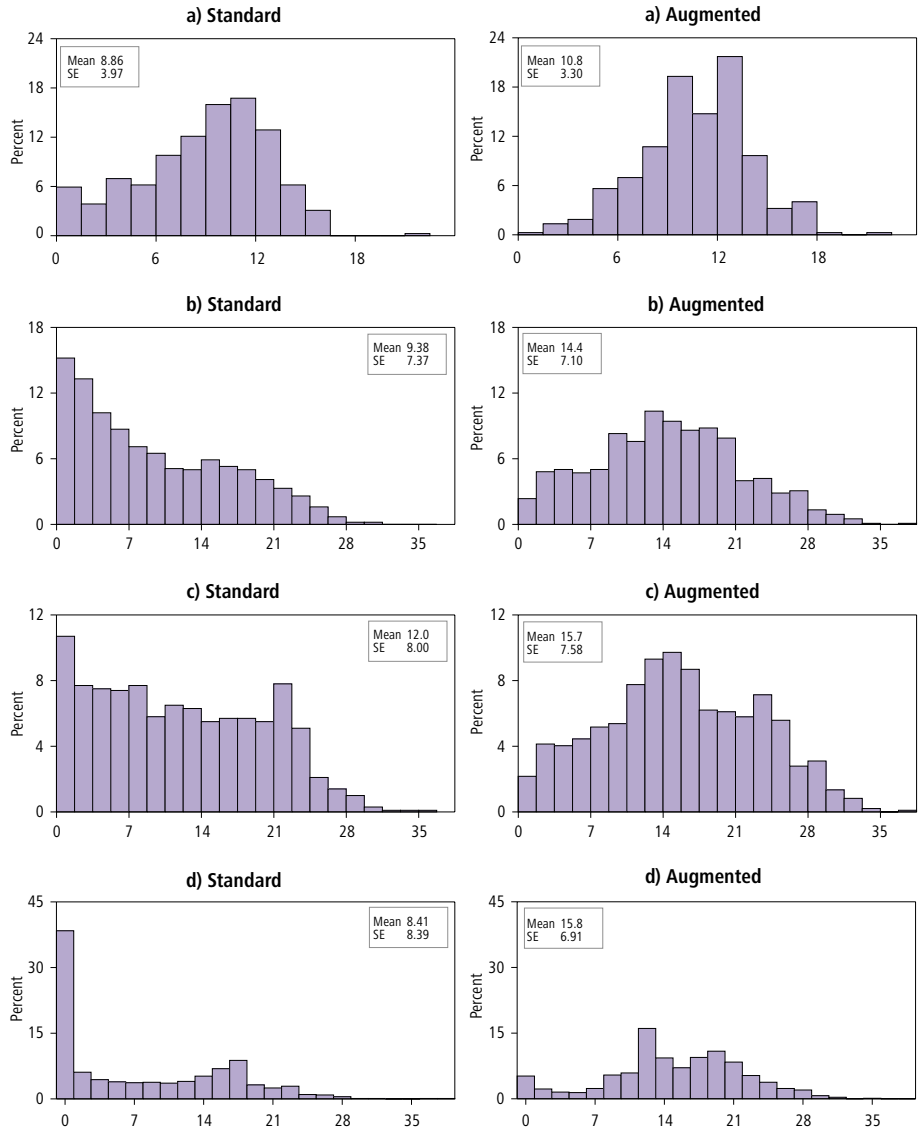


Figure A.IV.2 Histograms of Standard and Augmented variance estimators for GREG-log; SRSWOR design.

a) n=2000, set=8, domain=1 (minor). True SE 11.4.
b) n=1000, set=7, domain=3 (minor). True SE 15.9.
c) n=1000, set=7, domain=2 (minor). True SE 16.6.
d) n=1000, set=8, domain=1 (minor). True SE 16.8.

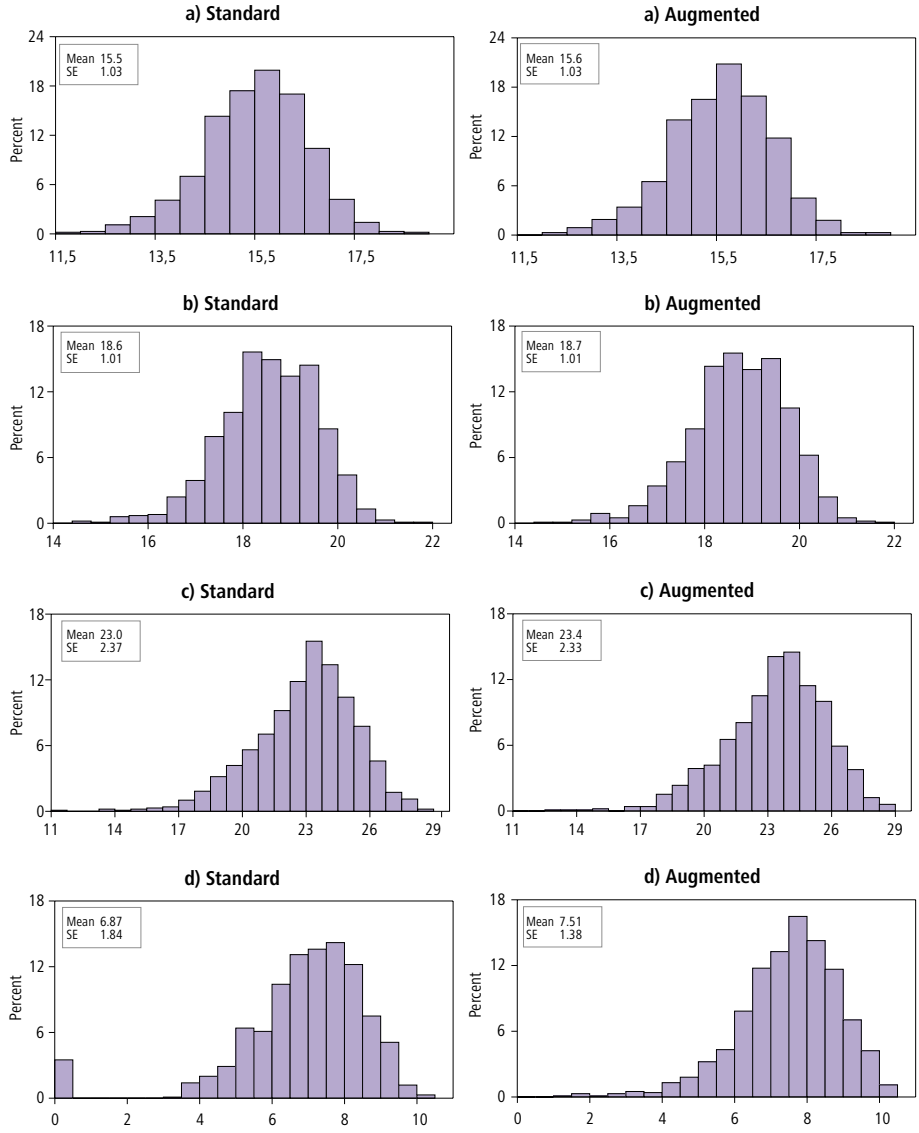


Figure A.IV.3 Histograms of Standard and Augmented variance estimators for GREG-log; πPS design.

a) n=80, size variable z1, set=1, domain=2 (major). True SE 15.6.
b) n=80, size variable z2, set=3, domain=3 (population). True SE 20.6.
a) n=40, size variable z1, set=4, domain=2 (major). Trues SE 22.9.
a) n=80, size variable z2, set=8, domain=2 (major). True SE 7.62.

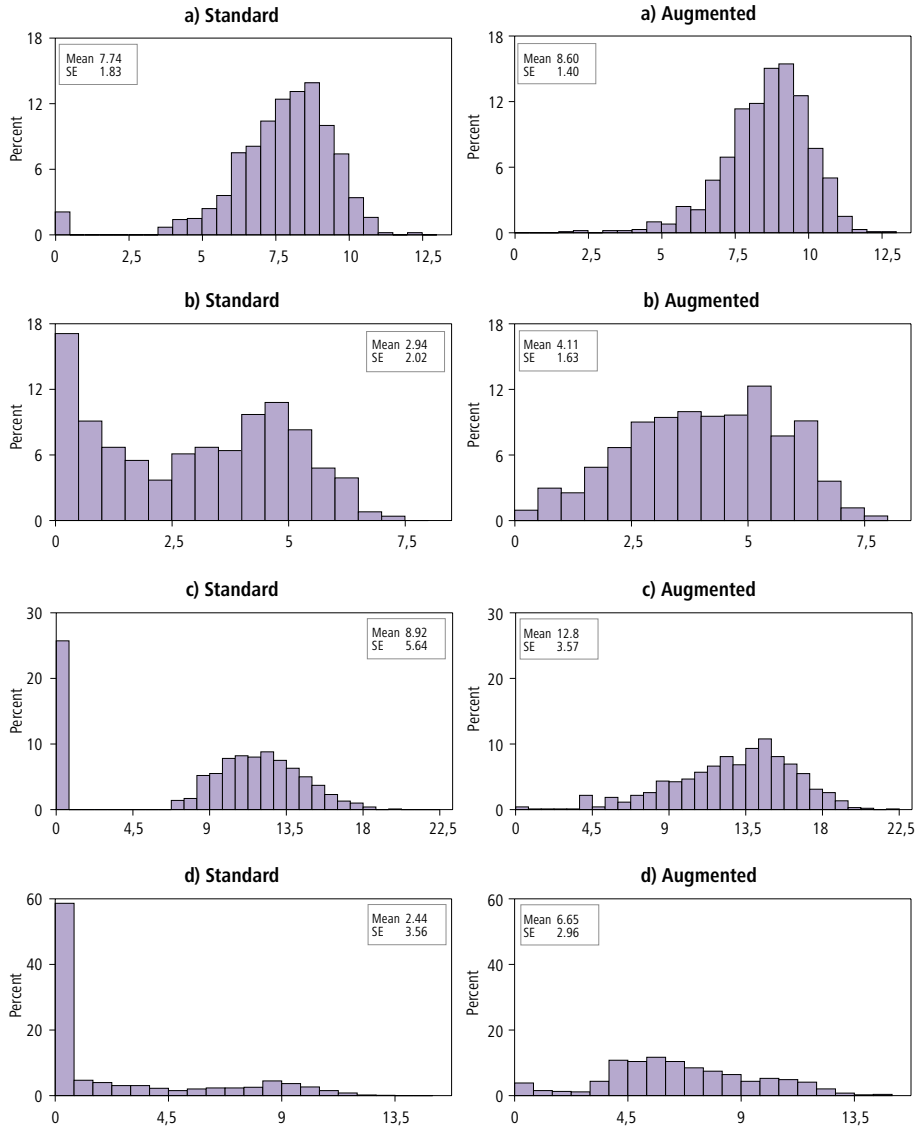


Figure A.IV.4 Histograms of Standard and Augmented variance estimators for GREG-log; πPS design.

a) n=80, size variable z1, set=7, domain=3 (population). True SE 9.04
b) n=80, size variable z2, set=7, domain=1 (minor). True SE 5.23.
a) n=40, size variable z2, set=7, domain=3 (population). True SE 15.7
a) n=40, size variable z1, set=7, domain=1 (minor). True SE 9.37.

TUTKIMUKSIA-SARJA
RESEARCH REPORTS SERIES

Tilastokeskus on julkaissut Tutkimuksia v. 1966 alkaen, v. 1990 lähtien ovat ilmestyneet seuraavat:

164. **Henry Takala,** Kunnat ja kuntainliitot kansantalouden tilinpidossa. Tammikuu 1990. 60 s.

165. **Jarmo Hyrkkö,** Palkansaajien ansiotasoindeksi 1985=100. Tammikuu 1990. 66 s.

166. **Pekka Rytkönen,** Siivouspalvelu, ympäristöhuolto ja pesulapalvelu 1980-luvulla. Tammikuu 1990. 70 s.

167. **Jukka Muukkonen,** Luonnonvaratilinpito kestävän kehityksen kuvaajana. 1990. 119 s.

168. **Juha-Pekka Ollila,** Tieliikenteen tavarankuljetus 1980-luvulla. Helmikuu 1990. 45 s.

169. **Tuovi Allén – Seppo Laaksonen – Päivi Keinänen – Seija Ilmakunnas,** Palkkaa työstä ja sukupuolesta. Huhtikuu 1990. 90 s.

170. **Ari Tyrkkö,** Asuinolotiedot väestölaskennassa ja kotitaloustiedustelussa. Huhtikuu 1990. 63 s.

171. **Hannu Isoaho – Osmo Kivinen – Risto Rinne,** Nuorten koulutus ja kotitausta. Toukokuu 1990. 115 s.

171b. **Hannu Isoaho – Osmo Kivinen – Risto Rinne,** Education and the family background of the young in Finland. 1990. 115 pp.

172. **Tapani Valkonen – Tuija Martelin – Arja Rimpelä,** Eriarvoisuus kuoleman edessä. Sosioekonomiset kuolleisuuserot Suomessa 1971–85. Kesäkuu 1990. 145 s.

173. **Jukka Muukkonen,** Sustainable development and natural resource accounting. August 1990. 96 pp.

174. **Iiris Niemi – Hannu Pääkkönen,** Time use changes in Finland in the 1980s. August 1990. 118 pp.

175. **Väinö Kannisto,** Mortality of the elderly in late 19th and early 20th century Finland. August 1990. 50 pp.

176. **Tapani Valkonen – Tuija Martelin – Arja Rimpelä,** Socio-economic mortality differences in Finland 1971–85. December 1990. 108 pp.

177. **Jaana Lähteenmaa – Lasse Siurala,** Nuoret ja muutos. Tammikuu 1991. 211 s.

178. **Tuomo Martikainen – Risto Yrjönen,** Vaalit, puolueet ja yhteiskunnan muutos. Maaliskuu 1991. 120 s.

179. **Seppo Laaksonen,** Comparative Adjustments for Missingness in Short-term Panels. April 1991. 74 pp.

180. **Ágnes Babarczy – István Harcsa – Hannu Pääkkönen,** Time use trends in Finland and in Hungary. April 1991. 72 pp.

181. **Timo Matala,** Asumisen tuki 1988. Kesäkuu 1991. 64 s.

182. **Iiris Niemi – Parsla Eglite – Algimantas Mitrikas – V.D. Patrushev – Hannu Pääkkönen,** Time Use in Finland, Latvia, Lithuania and Russia. July 1991. 80 pp.

183. **Iiris Niemi – Hannu Pääkkönen,** Vuotuinen ajankäyttö. Joulukuu 1992. 83 s.

183b. **Iiris Niemi – Hannu Pääkkönen – Veli Rajaniemi – Seppo Laaksonen – Jarmo Lauri,** Vuotuinen ajankäyttö. Ajankäyttötutkimuksen 1987–88 taulukot. Elokuu 1991. 116 s.

184. **Ari Leppälahti – Mikael Åkerblom,** Industrial Innovation in Finland. August 1991. 82 pp.

185. **Maarit Säynevirta,** Indeksiteoria ja ansiotasoindeksi. Lokakuu 1991. 95 s.

186. **Ari Tyrkkö,** Ahtaasti asuvat. Syyskuu 1991. 134 s.

187. **Tuomo Martikainen – Risto Yrjönen,** Voting, parties and social change in Finland. October 1991. 108 pp.

188. **Timo Kolu,** Työelämän laatu 1977–1990. Työn ja hyvinvoinnin koettuja muutoksia. Tammikuu 1992. 194 s.

189. **Anna-Maija Lehto,** Työelämän laatu ja tasa-arvo. Tammikuu 1992. 196 s.

190. **Tuovi Allén – Päivi Keinänen – Seppo Laaksonen – Seija Ilmakunnas,** Wage from Work and Gender. A Study on Wage Differentials in Finland in 1985. 88 pp.

191. **Kirsti Ahlqvist,** Kodinomistajaksi velalla. Maaliskuu 1992. 98 s.

192. **Matti Simpanen – Irja Blomqvist,** Aikuiskoulutukseen osallistuminen. Aikuiskoulutustutkimus 1990. Toukokuu 1992. 135 s.

193. **Leena M. Kirjavainen – Bistra Anachkova – Seppo Laaksonen – Iiris Niemi – Hannu Pääkkönen** – Zahari Staikov, Housework Time in Bulgaria and Finland. June 1992. 131 pp.

194. **Pekka Haapala – Seppo Kouvonen,** Kuntasektorin työvoimakustannukset. Kesäkuu 1992. 70 s.

195. **Pirkko Aulin-Ahmavaara,** The Productivity of a Nation. November 1992. 72 pp.

196. **Tuula Melkas,** Valtion ja markkinoiden tuolla puolen. Kanssaihmisten apu Suomessa 1980-luvun lopulla. Joulukuu 1992. 150 s.

197. **Fjalar Finnäs,** Formation of unions and families in Finnish cohorts born 1938–67. April 1993. 58 pp.

198. **Antti Siikanen – Ari Tyrkkö,** Koti – Talous – Asuntomarkkinat. Kesäkuu 1993. 167 s.

199. **Timo Matala,** Asumisen tuki ja aravavuokralaiset. Kesäkuu 1993. 84 s.

200. **Arja Kinnunen,** Kuluttajahintaindeksi 1990=100. Menetelmät ja käytäntö. Elokuu 1993. 89 s.

201. **Matti Simpanen,** Aikuiskoulutus ja työelämä. Aikuiskoulutustutkimus 1990. Syyskuu 1993. 150 s.

202. **Martti Puohiniemi,** Suomalaisten arvot ja tulevaisuus. Lokakuu 1993. 100 s.

203. **Juha Kivinen – Ari Mäkinen,** Suomen elintarvike- ja metallituoteteollisuuden rakenteen, kannattavuuden ja suhdannevaihteluiden yhteys; ekonometrinen analyysi vuosilta 1974 – 1990. Marraskuu 1993. 92 s.

204. **Juha Nurmela,** Kotitalouksien energian kokonaiskulutus 1990. Marraskuu 1993. 108 s.

205a. **Georg Luther,** Suomen tilastotoimen historia vuoteen 1970. Joulukuu 1993. 382 s.

205b. **Georg Luther,** Statistikens historia i Finland till 1970. December 1993. 380 s.

206. **Riitta Harala – Eva Hänninen-Salmelin – Kaisa Kauppinen-Toropainen – Päivi Keinänen – Tuulikki Petäjäniemi – Sinikka Vanhala,** Naiset huipulla. Huhtikuu 1994. 64 s.

207. **Wangqiu Song,** Hedoninen regressioanalyysi kuluttajahintaindeksissä. Huhtikuu 1994. 100 s.

208. **Anne Koponen,** Työolot ja ammattillinen aikuiskoulutus 1990. Toukokuu 1994. 118 s.

209. **Fjalar Finnäs,** Language Shifts and Migration. May 1994. 37 pp.

210. **Erkki Pahkinen – Veijo Ritola,** Suhdannekäänne ja taloudelliset aikasarjat. Kesäkuu 1994. 200 s.

211. **Riitta Harala – Eva Hänninen-Salmelin – Kaisa Kauppinen-Toropainen – Päivi Keinänen – Tuulikki Petäjäniemi – Sinikka Vanhala,** Women at the Top. July 1994. 66 pp.

212. **Olavi Lehtoranta,** Teollisuuden tuottavuuskehityksen mittaminen toimialatasolla. Tammikuu 1995. 73 s.

213. **Kristiina Manderbacka,** Terveydentilan mittarit. Syyskuu 1995. 121 s.

214. **Andres Vikat,** Perheellistyminen Virossa ja Suomessa. Joulukuu 1995. 52 s.

215. **Mika Maliranta,** Suomen tehdasteollisuuden tuottavuus. Helmikuu 1996. 189 s.

216. **Juha Nurmela,** Kotitaloudet ja energia vuonna 2015. Huhtikuu 1996. 285 s.

217. **Rauno Sairinen,** Suomalaiset ja ympäristöpolitiikka. Elokuu 1996. 179 s.

218. **Johanna Moisander,** Attitudes and Ecologically Responsible Consumption. August 1996. 159 pp.

219. **Seppo Laaksonen** (ed.), International Perspectives on Nonresponse. Proceedings of the Sixth International Workshop on Household Survey Nonresponse. December 1996. 240 pp.

220. **Jukka Hoffrén,** Metsien ekologisen laadun mittaaminen. Elokuu 1996. 79 s.

221. **Jarmo Rusanen – Arvo Naukkarinen** – Alfred Colpaert – Toivo Muilu, Differences in the Spatial Structure of the Population Between Finland and Sweden in 1995 – a GIS viewpoint. March 1997. 46 pp.

222. **Anna-Maija Lehto,** Työolot tutkimuskohteena. Marraskuu 1996. 289 s.

223. **Seppo Laaksonen** (ed.), The Evolution of Firms and Industries. June 1997. 505 pp.

224. **Jukka Hoffrén,** Finnish Forest Resource Accounting and Ecological Sustainability. June 1997. 132 pp.

225. **Eero Tanskanen**, Suomalaiset ja ympäristö kansainvälisestä näkökulmasta. Elokuu 1997. 153 s.

226. **Jukka Hoffrén**, Talous hyvinvoinnin ja ympäristöhaittojen tuottajana –Suomen ekotehokkuuden mittaaminen. Toukokuu 1999. 154 s.

227. **Sirpa Kolehmainen,** Naisten ja miesten työt. Työmarkkinoiden segregoituminen Suomessa 1970–1990. Lokakuu 1999. 321 s.

228. **Seppo Paananen**, Suomalaisuuden armoilla. Ulkomaalaisten työnhakijoiden luokittelu. Lokakuu 1999. 152 s.

229. **Jukka Hoffrén,** Measuring the Eco-efficiency of the Finnish Economy. October 1999. 80 pp.

230. **Anna-Maija Lehto – Noora Järnefelt** (toim.), Jaksaen ja joustaen. Artikkeleita työolotutkimuksesta. Joulukuu 2000. 264 s.

231. **Kari Djerf,** Properties of some estimators under unit nonresponse. January 2001. 76 pp.

232. **Ismo Teikari,** Poisson mixture sampling in controlling the distribution of response burden in longitudinal and cross section business surveys. March 2001. 120 pp.

233. **Jukka Hoffrén,** Measuring the Eco-efficiency of Welfare Generation in a National Economy. The Case of Finland. November 2001. 199 pp.

234. **Pia Pulkkinen,** ”Vähän enemmän arvoinen” Tutkimus tasa-arvokokemuksista työpaikoilla. Tammikuu 2002. 154 s.

235. **Noora Järnefelt – Anna-Maija Lehto,** Työhulluja vai hulluja töitä? Tutkimus kiirekokemuksista työpaikoilla. Huhtikuu 2002. 130 s.

236. **Markku Heiskanen,** Väkivalta, pelko, turvattomuus. Surveytutkimusten näkökulmia suomalaisten turvallisuuteen. Huhtikuu 2002. 323 s.

237. **Tuula Melkas,** Sosiaalisesta muodosta toiseen. Suomalaisten yksityiselämän sosiaalisuuden tarkastelua vuosilta 1986 ja 1994. Huhtikuu 2003. 195 s.

238. **Rune Höglund – Markus Jäntti – Gunnar Rosenqvist (eds.),** Statistics, econometrics and society: Essays in honour of Leif Nordberg. April 2003. 260 pp.

239. **Johanna Laiho – Tarja Nieminen (toim.)**, Terveys 2000 -tutkimus. Aikuisväestön haastatteluaineiston tilastollinen laatu. Otanta-asetelma, tiedonkeruu, vastauskato ja estimointi- ja analyysiasetelma. Maaliskuu 2004. 95 s.

240. **Pauli Ollila,** A Theoretical Overview for Variance Estimation in Sampling Theory with Some New Techniques for Complex Estimators. September 2004. 151 pp.

241. **Minna Piispa.** Väkivalta ja parisuhde. Nuorten naisten kokeman parisuhdeväkivallan määrittely surveytutkimuksessa. Syyskuu 2004. 216 s.

242. **Eugen Koev.** Combining Classification and Hedonic Quality Adjustment in Constructing a House Price Index. (Tulossa).

243. **Henna Isoniemi – Irmeli Penttilä (toim.)**, Perheiden muuttuvat elinolot. Artikkeleita lapsiperheiden elämänmuutoksista. Syyskuu 2005. 168 s.

244. **Anna-Maija Lehto – Hanna Sutela – Arto Miettinen (toim.)**, Kaikilla mausteilla. Artikkeleita työolotutkimuksesta. Toukokuu 2006. 385 s.

245. **Jukka Jalava – Jari Eloranta – Jari Ojala (toim.)** Muutoksen merkit – Kvantitatiivisia perspektiivejä Suomen taloushistoriaan. Tammikuu 2007. 373 s.

246. **Jari Kauppila.** The Structure and Short-Term Development of Finnish Industries in the 1920s and 1930s. An Input-output Approach. Elokuu 2007. 274 s.

247. **Mikko Myrskylä.** Generalised Regression Estimation for Domain Class Frequencies. Elokuu 2007. 137 s.

The Research Reports series describes Finnish society in the light of up-to-date research results. Scientific studies that are carried out at Statistics Finland or are based on the datasets of Statistics Finland are published in the series.

There is an increasing demand for accurate statistics on sub-groups or domains of a population. Many of these statistics are class frequencies and proportions, such as the numbers of people in various labor market statuses by region or disease prevalence statistics by demographic group. This report is concerned with estimation of such statistics.

The report reviews standard design-based model-assisted estimation methods and modern estimators which use non-linear models as assisting tools in estimation. These modern estimators assume availability of unit-level auxiliary information, which is now commonly available, for example, from administrative registers. However these estimators are rarely used, possibly because little is known on their properties. This report studies extensively the properties of these modern estimators, contrasts them with the standard methods, characterises the conditions under which the modern methods are more efficient, and discusses variance estimation for the modern estimators.

Tilastokeskus, myyntipalvelu
PL 4C
00022 TILASTOKESKUS
puh. (09) 1734 2011
faksi (09) 1734 2500
myynti@tilastokeskus.fi
www.tilastokeskus.fi

Statistikcentralen, försäljning
PB 4C
00022 STATISTIKCENTRALEN
tfn (09) 1734 2011
fax (09) 1734 2500
myynti@stat.fi
www.stat.fi

Statistics Finland, Sales Services
P.O.Box 4C
FI-00022 STATISTICS FINLAND
Tel. +358 9 1734 2011
Fax +358 9 1734 2500
sales@stat.fi
www.stat.fi

ISSN 0355–2071
= Tutkimuksia
ISBN 972–952–467–712–7