

## ABSTRACT

Title of Document:

TREATMENT OF INFLUENTIAL  
OBSERVATIONS IN THE CURRENT  
EMPLOYMENT STATISTICS SURVEY

Julie B. Gershunskaya, Doctor of Philosophy,  
2011

Directed By:

Professor Partha Lahiri, Joint Program in Survey  
Methodology

It is common for many establishment surveys that a sample contains a fraction of observations that may seriously affect survey estimates. Influential observations may appear in the sample due to imperfections of the survey design that cannot fully account for the dynamic and heterogeneous nature of the population of businesses. An observation may become influential due to a relatively large survey weight, extreme value, or combination of the weight and value.

We propose a Winsorized estimator with a choice of cutoff points that guarantees that the resulting mean squared error is lower than the variance of the original survey weighted estimator. This estimator is based on very un-restrictive modeling assumptions and can be safely used when the sample is sufficiently large.

We consider a different approach when the sample is small. Estimation from small samples generally relies on strict model assumptions. Robustness here is understood as insensitivity of an estimator to model misspecification or to appearance of outliers.

The proposed approach is a slight modification of the classical linear mixed model application to small area estimation. The underlying distribution of the random error term is a scale mixture of two normal distributions. This setup can describe outliers in individual observations. It is also suitable for a more general situation where units from two distinct populations are put together for estimation.

The mixture group indicator is not observed. The probabilities of observations coming from a group with a smaller or larger variance are estimated from the data. These conditional probabilities can serve as the basis for a formal test on outlyingness at the area level.

Simulations are carried out to compare several alternative estimators under different scenarios. Performance of the bootstrap method for prediction confidence intervals is investigated using simulations. We also compare the proposed method with alternative existing methods in a study using data from the Current Employment Statistics Survey conducted by the U.S. Bureau of Labor Statistics.

TREATMENT OF INFLUENTIAL OBSERVATIONS IN THE CURRENT  
EMPLOYMENT STATISTICS SURVEY

By

Julie Borisovna Gershunskaya

Dissertation submitted to the Faculty of the Graduate School of the  
University of Maryland, College Park, in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
2011

Advisory Committee:  
Professor Partha Lahiri, Chair/Advisor  
Dr. Alan Dorfman  
Dr. John Eltinge  
Professor Eric Slud  
Professor Richard Valliant

Disclaimer: Any opinions expressed in this Thesis are those of the author and do not constitute policy of the U.S. Bureau of Labor Statistics

## Dedication

*To my mother and in memory of my father*

## Acknowledgements

I would like to express gratitude to my advisor Professor Partha Lahiri. I am thankful to Partha for his wise and careful guidance of this work, for his infinite patience and ability to listen and understand ideas that were often vague and ambiguous in the beginning.

I am grateful to my Committee Members for investing the time to read over my work and for their numerous very useful comments. Their suggestions greatly improved the narrative, helped clarify many points and led to corrections in certain details of the dissertation.

I would like to use this opportunity to avow my appreciation of the JPSM faculty for creating a learning environment that is friendly and serious at the same time, the kind of a happy atmosphere that facilitates the desire to study and discover. I already miss taking classes, doing homework, and experiencing the everyday camaraderie of my fellow students.

My sincere thanks go to Rupa Jethwa Eapen and Sarah Gebremicael, the wonderful JPSM staff, for always being so joyfully helpful and friendly, to Duane Gilbert for having fast and professional responses to each of my technical problems or questions.

Finally, I would like to thank my family, my husband Vlad and my dear sons Sasha and Greg, for their love and for the chivalrous enduring of my (at times bothersome) interest in philosophy.

# Table of Contents

List of Tables .....	vii
List of Figures .....	viii
Chapter 1: Introduction and Literature Review .....	1
1.1 Introduction .....	1
1.2 A brief overview of the Current Employment Statistics survey .....	3
1.2.1 The CES sample design .....	3
1.2.2 CES estimator of relative employment growth .....	4
1.2.3 Influential observations in CES .....	4
1.3 Approaches to robust estimation in survey sampling .....	7
1.3.1 Descriptive population quantities .....	7
1.3.2 Models at different stages of survey sampling .....	8
1.3.3 The role of sampling weights in robust estimation, methods for dealing with extreme weights .....	9
1.3.4 Survey weights as random variables .....	12
1.3.5 Treatment of extreme observations in surveys .....	14
1.4 Estimation under informative sampling (Pfeffermann and Sverchkov approach) .....	18
1.4.1 Prediction of the nonsampled values based on the sample-complement distribution .....	21
1.5 The influence function approach .....	26
1.5.1 The Gâteaux derivative and the first order von Mises expansion .....	26
1.5.2 Examples of influence functions .....	28
1.5.3 Applications of the influence function approach in surveys .....	32
1.6 Robust small area estimation .....	35
Chapter 2: Robust Estimation in Moderately Large Samples .....	44
2.1 Linearization of a finite population quantity .....	45
2.2 Application: Estimation from large samples in CES .....	48
2.3 On the choice of cutoff values for the Winsorized mean .....	51
2.3.1 Some simple illustrative examples .....	54
2.3.2 Accounting for survey design when choosing the cutoff points .....	59
2.3.3 Using information not included in the sampling design .....	63
2.4 Simulation study .....	65
2.5 Mean squared error estimation using the bootstrap (simulation study) .....	72
Summary .....	73
Chapter 3: Robust Small Area Estimation .....	75
3.1 The proposed model .....	76
3.2 Identifiability of the model parameters .....	82
3.3 EM algorithm .....	85
3.3.1 Approximate computation of the first two conditional moments of the random effects .....	93
3.4 Parametric bootstrap for prediction confidence intervals .....	99
3.5 Bias correction .....	101

3.6 Simulation study .....	107
3.7 Properties of the parameter estimates when the number of areas increases .....	119
3.8 Simulations for prediction confidence intervals using the parametric bootstrap	121
3.9 Linearization of a finite population target in small area estimation, with application to the CES survey .....	123
Summary .....	138
 Chapter 4: Concluding Remarks and Future Research .....	 140
 Appendix A. The proof of Result 1 from Section 2.3 .....	 142
Appendix B. R code for the Winsorization example of Section 2.3.1 .....	145
Appendix C: EM algorithm for the scale mixture-fixed effects model WN2F from Section 2.4 .....	147
Appendix D: On the maximum likelihood estimator of $\beta$ .....	150
 References .....	 151



## List of Tables

Table 1. Normal $N(0,1)$ distribution. Bias and Standard Error, in hundreds; RMSE ratio as percentage.....	57
Table 2. Symmetrical mixture: $0.97N(0,1) + 0.03N(0,10)$ . Bias and Standard Error, in hundreds; RMSE ratio as percentage.....	57
Table 3. Asymmetrical mixture: $0.97N(0,1) + 0.03N(3,10)$ . Bias and Standard Error, in hundreds; RMSE ratio as percentage.....	58
Table 4. Lognormal distribution, $\log(y_j) \stackrel{iid}{\sim} N(0,1)$ . Bias and Standard Error, in hundreds; RMSE ratio as percentage.....	58
Table 5. PPS sampling, Bias and Standard Error, in hundreds; RMSE ratio as percentage; 5000 simulation runs .....	61
Table 6. Population $0.7N(0,1) + 0.3N(12,9)$ , 5000 simulation runs, (in hundreds)...	63
Table 7. $0.7N(0,1) + 0.3N(4,9)$ , 5000 simulation runs, (in hundreds).....	65
Table 8. Description of the simulation.....	66
Table 9. Bias and standard errors of estimators for the three finite populations. The last column is the RMSE ratio to the baseline RMSE of the HT estimator.....	71
Table 10. True MSE based on 300 samples from a finite population and estimated BWO MSE averages and standard errors (in parentheses) based on 300 estimates of MSE, each derived from 500 bootstrap iterations. ....	73
Table 11. Simulation results for scenarios 1-4 (250 runs) Median values of relative biases, expressed as a percentage.....	114
Table 12. Simulation results for scenarios (1)-(4). Median values of relative root mean squared errors, expressed as a percentage.....	115
Table 13. Simulation results for scenario 5, [70/30]. Median values of relative biases and relative root mean squared errors, expressed as a percentage.....	116
Table 14. Simulation results for scenarios 6-8. Median values of relative biases and relative root mean squared errors, expressed as a percentage. ....	118
Table 15. Mean estimates and the simulation standard errors (in parentheses) for scenario with $p = 0.03$ (a method used in the EM algorithm is indicated in parentheses next to N2).....	120
Table 16. Mean estimates and the simulation standard errors (in parentheses) for scenario with $p = 0.30$ (a method used in the EM algorithm is indicated in parentheses next to N2).....	121
Table 17. Average coverage and median length of confidence intervals (nominal coverage 95%) using the NER model and the N2 mixture model, for different population patterns.....	123
Table 18. Alabama, by Industry, Empirical Root Mean Squared Error .....	131
Table 19. Alabama, by Industry, 75th Percentile Absolute Error.....	131
Table 20. California, by Industry, Empirical Root Mean Squared Error, % .....	132
Table 21. California, by Industry, 75th Percentile Absolute Error.....	132
Table 22. Florida, by Industry, Empirical Root Mean Squared Error .....	133
Table 23. Florida, by Industry, 75th Percentile Absolute Error.....	133
Table 24. Pennsylvania, by Industry, Empirical Root Mean Squared Error.....	134
Table 25. Pennsylvania, by Industry, 75th Percentile Absolute Error.....	134

## List of Figures

Figure 1: Two examples of the weighted current month versus previous month employment plots. The red line shows the survey weighted estimate of the relative change .....	5
Figure 2. a. Histogram of weighted over-the-month changes in employment overlaid by the density of the normal distribution. b. A normal Q-Q plot of weighted over-the-month employment changes. ....	6
Figure 3. Density plots of (1) normal $N(0,1)$ ; (2) symmetrical mixture 97% $N(0,1)$ and 3% $N(0,10)$ ; (3) asymmetrical mixture 97% $N(0,1)$ and 3% $N(3,10)$ ; (4) lognormal distribution.....	56
Figure 4. Relative errors for scenarios 1-4, areas are sorted in ascending order of the sample size: (1) $[0,0]$ scenario; (2) $[0,u]$ scenario; (3) $[e,0]$ scenario; (4) $[e,u]$ scenario. ....	113
Figure 5. Relative errors for scenarios 5-8, areas are sorted in ascending order of the sample size: (5) $[70/30]$ scenario (see Table 13); (6) $[et,0]$ scenario; (7) $[0,ut]$ scenario; (8) $[et,ut]$ scenario. ....	117
Figure 6. California, Wholesale Trade (industry 41) deviations from true population values (in hundreds) of the relative employment change estimates, by areas. ....	135
Figure 7. California, Retail Trade (industry 42) deviations from true population values (in hundreds) of the relative employment change estimates, by areas. ....	136
Figure 8. Pennsylvania, Transportation and Utilities (industry 43) deviations from true population values (in hundreds) of the relative employment change estimates, by areas. ....	137

# Chapter 1: Introduction and Literature Review

## 1.1 Introduction

It is common for many establishment surveys that a sample contains a fraction of observations that may seriously affect survey estimates. An observation may become influential due to a relatively large survey weight, extreme value, or combination of the weight and value.

Establishments in the target population vary greatly by size. The population consists of a relatively small number of large companies, while most of the national employment is situated in small-size enterprises. Businesses are selected into sample with different probabilities, and the resulting survey weights are highly variable; even though an effort is taken at the design stage of a survey to minimize the variance of the survey weighted estimator, the estimates may still be very unstable.

Another aspect of a survey of businesses is the potential of change in the establishment attributes that are used for sample selection, as well as possible changes in the units' composition. For example, industrial allocation or the establishment employment level may change after a sample has been selected. As a result, it may happen that a larger (than expected at the time of sampling) employment size becomes associated with a large survey weight creating predisposition for the influential observation. (The problem of "stratum jumpers" is discussed in Rivest 1999).

We call the units that have large impact on estimation *the influential observations*; the effect caused by these observations may be due to imperfections of the survey design that cannot fully account for the dynamic and heterogeneous nature of a population of establishments. In the current research, we devise a model-based estimation procedure that takes into account the survey design and features of the probability distribution of employment in a population of businesses, leading to an estimator that is robust to model misspecifications and is more efficient than a pure survey weighted estimator.

For estimation from moderately large samples, we propose a Winsorization based estimator with a choice of the cutoff points that guarantees that the resulting mean squared error is lower than the variance of the original survey weighted estimator. This estimator is based on mild modeling assumptions; thus it can be safely used when the sample is sufficiently large.

We consider a different approach when the sample is small. Estimation from small samples generally relies on strict model assumptions. Robustness here is understood as insensitivity of an estimator to model misspecification or to appearance of outliers. The proposed approach is a slight modification of a classical linear mixed model application to small area estimation. The underlying distribution of the random error term is a scale mixture of two normal distributions. This setup can describe outliers in individual observations. It is also suitable for a more general situation where units from two distinct populations are put together for estimation.

The techniques are evaluated using simulation studies. The bootstrap is used to measure uncertainty of the estimator. A study involving real data from the CES sample is also presented.

## 1.2 A brief overview of the Current Employment Statistics survey

To facilitate the discussion, we describe briefly relevant details of the CES sample selection and estimation methods. While referring to CES throughout the discussion, we strive to produce a general method for robust estimation that can be adapted to other surveys.

### *1.2.1 The CES sample design*

The CES sample is selected once a year from a frame based on the Quarterly Census of Employment and Wages (QCEW) data file. This is an administrative dataset containing records of employment and wages for nearly every U.S. establishment covered by the States' unemployment insurance (UI) laws. (The QCEW file becomes available to BLS on a lagged basis and is important for the CES survey in many respects, see *BLS Handbook of Methods*, 2004, for more information about QCEW).

Strata on the frame are defined by State, the industrial supersector based on the North American Industrial Classification System (NAICS) and on the total employment size of establishments within a UI account. A stratified simple random sample of UI accounts is selected using optimal allocation to minimize, for a given cost per State, a State level variance of the monthly employment change estimate.

### 1.2.2 CES estimator of relative employment growth

Relative growth of employment from a previous to current month is estimated using a set of the establishments reporting positive employment in both adjacent months, the so called matched sample of establishments, denoted by  $S_t$ . The weighted link relative (WLR) estimator is

$$\hat{R}_t = \frac{\sum_{j \in S_t} w_j y_{j:t}}{\sum_{j \in S_t} w_j y_{j:t-1}}, \quad (1.2.1)$$

where  $j$  denotes an establishment,  $t$  is a current month.

The numerator of the ratio is the survey weighted sum of the current month reported employment; the denominator is the survey weighted sum of the previous month employment. See the *BLS Handbook of Methods* (2004, Chapter 2) for further details on the CES estimation procedures.

### 1.2.3 Influential observations in CES

A definition for an influential observation must be tied to the form of an estimator. In a given month, CES estimates relative employment growth, the ratio of the two survey weighted sums, as shown in (1.2.1). For this type of an estimator, a report having a relatively large survey weight or a large change in the size of its employment may become influential. Combination of a moderately large weight with a moderately large employment change may also produce an influential report.

In Figure 1, we display examples of the weighted employment at month  $t$  plotted against the weighted employment at the previous month,  $t-1$ . Generally, in any given month, there is a handful of observations that stand apart from the rest of the sample.

One reason is the form of the distribution of the employment change: there is a large number of establishments that do not change employment; many establishments have very little change in their employment. However, there are always units having a substantial change in employment and at times they also have a large survey weight. The histogram of the establishments employment change has a spike around zero and long tails (see a typical histogram and a normal Q-Q plot in Figure 2). A sample is prone to outliers in the sense that there is a high probability that a handful of observations from the tails of the distribution are present in the sample.

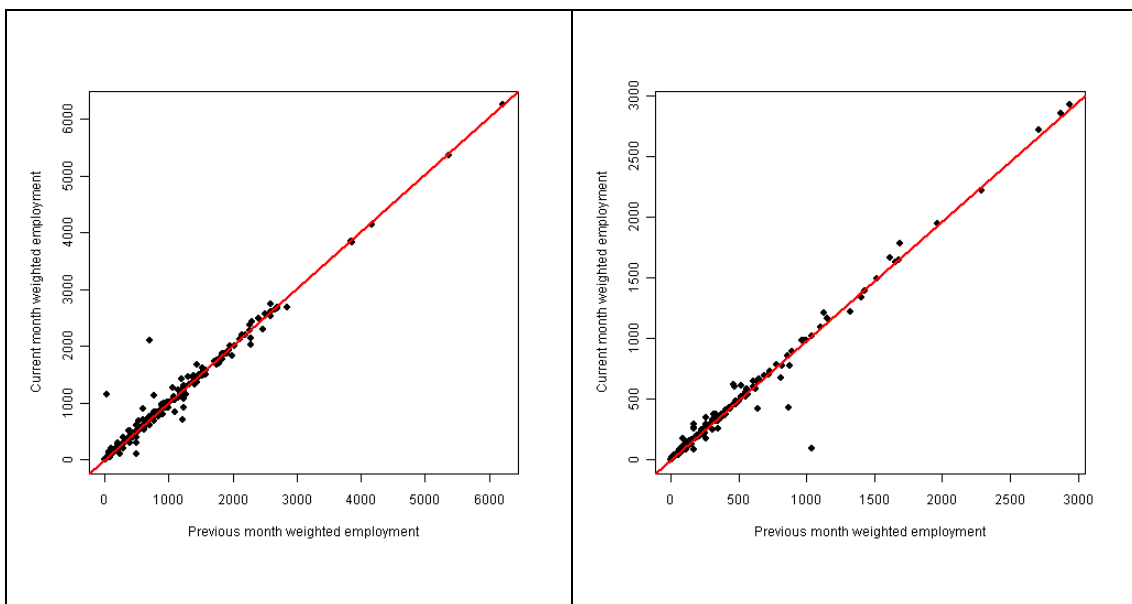


Figure 1: Two examples of the weighted current month versus previous month employment plots. The red line shows the survey weighted estimate of the relative change

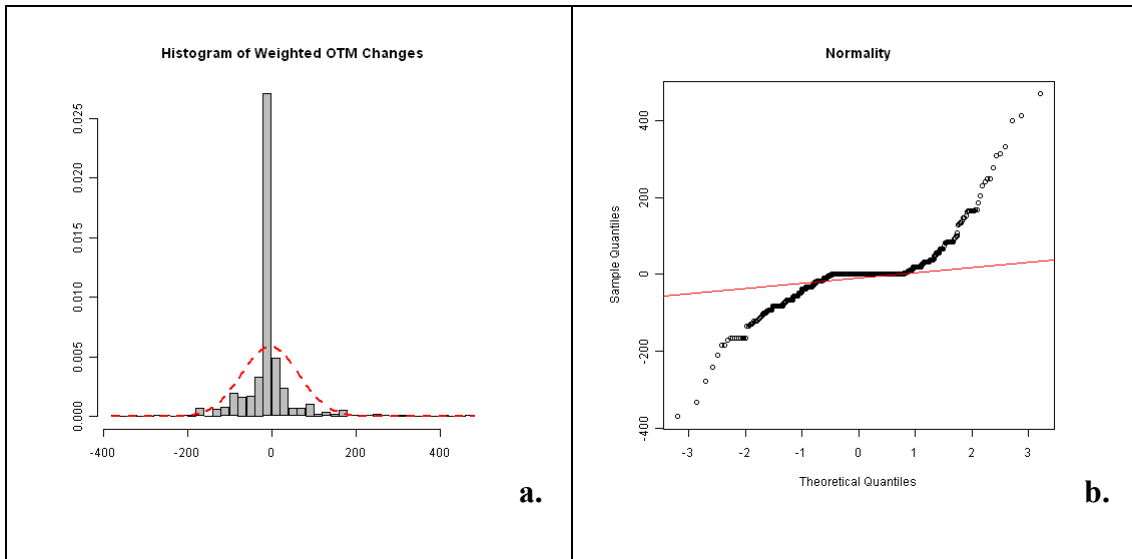


Figure 2. a. Histogram of weighted over-the-month changes in employment overlaid by the density of the normal distribution. b. A normal Q-Q plot of weighted over-the-month employment changes.

Another source of outliers, as mentioned in the introduction, is the dynamic nature of a population of establishments, which often causes “misclassification” of establishments; e.g., changes in the industrial classification or an employment size class after the sample has been selected. These changes may cause problems in estimation, especially in smaller samples.

Estimation of the National and State level employment is of central importance in CES. However, there is also a lot of interest in publication of estimates for many smaller domains defined at a finer industrial and geographical detail. At these levels, the sample is often scarce and a single influential observation, if left untreated, may ruin the resulting estimates.



## 1.3 Approaches to robust estimation in survey sampling

“Robustness is usually understood to mean that inferences made from a sample are insensitive to violations of the assumptions that have been made.” (Hansen, Madow, and Tepping, 1983). Before reviewing the methods of robust estimation, we survey the literature to gain understanding about what kind of assumptions are made in survey sampling, in particular, when a descriptive population quantity is of interest.

### *1.3.1 Descriptive population quantities*

In a large-scale government survey, we are usually interested in estimating certain descriptive statistics of the finite population, such as smooth functions of population means or totals; for example, the relative change in employment can be viewed as the ratio of two means; various forms of price indexes provide a somewhat more complex set of examples. A *descriptive statistic* can be defined as a “known function of the finite population values.” (Pfeffermann 1993). Motivation for the form of such a function does not necessarily come from a stochastic model. For example, although the stochastic approach to defining index numbers has a long history (Clements *et al.* 2006), definitions of price indexes in common use are often motivated using *deterministic* approaches coming from economic or axiomatic theories (see Diewert 1981; Balk 1995).

When analytic inference is required, Pfeffermann (1993) invokes the notion of “corresponding descriptive population quantity (CDPQ)” defined as a solution to a set of population estimating equations for an unknown parameter. Thus, descriptive

population quantities also play an important role in analytic inference about model parameters.

In the present research, we are concerned only with the former situation, where a target is given in a *pre-specified* form and *without* reference to a particular model.

**Remark.** Models are usually formulated for the finite population: i.e., the finite population measurements are assumed to be realizations from some ideal distribution, a “superpopulation”. Such models are formulated *a priori*, in the sense that the finite population is not observed directly and, therefore, modeling assumptions cannot be checked using finite population observations; thus, the resulting CDPQ also can be viewed as a *pre-specified* target that needs to be estimated from the observed sample.

### *1.3.2 Models at different stages of survey sampling*

There is a long-standing discussion on approaches to inference from survey sampling. Inferences can be made with respect to an assumed model (the model-based or prediction approach) or with respect to the randomization distribution induced by the hypothetical repeated sampling from a finite population (the randomization or design-based approach).

Adherents of either approach agree that models are important in designing an optimal sampling procedure and in deriving an efficient estimator (see Hansen, Madow, and Tepping, 1983, and the discussion; Särndal, Swensson, and Wretman, 1992; Valliant, Dorfman, and Royall, 2000). Thus, even when a stochastic model is not required for the definition of a target, a *working stochastic model* is often formulated or at least used implicitly.

The word “working”, in the reference to a model, suggests that the model is expected to be only approximately correct: it is not indeed possible to know what the “true” model is when dealing with real-life data. Model assumptions may not hold and criteria are needed to assure *robustness* of inferences *to model misspecifications*. The notion of *design consistency* provides such criteria. Design consistency is an asymptotic property, which becomes important when the sample is sufficiently large. It assures that the estimator of a finite population parameter indeed targets the parameter rather than something else. An estimator is design consistent if, as the sample and population sizes become infinitely large (according to a certain well-defined rule, see Isaki and Fuller 1982), the estimator approaches the target population quantity in probability under the randomization distribution. A good model estimator would be one from the class of design consistent estimators (see Hansen, Madow, and Tepping 1983; Little 1983; Pfeffermann 1993).

### *1.3.3 The role of sampling weights in robust estimation, methods for dealing with extreme weights*

Naturally, with the design-based approach, where inferences are made with respect to the sample selection probabilities, the sampling weights, defined as the inverse values of the selection probabilities, play an essential role in estimation.

From the model-based perspective, due to the complex design involving sampling with unequal probabilities, the distribution of sample values, in general, is different from the population distribution. This difference should be taken into account when constructing estimators; in other words, it is important for an estimation procedure to

be *ignorable* to the sample design (see Pfeffermann, 1993, and references therein). Violations of this principle may lead to considerable bias and meaningless results. Conditioning on available design information is one way to account for the sample design: examining graphs of model residuals versus the sampling weights is helpful in determining if ignorability of the design is achieved. Any pattern found on this graph would indicate that the design is nonignorable. More rigorous methods of testing for ignorability also exist (e.g., Sverchkov and Pfeffermann 2004). In practice, not all design variables may be available to an analyst, or their inclusion in the model may become cumbersome. Sampling weights are often used as surrogates of design information to protect against estimation bias.

A survey-weighted estimator may be very inefficient when the survey design is not optimal for a given data item (for example, in a multi-purpose survey where design is tailored for different or multiple characteristics of interest) or due to the cost constraints associated with the sample collection. In other words, when *the working model* used in the sample design does not hold for estimation of a particular characteristic of interest, the design-variance of the survey weighted estimator may be high.

Thus, even when an explicit model is not specified for the purpose of estimation, it is desirable for an estimator to be insensitive to possible non-optimality of the assumptions made at the sampling design stage.

For optimal designs, variation in survey weights increases precision of the survey estimates. However, in cases when design is not optimal for a given data item or analysis, using widely dispersed weights may significantly inflate the variance. One

way of trading off between the design-based bias and variance is to control extreme survey weights. Potter (1988) reviews some of the methods. Weight trimming procedures entail modifying the extreme weights by setting them equal to some lower value while the untrimmed weights are adjusted upward to compensate for the trimmed portion of the weights (Potter, 1988; Potter, 1990). The choice of the trimming point is often arbitrary. Distribution of the survey weights is usually controlled in multi-purpose surveys even before examining the actual effect on an estimator, so that the procedure is not data driven. This is usually justified by operational simplicity: for example, in a multi-purpose survey, it is sometimes convenient to keep one set of weights for many survey variables. A more efficient method would explicitly take into account the effect of the trimmed weights on a survey estimator and proceed to minimize the mean squared error of the estimator by trading off the reduced variance and possible bias resulting from altering the survey weights. One disadvantage of this approach is that the resulting cut-off level for the weight trimming may be different for different data items.

A model-based approach to weight trimming was proposed by Elliott and Little (2000) in the context of estimating finite population means. First, a sample is divided into strata by distinct values of the weights. Models are considered for the survey variables within each stratum: a common mean is imposed on the strata having extreme weights and separate means are considered for each of the lower weight strata. Thus, the weight trimming is accomplished by pooling together the highest weights' strata. The assumption for these pooled strata is that their data are exchangeable; if this assumption fails (for example, when the mean of the highest

weight stratum is considerably different from the other strata), then trimming the highest weights may result in a substantial bias.

An alternative way to modify survey weights, proposed in the same paper (Elliott and Little, 2000), is by using the weight smoothing models. These models treat the strata means as random effects, and the resulting estimate is a compromise, in the form of a composite estimator, of the survey weighted and the unweighted means.

The weight trimming procedures are aimed at reducing the variation in weights; however, they do not protect against effects of extreme sample observations that sometimes occur in surveys.

#### *1.3.4 Survey weights as random variables*

The variables used to design a survey determine the probabilities of inclusion in the sample. In most surveys, at the design stage, the design variables are known for all population units, so they can be regarded as non-random variables for a given fixed finite population. For example, in stratified simple random sampling, strata information is available for all population units, and sample inclusion probabilities are determined by a known number of the population and sample units in each stratum. In such a case, after conditioning on the design variables, the survey weights can be viewed as nonrandom.

At the estimation and analysis stage, it is often the case that the survey design variables are not available for all population units and the inclusion probabilities are often only available for the sampled part of the population. In this case, the design

variables can be viewed as random quantities and so are the inclusion probabilities and the survey weights.

There are reasons to view weights as random quantities even in a simple case like the stratified simple random sampling where the design variables (i.e., strata indicators) are known for all population units and can be regarded as fixed quantities. New information often becomes available after the sample is collected and this information can be more efficiently taken into account if weights are viewed as random variables.

In addition, in the presence of nonresponse, it is not possible to know with certainty the factors that determine the probability of response.

We use the CES survey to further motivate the discussion.

First, as noted earlier, the population of businesses is very dynamic. The snapshot of a population at the time of sample selection is only suggestive of the status of the population at the time of estimation. The variables involved in the CES survey design change: establishments constantly grow or contract and sometimes they also change their industrial classification or geographical location. In particular, the number of population units is not fixed, it continuously changes over time: thus, it is not possible to know the number of units in individual strata, and even the total number of the current population units is not available.

Second, extreme survey weights may cause observations to become highly influential and have a detrimental effect on the estimate, thus prompting the search for a procedure that would reduce the weights of such units.

The above two points are related. The weight depends on a unit's size class at the time of sample selection. However, the unit may belong to a different size class at the time of estimation or the content of the original size stratum may change.

A procedure of weight reduction or "smoothing" can be properly justified by regarding the weights as random rather than fixed non-random quantities. Therefore, we assume a general model-based framework that views both the study variables and their survey weights as random quantities. Such approach to inferences from survey sampling was introduced by Pfeffermann and Sverchkov (PS or SP, hereafter) in a series of papers (PS 1999, 2003, 2007, 2009; SP 2004). This is a model-based approach in the sense that the finite population values are viewed as random variables from a superpopulation distribution. The weights are incorporated into the estimation to account for informativeness of the design.

### *1.3.5 Treatment of extreme observations in surveys*

There is a difference in what is usually called an outlying observation in survey sampling from that in other fields of statistics where the inference is made with respect to an assumed model. In general regression analysis, outliers may occur in values of the analysis variable (i.e., y-values). An outlying value may be interpreted as a gross error in measurement or as a valid observation that comes from a somewhat different parametric distribution than the bulk of the sample. Outliers are also possible in the values of the explanatory variables (i.e., x-values). This sort of outlying observations is usually called *influential points* rather than outliers. The name originates from regression analysis where inference is made conditional on the



values of the explanatory variables; in experimental design, for example, the values of the design variables are not random and the outlying x-values are not necessarily “bad” points but are deliberately chosen by a scientist to reduce the variance of estimates of the model parameters. (Similarly, the sampling weights can be treated as a sort of “design variables”.)

Strictly speaking, following this logic, from the design-based perspective, the notion of outlying values in survey estimation is meaningless (unless it is a reporting error) because the measurements under the design-based approach are viewed as non-random quantities. The suitable alternative is to call the unusual observations *influential* points. Nevertheless, the word “outlier” is routinely used in surveys even when inferences are based on the randomization distribution.

In their discussion on foundations of survey sampling, Hansen, Madow, and Tepping (1983) suggested that an outlier, from the design-based perspective, should be either removed from the sample (presumably, for a reporting error) or that its weight must be reduced. However, as noted in the paper, with such intervention “sampling error is not readily interpreted.” Indeed, basing solely on the design-based theory, there is no justification for either of these actions. A rationalization of such adjustments would involve certain *modifications* to prior *assumptions*, thus, making these assumptions explicit. This means that the purely design-based approach, that considers population values to be fixed quantities, is unsuitable for inferences from a procedure involving treatment of extreme observations.

The model-based oriented authors, on the other hand, recognize the importance of outliers for finite population inferences. Chambers (1986) distinguishes between

*representative* and *nonrepresentative* outliers. Nonrepresentative outliers may be caused by an error in measurements or they may be genuine values that are not representative of other units in the population. Representative outliers are true reports that may be *similar to other population units not present in the sample*. There is an overtone in this definition hinting on imperfections of the sampling design that have lead to the observed sample. Indeed, one proposed scenario to deal with this problem is to assume that the outlying observations come from a separate stratum with a higher variance than the rest of the sample (see also Box and Tiao 1968, Huber 1981). We can read it as the call for *weight adjustment*.

We now stop this philosophical-linguistic digression and briefly review some methods for dealing with extreme sample values in surveys.

Lee (1995) describes two general reasons that an observation may be called outlier in survey sampling: it may have an extreme reported value, as compared to the bulk of the sample, or it may have a large sampling weight even though its reported value may not be extreme. In either event, an observation may not automatically be influential for a given survey estimator. First of all, the influence varies depending on the form of the estimator; second, it is the combined effect of an observation value and its survey weight that determines the influence of a given observation.

It is well known that the survey weighted estimator is design-unbiased (e.g., the Horvitz-Thompson estimator of the population mean) or nearly design-unbiased (e.g., the ratio estimator). However, its design-variance may be inflated because of a few influential observations. Downweighting the extreme points may introduce some bias, but it is usually aimed to reduce the design-variance, such that this “variance-bias

trade-off” strategy aims to reduce the mean squared error (MSE) of the estimator (see also discussion in Lee 1995).

Winsorization is a technique often used to reduce the effect of extreme units. In this approach, the values of observations on the tail of the distribution are reduced to a point between their original value and some predefined cutoff level. Kokic and Bell (1994) applied this approach in the stratified sampling design to reduce the influence of the outlying observations on the expansion estimator of the population total. They developed a method of choosing a set of cutoff values for each stratum which is optimal with respect to the design MSE of the resulting estimator.

Chambers (1986) considered estimation of the finite population total using the best linear unbiased estimator (BLUE) under the linear regression model. The approach does not use survey weights; instead, the model uses the auxiliary information associated with the population elements, including the survey design information, such as, for example, their measure of size. Robust estimation methods (see Huber 1981) developed for samples from an infinite population can be adapted to estimation of the finite population parameters. However, in contrast with the classic infinite population theory, when dealing with the finite population prediction, one has to account for the possibility that the finite population itself contains outliers with respect to the model under consideration. This is an important distinction of the finite population estimation: “effectively, in a finite population problem, we need to predict extreme as well as typical observations” (Welsh and Ronchetti 1998).

Since the outlying units encountered in the sample may be similar to some of the extreme non-sampled units, it is more sensible to give an outlying sample observation

a smaller weight rather than simply discard it. To accomplish this, Chambers (1986) proposed a decomposition of non-sampled prediction into two parts. The first term corresponds to an estimate from the model assuming it holds; the model parameters are estimated using some robust method under the assumed model. The second term is an estimate of the difference between the true total of the non-sampled part and its expectation under the model. The degree of constraint put on the outlying observations depends on the choice of the estimator for this difference. Chambers (1986) considered possible strategies in choosing the estimator.

## 1.4 Estimation under informative sampling (Pfeffermann and Sverchkov approach)

In this Section, we briefly review the details relevant to the application of the approach developed by Pfeffermann and Sverchkov to prediction of the finite population means.

The finite population values  $\{y_j, \mathbf{x}_j, \mathbf{z}_j, j=1, \dots, N\}$  are realizations of vectors of random variables having the probability density function (pdf)  $f_U(y_j, \mathbf{x}_j, \mathbf{z}_j)$ , where  $y_j$  is a study variable,  $\mathbf{x}_j$  is a vector of auxiliary variables, and  $\mathbf{z}_j$  is a vector of design variables; the subscript  $U$  signifies the superpopulation distribution. The sample values of the study variable  $y_j$  have conditional pdf  $f_S(y_j | \mathbf{x}_j) = f_U(y_j | \mathbf{x}_j, j \in S)$ , where  $S$  denotes the set of the sample units. This conditional (on the inclusion into sample) pdf may differ from the population pdf

$f_U(y_j | \mathbf{x}_j)$ . The relationship between the two pdf's can be obtained using Bayes formula:

$$f_S(y_j | \mathbf{x}_j) = \frac{\Pr(I_j = 1 | y_j, \mathbf{x}_j) f_U(y_j | \mathbf{x}_j)}{\Pr(I_j = 1 | \mathbf{x}_j)}, \quad (1.4.1)$$

where  $I_j = 1$  if  $j \in S$  and  $I_j = 0$  if  $j \notin S$ .

Let us examine the relationship (1.4.1). Under a model over the population units, the goal is to predict the parameters of interest of the distribution  $f_U(y_j | \mathbf{x}_j)$  given the available data. One could estimate the parameters using the sample data as if the same model were true for the units in the sample. However, unless the probabilities  $\Pr(I_j = 1 | y_j, \mathbf{x}_j)$  and  $\Pr(I_j = 1 | \mathbf{x}_j)$  are the same for all  $y_j$ 's, the two distributions,  $f_S(y_j | \mathbf{x}_j)$  and  $f_U(y_j | \mathbf{x}_j)$ , are different. The factor  $g(y_j, \mathbf{x}_j) = \frac{\Pr(I_j = 1 | y_j, \mathbf{x}_j)}{\Pr(I_j = 1 | \mathbf{x}_j)}$  provides a *mapping* between the sample and population pdf's.

**Remark.** Note that the vector of design variables  $\mathbf{z}_j$  is not used in the formula (1.4.1). The design variables, in general, are not intended to be used for inference. They are used at the design stage but, for various reasons, may not be available to the analyst at the estimation stage. For example, they may be masked due to confidentiality constraints. See also the relevant discussion in PS (2009). Note, however, that the auxiliary variables  $\mathbf{x}_j$  may account for some or all of the design information. If the design variables were known for all units in the population, then conditioning on  $Z_U = \{\mathbf{z}_j, j = 1, \dots, N\}$  would fully determine the values of probabilities

in (1.4.1),  $\Pr(I_j = 1 | y_j, \mathbf{x}_j, Z_U) = \Pr(I_j = 1 | \mathbf{x}_j, Z_U) = \Pr(I_j = 1 | Z_U) = \pi_j$ , where  $\pi_j$ 's are the sample inclusion probabilities. In such a case, the design process is ignorable for estimation:

$$f_S(y_j | \mathbf{x}_j, Z_U) = f_U(y_j | \mathbf{x}_j, Z_U).$$

Whether the design information is known or not, it is convenient to use a general approach and regard the inclusion probabilities  $\pi_j$ 's as random under a superpopulation model, with pdf  $f_U(\pi_j)$ .

PS (1999) showed that the marginal probabilities are equal to the respective conditional expectations,  $\Pr(I_j = 1 | y_j, \mathbf{x}_j) = E_U(\pi_j | y_j, \mathbf{x}_j)$  and

$\Pr(I_j = 1 | \mathbf{x}_j) = E_U(\pi_j | \mathbf{x}_j)$ . The formula (1.4.1) becomes

$$f_S(y_j | \mathbf{x}_j) = \frac{E_U(\pi_j | y_j, \mathbf{x}_j) f_U(y_j | \mathbf{x}_j)}{E_U(\pi_j | \mathbf{x}_j)} \quad (1.4.2)$$

The formula

$$E_U(y_j | \mathbf{x}_j) = \frac{E_S(w_j y_j | \mathbf{x}_j)}{E_S(w_j | \mathbf{x}_j)}, \quad (1.4.3)$$

where  $w_j = 1/\pi_j$ , relates the expectations over the population and sample distributions (PS 1999).

**Remark.** It is important to emphasize that this is a model-based approach; in particular, the sample distribution is not the distribution over all possible samples as in the randomization approach in surveys. The sample distribution is obtained from

the superpopulation distribution by conditioning on the event of inclusion into the sample. The sample measurements as well as the inclusion probabilities (and the survey weights) are considered random variables and can be described using a model. See discussion on this point in PS (2009).

#### *1.4.1 Prediction of the nonsampled values based on the sample-complement distribution*

The prediction approach in survey sampling uses a model that holds for the sample units to predict the study variables for units outside the sample. If the sampling is informative, however, the distribution in the non-sampled part of the population (sample-complement) has pdf  $f_C(y_j | \mathbf{x}_j) = f_U(y_j | \mathbf{x}_j, j \notin S)$  that is, in general, different from the distribution in the sample (the subscript  $C$  signifies that distribution is over the sample-complement).

This difference must be accounted for. The following formula relates expectations over the sample and sample-complement parts of the population:

$$E_C(y_j | \mathbf{x}_j) = \frac{E_S([\![w_j - 1]\!] y_j | \mathbf{x}_j)}{E_S(w_j - 1 | \mathbf{x}_j)} \quad (1.4.4)$$

(SP 2004).

Suppose the target quantity is a finite population mean  $\bar{Y} = N^{-1} \sum_{j=1}^N y_j$ . SP (2004)

showed that the expectation  $E_U(\bar{Y} | D_S)$  of  $\bar{Y}$  over the population pdf given the data

$D_S = \{(y_i, w_i), i \in S \text{ and } (\mathbf{x}_j, I_j), j \in U\}$ , is the optimal predictor (minimizing the mean squared error with respect to the population pdf given the data) and

$$E_U(\bar{Y} | D_S) = f \frac{1}{n} \sum_{j \in S} y_j + (1-f) \frac{1}{N-n} \sum_{j \notin S} E_C(y_j | \mathbf{x}_j), \text{ where } f = \frac{n}{N} \quad (1.4.5)$$

where  $f = \frac{n}{N}$ , (follows from SP 2004, eq. 3.2)

Using the identity (1.4.4),

$$E_U(\bar{Y} | D_S) = f \frac{1}{n} \sum_{j \in S} y_j + (1-f) \frac{1}{N-n} \sum_{j \notin S} \frac{E_S([w_j - 1]y_j | \mathbf{x}_j)}{E_S(w_j - 1 | \mathbf{x}_j)} \quad (1.4.6)$$

Equation (1.4.6) suggests that the finite population mean can be predicted using a model over the sample units.

**Example 1** In the absence of auxiliary information  $\mathbf{x}_j$  for the non-sampled units, (1.4.6) can be estimated from the sample using the sample mean as an estimate of the expectation  $E_S$ :

$$\hat{\bar{Y}} = f \frac{1}{n} \sum_{j \in S} y_j + (1-f) \frac{\sum_{j=1}^n (w_j - 1) y_j}{\sum_{j=1}^n (w_j - 1)} \quad (1.4.7)$$

(see SP 2004, eq. 5.2)



**Example 2** Consider stratified simple random sampling and suppose vector  $\mathbf{x}_j$  consists of the design information for all population units, that is,  $\mathbf{x}_j = h_j$ , where

$h_j = 1, \dots, H, j = 1, \dots, N$ . The sample weights are  $w_{hj} = \frac{N_h}{n_h}, h = 1, \dots, H$

In this special case, the formula (1.4.6) becomes

$$E_U(\bar{Y} | D_S) = \frac{1}{N} \sum_{h=1}^H \left[ n_h \bar{y}_h + (N_h - n_h) \frac{E_S([w_{hj} - 1] y_{hj} | h_j = h)}{E_S(w_{hj} - 1 | h_j = h)} \right]$$

$$= \frac{1}{N} \sum_{h=1}^H \left[ n_h \bar{y}_h + (N_h - n_h) E_S(y_{hj} | h_j = h) \right], \quad (1.4.8)$$

where  $\bar{y}_h = n_h^{-1} \sum_{j=1}^{n_h} y_{hj}$ . We used the fact that

$$E_S([w_{hj} - 1] y_{hj} | h_j = h) = (w_{hj} - 1) E_S(y_{hj} | h_j = h) \text{ and } E_S(w_{hj} - 1 | h_j = h) = w_{hj} - 1.$$

We can estimate expectation  $E_S(y_{hj} | h_j = h)$  by the sample average in stratum  $h$ .

Then, the estimate of (1.4.8) is

$$\hat{\bar{Y}} = \frac{1}{N} \sum_{h=1}^H \left[ n_h \bar{y}_h + (N_h - n_h) \bar{y}_h \right] = \frac{1}{N} \sum_{h=1}^H \frac{N_h}{n_h} \sum_{j=1}^{n_h} y_j,$$

which is the standard estimator of the population mean for a stratified simple random sampling design.

**Example 3** (Example 2 continued) We can approach the estimation of (1.4.8) by using a model assumption to obtain the estimate of  $E_S(y_{hj} | h_j = h)$ . Since the

sampling is noninformative inside strata, the same model holds for the population and sample data. For example, we can assume a two-level model:

$$\begin{aligned} y_{hj} | \mu_h &\overset{ind}{\sim} N(\mu_h, \sigma^2) \\ \mu_h &\sim N(\mu, \tau^2), \end{aligned} \tag{1.4.9}$$

$$h = 1, \dots, H.$$

This model leads to the following estimate:

$$\hat{\bar{Y}} = N^{-1} \sum_{h=1}^H (n_h \bar{y}_h + (N_h - n_h) \hat{\mu}_h), \tag{1.4.10}$$

where

$$\hat{\mu}_h = \gamma_h \bar{y}_h + (1 - \gamma_h) \tilde{y}_h; \gamma_h = \frac{\tau^2}{\tau^2 + n_h^{-1} \sigma^2}, \tilde{y}_h = \left[ \sum_h \left( \frac{1}{\tau^2 + n_h^{-1} \sigma^2} \right) \right]^{-1} \left[ \sum_h \left( \frac{1}{\tau^2 + n_h^{-1} \sigma^2} \right) \bar{y}_h \right]$$

Note that (1.4.10) is the same estimate as the one used in Ghosh and Meeden (1986) or Elliott and Little (2000) (see the exchangeable random effects model, other models are also possible). Ghosh and Lahiri (1987) obtained the same formula without the normality assumption.

*Discussion:* The survey design may not be efficient for a variable of interest for several reasons (e.g., cost constraints, multipurpose survey designed to meet several goals, or simply because at the design stage the actual values of the study variables,  $y_{hj}$ 's, are not observed). At the estimation stage, after  $y_{hj}$ 's have been observed, it may be desirable to adjust the inclusion probabilities (and survey weights) for a more efficient estimation.

With this in mind, result (1.4.10) can be re-written as

$$\hat{Y} = N^{-1} \sum_h \hat{w}_h \sum_{i=1}^{n_h} y_{hj}. \quad (1.4.11)$$

The weights in (1.4.11) have a composite form  $\hat{w}_h = \gamma_h w_h + (1 - \gamma_h) \tilde{w}$ , where

$$\tilde{w} = \left[ \sum_h \left( \frac{1}{\tau^2 + n_h^{-1} \sigma^2} \right) \right]^{-1} \left[ \sum_h \left( \frac{1}{\tau^2 + n_h^{-1} \sigma^2} \right) w_h \right].$$

They depend on the distribution of  $y_{hj}$ 's through the parameters  $\tau^2$  and  $\sigma^2$ . Thus, the original weights are modified based on the observed values of the study variable  $y_{hj}$ . One must be careful in making the modeling assumptions, however, as they may lead to biased estimates. For example, if variances are different across strata, the estimator based on model (1.4.9) that assumes equal variances may be badly biased.

The following strategy is often used by survey practitioners: obtain two versions of estimates, with and without weights. If the results are close, it is usually suggested to use the unweighted version because it is less variable. This method is somewhat ad hoc and it does not lend itself to an intermediate solution. Treating the survey weights as random variables allows for a more systematic way to test if weights are required, to adjust weights by regressing them on the auxiliary information (PS 1999), and, in general, to use modeling of the weights for a more efficient estimation.

## 1.5 The influence function approach

The approach to robust estimation proposed in this paper is based on the influence function and the first order von Mises expansion. This subsection contains some exposition of the related theory. We include several simple examples of the influence function and discuss the ways it can be used in surveys.

Hampel (1968, 1974) introduced the notion of *the influence function* in infinite population settings. It measures the effect that small changes in the underlying distribution have on the estimator. Important properties related to robustness of an estimator can be derived from the influence function and a robust estimator can be constructed by imposing constraints on the behavior of the influence function.

As noted earlier, the definition of what observation is to be considered influential depends on the form of the estimator. For example, an observation may be considered influential when the estimator is the ratio of two means and not influential when the estimator is a simple mean. The advantage of the influence function approach is also in that it provides a way to assess the effect of an observation taking into account the specific form of the estimator.

### *1.5.1 The Gâteaux derivative and the first order von Mises expansion*

Let  $Y$  denote a random variable having the probability distribution function  $F$ . Consider a real-valued functional  $T(F)$  defined on the space  $\mathcal{F}$  of probability distribution functions. Let  $H$  be another probability distribution function defined on  $\mathcal{F}$ .

The *Gâteaux derivative* of  $T$  at  $F$  in the direction of  $H$  is defined as

$$\lim_{\varepsilon \rightarrow 0} \frac{T((1-\varepsilon)F + \varepsilon H) - T(F)}{\varepsilon} = L_F(H - F). \quad (1.5.1)$$

Assume that  $L_F$  exists and can be represented as

$$L_F(H - F) = \int \psi_F dF, \quad (1.5.2)$$

for some real function  $\psi_F$ , and let  $\int \psi_F dF = 0$ .

Denote by  $\delta_{y_j}$  a probability measure that gives mass 1 to a given point  $y_j$ . If we choose  $H = \delta_{y_j}$  in (1.5.1) then, using representation (1.5.2), we find that the derivative would be  $\psi_F(y_j)$ . The influence function is defined as a derivative of  $T$  at  $F$  in the direction of  $\delta_{y_j}$  as

$$IF(y_j, F, T) = \lim_{\varepsilon \rightarrow 0} \frac{T((1-\varepsilon)F + \varepsilon \delta_{y_j}) - T(F)}{\varepsilon}. \quad (1.5.3)$$

It measures the sensitivity of  $T$  to inclusion of an observation with the value  $y_j$  in a very large sample. Accordingly, it may be more suitably denoted by  $IF_{F,T}(y_j)$ .

See Huber (1981), pp. 37-38.

Let a vector of measurements  $\mathbf{y} = (y_1, \dots, y_N)$  be a set of  $N$  independent realizations (possibly vector-valued) from the probability distribution  $F$ . Suppose a finite population quantity can be viewed as a real-valued functional  $T(F_N)$ , where  $F_N$  is the empirical distribution function (edf) corresponding to  $\mathbf{y}$ ; the value of the

functional  $T$  at  $F$ , i.e.,  $T(F)$ , is the ideal infinite population (i.e., superpopulation) parameter.

The first order Taylor expansion of  $T(F_N)$  in the neighborhood of  $F$ , using Gâteaux derivatives (this particular form of the Taylor expansion is called the von Mises expansion), is

$$T(F_N) = T(F) + N^{-1} \sum_{j=1}^N IF(y_j, F, T) + R_N. \quad (1.5.4)$$

Under suitable regularity conditions, the remainder term  $R_N$  in the above expansion is expected to be negligible (see discussion in Hampel *et al.* 1986, page 85). An outline of the proof that the order of the remainder term is  $O_p(N^{-1})$  can be found in Cox and Hinkley (1974). While it seems possible to make the statement rigorous for certain statistical functionals using results presented in Serfling (1976; Problem 3, page 241, Lemma 6.3.2B, page 223), we did not attempt to do so in this dissertation.

### *1.5.2 Examples of influence functions*

We now present examples of derivation of the influence function in cases of the linear functional, smooth functions of linear functionals, and for the quantiles of the probability distributions.

#### ***Example 1. The influence function of the linear functional.***

Let  $T$  be a linear functional. The derivative of the continuous linear functional is the functional itself. Indeed,

$$\frac{T((1-\varepsilon)F + \varepsilon H) - T(F)}{\varepsilon} = T(H - F).$$

For example, the derivative of  $\mu = T(F) = \int y dF(y)$  is

$$T(H - F) = \int y dH(y) - \mu = \int (y - \mu) dH(y), \text{ and the influence function is}$$

$$IF(y, F, T) = y - \mu.$$

If  $H(y) = F_N(y)$ , then the derivative is  $T(F_N - F) = \frac{1}{N} \sum_{j=1}^N (y_j - \mu)$ , and the

influence function is  $IF(y_j, F, T) = y_j - \mu$  (see also Hampel 1974).

**Example 2 A smooth function of linear functionals.**

For a smooth function of linear estimators, the influence function can be obtained as the usual derivative of the composite function. Let us, for example, derive the influence function for the ratio of two means.

Let the finite population values  $(y_{11}, \dots, y_{1N})$  and  $(y_{21}, \dots, y_{2N})$  be realizations of random variables with cdf's  $F_1$  and  $F_2$ , respectively; let  $\mu_1$  and  $\mu_2$  be their respective first moments. Consider estimation of  $R = \frac{\mu_1}{\mu_2}$ . It can be viewed as the

ratio of two linear functionals, i.e., as the composite functional

$$G(T(F_1), T(F_2)) = \frac{T(F_1)}{T(F_2)}, \text{ where } T(F_1) \equiv \mu_1 \text{ and } T(F_2) \equiv \mu_2.$$

We write,

$$R' = \frac{\partial G}{\partial \mu_1} T'(F_1) + \frac{\partial G}{\partial \mu_2} T'(F_2), \text{ where } \frac{\partial G}{\partial \mu_1} = \frac{1}{\mu_2}, \frac{\partial G}{\partial \mu_2} = -\frac{\mu_1}{\mu_2^2}.$$

For empirical distribution functions, as in *Example 1*, we have

$$T'(F_1) = \frac{1}{N} \sum_{j=1}^N (y_{1j} - \mu_1) \text{ and } T'(F_2) = \frac{1}{N} \sum_{j=1}^N (y_{2j} - \mu_2)$$

Therefore,

$$R' = \frac{1}{\mu_2} \frac{1}{N} \sum_{j=1}^N (y_{1j} - \mu_1) - \frac{\mu_1}{\mu_2^2} \frac{1}{N} \sum_{j=1}^N (y_{2j} - \mu_2) = \frac{1}{N} \sum_{j=1}^N \frac{1}{\mu_2} \left( y_{1j} - \frac{\mu_1}{\mu_2} y_{2j} \right)$$

and the influence function is  $\frac{1}{\mu_2} \left( y_{1j} - \frac{\mu_1}{\mu_2} y_{2j} \right)$ .

**Example 3 The influence function for quantiles.**

By definition, let  $q_\alpha = T(F) = F^{-1}(\alpha) = \inf \{y : F(y) \geq \alpha\}$  be the quantile at level  $\alpha$ ,

for some cdf  $F$ . Assume the positive density  $f = F'$  exists in a neighborhood of  $q_\alpha$ .

Let  $F_\varepsilon = (1 - \varepsilon)F + \varepsilon H$  be a perturbed cdf.

$$\begin{aligned} T(F_\varepsilon) &= \inf \{y : F_\varepsilon(y) \geq \alpha\} = \inf \{y : (1 - \varepsilon)F(y) + \varepsilon H(y) \geq \alpha\} \\ &= \inf \left\{ y : F(y) \geq \frac{\alpha - \varepsilon H(y)}{1 - \varepsilon} \right\} \end{aligned}$$

Therefore,  $T(F_\varepsilon) = F^{-1} \left( \frac{\alpha - \varepsilon H}{1 - \varepsilon} \right)$

Let us find the derivative:

$$\left. \frac{\partial T(F_\varepsilon)}{\partial \varepsilon} \right|_{\varepsilon=0} = \left. \frac{\partial}{\partial \varepsilon} \left( F^{-1} \left( \frac{\alpha - \varepsilon H}{1 - \varepsilon} \right) \right) \right|_{\varepsilon=0} = \frac{\alpha - H(F^{-1}(\alpha))}{f(F^{-1}(\alpha))} = \frac{\alpha - H(q_\alpha)}{f(q_\alpha)}$$

Let



$$H(q_\alpha) = \begin{cases} 1, & y \leq q_\alpha \\ 0, & y > q_\alpha \end{cases} .$$

Thus, the influence function is

$$IF(y, F, T) = \begin{cases} \frac{\alpha - 1}{f(q_\alpha)}, & y \leq q_\alpha \\ \frac{\alpha}{f(q_\alpha)}, & y > q_\alpha \end{cases}$$

$$\text{For the median } (\alpha = 0.5), IF(y, F, T) = \begin{cases} -\frac{1}{2} \frac{1}{f(q_{0.5})}, & y \leq q_{0.5} \\ \frac{1}{2} \frac{1}{f(q_{0.5})}, & y > q_{0.5} \end{cases}$$

(see also Huber 2004, pp.56-57).

The influence function for the median is bounded (as long as  $f(q_\alpha) > 0$ ), thus the median is a robust estimator of the location parameter. Note the distinction between *estimation* of the finite population median (and the population quantiles, in general) and using the median as an *estimator* of the location parameter under a model. In surveys with unequal weighting, the estimator of the finite population median depends on the distribution of the weights. The survey weighted estimate of the influence function would involve weights of the observations with values  $y \leq q_{0.5}$ , thus, the estimator is not robust to appearance of extreme weights.

### 1.5.3 Applications of the influence function approach in surveys

Finite sample versions of the influence function exist. One way to obtain the sample version of the influence function is by replacing  $F$  in (1.5.3) with  $F_{n-1}$  and  $\varepsilon$  with  $1/n$ . This version is called the *sensitivity curve* (Tukey 1970).

Hulliger (1995) defined the sensitivity curve for a sample drawn with unequal inclusion probabilities. He considers a class of M-estimators to robustify the Horvitz-Thompson (HT) estimator of the finite population mean. To this end, he describes a superpopulation linear model that is implicit for the HT estimator. This model involves an auxiliary variable  $x_j$  that represents a size measure used to define the inclusion probabilities  $\pi_j$ . The HT estimator is viewed as a functional of the sample empirical distribution function (edf). The sample edf is itself an estimate that employs the sampling weights; hence, the sensitivity curve includes the sampling weights (that are non-random). The influence of an observation on the HT estimator depends on the residual  $y_j - \beta x_j$ . Properties of the HT estimator and its robustified version may be studied using the sensitivity curve. The sensitivity curve is also used to derive the approximate variance of the estimator.

The approach considered by Hulliger (1995) is not a prediction approach but is intended to “establish the link to classical robust statistics” in order to robustify the HT estimator. In the current paper, instead of establishing the “link”, we strive to produce a model-based approach to robust estimation.

Zaslavsky, Schenker and Belin (2001) (ZSB, hereafter) used the influence function approach in a cluster sample for the 1990 Post Enumeration Survey (PES). They define the influence function for the finite population and use the empirical influence function for the sample units. The population is defined as a set of vectors  $\{(y_j, w_j), j = 1, \dots, N\}$  having distribution  $G$  that assigns mass  $\pi_j = 1/w_j$  to each unit  $j$  in the finite population. The influence function of a unit  $(y_j, w_j)$  on a functional  $Q$  on  $G$  is defined by analogy to (1.5.3) as

$$IF\left((y_j, w_j), G, Q\right) = \lim_{\varepsilon \rightarrow 0} \frac{Q\left((1-\varepsilon)G + \varepsilon\delta_{(y_j, w_j)}\right) - Q(G)}{\varepsilon}.$$

The corresponding sample version of the influence function is obtained by replacing  $\varepsilon$  with  $1/n$  and  $G$  with  $G_n$ , where  $G_n$  assigns mass  $1/n$  to each  $(y_j, w_j)$  in the sample.

Consider a finite population quantity  $T(F_N)$ , where the distribution  $F_N$  puts mass  $1/N$  on the finite population values  $y_j, j = 1, \dots, N$ . Let  $F_n$  be the weighted empirical cumulative distribution function that assigns mass  $w_j / \sum_{j \in S} w_j$  to a sampled value  $y_j, j \in S$ . The goal is to define the influence of a unit  $j$  on the survey weighted estimator  $T(F_n)$  of the target  $T(F_N)$ . ZSB note that  $F_N$  maps to  $F_n$  by the same mapping as  $G$  maps to  $G_n$ . Therefore, the influence of a sample unit  $j$  on an

estimator  $T(F_n)$  of  $T(F_N)$  is the same as the influence of the unit on an estimator  $Q(G_n)$  for  $Q(G)$ .

The form of the empirical influence function considered by ZSB is similar to the form of the estimated influence function proposed in the current paper. However, our justification of the method is different. Our approach differs from Hulliger (1995) or Zaslavsky *et al.* (2001) in that we do not use a finite version of the influence function (such as the sensitivity curve): we view the finite population quantity of interest as a functional of the finite population edf, derive the influence function with reference to the ideal superpopulation (infinite) distribution function and then estimate it using a sample.

As suggested by Hampel (1968, 1974), there is a close tie between the influence function and M-estimators and “this opens many possibilities of defining new estimators with prescribed properties.” ZSB fit a long-tailed distribution to the influence statistics, thus determining the adjustment factors to reduce the effect of the influential clusters. They derive a robust estimator directly from the analysis of the influence function by employing the t-distribution and M-estimation (Huber 1981).

The use of the multilevel or hierarchical Bayes modeling, e.g., utilizing mixture models, may be a good way to approach the estimation. Hampel *et al.* (1986) define the robust estimation approach as lying between two extremes, the fully parametric and the non-parametric approaches to estimation. Although a model should be explicitly stated in a mixture model approach, the flexibility of the mixture modeling places it “between the extremes.”

In their modeling, ZSB deal with the combined effect of the survey weights and unusual data values. This is important in samples with differential weights because it is often a combination of moderately high weight and outlying data that creates an influential point. We explore the same idea of combining the survey weights and the sample measurements and use the product as a one-dimensional random variable. It is also possible, within the same general framework, to approach the problem by analyzing the two-dimensional variable that includes data value and weight as the two components of the random vector. This approach is not pursued in this dissertation.

## 1.6 Robust small area estimation

Complex surveys are usually designed to collect enough sample units from a population of interest and make estimates of population quantities based on this sample with a satisfactory precision; however, at a progressively finer level of detail, where the sample is sparse, direct sample based estimates are not reliable anymore. The problem of estimation at such detailed levels is known as the small area estimation (SAE) problem.

Small area estimation generally relies on some, implicit or explicit, modeling assumptions. Robustness here is often understood as insensitivity of the estimator to model misspecification. For example, Ghosh and Lahiri (1987) obtained the linear Bayes estimator of a small area mean using the so called posterior linearity assumption under which the posterior mean is a linear function of the observation. The resulting linear empirical Bayes estimator of the small area mean is, however, identical to the normality-based empirical Bayes estimator demonstrating its

robustness within the class of Bayesian models covered by the posterior linearity assumption.

It may happen that a model explains well the bulk of the data, yet a few observations do not fit into the model. Such observations may adversely affect estimation of the model parameters. This calls for development of methods of estimation that are robust to the appearance of outliers, and several outlier resistant methods have been proposed in the SAE literature in recent years. Heavy tailed distributions, such as Cauchy or  $t$ -distributions, offer some protection against outliers. In area-level settings, Datta and Lahiri (1995) considered a general case of the scale mixture of normal distributions and studied the behavior of the Bayes estimator asymptotically, when a single outlier is extremely large. They showed that the Bayes estimator for an outlying area approaches the direct estimator for that area while retaining shrinkage for the non-outlying areas. Robust area-level models involving the  $t$ -distribution were also considered by Xie *et al.* (2005), Huang and Bell (2006). Ghosh *et al.* (2008) introduced a robust approach using the influence function in the context of area-level models.

We consider an outlier robust approach to estimation in unit-level models. In a simulation study in Chapter 3, we compare the proposed model with approaches of Fellner (1986), Chambers and Tzavidis (2006), and Sinha and Rao (2008). We now review in some detail these methods of robust small area estimation.

Under the prediction approach to surveys, an estimator of  $\bar{Y}_m$ , the small area  $m$  mean, is given by

$$\hat{Y}_m = f_m \bar{y}_m + (1 - f_m) \hat{Y}_{mr}, \quad (1.6.1)$$

where  $m=1, \dots, M$ ;  $\bar{y}_m = n_m^{-1} \sum_{j=1}^{n_m} y_{mj}$  is the sample mean, index  $mj$  denotes observation  $j$  from area  $m$ ,  $f_m = N_m^{-1} n_m$ ,  $N_m$  and  $n_m$  are the number of area  $m$  population and sample units,  $\sum_{m=1}^M N_m = N$ ;  $\sum_{m=1}^M n_m = n$ ;  $\hat{Y}_{mr}$  is a model-dependent predictor of the mean of the non-sampled part of area  $m$ .

In particular, the predictor  $\hat{Y}_{mr}$  can be obtained based on linear mixed model assumptions. The linear mixed model for the vector of observations  $\mathbf{y}$  is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \quad (1.6.2)$$

where  $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_M)^T$ ,  $\mathbf{y}_m = (y_{m1}, \dots, y_{m_{n_m}})^T$ ;  $m=1, \dots, M$ ;  $\sum_{m=1}^M n_m = n$ ;  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$  is a vector of parameters;  $\mathbf{u} = (u_1, \dots, u_M)^T$  is a vector of random effects;  $\mathbf{u} \sim N(0, \mathbf{D})$ ;  $\mathbf{e} \sim N(0, \mathbf{R})$ ;  $\mathbf{u}$  and  $\mathbf{e}$  are assumed to be mutually independent;  $\mathbf{X}$  is an  $n \times p$  matrix of known auxiliary variables;  $\mathbf{Z}$  is an  $n \times M$  known design matrix for the random effects;  $\boldsymbol{\Sigma} = \mathbf{R} + \mathbf{Z}\mathbf{D}\mathbf{Z}^T$  is the variance-covariance matrix of  $\mathbf{y}$ . It is assumed that the variance-covariance matrix is parameterized by some vector of variance components  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_L)^T$ ,  $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ .

An important special case of the linear mixed model is the nested-error regression (NER) model considered by Battese, Harter, and Fuller (1988). In the case of NER,  $\mathbf{R} = \sigma^2 \mathbf{I}_n$  and  $\mathbf{D} = \tau^2 \mathbf{I}_M$ . The model can be written as follows:

$$y_{mj} = \mathbf{x}_{mj}^T \boldsymbol{\beta} + u_m + \varepsilon_{mj}, \quad (1.6.3)$$

$$u_m \stackrel{iid}{\sim} N(0, \tau^2) \text{ and } \varepsilon_{mj} \stackrel{iid}{\sim} N(0, \sigma^2), \quad (1.6.4)$$

$$j = 1, \dots, n_m, m = 1, \dots, M,$$

where  $\mathbf{x}_{mj}$  is a vector of auxiliary variables for an observation  $(mj)$ ,  $\boldsymbol{\beta}$  is the corresponding vector of parameters. The distribution of the random effects  $u_m$  describes deviations of the area means from values  $\mathbf{x}_{mj}^T \boldsymbol{\beta}$ ;  $\varepsilon_{mj}$  are errors in individual observations.

Assume that sampling is non-informative for the distribution of measurements  $\mathbf{y}$ , given the auxiliary information  $\mathbf{X}$ . The best linear unbiased predictor (BLUP) of  $\bar{Y}_{mr}$  under the linear mixed model has the form

$$\hat{\bar{Y}}_{mr} = \bar{\mathbf{x}}_{mr}^T \hat{\boldsymbol{\beta}} + \hat{u}_m, \quad (1.6.5)$$

where  $\bar{\mathbf{x}}_{mr}^T = (N_m - n_m)^{-1} \sum_{j=n_m+1}^{N_m} \mathbf{x}_{mj}^T$ ;  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{y}$  is the best linear unbiased estimator (BLUE) of  $\boldsymbol{\beta}$ ; the best linear unbiased predictor (BLUP)  $\hat{\mathbf{u}}$  of  $\mathbf{u}$  is given by  $\hat{\mathbf{u}} = \mathbf{DZ}^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ . In the case of NER, BLUP of  $u_m$  is spelled out as

$$\hat{u}_m = \frac{\tau^2}{\sigma^2/n_m + \tau^2} (\bar{y}_m - \bar{\mathbf{x}}_m^T \hat{\boldsymbol{\beta}}). \quad (1.6.6)$$

When the variance  $\boldsymbol{\Sigma}$  is unknown, it is estimated from the data producing the empirical best linear unbiased predictor (EBLUP) for  $\bar{Y}_{mr}$ .



There may be some areas that do not fit the assumption on the random effects  $u_m$ . We will call such areas *outlying areas*. There may also be individual observations that are not well described by the model assumption on the error terms  $\varepsilon_{mj}$ . Such observations will be called *individual outliers*. The influence of outliers on estimation of the model parameters can be reduced by using bounded influence functions for the corresponding residual terms when fitting the model estimating equations. This approach was taken by Fellner (1986); a modification of Fellner's approach, also involving the bounded influence functions, was proposed by Sinha and Rao (2008).

Fellner (1986) proposed to solve simultaneously the following set of estimating equations:

$$\mathbf{X}^T \mathbf{R}^{-1/2} \boldsymbol{\Psi}(\mathbf{R}^{-1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})) = \mathbf{0} \quad (1.6.7)$$

$$\mathbf{Z}^T \mathbf{R}^{-1/2} \boldsymbol{\Psi}(\mathbf{R}^{-1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})) - \mathbf{D}^{-1/2} \boldsymbol{\Psi}(\mathbf{D}^{-1/2}\mathbf{u}) = \mathbf{0}, \quad (1.6.8)$$

where  $\boldsymbol{\Psi}(\mathbf{u}) = (\psi_b(u_1), \dots, \psi_b(u_M))^T$ ,  $\psi_b(u_m)$  is a bounded function; for example, it can be Huber's function  $\psi_b(u_m) = \min(b, \max(-b, u_m))$ , where  $b$  is a tuning parameter, with a usual choice of  $b = 1.345$ . The equations are solved iteratively. Consider, for simplicity, the case of the nested-error regression model, where the variance components are  $\boldsymbol{\theta} = (\sigma^2, \tau^2)^T$ . The variance components are estimated at each iteration step  $k+1$  as

$$\hat{\tau}^{2(k+1)} = \hat{\tau}^{2(k)} \boldsymbol{\Psi}^T(\hat{\tau}^{-1(k)} \hat{\mathbf{u}}^{(k)}) \boldsymbol{\Psi}(\hat{\tau}^{-1(k)} \hat{\mathbf{u}}^{(k)}) / h(M - \nu^{(k)}) \quad (1.6.9)$$

$$\hat{\sigma}^{2(k+1)} = \hat{\sigma}^{2(k)} \Psi^T \left( \hat{\sigma}^{-1(k)} \hat{\mathbf{e}}^{(k)} \right) \Psi \left( \hat{\sigma}^{-1(k)} \hat{\mathbf{e}}^{(k)} \right) / h \left( N - p - \left( M - v^{(k)} \right) \right), \quad (1.6.10)$$

where  $\hat{\mathbf{e}}^{(k)} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(k)} - \mathbf{Z}\hat{\mathbf{u}}^{(k)}$ ,  $v^{(k)} = \frac{tr(\mathbf{T}^{(k)})}{\hat{\tau}^{2(k)}}$ ,  $\mathbf{T}^{(k)}$  is formed by  $M$  last rows and

columns of the matrix  $(\mathbf{C}^T \mathbf{C})^{-1}$ , where  $\mathbf{C} = \begin{pmatrix} \mathbf{R}^{-1/2} \mathbf{X} & \mathbf{R}^{-1/2} \mathbf{Z} \\ \mathbf{0} & \mathbf{D}^{-1/2} \end{pmatrix}$ ; the constant  $h$  is

$h = E[\psi_b^2(u)]$ , where  $u \sim N(0,1)$ . It depends on the tuning parameter  $b$  and is

$$h(b) = F_{\chi_3^2}(b^2) + b^2 \left( 1 - F_{\chi_1^2}(b^2) \right).$$

Sinha and Rao (2008) propose to start from obtaining the estimates of the variance components from a marginal model. For this, the following estimating equations are to be solved:

$$\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{U}^{1/2} \boldsymbol{\Psi}(\mathbf{r}) = \mathbf{0} \quad (1.6.11)$$

$$\boldsymbol{\Psi}^T(\mathbf{r}) \mathbf{U}^{1/2} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_l} \boldsymbol{\Sigma}^{-1} \mathbf{U}^{1/2} \boldsymbol{\Psi}(\mathbf{r}) - tr \left( \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_l} \mathbf{K} \right) = 0, \quad (1.6.12)$$

where  $\mathbf{r} = \mathbf{U}^{-1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ ,  $\mathbf{U} = \text{diag}(\boldsymbol{\Sigma})$ ,  $\boldsymbol{\Psi}(\mathbf{r})$  is defined similar to the bounded function of Fellner's approach, and  $\mathbf{K} = h\mathbf{I}_n$  with  $h$  as in Fellner's approach. The equations (1.6.11) and (1.6.12) are solved using the Newton-Raphson algorithm. After the robust estimates of parameters are obtained, they are plugged into the equation (1.6.8) to obtain the robust prediction of the random effects using the Newton-Raphson algorithm.

The predictor for  $\bar{Y}_{mr}$  based on such a robustified fitting of the linear mixed model (using either Fellner's or Sinha and Rao's algorithm) is called the Robust Empirical Best Linear Unbiased Predictor (REBLUP):

$$\hat{\bar{Y}}_{mr}^{REBLUP} = \bar{\mathbf{x}}_{mr}^T \hat{\boldsymbol{\beta}}^{REBLUP} + \hat{u}_m^{REBLUP}. \quad (1.6.13)$$

An alternative to the mixed model approach to robust SAE is based on M-quantile regression, which is a generalization of the quantile regression technique. This approach was proposed by Chambers and Tzavidis (2006).

In M-quantile regression, a separate set of linear regression parameters is considered for quantiles  $q$  of the conditional distribution of  $\mathbf{y}$  given  $\mathbf{X}$ . The M-estimator of the vector  $\boldsymbol{\beta}_q$  of the  $q$  th quantile regression coefficients is a solution to estimating equations of the form

$$\sum_{m=1}^M \sum_{j=1}^{n_m} \psi_q(r_{mj,q}) \mathbf{x}_{mj} = 0, \quad (1.6.14)$$

where  $r_{mj,q} = y_{mj} - \mathbf{x}_{mj}^T \boldsymbol{\beta}_q$  are residuals,

$\psi_q(r_{mj,q}) = 2\psi(s_q^{-1}r_{mj,q})\{qI(r_{mj,q} > 0) + (1-q)I(r_{mj,q} \leq 0)\}$ ,  $\psi$  is a bounded influence function,  $s_q$  is a robust estimate of scale. For example,

$s_q = \text{med} \left| r_{jq} - \text{med}(r_{jq}) \right| / 0.6745$ . Denote the quantile of an observation ( $mj$ ) by  $q_{mj}$ .

The second step consists of finding the average quantile of the observations in each area  $m$  as  $\bar{q}_m = n_m^{-1} \sum_{j=1}^{n_m} q_{mj}$ . The estimate of each area's slope  $\boldsymbol{\beta}_m$  is determined by the

value of the area's average quantile  $\bar{q}_m$ ,  $\hat{\beta}_m^{MQ} = \hat{\beta}_{q=\bar{q}_m}$ . The M-quantile estimator of  $\bar{Y}_{mr}$  is given by

$$\hat{Y}_{mr}^{MQ} = \bar{\mathbf{x}}_{mr}^T \hat{\beta}_m^{MQ}. \quad (1.6.15)$$

Outliers may suggest a real finite population structure that is not described by the assumed base model. Such representative outliers (using Chambers' 1986 terminology) carry important information and it would be unwise to ignore it and rely only on the base model. In the non-SAE settings, Chambers (1986) proposed to apply a bias correction to the initial estimator, where the initial estimator is based firmly on the assumed working model while the bias correction is an estimated mean of residuals after relaxing the modeling assumptions. The bias correction idea in application to SAE is to add separate bias correction terms to the initial predictors for each area, a method explored by Chambers *et al.* (2009). The drawback of such adaptation of the non-SAE methodology is that inevitably the estimation of the bias correction terms for small areas would be based on small samples, potentially leading to inefficient estimates.

We next describe the bias correction approach proposed by Chambers *et al.* (2009). The estimation consists of two steps. First, find robust estimates using any outlier robust estimation method, for example, one of the approaches described above. Second, estimate the bias of the initial robust estimate using, again, an outlier robust approach but with different tuning parameters in the bounded influence functions. The purpose of the second step is to "undo" the effect of a possible model misspecification imposed at step one. The second step boundaries for the influence

function should be wide enough, promoting more reliance on the data rather than on the model assumptions. The final estimate is the sum of the robust estimate computed at the first step and the bias correction term computed at the second step.

Let  $\phi(\cdot)$  be some bounded function. It can be Huber's function  $\phi_c(r) = \min(c, \max(-c, r))$ , where the tuning parameter  $c$  is chosen to be relatively large; for example,  $c = 3$ .

The bias-corrected version of REBLUP (either Fellner's or Sinha and Rao's approach) is

$$\hat{Y}_{mr}^{REBLUP+BC} = \hat{Y}_{mr}^{REBLUP} + n_m^{-1} \sum_{j=1}^{n_m} s_m^{REBLUP} \phi \left( \frac{y_{mj} - \mathbf{x}_{mj}^T \hat{\boldsymbol{\beta}}^{REBLUP} - \hat{u}_m^{REBLUP}}{s_m^{REBLUP}} \right). \quad (1.6.16)$$

The bias-corrected version of Chambers and Tzavidis' approach is

$$\hat{Y}_{mr}^{MQ+BC} = \hat{Y}_{mr}^{MQ} + n_m^{-1} \sum_{j=1}^{n_m} s_m^{MQ} \phi \left( \frac{y_{mj} - \mathbf{x}_{mj}^T \hat{\boldsymbol{\beta}}^{MQ}}{s_m^{MQ}} \right). \quad (1.6.17)$$

Here  $s_m^{REBLUP}$  and  $s_m^{MQ}$  are some robust estimates of scale for the respective sets of residuals in area  $m$ . For example, for the REBLUP residuals

$e_{mj}^{REBLUP} = y_{mj} - \mathbf{x}_{mj}^T \hat{\boldsymbol{\beta}}^{REBLUP} - \hat{u}_m^{REBLUP}$ , the estimator for the scale can be

$s_m^{REBLUP} = \text{med} \left| e_{mj}^{REBLUP} - \text{med} \left( e_{mj}^{REBLUP} \right) \right| / 0.6745$ ; for the MQ residuals

$e_{mj}^{MQ} = y_{mj} - \mathbf{x}_{mj}^T \hat{\boldsymbol{\beta}}^{MQ}$ , one can use the estimator  $s_m^{MQ} = \text{med} \left| e_{mj}^{MQ} - \text{med} \left( e_{mj}^{MQ} \right) \right| / 0.6745$ .

## Chapter 2: Robust Estimation in Moderately Large Samples

In this chapter, we consider estimation of the finite population parameters in moderately large samples. At the first step, the finite population quantity of interest is approximated using a first order Taylor expansion. We would like to emphasize that at the first step we deal only with the finite population rather than the sample: thus, the precision of this approximation depends only on the size of the population and does not depend on the sample size. The finite population is usually large enough to ensure that the linearization provides a good approximation of the target quantity.

We also note that even though we view the population units as random realizations from superpopulation distribution, at the first step, we do not assume any specific model and only require that the finite population quantity of interest be sufficiently regular to admit a Taylor expansion and the population measurements should be independent.

At the second step, we re-express the quantity of interest in terms of the expectation under the sample distribution, see (2.1.6) below. After that point, we start making modeling assumptions based on the observed sample.

In Section 2.1 we discuss the idea of linearization in its general form and in Section 2.2 we apply it to the CES estimator. In Section 2.3, we discuss Winsorization and the way to choose the cutoff points leading to an estimator with a reduced mean square error. We present several simulation examples that demonstrate the importance of taking into account the sampling design. Simulation results using several scenarios

for the finite population distribution are presented in Section 2.4 and mean squared error estimation is considered in Section 2.5.

## 2.1 Linearization of a finite population quantity

Assume that a vector of population measurements  $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$  is a realization from some probability distribution  $F$  (a superpopulation distribution) (in general, each  $\mathbf{y}_j$  is a vector of measurements on a unit  $j$ );  $P$  is a set of population units and  $S$  is a set of sampled units;  $N$  and  $n$  are the numbers of units in the population  $P$  and the sample  $S$ , respectively.

Let  $F_N$  denote the edf of the finite population  $P$ . Suppose we are interested in estimating the finite population quantity  $T(F_N)$  defined as a smooth function of the finite population means.  $T(F_N)$  is assumed to be sufficiently regular to be linearized near  $F$  using the Taylor expansion

$$T(F_N) = T(F) + N^{-1} \sum_{i=1}^N IF_{F,T}(\mathbf{y}_i) + R_N, \quad (2.1.1)$$

where  $T(F)$  is a superpopulation parameter and  $IF_{F,T}(\mathbf{y}_j)$  is the influence function of the functional  $T$  (see (1.5.4)).

As noted in Section 1.5.1, under suitable regularity conditions, the remainder term is small. Let us drop the remainder term in (2.1.1) and *redefine* the population quantity that we target in our estimation to be

$$\tilde{T}(F_N) = T(F) + N^{-1} \sum_{j=1}^N IF_{F,T}(\mathbf{y}_j). \quad (2.1.2)$$

Given the *population* size  $N$  is large, this quantity differs from the ideal target,  $T(F)$ , by a small value.

For the moment, we suppose that the parameter  $T(F)$  is known. The terms  $IF_{F,T}(\mathbf{y}_j)$  can be viewed as generalized residuals, representing the difference between the population observation  $\mathbf{y}_j$  and the parameter  $T(F)$ . Thus, the second term on the right hand side of (2.1.2) represents the mean difference between the population units and the value of  $T(F)$ . This difference is to be estimated using some robust method (for example, the Winsorization approach is considered in Section 2.3.)

To estimate  $\tilde{T}(F_N)$ , let us equivalently re-write (2.1.2) as

$$\tilde{T}(F_N) = T(F) + \frac{n}{N} \bar{U}_S + \frac{N-n}{N} \bar{U}_C, \quad (2.1.3)$$

where

$$\bar{U}_S = \frac{1}{n} \sum_{j \in S} IF_{F,T}(\mathbf{y}_j) \quad (2.1.4)$$

and

$$\bar{U}_C = \frac{1}{N-n} \sum_{j \in C} IF_{F,T}(\mathbf{y}_j) \quad (2.1.5)$$

are means of the influence function for the observations that are included, (2.1.4), and not included, (2.1.5), in the sample.



Under the prediction approach to inference in sampling from finite populations, the goal is to predict values in the sample-complement part of the population,  $C$ , using the sample measurements. In our formulation, the problem is to predict the value of  $\bar{U}_C$ .

The distribution of the sample measurements may differ from the distribution of the population measurements. If this is the case, it is important to account for the difference in order to avoid estimation bias. We employ the general result (1.4.4) to predict  $\bar{U}_C$  by estimating the sample-complement expectation  $E_C[IF_{F,T}(\mathbf{y}_j)]$  from the sample.

Using (1.4.4), the population quantity can be expressed as

$$\tilde{T}(F_N) = T(F) + f\bar{U}_C + (1-f)E_S \left[ \frac{(w_j - 1)}{E_S[w_j - 1]} IF_{F,T}(\mathbf{y}_j) \right] \quad (2.1.6)$$

The estimator of  $\tilde{T}(F_N)$  takes the form

$$\hat{\tilde{T}}(F_N) = \hat{T}(F_N) + f \frac{1}{n} \sum_{j=1}^n \hat{u}_j + (1-f) \frac{\hat{E}_S \left[ (w_j - 1) \hat{u}_j \right]}{\hat{E}_S[w_j - 1]}, \quad (2.1.7)$$

where  $\hat{u}_j$  is an estimator of  $IF_{F,T}(y_j)$  and depends on a choice of  $\hat{T}(F_N)$ ;  $\hat{E}_S[\cdot]$  denotes an estimator of the expectation  $E_S[\cdot]$ .

**Remark.** Chambers (1986) used decomposition somewhat similar to (2.1.3) in estimation of finite population totals under linear model assumptions (in this case,  $T(F_N) = \tilde{T}(F_N)$ ). The first part of the decomposition involves an outlier-robust

estimator under the assumption that the model holds exactly. The estimator that reduces the effect of outlying observations may be biased if the finite population itself contains outliers, in other words, if the assumed model does not hold exactly. Since the sample outliers may be representative of similar units in the finite population, they should not be completely neglected. Thus, the last term of the decomposition estimated the difference between the “true” population value and the corresponding expectation under the model. It can be interpreted as a bias correction term, where the “bias” is understood as a difference between “truth” and the model assumptions.

When the sample is (moderately) large, we may use a survey weighted estimator (or some design consistent estimator) for  $T(F)$ . As noted earlier, often this estimator is sensitive to outliers. The outliers may be viewed as indicators that the implicit model underlying the use of the survey weighted estimator is “misspecified”. From this point of view, the other terms in (2.1.7) are meant to correct for bias.

The expectation  $E_s[\cdot]$  does not have to be estimated as a sample arithmetic average. Some modeling methods can be used to find an estimator for the last term in (2.1.7). Auxiliary information, if available, can also be used in modeling the last term in (2.1.7).

## 2.2 Application: Estimation from large samples in CES

We now consider in some detail the CES estimator of the relative monthly employment growth. The target finite population quantity is

$$R_t = \frac{\sum_{j \in P_t} y_{j,t}}{\sum_{j \in P_t} y_{j,t-1}} \quad (2.2.1)$$

where

$P_t$  is a set of a population establishments having non-zero employment in both previous  $t-1$  and current  $t$  months;  $y_{j,t-1}, y_{j,t}$  are the previous and current month's levels of employment in establishment  $j$ , respectively.

Consider a superpopulation parameterization. For a given month  $t$ , consider the set of finite population observations  $\{(y_{j,t-1}, y_{j,t}) \mid j \in P_t\}$  to be independent realizations of a random vector  $(Y_{t-1}, Y_t)$  having some unspecified probability distribution  $F_{t-1,t}$ .

Denote by  $(\theta_{t-1}, \theta_t)$  a vector of means of  $(Y_{t-1}, Y_t)$ . The superpopulation parameter of interest is a function of the superpopulation means  $(\theta_{t-1}, \theta_t)$ :

$$T(F) = T(\theta_{t-1}, \theta_t; F) = \frac{\theta_t}{\theta_{t-1}}.$$

The influence function is

$$u_{j,t} \equiv IF_{F,T}(y_{j,t-1}, y_{j,t}) = \frac{1}{\theta_{t-1}} \left( y_{j,t} - \frac{\theta_t}{\theta_{t-1}} y_{j,t-1} \right) \quad (2.2.2)$$

(see *Example 2* in Section 1.5.2).

Note that  $N$  is unknown and is estimated as  $\hat{N} = \sum_{j \in S_t} w_j$ . The sampling fraction is

estimated as  $\hat{f} = \frac{n}{\hat{N}}$ . Let  $\hat{R}_t = \frac{\sum_{j \in S_t} w_j y_{j,t}}{\sum_{j \in S_t} w_j y_{j,t-1}}$  be the usual WLR estimator (as defined in

Section 1.2.2) and let

$$\hat{\theta}_{t-1} = \frac{\sum_{j \in S_t} w_j y_{j,t-1}}{\sum_{j \in S_t} w_j} \quad \text{and} \quad \hat{\theta}_t = \frac{\sum_{j \in S_t} w_j y_{j,t}}{\sum_{j \in S_t} w_j}.$$

Let

$$\hat{u}_j = \frac{1}{\hat{\theta}_{t-1}} (y_{j,t} - \hat{R}_t y_{j,t-1}) \quad \text{and} \quad \hat{w} = \frac{1}{\tilde{n}} \sum_{\substack{j \in S_t \\ w_j \neq 1}} w_j, \quad \tilde{n} = \sum_{j \in S_t} I_{w_j \neq 1}. \quad (2.2.3)$$

Following (2.1.7), the estimator, in its general form, is

$$\hat{\hat{R}}_t = \hat{R}_t + \hat{f} \frac{1}{n} \sum_{j=1}^n \hat{u}_j + (1 - \hat{f}) \frac{\hat{E}_S \left[ (w_j - 1) \hat{u}_j \right]}{\hat{w} - 1}, \quad (2.2.4)$$

**Remark.** Until now we made no specific assumptions about the underlying distribution. In fact, this formula can be viewed as a pure identity: in the case where the expectation  $\hat{E}_S \left[ (w_j - 1) \hat{u}_j \right]$  is estimated as a sample arithmetic average, the usual WLR estimator is recovered.

The expectation does not have to be estimated as an arithmetic average. Similar to Zaslavsky *et al.* (2001), we combine  $\hat{u}_j$  and  $w_j$  into one variable  $\hat{u}_j^w = (w_j - 1) \hat{u}_j$ . Modeling assumptions over  $\hat{u}_j^w$  allow for a simple and simultaneous treatment of extreme survey weights, measurements, or their combined effect. Another possibility

is to model the weights as proposed by Beaumont (2008) (see also Pfeffermann and Sverchkov 2007).

In the next subsection we describe the Winsorization approach that is based on very weak modeling assumptions; thus it is suitable for estimation in larger samples.

### 2.3 On the choice of cutoff values for the Winsorized mean

Censoring of extreme sample measurements has been used in statistics for a long time. In this subsection, we discuss Winsorization, a method named after C.P. Winsor who was one of its first proponents back in the 1940's. The Winsorized mean is obtained by, first, replacing the sample measurements exceeding a certain cutoff point by a value closer or equal to the cutoff, and then taking the arithmetic mean of these modified sample values.

For symmetric distributions with long tails, the Winsorized mean is a good alternative to the estimator based on the original un-augmented data (Tukey and McLaughlin 1963). However, when the distribution is asymmetric, editing of the extreme values may lead to a biased estimator. The goal of Winsorization is to reduce the mean squared error of an estimator, while accepting some bias. For skewed distributions, Searls (1966) proved the existence of a region on the longer tail of a distribution with the property that the cutoff values chosen from this region lead to an estimator with a reduced mean squared error; for specific skewed distributions, the optimal cutoff points can be found via a simple algorithm (see an example in Searls 1966).

When the true underlying distribution is not assumed known, the algorithm proposed by Searls (1966) cannot be applied. A popular practical estimator is the “classical”

once-Winsorized mean. It is a particular form of Winsorization where the second extreme value is used as the cutoff point. However, when a distribution does not possess a “sufficiently long” tail, the once-Winsorized mean is less efficient than the original mean. Thus, it may be beneficial to test if the tails are “long enough” to warrant Winsorization. Fuller (1991) suggested that the right (or longer) tail of a sample distribution can be approximated by the right tail of a Weibull distribution. A test on the shape parameter of the Weibull provides an answer on the question about the advantages of Winsorization for a given sample. If the test suggests that the shape parameter is greater than one, then the once-Winsorized mean has a smaller mean squared error than the variance of the original sample mean. (It is also possible to consider other versions of Winsorizing cutoff points, depending on the result of the test.)

The above two paragraphs suggest an apparent tension between the ease of implementation and the difficulty in justification for implementation of the Winsorized mean. We would like to avoid the inconvenience of the latter. Instead of relying on assumptions about the form of a distribution or on results from testing, we assume “no more” than that the true mean is known. It could be argued that such an assumption amounts to tautology since we assert the knowledge of the parameter of interest. In reality, using a “guess” value that is “reasonably close” to the truth is a well established practice in statistics, the technique rooted in standard differential calculus.

To introduce the idea, consider a simple case of  $n$  independent observations  $u_1, \dots, u_n$ .

Assume  $E(u_j) = 0$ ,  $Var(u_j) < \infty$ .

Let  $u_j(K, L)$  denote the Winsorized value of  $u_j$ , such that

$$u_j(K, L) = u_j + (K - u_j)J_j + (L - u_j)I_j, \quad (2.3.1)$$

where

$$J_j = \begin{cases} 1, & u_j \geq K \\ 0, & u_j < K \end{cases} \text{ and } I_j = \begin{cases} 1, & u_j < L \\ 0, & u_j \geq L \end{cases}, \text{ for } j = 1, \dots, n.$$

Let  $\bar{u} = n^{-1} \sum_{j=1}^n u_j$  be the original mean and  $\bar{u}(K, L) = n^{-1} \sum_{j=1}^n u_j(K, L)$  the Winsorized mean.

**Result 1:** Let  $K$  and  $L$  satisfy, respectively, the equations

$$K + \sum_{j=1}^n (K - u_j)J_j = 0 \quad (2.3.2)$$

and

$$L + \sum_{j=1}^n (L - u_j)I_j = 0. \quad (2.3.3)$$

Then  $MSE[\bar{u}(K, L)] \leq Var(\bar{u})$ .

The proof of the above statement is given in Appendix A.

**Remark.** This result was inspired by the works of Searls (1966) and Kokic and Bell (1994). One distinction is that these papers are restricted to one-sided Winsorization,

in which the mean  $\bar{u}(K) = n^{-1} \sum_{j=1}^n u_j(K)$  is based on the Winsorized values

$u_j(K) = u_j + (K - u_j)J_j$ . Second, these papers are tasked with finding optimal values minimizing, rather than merely reducing, the mean squared error of an estimator. However, it is not possible to claim optimality without additional assumptions about the underlying distributions. For example, a solution to the equation

$$K + \left(1 - \frac{1}{n}\right) \sum_{j=1}^n E(K - u_j)J_j = 0$$

would provide the optimal value  $K$  for  $\bar{u}(K)$ . However, it includes the expectation taken at the tail of a distribution. To estimate the expectation, one approach is to assume a specific form of the distribution, as in Searls (1966). The alternative is to assume similarity of the current sample to samples from previous years of the same survey and estimate the expectation from the previous years, the approach of Kopic and Bell (1994). It is worth mentioning that the assertion of optimality for two-sided Winsorization would require more stringent assumptions. Result 1 does not state optimality, but it guarantees, without additional assumptions, reduction in the mean squared error.

### *2.3.1 Some simple illustrative examples*

In this subsection, we present simple simulation examples that show how the Result 1 works. The R code is provided in Appendix B. It can be easily modified to explore how the theory works with other distributions and with alternative initial values for the mean.

We considered four scenarios, for three sample sizes of 50, 100, or 500 observations:



(1) normal  $N(0,1)$  distribution;

(2) symmetrical mixture of two normal distributions with common mean  $\mu = 0$  and different variances: 97%  $N(0,1)$  and 3%  $N(0,10)$ ;

(3) asymmetrical mixture of two normal distributions with different means and variances: 97%  $N(0,1)$  and 3%  $N(3,10)$  (thus, the true mean is  $\mu = 0.03 * 3 = 0.09$ );

(4) lognormal distribution:  $\log(y_j) \stackrel{iid}{\sim} N(0,1)$  (thus, the true mean is  $\mu = \exp(0.5)$ ).

The density plots of the distributions are shown in Figure 3.

For the estimation of  $\mu$ , we used either the sample mean  $\bar{y} = n^{-1} \sum_{j=1}^n y_j$  or the

Winsorized mean  $\bar{y}(K, L) = n^{-1} \sum_{j=1}^n [\mu_0 + u_j(K, L)]$ , where  $u_j(K, L)$  is a

Winsorized value of  $u_j = y_j - \mu_0$ . For the “guess value”  $\mu_0$  for the mean  $\mu$ , we used (i) the true parameter or (ii) the sample average.

We used  $R = 5000$  simulation runs and computed bias, standard error, and root mean squared error as

$$Bias(\hat{\mu}) = 100 \frac{1}{R} \sum_{r=1}^R (\hat{\mu}^{(r)} - \mu),$$

$$SE(\hat{\mu}) = 100 \frac{1}{R-1} \sum_{r=1}^R (\hat{\mu}^{(r)} - \bar{\hat{\mu}})^2,$$

$$RMSE(\hat{\mu}) = 100 \sqrt{\frac{1}{R} \sum_{r=1}^R (\hat{\mu}^{(r)} - \mu)^2},$$

where  $\hat{\mu}^{(r)}$  is the r-th simulation run of the corresponding estimator (either  $\bar{y}$  or

$\bar{y}(K,L)$ ) and  $\bar{\hat{\mu}} = \frac{1}{R} \sum_{r=1}^R \hat{\mu}^{(r)}$ . Also reported are average values of the estimated

cutoff points:  $K = \frac{1}{R} \sum_{r=1}^R K^{(r)}$  and  $L = \frac{1}{R} \sum_{r=1}^R L^{(r)}$ .

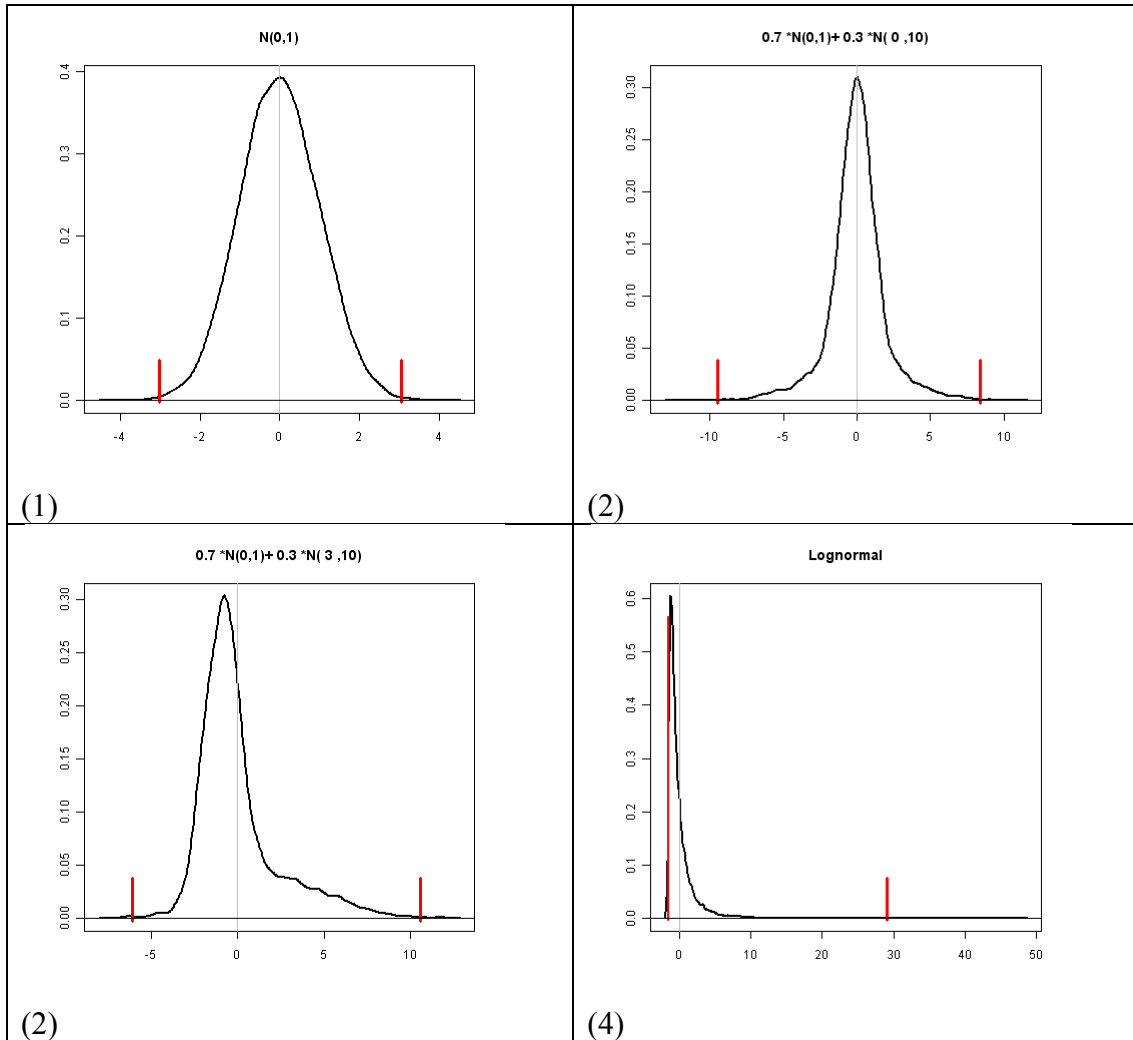


Figure 3. Density plots of (1) normal  $N(0,1)$ ; (2) symmetrical mixture 97%  $N(0,1)$  and 3%  $N(0,10)$ ; (3) asymmetrical mixture 97%  $N(0,1)$  and 3%  $N(3,10)$ ; (4) lognormal distribution.

Results are presented in Table 1- Table 4.

<i>True mean as initial guess</i>							
Sample size	Bias		SE		$\frac{RMSE[\bar{y}(K,L)]}{RMSE[\bar{y}]}$	Cutoff values	
	$\bar{y}$	$\bar{y}(K,L)$	$\bar{y}$	$\bar{y}(K,L)$		<i>L</i>	<i>K</i>
<b>50</b>	0.18	0.17	14.19	13.76	96.97	-1.46	1.46
<b>100</b>	0.15	0.15	9.97	9.81	98.43	-1.70	1.70
<b>500</b>	0.00	0.00	4.43	4.42	99.67	-2.22	2.22
<i>Estimated mean as initial guess</i>							
Sample size	Bias		SE		$\frac{RMSE[\bar{y}(K,L)]}{RMSE[\bar{y}]}$	Cutoff values	
	$\bar{y}$	$\bar{y}(K,L)$	$\bar{y}$	$\bar{y}(K,L)$		<i>L</i>	<i>K</i>
<b>50</b>	0.18	0.18	14.19	14.20	100.10	-1.46	1.46
<b>100</b>	0.15	0.15	9.97	9.97	100.06	-1.70	1.70
<b>500</b>	0.00	0.00	4.43	4.43	100.02	-2.22	2.22

Table 1. Normal  $N(0,1)$  distribution. Bias and Standard Error, in hundreds; RMSE ratio as percentage

<i>True mean as initial guess</i>							
Sample size	Bias		SE		$\frac{RMSE[\bar{y}(K,L)]}{RMSE[\bar{y}]}$	Cutoff values	
	$\bar{y}$	$\bar{y}(K,L)$	$\bar{y}$	$\bar{y}(K,L)$		<i>L</i>	<i>K</i>
<b>50</b>	0.28	0.27	15.86	14.91	94.05	-1.73	1.73
<b>100</b>	0.23	0.22	11.19	10.72	95.77	-2.12	2.13
<b>500</b>	-0.09	-0.08	5.07	4.98	98.29	-3.35	3.35
<i>Estimated mean as initial guess</i>							
Sample size	Bias		SE		$\frac{RMSE[\bar{y}(K,L)]}{RMSE[\bar{y}]}$	Cutoff values	
	$\bar{y}$	$\bar{y}(K,L)$	$\bar{y}$	$\bar{y}(K,L)$		<i>L</i>	<i>K</i>
<b>50</b>	0.28	0.28	15.86	15.38	96.98	-1.73	1.73
<b>100</b>	0.23	0.22	11.19	10.89	97.27	-2.12	2.13
<b>500</b>	-0.09	-0.08	5.07	5.00	98.59	-3.35	3.35

Table 2. Symmetrical mixture:  $0.97N(0,1) + 0.03N(0,10)$ . Bias and Standard Error, in hundreds; RMSE ratio as percentage

Normal distribution without contamination does not favor Winsorization. Yet, even in this case, according to the theory, the Winsorized mean has smaller RMSE when the true value is used as the initial guess. When the distribution is asymmetric, Winsorization causes a bias, still RMSE is reduced. There is some loss in efficiency if the guess value  $\mu_0$  is estimated from the same data; as a result, RMSE in the

estimated mean case is somewhat larger. In the non-normal cases considered in scenarios 2-4, Winsorization works well even with the estimated initial value for  $\mu_0$ .

<i>True mean as initial guess</i>							
Sample size	Bias		SE		$\frac{RMSE[\bar{y}(K,L)]}{RMSE[\bar{y}]}$	Cutoff values	
	$\bar{y}$	$\bar{y}(K,L)$	$\bar{y}$	$\bar{y}(K,L)$		L	K
<b>50</b>	0.37	-1.20	17.38	15.94	91.93	-1.56	2.35
<b>100</b>	0.32	-0.95	12.26	11.55	94.44	-1.83	3.10
<b>500</b>	-0.05	-0.63	5.56	5.46	98.94	-2.46	5.33
<i>Estimated mean as initial guess</i>							
Sample size	Bias		SE		$\frac{RMSE[\bar{y}(K,L)]}{RMSE[\bar{y}]}$	Cutoff values	
	$\bar{y}$	$\bar{y}(K,L)$	$\bar{y}$	$\bar{y}(K,L)$		L	K
<b>50</b>	0.37	-1.20	17.38	16.44	94.83	-1.57	2.35
<b>100</b>	0.32	-0.95	12.26	11.73	95.93	-1.83	3.10
<b>500</b>	-0.05	-0.63	5.56	5.48	99.26	-2.46	5.33

Table 3. Asymmetrical mixture:  $0.97N(0,1) + 0.03N(3,10)$ . Bias and Standard Error, in hundreds; RMSE ratio as percentage

<i>True mean as initial guess</i>							
Sample size	Bias		SE		$\frac{RMSE[\bar{y}(K,L)]}{RMSE[\bar{y}]}$	Cutoff values	
	$\bar{y}$	$\bar{y}(K,L)$	$\bar{y}$	$\bar{y}(K,L)$		L	K
<b>50</b>	0.34	-7.22	30.76	26.40	88.97	-1.25	4.99
<b>100</b>	0.33	-5.12	21.64	19.42	92.76	-1.33	6.74
<b>500</b>	0.08	-2.09	9.53	9.09	97.79	-1.45	12.28
<i>Estimated mean as initial guess</i>							
Sample size	Bias		SE		$\frac{RMSE[\bar{y}(K,L)]}{RMSE[\bar{y}]}$	Cutoff values	
	$\bar{y}$	$\bar{y}(K,L)$	$\bar{y}$	$\bar{y}(K,L)$		L	K
<b>50</b>	0.34	-7.22	30.76	27.33	91.90	-1.25	4.99
<b>100</b>	0.33	-5.12	21.64	19.76	94.29	-1.33	6.74
<b>500</b>	0.08	-2.09	9.53	9.12	98.12	-1.46	12.28

Table 4. Lognormal distribution,  $\log(y_j) \stackrel{iid}{\sim} N(0,1)$ . Bias and Standard Error, in hundreds; RMSE ratio as percentage

### 2.3.2 Accounting for survey design when choosing the cutoff points

We now consider the finite population setting and see how Result 1 can be applied in more complex situations.

In many surveys, the finite population measurements can be viewed as independent realizations from a superpopulation distribution. Then, according to Theorem 1 of Pfeffermann *et al.* (1998), under general regularity conditions, for many common sampling plans, the sample observations are asymptotically independent with respect to the sample distribution. (The asymptotic setup requires that the population size increases to infinity, while the sample size is fixed.) Thus, for many common situations, the assumptions of Result 1 hold.

**Example 1. Probability proportional to size (pps) sampling.** Suppose the finite population values  $y_j$ ,  $j=1, \dots, N$ , are generated as  $y_j = 100 + 5z_j + \varepsilon_j z_j^{1/2}$  for some vector  $\mathbf{z} = (z_1, \dots, z_N)$  (a “size” variable) and  $\varepsilon_j \stackrel{iid}{\sim} N(0, 9)$ . The target quantity is the

finite population mean  $\bar{Y} = N^{-1} \sum_{j=1}^N y_j$ . A sample of size  $n$  is selected using

probabilities proportional to size  $z_j$ . Let  $P\{j \in S \mid \mathbf{z}\} = \pi_j = \text{const} \cdot z_j$  and

$$E_U(\pi_j) = N^{-1}n.$$

Consider the Hájek estimator

$$\hat{\bar{y}} = \frac{\sum_{j \in S} w_j y_j}{\sum_{j \in S} w_j},$$

where  $w_j = \pi_j^{-1}$ .

The influence of an individual sample observation on the above estimator can be expressed as

$$u_j = \frac{1}{E_s(w_j)} w_j (y_j - \mu_y),$$

where  $\mu_y$  is the superpopulation mean of  $y_j$  and  $E_s(w_j) = [E_U(\pi_j)]^{-1} = n^{-1}N$ .

Thus,  $u_j = nN^{-1}w_j(y_j - \mu_y)$ .

The Winsorized estimator is  $\hat{y}(K, L) = \mu_y + \bar{u}(K, L)$ .

At each round of the simulation experiment presented here, values  $z_j$  were generated independently from the lognormal distribution with  $\log(z_j) \sim N(0, 1)$ ,  $j=1, \dots, N$ . In defining  $u_j$ , we used the true superpopulation value  $\mu_y = 100 + 5\mu_z$ , where  $\mu_z = \exp(0.5)$  is the mean of  $z_j$ . We considered two choices of the population and sample sizes: (1)  $N = 3000$ ,  $n = 30$  and (2)  $N = 10000$ ,  $n = 100$ . Table 5 displays results from  $R=5000$  simulation runs.

Bias was calculated as

$$Bias(\hat{y}) = 100 \left[ \frac{1}{R} \sum_{r=1}^R (\hat{y}^{(r)} - \bar{Y}^{(r)}) \right], \quad (2.3.4)$$

where  $\hat{y}$  denotes one of the estimators considered and the index  $r$  signifies the result of the  $r$ -th simulation run. The square root of the mean squared error (RMSE) is

$$RMSE = 100 \sqrt{Var(\hat{y}) + Bias(\hat{y})^2}, \quad (2.3.5)$$

where  $Var(\hat{y})$  is the variance over all simulation runs; the standard error is  $se(\hat{y}) = 100\sqrt{Var(\hat{y})}$ .

	<i>N=3000, n=30</i>			<i>N=10000, n=100</i>		
	<b>Bias</b>	<b>SE</b>	$\frac{RMSE[\hat{y}(K, L)]}{RMSE[\hat{y}]}$	<b>Bias</b>	<b>SE</b>	$\frac{RMSE[\hat{y}(K, L)]}{RMSE[\hat{y}]}$
$\hat{y}$	46.8	211.7	-	14.6	114.1	-
$\hat{y}(K, L)$	48.1	177.5	84.7	28.3	103.5	93.3

Table 5. PPS sampling, Bias and Standard Error, in hundreds; RMSE ratio as percentage; 5000 simulation runs

In this example, the Winsorized estimator performs better than the estimator based on the original data.

Next, we discuss the case of stratified sampling. If stratification is not properly accounted for and the conditions of Result 1 are not satisfied, the Winsorized estimator may perform poorly. When a population is deliberately divided into separate strata based on the information related to the variable of interest, the values  $u_1, \dots, u_n$  are to be obtained by subtracting corresponding strata means from the original sample values. The following simulation example demonstrates this point.

**Example 2 Stratified simple random sampling (STSRs).** Suppose the finite population measurements are independent realizations from a mixture of two normal

distributions  $y_j \stackrel{ind}{\sim} 0.7N(0,1) + 0.3N(12,9)$ . The goal is to estimate the finite

population mean  $\bar{Y} = N^{-1} \sum_{j=1}^N y_j$ . The population is divided into two strata

corresponding to the parts of the mixture. A sample of size n is drawn using a stratified simple random sampling design with equal probabilities of selection.

Consider two possibilities for forming the  $u$ -variables: subtract the common mean  $\mu$  from each observation,  $u_j^{(c)} = y_j - \mu$  or subtract separate strata means,  $u_j^{(s)} = y_j - \mu_h$ ,  $j \in h$ , for strata  $h = 1, 2$ . The variables  $u_j^{(c)}$  do not have mean zero, unless the strata means are equal.

The estimator based on the original data is  $\bar{y} = n^{-1} \sum_{j \in S} y_j$ . The Winsorized mean is

$$\bar{y}(K, L) = n^{-1} \sum_{j \in S} y_j(K, L),$$

for a choice of cutoffs  $(K, L)$ . Denote the cutoffs  $(K^{(c)}, L^{(c)})$  or  $(K^{(s)}, L^{(s)})$ , depending on the way of constructing  $u_j$ . Correspondingly,  $y_j(K^{(c)}, L^{(c)}) = u_j^{(c)}(K^{(c)}, L^{(c)}) + \mu$  or  $y_j(K^{(s)}, L^{(s)}) = u_j^{(s)}(K^{(s)}, L^{(s)}) + \mu_h$ ,  $j \in h$ ,  $h = 1, 2$ .

We used  $R = 5000$  simulation runs for each of the two choices of the population and sample sizes: (1)  $N = 3000$ ,  $n = 30$  and (2)  $N = 10000$ ,  $n = 100$ ; the strata sizes are  $N_1 = 0.7N$  and  $N_2 = 0.3N$ . (For this simulation example, we use the true superpopulation values of the parameters  $\mu$ ,  $\mu_1$ , and  $\mu_2$  when forming the  $u$ -values. The strata means are  $\mu_1 = 0$ ,  $\mu_2 = 12$  and the overall mean is  $\mu = 0.7\mu_1 + 0.3\mu_2 = 3.6$ . In reality, these values need to be “guessed” or estimated from the data.)

The simulation results are shown in Table 6. The bias and RMSE were calculated using formulas (2.3.4) and (2.3.5).



	<i>N=3000, n=30</i>			<i>N=10000, n=100</i>		
	<b>Bias</b>	<b>SE</b>	$\frac{\text{RMSE}}{\text{RMSE}(\bar{y})}$	<b>Bias</b>	<b>SE</b>	$\frac{\text{RMSE}}{\text{RMSE}(\bar{y})}$
$\bar{y}$	-0.3	33.5	-	0.0	18.6	-
$\bar{y}(K^{(c)}, L^{(c)})$	-16.6	30.8	104.5	-6.2	18.2	103.1
$\bar{y}(K^{(s)}, L^{(s)})$	-0.3	30.1	90.0	-0.1	18.0	96.3

Table 6. Population  $0.7N(0,1)+0.3N(12,9)$ , 5000 simulation runs, (in hundreds)

The results demonstrate that subtracting the overall mean is not the proper way to form the u-values, because the values defined in such a way are not uncorrelated under the stratified sampling design. Bias and RMSE of the resulting estimator  $\bar{y}(K^{(c)}, L^{(c)})$  are high compared to the original estimator  $\bar{y}$ . On the other hand,  $\bar{y}(K^{(s)}, L^{(s)})$  is clearly an improvement over  $\bar{y}$ .

### 2.3.3 Using information not included in the sampling design

Although in many common situations the Winsorized mean of Result 1 is better than the original mean in terms of the mean squared error, it is important to bear in mind the possibility of a bias that may incur due to Winsorization. Bias may accumulate if several biased estimates are aggregated to obtain a higher level estimate. In CES, bias also may build up over several months of estimation. It is desirable to avoid or reduce the bias of the Winsorized mean.

Large samples usually consist of a mixture of more homogeneous parts. To reduce the bias, it may be useful to incorporate available information that can explain the complexity of the observed sample distribution. For example, the bias may be

reduced if subpopulation means are subtracted from the sample measurements. The following example is designed to illustrate this situation.

**Example 3. Simple random sampling.** Similar to Example 2, the finite population measurements come from a mixture of two normal distributions. However, the mixture parts are not as clearly separated, in terms of the means of the mixture parts:

$y_j \stackrel{ind}{\sim} 0.7N(0,1) + 0.3N(4,9)$ . The sample is selected using simple random sampling

with replacement and the mixture parts represent poststrata. The estimator of the

population mean, based on the original data, is  $\bar{y} = n^{-1} \sum_{j=1}^n y_j$ .

Similar to Example 2, consider two versions of Winsorization. First, form the  $u$ -

values by subtracting the common mean  $\mu$  from each observation,  $u_j^{(c)} = y_j - \mu$ ; in

the second version, form the  $u$ -values by subtracting separate poststrata means,

$u_j^{(s)} = y_j - \mu_h$ ,  $j \in h$ , for poststrata  $h = 1, 2$ . In the case of simple random sampling,

both sets of  $u$ -values contain independent observations. Thus, in each case the mean

squared error of the Winsorized mean is expected to be lower than the variance of the

original mean.

The simulation results are presented in Table 7 for two choices of the population and

sample sizes: (1)  $N = 3000$ ,  $n = 30$  and (2)  $N = 10000$ ,  $n = 100$ ; the subpopulation

sizes are  $N_1 = 0.7N$  and  $N_2 = 0.3N$ . There were  $R = 5000$  simulation runs.

The bias and root mean squared error were calculated using formulas (2.3.4) and

(2.3.5).

	<i>N=3000, n=30</i>			<i>N=10000, n=100</i>		
	<b>Bias</b>	<b>SE</b>	$\frac{\text{RMSE}}{\text{RMSE}(\bar{y})}$	<b>Bias</b>	<b>SE</b>	$\frac{\text{RMSE}}{\text{RMSE}(\bar{y})}$
$\bar{y}$	-0.3	47.8	-	-0.1	26.1	-
$\bar{y}(\mathbf{K}^{(c)}, \mathbf{L}^{(c)})$	-7.8	44.9	95.4	-3.4	25.6	98.9
$\bar{y}(\mathbf{K}^{(s)}, \mathbf{L}^{(s)})$	-0.3	45.5	95.3	-0.1	25.6	98.0

Table 7 .  $0.7N(0,1) + 0.3N(4,9)$ , 5000 simulation runs, (in hundreds)

Taking into account the subpopulation means reduces the bias. However, it does not necessarily lead to decreased RMSE.

Another way to reduce bias is to use Winsorization only when the benefits are evident. For example, suppose certain critical bounds for the estimate can be established based on the previous years of the same survey. Then Winsorization can be used only when the original estimate does not conform to the bounds. This approach has proved to be useful for the CES estimates.

## 2.4 Simulation study

The simulation study shows performances of several estimators under different scenarios. Winsorization may not be the most efficient estimator, yet it is safer to use than some model-based alternatives.

A stratified simple random sample is selected from  $H=4$  strata of a finite population  $P$ , with the differential selection probabilities across strata.

The goal is to estimate the finite population mean  $\bar{Y} = \frac{1}{N} \sum_{h=1}^H \sum_{j=1}^{n_h} y_{jh}$ , where  $y_{jh}$  is the

value of unit  $j$  from stratum  $h$ ;  $N$  is the total number of units in the population.

We generate the finite population using the following procedure. At the first step, the strata means  $m_h$  are generated from the normal distribution with expectation 0 and standard deviation 30. At the second step, we generate  $y_{jh}$  values for each stratum  $h$ .

The population and sample sizes for each stratum, the superpopulation means, and the sampling weights are given in Table 8.

<i>Stratum, h</i>	<i>Population size, <math>N_h</math></i>	<i>Sample size, <math>n_h</math></i>	<i>Superpopulation mean, <math>m_h</math></i>	<i>Sample weight, <math>w_h = N_h/n_h</math></i>
1	15000	150	-17.03	100.00
2	5000	150	-24.44	33.33
3	1500	500	-14.82	3.00
4	500	400	0.05	1.25

Table 8. Description of the simulation

Consider several possibilities:

- 1) “Best Case (BC) scenario”: the population values  $y_{jh}$  come from the normal distribution with mean  $m_h$  and standard deviation  $\sigma_h$ , and the sample inclusion probabilities,  $\pi_h$ , are such that  $\sigma_h = 150\pi_h = 150/w_h$ :

$$y_{jh} \sim N(m_h, \sigma_h^2), h = 1, \dots, H$$

- 2) “Stratum Jumpers (SJ) scenario”: suppose some units change their stratum after the sample has been selected. To simulate this situation, the population values are generated exactly as under the BC scenario, however, a small fraction (less than 0.1%) of the units’ values are generated as if the units belonged to a “foreign” stratum, as follows:

- for units in strata 1 and 2, 0.1 per cent of the units are generated from the distribution of the strata 3 and 4, respectively:

$$0.1\% \text{ with } y_{jh} \sim N(m_{h+2}, \sigma_{h+2}^2), h = 1, 2$$

- for units in stratum 3, 0.05 per cent of the units are generated from the distribution of the stratum 1 and another 0.05 per cent of the units are generated from the distribution of the stratum 4:

$$0.05\% \text{ with } y_{j3} \sim N(m_2, \sigma_2^2) \text{ and } 0.05\% \text{ with } y_{j3} \sim N(m_4, \sigma_4^2)$$

- for units in stratum 4: 0.05 per cent of the units are generated from the distribution of the stratum 3.

$$0.05\% \text{ with } y_{j4} \sim N(m_3, \sigma_3^2)$$

- 3) “Spike at Center (SC) scenario”: in each stratum, 90 per cent of the data are generated from the normal distribution with the standard deviation that is significantly (100 times) smaller than the other 10 per cent:

$$90\% \text{ with } y_{jh} \sim N(m_h, (\sigma_h/10)^2), 10\% \text{ with } y_{jh} \sim N(m_h, (10\sigma_h)^2),$$

$$h = 1, \dots, H$$

- 4) “Spike and Shift (SH) scenario”: in each stratum, 90 per cent of the data are generated from the normal distribution with the standard deviation that is significantly (100 times) smaller than the other 10 per cent and the mean is shifted:

90% with  $y_{jh} \sim N\left(m_h, (\sigma_h/10)^2\right)$ , 10% with  $y_{jh} \sim N\left(m_h + 10, (10\sigma_h)^2\right)$ ,

$h = 1, \dots, H$

From each of the four populations, we selected 300 random samples using a stratified simple random sampling design, with probabilities  $\pi_h = 1/w_h$ ,  $h=1, \dots, 4$ ,  $w_h$  as in Table 8. From each sample, we calculated estimates based on the following four types of estimators: (1) Horvitz-Thompson (HT) estimator,  $\hat{Y}_{HT}$ ; (2) Exchangeable random effects (WN1) for weighted residuals model,  $\hat{Y}_{WN1}$ ; (3) Scale mixture of two normal distributions (WN2F) for weighted residuals,  $\hat{Y}_{WN2F}$ ; (4) Winsorization cutoffs estimator,  $\hat{Y}_{Wz}$ .

Specification details of the estimators follow.

(1) The formula for the HT estimator is

$$\hat{Y}_{HT} = \frac{1}{N} \sum_{h=1}^H \frac{N_h}{n_h} \sum_{j=1}^{n_h} y_{jh},$$

where  $N = \sum_{h=1}^H N_h$ .

Next, denote  $e_{jh}^w = (w_h - 1)e_{jh}$ , where  $e_{jh} = y_{jh} - \hat{Y}_{HT}$ .

For WN1 and WN2F cases, the estimator has the form

$$\hat{Y} = \hat{Y}_{HT} + \frac{n}{N} \bar{e} + \frac{1}{N} \sum_{h=1}^H n_h \hat{\mu}_h^{we},$$

where  $\hat{\mu}_h^{we}$  is derived from a model.

(2) The WN1 model is formulated as a two level model by the following statements:

$$\text{Level 1: } e_{jh}^w \stackrel{ind}{\sim} N(\mu_h^{we}, \sigma^2), \quad (2.4.1)$$

$$\text{Level 2: } \mu_h^{we} \stackrel{iid}{\sim} N(\mu^{we}, \tau^2), \quad (2.4.2)$$

$$h = 1, \dots, H$$

(3) The WN2F model is described as follows:

$$e_{jh}^w | z_{kj} = 1 \stackrel{ind}{\sim} N(\mu_h^{we}, \sigma_k^2), \quad (2.4.3)$$

where  $k = 1, 2; j = 1, \dots, n_h; h = 1, \dots, H; \sum_{h=1}^H n_h = n;$

$z_{kj}$  is a mixture class indicator for an observation  $j$  and class  $k = 1, 2;$

$\sigma_k^2$  is a variance parameter of the  $k^{th}$  component of the mixture.

We applied the EM algorithm described in Appendix C to fit the WN2F model.

(4) The Winsorization estimator description.

Suppose  $\sum_{j=1}^n w_j = N$ . Note that the HT estimator equivalently can be written

as

$$\hat{Y} = \hat{Y}_{HT} + \mu + f\bar{u} + (1-f)\bar{u}^w, \quad (2.4.4)$$

where  $\mu = (1-f)n^{-1} \sum_{j=1}^n \frac{(w_j-1)\mu_j}{\bar{w}-1}$ ,  $\mu_j = \hat{E}[u_j | h] = n_h^{-1} \sum_{j=1}^{n_h} u_j$ ,  $u_j = y_j - \hat{Y}_{HT}$ ;

$$\bar{u}^w = n^{-1} \sum_{j=1}^n u_j^w, \quad u_j^w = \frac{(w_j-1)(u_j - \mu_j)}{\bar{w}-1}, \quad \bar{u} = n^{-1} \sum_{j=1}^n u_j, \quad \bar{w} = n^{-1} \sum_{j=1}^n w_j = f^{-1}.$$

Let the adjusted value for  $u_j^w$  be

$$u_j^w(K, L) = u_j^w + (K - u_j^w)J_j + (L - u_j^w)I_j \quad (2.4.5)$$

where  $J_j = 1$  if  $u_j^w \geq K$  and  $J_j = 0$ , otherwise;  $I_j = 1$  if  $u_j^w < L$  and  $I_j = 0$ , otherwise.

Then the Winsorized estimator is defined by

$$\hat{Y}^{adj} = \hat{Y}_{HT} + \mu + f\bar{u} + (1-f)\bar{u}^w(K, L), \quad (2.4.6)$$

where  $\bar{u}^w(K, L) = n^{-1} \sum_{j=1}^n u_j^w(K, L)$  and the values for  $K$  and  $L$  are obtained by

solving the equations (2.3.2) and (2.3.3), as outlined in Kocic and Bell (1994).

To evaluate each estimator, we calculated the empirical bias and root mean squared errors:

$$Bias = \frac{1}{300} \sum_{r=1}^{300} (\hat{Y}_r - \bar{Y}) \quad \text{and} \quad RMSE = \sqrt{\frac{1}{300} \sum_{r=1}^{300} (\hat{Y}_r - \bar{Y})^2},$$

where  $\hat{Y}_r$  are estimates derived from sample  $r$ ,  $r = 1, \dots, 300$ .

The results for each of the three types of the finite population are reported in Table 9.



<b>“Best Case” Population</b>				
	<b>Bias</b>	<b>SE</b>	<b>RMSE</b>	$\frac{RMSE}{RMSE_{HT}}$
<b>HT</b>	-0.01	0.18	0.18	-
<b>WN1</b>	-0.01	0.18	0.18	1.00
<b>WN2F</b>	-0.07	0.22	0.23	1.25
<b>Wz</b>	0.01	0.18	0.18	0.99
<b>“Stratum Jumpers” Population</b>				
	<b>Bias</b>	<b>SE</b>	<b>RMSE</b>	$\frac{RMSE}{RMSE_{HT}}$
<b>HT</b>	0.00	0.42	0.42	-
<b>WN1</b>	0.00	0.42	0.42	1.00
<b>WN2F</b>	-0.03	0.20	0.20	0.47
<b>Wz</b>	0.01	0.32	0.32	0.76
<b>“Spike at Center” Population</b>				
	<b>Bias</b>	<b>SE</b>	<b>RMSE</b>	$\frac{RMSE}{RMSE_{HT}}$
<b>HT</b>	0.01	0.58	0.58	-
<b>WN1</b>	0.02	0.59	0.59	1.01
<b>WN2F</b>	0.04	0.20	0.20	0.35
<b>Wz</b>	0.02	0.57	0.57	0.98
<b>“Spike and Shift” Population</b>				
	<b>Bias</b>	<b>SE</b>	<b>RMSE</b>	$\frac{RMSE}{RMSE_{HT}}$
<b>HT</b>	0.01	0.61	0.61	-
<b>WN1</b>	0.02	0.62	0.62	1.01
<b>WN2F</b>	-0.90	0.20	0.93	1.52
<b>Wz</b>	-0.01	0.60	0.60	0.98

Table 9. Bias and standard errors of estimators for the three finite populations. The last column is the RMSE ratio to the baseline RMSE of the HT estimator.

The performances depend on the underlying distribution in the finite population. HT and WN1 are very similar for any of the scenarios. The Winsorization (Wz) estimator is conservative in that the model assumptions are very weak. It performs slightly better than HT or WN1 under any scenario. Under SJ or SC, Wz does not provide as much gain in efficiency compared to WN2F. Wz is more efficient than HT or WN1

and it is safer to use than WN2F, in case the model does not hold. For example, under BC or SH, WN2F is not as good as the other estimators.

## 2.5 Mean squared error estimation using the bootstrap

### (simulation study)

In this simulation study, we used the finite population generated from the superpopulation model under the four scenarios BC, SJ, SC, and SH (see description in the previous subsection) to obtain mean squared error estimates for the Horvitz-Thompson (HT) and Winzorization (Wz) estimators of the mean.

The sample fractions in different strata vary from negligible to fairly large. It is desirable to account for non-negligible sample fraction in estimation. Gross (1980) proposed a variant of bootstrap known as the without-replacement bootstrap (BWO). A generalization of the procedure was proposed in Sverchkov and Pfeffermann (2004). Following Sverchkov and Pfeffermann (2004), we assume that the sample observations are uncorrelated and are independent from the sample-complement part of the universe. Detailed description of the bootstrap procedure follows.

Independently from each stratum  $h$ , select a pseudo-population of size  $N_h$  out of  $n_h$  sample units, using a simple random sampling with replacement (SRSWR) procedure. Select  $B=500$  stratified simple random samples using the same sampling design as used for the original sample. Derive the bootstrap estimates by following the estimation steps as in the original sample. Use standard bootstrap formula to compute RMSE.

To assess biases and variances of these MSE estimates, we used Monte Carlo simulation. We simulated 300 different “original samples” from the finite populations with fixed superpopulation parameters and repeated each bootstrap procedure for these 300 different original samples. Thus, we obtained 300 estimates of MSE for each estimator. In Table 10, we show average of these 300 estimates, with the simulation standard error in parentheses.

	<i>True MSE</i>	<i>BWO MSE</i>
<i>“Best Case” Population</i>		
<b>HT</b>	0.034	0.033 (0.002)
<b>Wz</b>	0.033	0.032 (0.002)
<i>“Stratum Jumpers” Population</i>		
<b>HT</b>	0.178	0.167 (0.143)
<b>Wz</b>	0.104	0.122 (0.106)
<i>“Spike at Center” Population</i>		
<b>HT</b>	0.338	0.334 (0.064)
<b>Wz</b>	0.322	0.316 (0.061)
<i>“Spike and Shift” Population</i>		
<b>HT</b>	0.373	0.362 (0.070)
<b>Wz</b>	0.359	0.346 (0.068)

Table 10. True MSE based on 300 samples from a finite population and estimated BWO MSE averages and standard errors (in parentheses) based on 300 estimates of MSE, each derived from 500 bootstrap iterations.

## Summary

In this Chapter, we proposed a method of identifying influential observations when the target population quantity is a function of the finite population given in a predefined form. The first step is to linearize the target and to obtain the influence function. This reduces the problem to estimation of the mean of the influence function.

The second step is to find the prediction of the new linear target. This is done using the relationship between the sample and population distribution to account for the informativeness of the sampling design.

The efficiency of the estimator of the mean of the influence function can be improved by using Winsorization. We proved a general result that under mild conditions certain cutoff points guarantee that the mean squared error of the Winsorized estimator is smaller than the variance of the estimator based on the un-augmented data.

We demonstrated the effect of Winsorization using several simulation examples. The conclusion is that Winsorization provides modest improvement to an estimator. Stronger model assumptions may give much better results. However, they may also lead to disastrous results if the model assumptions do not hold. Winsorization is safe to use in most cases. However, the estimator will be biased if the underlying distribution is not symmetric. The bias may accumulate when estimates are aggregated to a higher level or over time (as in the CES series). In such a case, it is advisable to use Winsorization sparingly, only when the improvement is evident. The evaluation of the need for Winsorization can be based on the historical information.

## Chapter 3: Robust Small Area Estimation

Linear mixed models have proved to be very useful in small area estimation problems. In this chapter, we consider a slight modification of the classical model. In order to accommodate the possibility of outlying observations, we make the assumption that the underlying distribution of the sample measurements is a scale mixture of two normal distributions, where outliers come from a distribution with larger variance than “regular” observations.

It was perhaps Newcomb (1886) who first proposed using mixtures of normal distributions to “obtain the best result” since “the cases are quite exceptional in which the errors are found to really follow the law” (by “the law” was meant the normal distribution). Tukey (1960) used the scale mixture of two normal distributions to demonstrate the effect that a small fraction of contamination may have on the resulting estimates, and a mixture model of this type is commonly cited as Tukey’s gross error model. Huber (1981) used the gross error model example in the beginning of his book to motivate the development of estimation methods resistant to deviations from distributional assumptions.

It turns out that modeling the errors using a scale mixture distribution may be useful even when the fraction of units with larger variance is not small. In other words, the units with larger variance are not necessarily “outliers” but valid members of a distinct part of the population.

Mixture distributions are usually considered for the case of independent observations. However, observations in small area estimation problems are assumed to be

correlated within areas. The model developed in this chapter accounts for this more complicated data structure.

This Chapter is organized as follows. The model is formulated in Section 3.1. In Section 3.2 we discuss identifiability of the model parameters. The maximum likelihood estimates of the model parameters can be obtained using the EM iterative algorithm. The algorithm is described in Section 3.3. Parametric bootstrap can be used to construct prediction confidence intervals; the bootstrap algorithm is given in Section 3.4. Bias correction approaches are discussed in Section 3.5. Numerical comparison of several robust estimators considered in the literature is given in Section 3.6. In Section 3.7, we present a small simulation study that is aimed to explore how the mean squared error of the estimated model parameters, derived using the proposed EM algorithm, decreases with the increased sample size. Evaluation of the bootstrap performance in terms of the percent coverage and length of the confidence intervals is in Section 3.8. Finally, in Section 3.9 we consider application of the approach to the CES survey data. This section also includes application of the linearization technique, discussed in Section 2.1 of Chapter 2, to small area settings.

### 3.1 The proposed model

Consider a modification of the nested-error regression model, where the error terms come from mixture of two normal distributions (thus, the model is named N2) with common, zero, mean and different variances. The model is given by (3.1.1)-(3.1.3) below:

$$y_{mj} = \mathbf{x}_{mj}^T \boldsymbol{\beta} + u_m + \varepsilon_{mj}, \quad (3.1.1)$$

$$u_m \stackrel{iid}{\sim} N(0, \tau^2) \text{ and } \varepsilon_{mj} | z_{mj} \stackrel{iid}{\sim} (1 - z_{mj})N(0, \sigma_1^2) + z_{mj}N(0, \sigma_2^2), \quad (3.1.2)$$

$$j = 1, \dots, n_m, m = 1, \dots, M,$$

and the mixture part indicator is a binomial variable

$$z_{mj} | p \stackrel{iid}{\sim} \text{Bin}(1; p), \quad (3.1.3)$$

where

$p$  is the probability of belonging to mixture part 2 (where  $\sigma_2 \geq \sigma_1$ ).

Note that, conditional on the values of the mixture part indicators  $z_{mj}$ , the model is the usual mixed effects model. Alternatively, we can write the distribution function of the random vector  $\mathbf{y}$  in model (3.1.1)-(3.1.3) as a mixture of  $K$  multivariate normal distributions, as follows:

$$h(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}) = \sum_{k=1}^K \lambda_k f_k(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}_k), \quad (3.1.4)$$

where

$$f_k(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}_k) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}_k|^{1/2}} e^{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})},$$

$\sum_{k=1}^K \lambda_k = 1$ ,  $\boldsymbol{\Sigma}_k = \mathbf{R}_k + \mathbf{Z}\mathbf{D}\mathbf{Z}^T$ ,  $\mathbf{D} = \text{diag}(\tau^2)$ ,  $\mathbf{Z}$  is the  $n \times M$  design matrix for the

random effects; each diagonal matrix  $\mathbf{R}_k$  has entries  $\sigma_1^2$  and  $\sigma_2^2$  in a specific to a given component  $k$  order. Assume that the variance components  $\tau^2, \sigma_1^2, \sigma_2^2$  are

strongly positive and that there is a positive number of observations in at least one area.

The set of parameters  $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K\} \subset \Omega$ , where  $\Omega$  denotes the parameter space;  $\boldsymbol{\theta}_k$  denotes the set of parameters  $p$ ,  $\boldsymbol{\beta}$ , and the variance-covariance matrix  $\boldsymbol{\Sigma}_k$  that depends on the variance components,  $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_k(\tau^2, \sigma_1^2, \sigma_2^2)$ ;  $p$  is the probability of appearance of  $\sigma_2^2$  in the diagonal of  $\mathbf{R}_k$ ,  $0 < p < 1$ .

In the case where  $\sigma_1^2 = \sigma_2^2$ ,  $K = 1$  and we obtain the usual case of a mixed effects model. If  $\sigma_1^2 \neq \sigma_2^2$ , there are  $K = 2^n$  distinct matrices  $\boldsymbol{\Sigma}_k$ . Suppose the diagonal terms of a matrix  $\mathbf{R}_k$  contain  $n - l_k$  values  $\sigma_1^2$  and  $l_k$  values  $\sigma_2^2$ ,  $0 \leq l_k \leq n$ ;

$l_k = \sum_{j=1}^n z_{kj}$ , where  $z_{kj}$  is an indicator variable:  $z_{kj} = 1$ , when an observation  $j$  comes

from the distribution with  $\sigma_2^2$  value for the random error variance, and  $z_{kj} = 0$ ,

otherwise;  $j = 1, \dots, n$ . The probability that a random vector  $\mathbf{y}$  belongs to the  $k$ -th

mixture part is  $\lambda_k = p^{l_k} (1 - p)^{n - l_k}$ .

To understand the setup, let us first consider a hypothetical situation when all the model parameters are known.

Assuming the parameters are known, the model N2 predictor of the non-sampled part of the population mean in area  $m$  is given by

$$\bar{Y}_{mr}^{N2} = \bar{\mathbf{x}}_{mr}^T \boldsymbol{\beta} + u_m^{N2}, \quad (3.1.5)$$



where  $u_m^{N2}$  is a predictor of area  $m$  random effect and  $\bar{\mathbf{x}}_m^T = (N_m - n_m)^{-1} \sum_{j=n_m+1}^{N_m} \mathbf{x}_{mj}^T$  is

the mean of auxiliary variables over the non-sampled part of the population in area  $m$ .

If indicators  $z_{mj}$  were observed, the predictor for the random effect would be

$$E(u_m | \mathbf{y}_m, \mathbf{z}_m) = \frac{\tau^2}{\sigma_m^2 + \tau^2} (\bar{y}_m - \bar{\mathbf{x}}_m^T \boldsymbol{\beta}) \quad (3.1.6)$$

where

$$\sigma_m^2 = \left( \sum_{j=1}^{n_m} \sigma_{mj}^{-2} \right)^{-1}, \quad (3.1.7)$$

$$\sigma_{mj}^{-2} = (1 - z_{mj}) \sigma_1^{-2} + z_{mj} \sigma_2^{-2}, \quad (3.1.8)$$

$$\bar{y}_m = \sigma_m^2 \sum_{j=1}^{n_m} \sigma_{mj}^{-2} y_{mj}, \quad (3.1.9)$$

$$\bar{\mathbf{x}}_m = \sigma_m^2 \sum_{j=1}^{n_m} \sigma_{mj}^{-2} \mathbf{x}_{mj}, \quad (3.1.10)$$

Since indicators  $z_{mj}$  are not observed, the predictor is

$$u_m^{N2} = E \left[ \frac{\tau^2}{\sigma_m^2 + \tau^2} (\bar{y}_m - \bar{\mathbf{x}}_m^T \boldsymbol{\beta}) | \mathbf{y}_m \right], \quad (3.1.11)$$

where the expectation is taken over the conditional distribution of  $\mathbf{z}_m$  given  $\mathbf{y}_m$ .

Next, consider the variance of the predictor  $u_m^{N2}$ . For a given set of indicators, the variance is

$$\text{Var}(u_m | \mathbf{y}_m, \mathbf{z}_m) = \frac{\tau^2 \sigma_m^2}{\sigma_m^2 + \tau^2}. \quad (3.1.12)$$

Therefore, the total variance of  $u_m^{N2}$  is

$$\begin{aligned} v_m^2 &= \text{Var}(u_m | \mathbf{y}_m) = E(\text{Var}[u_m | \mathbf{y}_m, \mathbf{z}_m] | \mathbf{y}_m) + \text{Var}(E[u_m | \mathbf{y}_m, \mathbf{z}_m] | \mathbf{y}_m) \\ &= E\left[\frac{\tau^2 \sigma_m^2}{\sigma_m^2 + \tau^2} | \mathbf{y}_m\right] + \text{Var}\left(\frac{\tau^2}{\sigma_m^2 + \tau^2} (\bar{y}_m - \bar{\mathbf{x}}_m^T \boldsymbol{\beta}) | \mathbf{y}_m\right). \end{aligned} \quad (3.1.13)$$

The variance and expectation in the right hand side of (3.1.13) are taken over the conditional distribution of  $\mathbf{z}_m$  given  $\mathbf{y}_m$ .

Let us now discuss the formula for the conditional probability of an observation ( $m_j$ ) belonging to part 2 of the mixture. The probability is

$$p_{mj} = P\{z_{mj} = 1 | y_{mj}\} = \frac{p \varphi\left(\frac{y_{mj} - \mathbf{x}_{mj}^T \boldsymbol{\beta}}{\sqrt{\sigma_2^2 + \tau^2}}\right)}{(1-p) \varphi\left(\frac{y_{mj} - \mathbf{x}_{mj}^T \boldsymbol{\beta}}{\sqrt{\sigma_1^2 + \tau^2}}\right) + p \varphi\left(\frac{y_{mj} - \mathbf{x}_{mj}^T \boldsymbol{\beta}}{\sqrt{\sigma_2^2 + \tau^2}}\right)}, \quad (3.1.14)$$

where  $\varphi(\cdot)$  is the standard normal pdf. Suppose there is a fraction of extreme outliers

in the data. The absolute value of  $\frac{y_{mj} - \mathbf{x}_{mj}^T \boldsymbol{\beta}}{\sqrt{\sigma_1^2 + \tau^2}}$  of an outlier is large and the probability

$\varphi\left(\frac{y_{mj} - \mathbf{x}_{mj}^T \boldsymbol{\beta}}{\sqrt{\sigma_1^2 + \tau^2}}\right)$  is small. If  $\varphi\left(\frac{y_{mj} - \mathbf{x}_{mj}^T \boldsymbol{\beta}}{\sqrt{\sigma_1^2 + \tau^2}}\right)$  tends to zero, the value of  $p_{mj}$  for such unit

tends to 1. In such a case, the expected value for the inverse variance  $\sigma_{mj}^{-2}$  will be

close to  $\sigma_2^{-2}$ . Each unit has its conditional probability  $p_{mj}$  and its individual expected

variance, depending on relative distances from a common mean.

Next, consider a situation where the value of the parameter vector  $\boldsymbol{\beta}$  is unknown but all the other parameters are known. For a given set  $k$  of indicators, we have a usual linear mixed model, a component  $f_k(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}_k)$  in representation (3.1.4). Then the maximum likelihood estimator for  $\boldsymbol{\beta}$  is a solution to the estimating equations

$$\mathbf{X}^T \boldsymbol{\Sigma}_k^{-1} \mathbf{y} - \mathbf{X}^T \boldsymbol{\Sigma}_k^{-1} \mathbf{X} \boldsymbol{\beta} = 0. \quad (3.1.15)$$

Since the set of indicators is not known, the estimator for  $\boldsymbol{\beta}$  is a solution to the expectation of the expression (3.1.15) over the values of indicators. It is given by

$$\hat{\boldsymbol{\beta}} = \left( \mathbf{X}^T E[\boldsymbol{\Sigma}_k^{-1} | \mathbf{y}] \mathbf{X} \right)^{-1} \mathbf{X}^T E[\boldsymbol{\Sigma}_k^{-1} | \mathbf{y}] \mathbf{y}. \quad (3.1.16)$$

See Appendix D.

Note that  $\hat{\boldsymbol{\beta}}$  is not a linear estimator on  $\mathbf{y}$ . Correspondingly, the predictor for the random effects is not a linear predictor (it is still the best predictor (BP) with respect to the model.)

For the exposition, it is convenient to take a look at the estimator for  $\boldsymbol{\beta}$  when it is a step in an iterative procedure (like the EM algorithm considered in Section 3.4).

Suppose the value of  $u_m^{N2}$  is known from the previous step. The estimator for  $\boldsymbol{\beta}$  at the current step is

$$\hat{\boldsymbol{\beta}} = \left( \sum_{m=1}^M \sum_{j=1}^{n_m} \tilde{\sigma}_{mj}^{-2} \mathbf{x}_{mj} \mathbf{x}_{mj}^T \right)^{-1} \sum_{m=1}^M \sum_{j=1}^{n_m} \tilde{\sigma}_{mj}^{-2} \mathbf{x}_{mj}^T (y_{mj} - u_m^{N2}), \quad (3.1.17)$$

where  $\tilde{\sigma}_{mj}^{-2} = (1 - p_{mj}) \sigma_1^{-2} + p_{mj} \sigma_2^{-2}$  is the expected value of  $\sigma_{mj}^{-2}$ . Thus, each observation in the estimator for  $\boldsymbol{\beta}$  is “weighted” according to its probability of being

from part 2 of the mixture. Since for extreme outliers the expected value of  $\sigma_{mj}^{-2}$  is close to  $\sigma_2^{-2}$ , their impact on the estimate of  $\boldsymbol{\beta}$  is reduced compared to the nested error regression model, where  $\sigma_1^2 = \sigma_2^2$ . This makes the estimator of  $\boldsymbol{\beta}$  robust to outliers.

The “direct” estimator of  $\bar{y}_m$  given by (3.1.9) also accounts for outliers. In fact, it cannot be called a “direct” estimator because it depends on units from other areas through the estimates of variances and the probabilities of belonging to part 2 of the mixture.

### 3.2 Identifiability of the model parameters

In this section we discuss identifiability of the parameters in model N2. We will use representation (3.1.4).

Loosely speaking, a set of parameters is said to be identifiable when distinct sets of the parameter values determine distinct distributions. This sentence, of course, does not specify what is meant by a “distinct set”. For example, mixture distributions are invariant to permutations of their components (and thus, to permutations in the values of corresponding parameter vector). It makes sense to not disqualify mixture distributions as non-identifiable based on this simple fact. We now reproduce, with minor changes in notation, the definition of identifiability for mixture distributions, as stated in Yakowitz and Spragins (1968).

Let  $F = \{f(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}), \boldsymbol{\theta} \in \mathbb{R}^s, \mathbf{X} \in \mathbb{R}^{n \times p}, \mathbf{y} \in \mathbb{R}^n\}$  be a family of  $n$ -dimensional distribution functions. The set of all finite mixtures of a class  $F$  of distributions is the convex hull

$$H = \left\{ h(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}) : h(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}) = \sum_{k=1}^K \lambda_k f_k(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}_k), \lambda_k > 0, \sum_{k=1}^K \lambda_k = 1, f_k(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}_k) \in F, K = 1, 2, \dots \right\}$$

Uniqueness of representation property means that if

$$\sum_{k=1}^K \lambda_k f_k(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}_k) = \sum_{l=1}^M \lambda'_l f'_l(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}_l) \quad (3.2.1)$$

then (1)  $K = M$  and (2) for any  $1 \leq k \leq K$  there exist  $1 \leq l \leq K$ , such that  $\lambda_k = \lambda'_l$  and  $f_k(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}_k) = f'_l(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}_l)$ .

If the uniqueness of representation holds for class  $H$ , it is said that the family  $F$  generates identifiable finite mixtures  $H$ .

The finite mixtures generated by the family of  $n$ -dimensional multivariate normal distributions are identifiable, by Proposition 2 of Yakowitz and Spragins (1968).

The proof of identifiability is particularly straightforward in the case of model N2.

First, if  $\sigma_1^2 = \sigma_2^2$ , then  $f_k(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}_k) = f'_l(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}_l)$  for all  $k, l$ . Thus, all elements of

$H$  coincide with the original distribution. Next, suppose  $\sigma_1^2 \neq \sigma_2^2$ , and so  $K = 2^n$ .

We need to prove that the set of distributions  $\{f_k(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}_k); k = 1, \dots, 2^n\}$  involved in

(3.1.4) is a linearly independent set. The uniqueness of representation of a mixture distribution as a linear combination of the component pdf's follows immediately.

Below is the proof that the set of distributions  $\{f_k(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}_k); k = 1, \dots, 2^n\}$  is indeed a linearly independent set.

Suppose for a vector  $a = (a_1, \dots, a_{2^n})^T$ ,

$\sum_{k=1}^{2^n} a_k f_k(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}_k) = 0$  for almost all  $\mathbf{y}$  (i.e., for all  $\mathbf{y}$  except possibly for a set of measure 0).

Consider a linear combination of moment generating functions:

$$\sum_{k=1}^{2^n} a_k g_k(\mathbf{t}) = \sum_{k=1}^{2^n} a_k \int e^{t^T \mathbf{y}} f_k(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}_k) d\mathbf{y} = \int e^{t^T \mathbf{y}} \left\{ \sum_{k=1}^{2^n} a_k f_k(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}_k) \right\} d\mathbf{y} = 0.$$

Therefore,

$$\sum_{k=1}^{2^n} a_k g_k(\mathbf{t}) = \sum_{k=1}^{2^n} a_k e^{t^T \mathbf{x} \boldsymbol{\beta}_k + \frac{1}{2} t^T \boldsymbol{\Sigma}_k t} = 0 \text{ for all vectors } \mathbf{t}.$$

Taking the zero vector  $\mathbf{t} = (0, \dots, 0)$ , we obtain

$$\sum_{k=1}^{2^n} a_k = 0. \tag{3.2.2}$$

Let  $i(k)$  denote a permutation of indexes  $k$ . It follows from (3.2.2) that for any permutation  $i(k)$ ,

$$\sum_{k=1}^{2^{n-1}} a_{i(k)} = - \sum_{k=2^{n-1}+1}^{2^n} a_{i(k)}. \tag{3.2.3}$$

By using varying values of zero-one vectors  $\mathbf{t} = (0, \dots, 0, 1, 0, \dots, 0)$ , where 1 appears in turn at different places, we find that

$$e^{\frac{1}{2}\sigma_1^2} \sum_{k=1}^{2^{n-1}} a_{i(k)} + e^{\frac{1}{2}\sigma_2^2} \sum_{k=2^{n-1}+1}^{2^n} a_{i(k)} = 0. \quad (3.2.4)$$

Hence, from (3.2.3) and (3.2.4) (and since  $\sigma_1^2 \neq \sigma_2^2$ ), it follows that

$$\sum_{k=1}^{2^{n-1}} a_{i(k)} = \sum_{k=2^{n-1}+1}^{2^n} a_{i(k)} = 0. \quad (3.2.5)$$

The above statement is true for all possible permutations of indexes. This can only be possible when  $a_k = 0$  for all  $k$ .

Thus, the set of functions in (3.1.4) is linearly independent.

Suppose  $\sum_{k=1}^{2^n} c_k f_k(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}_k)$  and  $\sum_{k=1}^{2^n} d_k f_k(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}_k)$  are two representations of the same mixture distribution. Then  $\sum_{k=1}^{2^n} (c_k - d_k) f_k(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}_k) = 0$ , for almost all  $\mathbf{y}$ . This can only be true if  $c_k - d_k = 0$  for every  $k$ .

### 3.3 EM algorithm

The EM (“expectation-maximization”) iterative algorithm of Dempster *et al.* (1977) is a suitable way of finding the maximum of the log-likelihood for the case of mixture distributions.

Equation (3.1.4), when it is viewed as a function of the parameter vector  $\theta$ , defines the likelihood function for the random vector  $\mathbf{y}$ . The log-likelihood is (we omit  $X$ , for simplicity)

$$L(\mathbf{y}; \boldsymbol{\theta}) = \log \sum_{k=1}^{2^n} \lambda_k f_k(\mathbf{y} | \boldsymbol{\theta}_k). \quad (3.3.1)$$

This is not a convenient representation for the purpose of maximizing the likelihood function with respect to the vector of parameters. The efficacy of EM stems from a convenient form of the so-called “complete data” likelihood, which is the likelihood that handles the unobserved part of the augmented data vector as if it is being observed.

Iteration of the EM algorithm consists of the so-called “E-step”, finding of the expected value of the logarithm of the likelihood function of the complete data likelihood, given the “current” values of parameters. This step is followed by the “M-step”, which entails obtaining new values of the parameters as maximizers of this function.

Consider a random vector of indicators  $(i_1, \dots, i_K)$ , where  $i_k = 1$  when the realized vector  $\mathbf{y}$  comes from the  $k$ -th part of the mixture, and  $i_k = 0$ , otherwise;

$$\lambda_k = P\{i_k = 1\}.$$

From now on, let us denote the parameter vector by  $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_1, \sigma_2, \tau, p)$ .

Note that the conditional probability of  $i_k$ , given the data, is



$$P\{i_k = 1 | \mathbf{y}, \boldsymbol{\theta}\} = \frac{\lambda_k f_k(\mathbf{y} | \boldsymbol{\theta})}{h(\mathbf{y} | \boldsymbol{\theta})}.$$

The expected value of the logarithm of the complete data likelihood function (if the vector of random effects  $\mathbf{u}$  and indicator  $i_k$  would be observed, in addition to the data vector  $\mathbf{y}$ ), given current values of the parameters, is

$$Q(\boldsymbol{\theta} | \boldsymbol{\theta}^c) = E[\log h(\mathbf{y}, \mathbf{u}, i_k | \boldsymbol{\theta}) | \mathbf{y}, \boldsymbol{\theta}^c], \quad (3.3.2)$$

with

$$h(\mathbf{y}, \mathbf{u}, i_k | \boldsymbol{\theta}) = p^{i_k} (1-p)^{n-i_k} \frac{1}{(2\pi)^{M/2} |\mathbf{D}|^{1/2}} e^{-\frac{1}{2} \mathbf{u}^T \mathbf{D}^{-1} \mathbf{u}} \frac{1}{(2\pi)^{n/2} |\mathbf{R}_k|^{1/2}} e^{-\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})^T \mathbf{R}_k^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})}.$$

The expectation in (3.3.2) is taken over the joint conditional distribution of  $\mathbf{u}$  and  $i_k$  given  $\mathbf{y}$ .  $Q(\boldsymbol{\theta} | \boldsymbol{\theta}^c)$  can be presented as a sum of two components:

$$Q(\boldsymbol{\theta} | \boldsymbol{\theta}^c) = U(\boldsymbol{\theta} | \boldsymbol{\theta}^c) + \mathbf{V}(\boldsymbol{\theta} | \boldsymbol{\theta}^c), \quad (3.3.3)$$

where

$$\begin{aligned} U(\boldsymbol{\theta} | \boldsymbol{\theta}^c) &= E[\log \lambda_k | \mathbf{y}, \boldsymbol{\theta}^c] \\ &= \sum_{k=1}^{2^n} P\{i_k = 1 | \mathbf{y}, \boldsymbol{\theta}^c\} \log \lambda_k, \end{aligned} \quad (3.3.4)$$

$$\mathbf{V}(\boldsymbol{\theta} | \boldsymbol{\theta}^c) = E[\log h(\mathbf{y}, \mathbf{u} | i_k = 1, \boldsymbol{\theta}) | \mathbf{y}, \boldsymbol{\theta}^c]. \quad (3.3.5)$$

Recall that  $\lambda_k$  is a function of  $p$ , so the first term of (3.3.3) depends on  $p$ ; however, the second term of (3.3.3) does not depend on  $p$ . Thus, the maximum likelihood of the parameter  $p$  is based solely on the first part,  $U(\boldsymbol{\theta}|\boldsymbol{\theta}^c)$ .

Let us consider the term  $U(\boldsymbol{\theta}|\boldsymbol{\theta}^c)$  first. To say that  $i_k$  is observed is the same as to say that a vector  $\mathbf{z} = (z_1, \dots, z_n)$  is observed, where each component  $z_j$  is an indicator variable:  $z_j = 1$ , when an observation  $j$  comes from the distribution with the  $\sigma_2^2$  value for the random error variance, and  $z_j = 0$ , otherwise;  $j = 1, \dots, n$ .

$$U(\boldsymbol{\theta}|\boldsymbol{\theta}^c) = \sum_{k=1}^{2^n} P\{i_k = 1 | \mathbf{y}, \boldsymbol{\theta}^c\} \log \lambda_k = E[\log g(\mathbf{z}; p) | \mathbf{y}, \boldsymbol{\theta}^c],$$

where  $g(\mathbf{z}; p)$  is the likelihood function for  $p$ , when a  $\mathbf{z}$  is observed. (The expectation is over the conditional distribution of  $\mathbf{z}$  given  $\mathbf{y}$  and the current values of the parameters.)

$$g(\mathbf{z}; p) = \prod_{j=1}^n p^{z_j} (1-p)^{1-z_j}.$$

Thus,

$$U(\boldsymbol{\theta}|\boldsymbol{\theta}^c) = \sum_{j=1}^n \left[ E(z_j | \mathbf{y}, \boldsymbol{\theta}^c) \log p + (1 - E(z_j | \mathbf{y}, \boldsymbol{\theta}^c)) \log(1-p) \right].$$

It follows that the M-step maximizer with respect to  $p$  is

$$p^+ = \frac{1}{n} \sum_{j=1}^n E(z_j | \mathbf{y}, \boldsymbol{\theta}^c),$$

where the conditional probability of an observation coming from part 2 of a mixture is

$$p_j^c = E(z_j | \mathbf{y}, \boldsymbol{\theta}^c) = \frac{p^c \varphi\left(\frac{y_j - \mathbf{x}_j^T \boldsymbol{\beta}^c}{\sqrt{\sigma_2^{2c} + \tau^{2c}}}\right)}{(1-p^c) \varphi\left(\frac{y_j - \mathbf{x}_j^T \boldsymbol{\beta}^c}{\sqrt{\sigma_1^{2c} + \tau^{2c}}}\right) + p^c \varphi\left(\frac{y_j - \mathbf{x}_j^T \boldsymbol{\beta}^c}{\sqrt{\sigma_2^{2c} + \tau^{2c}}}\right)}, \quad (3.3.6)$$

$\varphi(\cdot)$  is the standard normal pdf.

Now consider the second term of (3.3.2). The complete data log-likelihood  $\log h(\mathbf{y}, \mathbf{u} | i_k = 1, \boldsymbol{\theta})$  has the form

$$\log h(\mathbf{y}, \mathbf{u} | i_k = 1, \boldsymbol{\theta}) = c + \frac{1}{2} \left\{ \log |\mathbf{D}^{-1}| - \mathbf{u}^T \mathbf{D}^{-1} \mathbf{u} + \log |\mathbf{R}_k^{-1}| - \mathbf{e}^T \mathbf{R}_k^{-1} \mathbf{e} \right\}, \quad (3.3.7)$$

where  $\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}$ ,  $\mathbf{R}_k = \text{diag}_n \left[ (1 - z_{kj}) \sigma_1^2 + z_{kj} \sigma_2^2 \right]$ , for a combination  $i_k = 1$  of the indicator vector,  $z_k = (z_{k1}, \dots, z_{kn})$ ;  $\mathbf{D} = \tau^2 \mathbf{I}_M$ ,  $\mathbf{I}_M$  is the identity matrix of size  $M$ , and  $c$  does not depend on the model parameters. (The inverse of  $\mathbf{R}_k$  can be written as  $\mathbf{R}_k^{-1} = \text{diag}_n \left[ (1 - z_{kj}) \sigma_1^{-2} + z_{kj} \sigma_2^{-2} \right]$ .)

For a given set of mixture indicators, the distribution is multivariate normal. The conditional expectation and variance of the random effects, given the data vector and the current values of the parameters are

$$\begin{aligned} E[\mathbf{u} | \mathbf{y}, \boldsymbol{\theta}^c] &= E \left[ E[\mathbf{u} | \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}^c] | \mathbf{y}, \boldsymbol{\theta}^c \right] \\ &= E \left\{ \left[ \mathbf{D}^{-1c} + \mathbf{Z}^T \mathbf{R}_k^{-1c} \mathbf{Z} \right]^{-1} \mathbf{Z}^T \mathbf{R}_k^{-1c} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^c) | \mathbf{y}, \boldsymbol{\theta}^c \right\} \end{aligned} \quad (3.3.8)$$

and

$$\begin{aligned}
\text{Var}[\mathbf{u} | \mathbf{y}, \boldsymbol{\theta}^c] &= E[\text{Var}[\mathbf{u} | \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}^c] | \mathbf{y}, \boldsymbol{\theta}^c] + \text{Var}[E[\mathbf{u} | \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}^c] | \mathbf{y}, \boldsymbol{\theta}^c] \\
&= E\left\{\left[\mathbf{D}^{-1c} + \mathbf{Z}^T \mathbf{R}_k^{-1c} \mathbf{Z}\right]^{-1} | \mathbf{y}, \boldsymbol{\theta}^c\right\} \\
&\quad + \text{Var}\left\{\left[\mathbf{D}^{-1c} + \mathbf{Z}^T \mathbf{R}_k^{-1c} \mathbf{Z}\right]^{-1} \mathbf{Z}^T \mathbf{R}_k^{-1c} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^c) | \mathbf{y}, \boldsymbol{\theta}^c\right\}. \tag{3.3.9}
\end{aligned}$$

(The expectations and variance on the right hand side of (3.3.8) and (3.3.9) are with respect to distribution of  $\mathbf{z}$  given  $\mathbf{y}$  and current values of the parameters.)

Apart from the cases of unrealistically small samples, the direct computation of the above expectations is unfeasible because it involves evaluation of the products of all possible combinations of the individual unit probabilities  $p_j^c = E(z_j | \mathbf{y}, \boldsymbol{\theta}^c)$ . We describe approximate methods for computation of (3.3.8) and (3.3.9) in Section 3.3.1 below. For now, let us suppose this problem is solved and we obtained the values for these expressions.

Denote  $\mathbf{R}^{-1} = E[\mathbf{R}_k^{-1} | \mathbf{y}, \boldsymbol{\theta}^c]$ . We have

$$\mathbf{R}^{-1} = \text{diag}(1 - p_j^c) \sigma_1^{-2} + \text{diag}(p_j^c) \sigma_2^{-2} \tag{3.3.10}$$

and

$$\begin{aligned}
E[\log |\mathbf{R}_k^{-1}| | \mathbf{y}, \boldsymbol{\theta}^c] &= n(1 - p^c) \log \sigma_1^{-2} + np^c \log \sigma_2^{-2} \\
&= E\left[\sum_{j=1}^n \left\{(1 - z_j) \log \sigma_1^{-2} + z_j \log \sigma_2^{-2}\right\} | \mathbf{y}, \boldsymbol{\theta}^c\right]
\end{aligned}$$

$$\begin{aligned}
&= \sum_{j=1}^n \left\{ E \left[ (1 - z_j) \mid \mathbf{y}, \boldsymbol{\theta}^c \right] \log \sigma_1^{-2} + E \left[ z_j \mid \mathbf{y}, \boldsymbol{\theta}^c \right] \log \sigma_2^{-2} \right\} \\
&= n(1 - p^c) \log \sigma_1^{-2} + np^c \log \sigma_2^{-2}.
\end{aligned} \tag{3.3.11}$$

Note also that

$$E[\mathbf{e}^T \mathbf{R}_k^{-1} \mathbf{e} \mid \mathbf{y}, \boldsymbol{\theta}^c] = E[E[\mathbf{e}^T \mathbf{R}_k^{-1} \mathbf{e} \mid \mathbf{y}, \mathbf{u}, \boldsymbol{\theta}^c] \mid \mathbf{y}, \boldsymbol{\theta}^c] = E[\mathbf{e}^T \mathbf{R}^{-1} \mathbf{e} \mid \mathbf{y}, \boldsymbol{\theta}^c] \tag{3.3.12}$$

We can write

$$\mathbf{V}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^c) = c + \Delta(\sigma_1^2, \sigma_2^2) + W(\boldsymbol{\theta} \mid \boldsymbol{\theta}^c), \tag{3.3.13}$$

where

$$W(\boldsymbol{\theta} \mid \boldsymbol{\theta}^c) = \frac{1}{2} E \left[ \log |\mathbf{D}^{-1}| - \mathbf{u}^T \mathbf{D}^{-1} \mathbf{u} + \log |\mathbf{R}^{-1}| - \mathbf{e}^T \mathbf{R}^{-1} \mathbf{e} \mid \mathbf{y}, \boldsymbol{\theta}^c \right], \tag{3.3.14}$$

and

$$\Delta(\sigma_1^2, \sigma_2^2) = \frac{1}{2} \left\{ E \left[ \log |\mathbf{R}_k^{-1}| \mid \mathbf{y}, \boldsymbol{\theta}^c \right] - \log |\mathbf{R}^{-1}| \right\}, \tag{3.3.15}$$

with

$$\log |\mathbf{R}^{-1}| = \sum_{j=1}^n \log \left( (1 - p_j^c) \sigma_1^{-2} + p_j^c \sigma_2^{-2} \right). \tag{3.3.16}$$

Note that  $W(\boldsymbol{\theta} \mid \boldsymbol{\theta}^c)$  has the form of an expectation of the complete data log-likelihood

function of a multivariate normal variable  $N_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$  with  $\boldsymbol{\Sigma} = \mathbf{R} + \mathbf{Z}\mathbf{D}\mathbf{Z}^T$ .

Proceed to maximize  $\mathbf{V}(\boldsymbol{\theta}|\boldsymbol{\theta}^c)$ . Since the term  $\Delta(\sigma_1^2, \sigma_2^2)$  of (3.3.13) does not involve parameters  $\mathbf{D}$  and  $\boldsymbol{\beta}$ , the usual linear mixed model  $\mathbf{D}$  and  $\boldsymbol{\beta}$  maximizers of  $W(\boldsymbol{\theta}|\boldsymbol{\theta}^c)$  also maximize  $\mathbf{V}(\boldsymbol{\theta}|\boldsymbol{\theta}^c)$ :

$$\boldsymbol{\beta}^+ = (\mathbf{X}^T \mathbf{R}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{R}^{-1} (\mathbf{y} - E[\mathbf{u} | \mathbf{y}, \boldsymbol{\theta}^c]), \quad (3.3.17)$$

$$\tau^{2+} = \frac{1}{M} E[\mathbf{u}^T \mathbf{u} | \mathbf{y}, \boldsymbol{\theta}^c] = \frac{1}{M} \left\{ E[\mathbf{u} | \mathbf{y}, \boldsymbol{\theta}^c]^T E[\mathbf{u} | \mathbf{y}, \boldsymbol{\theta}^c] + \text{Var}[\mathbf{u} | \mathbf{y}, \boldsymbol{\theta}^c] \right\} \quad (3.3.18)$$

The  $\sigma_1^2$  and  $\sigma_2^2$  maximizers of  $\mathbf{V}(\boldsymbol{\theta}|\boldsymbol{\theta}^c)$  also have a simple explicit form. The derivatives are

$$\frac{\partial \mathbf{V}(\boldsymbol{\theta}|\boldsymbol{\theta}^c)}{\partial (\sigma_1^2)} = \frac{1}{\sigma_1^2} \left[ \frac{1}{\sigma_1^2} E[\mathbf{e}^T \text{diag}(1 - p_j^c) \mathbf{e} | \mathbf{y}, \boldsymbol{\theta}^c] - n(1 - p^c) \right] = 0,$$

$$\frac{\partial \mathbf{V}(\boldsymbol{\theta}|\boldsymbol{\theta}^c)}{\partial (\sigma_2^2)} = \frac{1}{\sigma_2^2} \left[ \frac{1}{\sigma_2^2} E[\mathbf{e}^T \text{diag}(p_j^c) \mathbf{e} | \mathbf{y}, \boldsymbol{\theta}^c] - np^c \right] = 0.$$

(The expectations in the above formulas are with respect to the conditional distribution of  $\mathbf{u}$  given  $\mathbf{y}$ .)

So the M-step maximizers with respect to  $\sigma_1^2$  and  $\sigma_2^2$  are

$$\sigma_1^{2+} = \frac{1}{n(1 - p^c)} E[\mathbf{e}^T \text{diag}(1 - p_j^c) \mathbf{e} | \mathbf{y}, \boldsymbol{\theta}^c], \quad (3.3.19)$$

$$\sigma_2^{2+} = \frac{1}{np^c} E[\mathbf{e}^T \text{diag}(p_j^c) \mathbf{e} | \mathbf{y}, \boldsymbol{\theta}^c]. \quad (3.3.20)$$

Thus at an iteration of the EM algorithm, we find maximizers of the expected value of the complete data log-likelihood.

### 3.3.1 Approximate computation of the first two conditional moments of the random effects

We considered several possibilities for evaluation of expressions (3.3.8) and (3.3.9).

Method 1. Consider the following Monte Carlo approximation.

For an  $l$ -th Monte Carlo cycle, do the following:

1. Given the current values of the conditional probabilities  $p_j^c$ , draw a Poisson sample from the original data. Each observation is selected into the sample with probability  $p_j^c$ .
2. If an observation  $j$  is selected into the sample, assign it to part 2 of the mixture, i.e. let  $z_j^{(l)} = 1$ ; otherwise assign it to part 1 of the mixture, i.e. let  $z_j^{(l)} = 0$ .
3. Use the current values of the parameters and the values of indicators obtained in the above step to compute prediction for the random effects

$$\mathbf{u}^{(l)} = [\mathbf{D}^{-1c} + \mathbf{Z}^T \mathbf{R}_l^{-1c} \mathbf{Z}]^{-1} \mathbf{Z}^T \mathbf{R}_l^{-1c} (\mathbf{y} - \mathbf{X} \boldsymbol{\beta}^c) \quad (3.3.21)$$

and the variance

$$\mathbf{v}^{(l)} = [\mathbf{D}^{-1c} + \mathbf{Z}^T \mathbf{R}_l^{-1c} \mathbf{Z}]^{-1}. \quad (3.3.22)$$

Repeat this procedure  $L$  times and obtain the estimates of  $E[\mathbf{u} | \mathbf{y}, \boldsymbol{\theta}^c]$  as an average of the Monte Carlo based predictions,

$$\hat{\mathbf{u}} = L^{-1} \sum_{l=1}^L \mathbf{u}^{(l)}. \quad (3.3.23)$$

The variance of the prediction for the random effects is  $Var[\mathbf{u} | \mathbf{y}, \boldsymbol{\theta}^c] = E\{Var[\mathbf{u} | \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}^c]\} + Var\{E[\mathbf{u} | \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}^c]\}$ . The estimate of  $E\{Var[\mathbf{u} | \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}^c]\}$  is

$$\hat{\mathbf{v}}_1 = L^{-1} \sum_{l=1}^L \mathbf{v}^{(l)}. \quad (3.3.24)$$

The variance  $Var\{E[\mathbf{u} | \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}^c]\}$  of the predictions is estimated as

$$\hat{\mathbf{v}}_2 = (L-1)^{-1} \sum_{l=1}^L (\mathbf{u}^{(l)} - \hat{\mathbf{u}})^2. \quad (3.3.25)$$

The total variance  $Var[\mathbf{u} | \mathbf{y}, \boldsymbol{\theta}^c]$  is

$$\hat{\mathbf{v}} = \hat{\mathbf{v}}_1 + \hat{\mathbf{v}}_2. \quad (3.3.26)$$

We now provide justification for using the Poisson draws based on probabilities  $p_j^c$ .

We can write the quantity that we want to estimate by this procedure as

$$\begin{aligned} E[\mathbf{u} | \mathbf{y}, \boldsymbol{\theta}^c] &= E[E[\mathbf{u} | \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}^c] | \mathbf{y}, \boldsymbol{\theta}^c] \\ &= E[\mathbf{u}^{(k)} | \mathbf{y}, \boldsymbol{\theta}^c] \\ &= \sum_{k=1}^{2^n} \pi^{(k)} \mathbf{u}^{(k)} \end{aligned} \quad (3.3.27)$$



where  $\mathbf{u}^{(k)}$  is given by (3.3.21) and  $\pi^{(k)} = \prod_{j=1}^n p_j^{(k)}$ ;  $p_j^{(k)} = 1 - p_j^c$  if  $z_{kj} = 0$ , and

$p_j^{(k)} = p_j^c$  if  $z_{kj} = 1$ ;  $z_{kj}$  is indicator for position  $j$  at the  $k$ th mixture combination,

$1 \leq k \leq 2^n$ ; note also that  $\sum_{k=1}^{2^n} \pi^{(k)} = 1$ .

In order to estimate the above target for a population of  $2^n$  "units", we select a sample of size  $L$  (the number of the Monte Carlo runs), with replacement, and with probability proportional to "size", where the "size" variable is  $\pi^{(k)}$ . This is accomplished by drawing the Poisson sample at each step  $l$  of the  $L$  Monte Carlo runs. The estimator from this sample is

$$\hat{\mathbf{u}} = \frac{\sum_{l=1}^L w^{(l)} (\pi^{(l)} \mathbf{u}^{(l)})}{\sum_{l=1}^L w^{(l)} (\pi^{(l)})},$$

where  $w^{(l)} = 1/\pi^{(l)}$ . Thus,  $\hat{\mathbf{u}} = \frac{\sum_{l=1}^L \mathbf{u}^{(l)}}{L}$ .

Expressions (3.3.24) and (3.3.25) follow from similar considerations.

The above method works well and the algorithm converges fast when the probability  $p$  of being in part 2 of the mixture is small. Otherwise, the method may be unstable and would require many repetitions of the Monte Carlo steps.

## Method 2.

The idea is that we fit an area-level model corresponding to our unit-level model formulation and obtain prediction for the random effects from this area level model.

We also obtain the variance of the prediction from the same area level formulation. However, this variance does not account for the variability over the mixture indicators. The latter term is obtained using the Monte Carlo step.

Consider the following vector of adjusted residuals:

$$\bar{\mathbf{r}}^c = (\mathbf{Z}^T \mathbf{R}^{-1c} \mathbf{Z})^{-1} [\mathbf{Z}^T \mathbf{R}^{-1c} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^c)], \quad (3.3.28)$$

where  $\mathbf{R}^{-1c}$  is a current value of the diagonal matrix defined by (3.3.10).

Note that  $\bar{\mathbf{r}}^c$  is an area-level quantity that follows an area-level model with the same value of the random effects as the original unit level model (this is evident after noting that the multiplicative adjustments to the residuals  $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^c$  in (3.3.28) add up to one). Thus, the variance of the prediction of the random effect from this area-level model can be used as approximation to the expected variance of the prediction for random effect in the original unit-level model. This variance is

$$\tilde{\mathbf{v}}_1 = (\mathbf{D}^c + \mathbf{H}^c)^{-1} \mathbf{D}^c \mathbf{H}^c, \quad (3.3.29)$$

where  $\mathbf{H}^c = \text{diag}_M(h_m^c)$  is a diagonal matrix of the direct sample variances of  $\bar{\mathbf{r}}^c$ .

These variances are considered known in the area-level settings. We can approximate  $\mathbf{H}^c$  by plugging in a value derived from the data, i.e., the variance  $h_m^c$  for area  $m$  is computed as

$$h_m^c = n_m \left\{ \frac{1}{n_m - 1} \sum_{j=1}^{n_m} (r_{mj}^c - \bar{r}_m^c)^2 \right\}, \quad (3.3.30)$$

where  $r_{mj}^c = q_{mj}^c (y_{mj} - \mathbf{x}_{mj}^T \boldsymbol{\beta}^c)$ ,  $q_{mj}^c = \left( \sum_{j=1}^{n_m} w_{mj} \right)^{-1} w_{mj}$ ,  $w_{mj} = (1 - p_{mj}) \sigma_1^{-2} + p_{mj} \sigma_2^{-2}$ .

Prediction for the random effects is given by

$$\tilde{\mathbf{u}} = (\mathbf{D}^c + \mathbf{H}^c)^{-1} \mathbf{D}^c \bar{\mathbf{r}}^c. \quad (3.3.31)$$

The first term in the variance formula (3.3.9) is approximated by (3.3.29). Consider the second term now. We obtain it from the Monte Carlo simulations using formula (3.3.25). The total variance is

$$\tilde{\mathbf{v}} = \tilde{\mathbf{v}}_1 + \hat{\mathbf{v}}_2. \quad (3.3.32)$$

In addition, from the same Monte Carlo setup, we can estimate the bias of the prediction as

$$\tilde{\mathbf{b}} = L^{-1} \sum_{l=1}^L (\mathbf{u}^{(k)} - \tilde{\mathbf{u}}).$$

In estimation of the variance components, we use the mean squared error  $E[\mathbf{u}^T \mathbf{u} | \mathbf{y}, \boldsymbol{\theta}^c]$  of the random effect. We estimate it by adding up the terms:

$$\tilde{\mathbf{u}}^T \tilde{\mathbf{u}} + \tilde{\mathbf{v}} + \tilde{\mathbf{b}}^T \tilde{\mathbf{b}}.$$

Thus, we use the outcome from the area-level model as an approximation for the random effects and we use the Monte Carlo simulations to approximate the mean squared error of this estimate.

### Method 3.

This method is very simple and it works well when the probability  $p$  of being in the outlier part is small. In such a case, the conditional probabilities  $p_{mj}$  for the outliers are close to 1, while the probabilities of the other units are close to zero. Then, just plugging in the probabilities in place of the mixture indicators produces good estimates. Thus, for the prediction of the random effects and the corresponding variance we can use

$$\tilde{\mathbf{u}}^c = \left[ \mathbf{D}^{-1c} + \mathbf{Z}^T \mathbf{R}^{-1c} \mathbf{Z} \right]^{-1} \mathbf{Z}^T \mathbf{R}^{-1c} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^c) \quad (3.3.33)$$

and the variance

$$\tilde{\mathbf{v}} = \left[ \mathbf{D}^{-1c} + \mathbf{Z}^T \mathbf{R}^{-1c} \mathbf{Z} \right]^{-1}, \quad (3.3.34)$$

where  $\mathbf{R}^{-1c} = \text{diag}(1 - p_j^c) \sigma_1^{-2c} + \text{diag}(p_j^c) \sigma_2^{-2c}$ .

When  $p_{mj}$  of the sample units are either close to 1 or close to 0, the second part of the variance,  $\text{Var}\left\{E\left[\mathbf{u} \mid \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}^c\right]\right\}$ , is small.

Note that this method works well in roughly the same situation as Method 1, yet it is simpler than Method 1. Effectively, this method replaces the ML estimation of the mixture model parameters by a two-step procedure. At the first step, the conditional probabilities  $p_{mj}$  and corresponding expected values for the inverse variances  $\sigma_{mj}^{-2}$  are computed. At the second step, a multivariate normal model with variances  $\sigma_{mj}^2$  is fitted. To repeat, this “plug-in” procedure works well when the mixture is “well” separated and  $p_{mj}$  are either close to 1 or to 0, effectively declaring with some confidence the mixture membership of the units.

### 3.4 Parametric bootstrap for prediction confidence intervals

To obtain confidence intervals for the predictor of the random effects, we use the method analogous to the approach of Chatterjee, Lahiri, and Li (2008), henceforth, CLL.

In this section, we present the bootstrap algorithm. The simulation results are presented in Section 3.8.

To ease the notation we drop the superscript N2.

#### The Bootstrap Algorithm

The bootstrap is performed as follows. Define the ‘‘pivot’’ vector  $\hat{\eta} = (\hat{\eta}_1, \dots, \hat{\eta}_M)$ :

$$\hat{\eta} = \hat{\nu}^{-1} \left( \bar{Y}_r - \hat{Y}_r \right), \quad (3.4.1)$$

where  $\bar{Y}_r = (\bar{Y}_{1r}, \dots, \bar{Y}_{Mr})$ ,  $\hat{Y}_r = (\hat{Y}_{1r}, \dots, \hat{Y}_{Mr})$ ,  $\hat{\nu}^2 = \text{diag}_M(\hat{\nu}_m^2)$ , and  $\hat{\nu}_m^2$  is an estimate of variance (3.1.13) for area  $m$ .

Although the components of the vector  $\hat{\eta}$  are not normally distributed, the distribution can be approximated using the parametric bootstrap analogous to the CLL approach. For the case of the mixed mixture model N2, the algorithm is given by the following steps:

1. Generate  $u_m^* \sim N(0, \hat{\tau}^2)$  and  $z_{mj}^* \sim \text{Bin}(1; \hat{p})$ .
2. Generate  $e_{mj}^* \sim N(0, \hat{\sigma}_1^2)$ , if  $z_{mj}^* = 0$  and  $e_{mj}^* \sim N(0, \hat{\sigma}_2^2)$ , if  $z_{mj}^* = 1$ .
3. A set of bootstrap data  $y_{mj}^*$  is obtained as

$$y_{mj}^* = \mathbf{x}_{mj}^T \hat{\boldsymbol{\beta}} + u_m^* + e_{mj}^*,$$

where  $j = 1, \dots, n_m$ ,  $m = 1, \dots, M$ .

Let

$$\bar{Y}_{mr}^* = \bar{\mathbf{x}}_{mr}^T \hat{\boldsymbol{\beta}} + u_m^* \quad (3.4.2)$$

be bootstrap versions of the “true” population means.

4. From the bootstrap data  $y_{mj}^*$ , obtain the bootstrap estimates of the parameters

$$\left( \hat{p}^*, \hat{\boldsymbol{\beta}}^*, \hat{\tau}^*, \hat{\sigma}_1^*, \hat{\sigma}_2^* \right) \text{ using the same method as is used for the estimates } \left( \hat{p}, \hat{\boldsymbol{\beta}}, \hat{\tau}, \hat{\sigma}_1, \hat{\sigma}_2 \right).$$

Let

$$\hat{Y}_{mr}^* = \bar{\mathbf{x}}_{mr}^T \hat{\boldsymbol{\beta}}^* + \hat{u}_m^* \quad (3.4.3)$$

be a bootstrap estimate of  $\bar{Y}_{mr}^*$ .

5. The vector

$$\hat{\boldsymbol{\eta}}^* = \hat{\mathcal{V}}^{*-1} \left( \bar{Y}_r^* - \hat{Y}_r^* \right) \quad (3.4.4)$$

is a bootstrap approximation of  $\hat{\boldsymbol{\eta}}$ .

In the above,  $\hat{u}_m^*$  and the estimated parameters involved are bootstrap versions of the estimates of exactly the same form as the estimates based on the original sample.

The interval estimate for  $\bar{Y}_{mr}$  is given by  $\left(\hat{Y}_{mr} + q_{m1}\hat{\hat{v}}_m, \hat{Y}_{mr} + q_{m2}\hat{\hat{v}}_m\right)$ , where  $q_{m1}$  and  $q_{m2}$  are quantiles of the distribution of the bootstrap estimates  $\hat{\eta}_m^*$ .

### 3.5 Bias correction

In this section, we discuss two instances where the model assumptions may not hold. First, the outliers are assumed to appear randomly across areas. In fact, however, the outliers may be clustered in certain areas. This may lead to bias in the prediction of the area-level random effects. We propose an area-level bias correction method that is different from the one of Chambers *et al.* (2009): the proposed method attempts to preserve the efficiency of the initial model by introducing the corrections only to select areas, after these areas have been tested on possible outlyingness. Another potentially incorrect assumption is that the outliers are distributed symmetrically around a common mean. Failure of this assumption may lead to an overall bias across areas. The overall bias correction (OBC) can be based on the data combined from all areas, thus the initial modeling assumptions can be more safely relaxed to estimate the correction at this higher level.

If an area contains several units that have a high probability of belonging to the “outlier” part of the mixture, it is possible that the whole area would tend to be an outlier. Note that if outliers tend to be clustered in some areas, this would mean that the distribution of the mixture indicators depends on the area label, which would contradict the model assumption (3.1.3). The failure of the *random occurrence assumption* may lead to significant biases in the areas with a larger portion of the outlying observations. We propose a test to determine that an area is not an outlying

area and a simple method for area-level bias correction in areas where the test fails, as described below.

Consider the following “bias corrected” variations of  $\hat{Y}_{mr}^{N2}$ .

*Bias Correction 1 (BC1)*. Denote residuals  $e_{mj} = y_{mj} - \mathbf{x}_{mj}^T \hat{\boldsymbol{\beta}}$ .

For each area, find the estimate of the mean residual using a mixture of two normal distributions model and by treating areas *as fixed effects*:

$$e_{mj} = \mu_m + \varepsilon_{mj}, \quad (3.5.1)$$

$$\varepsilon_{mj} \mid z_{mj} \stackrel{iid}{\sim} (1 - z_{mj})N(0, \sigma_1^2) + z_{mj}N(0, \sigma_2^2), \quad (3.5.2)$$

$j = 1, \dots, n_m, m = 1, \dots, M$ , and

$$z_{mj} \mid p \stackrel{iid}{\sim} \text{Bin}(1; p). \quad (3.5.3)$$

The BC1 estimator is

$$\hat{Y}_{mr}^{N2+BC1} = \bar{\mathbf{x}}_{mr}^T \hat{\boldsymbol{\beta}} + \hat{\mu}_m^{BC1}, \quad (3.5.4)$$

where  $\hat{\mu}_m^{BC1}$  is the estimate of  $\mu_m$  from the above model.

*Bias Correction 2 (BC2)*. As a general rule, BC1 may be inefficient in areas where the estimates of  $\mu_m$  are based on a small number of observations. Therefore, we propose to use  $\hat{Y}_{mr}^{N2+BC1}$  only when we can demonstrate that an area  $m$  is an outlying area. Consider the following statistic:



$$\hat{P}_m = n_m^{-1} \sum_{j=1}^{n_m} \hat{P}_{mj}. \quad (3.5.5)$$

The distribution of the statistic  $\hat{p}_m$  under the random occurrence assumption can be simulated using the estimated model parameters. These simulated values can be used to obtain a threshold. If the actual estimated  $\hat{p}_m$  is greater than the threshold, the whole area is considered an outlier. The detailed procedure for an area  $m$  can be described by the following steps:

Generate  $\gamma \sim Bin(1; \hat{p})$  and  $\eta \sim \begin{cases} N(0, \hat{\sigma}_1^2 + \hat{\tau}^2) & \text{if } \gamma=0 \\ N(0, \hat{\sigma}_2^2 + \hat{\tau}^2) & \text{if } \gamma=1 \end{cases}$ .

Using the Bayes formula, find the probability of belonging to part 2 of the mixture, given the value of  $\eta$  :

$$\begin{aligned} p^{(a)} &= P\{z = 1 | \eta; \hat{\theta}\} \\ &= \frac{\hat{p}P\{\eta | z = 1; \hat{\theta}\}}{(1 - \hat{p})P\{\eta | z = 0; \hat{\theta}\} + \hat{p}P\{\eta | z = 1; \hat{\theta}\}} \\ &= \frac{\frac{\hat{p}}{\sqrt{\hat{\sigma}_2^2 + \hat{\tau}^2}} \exp\left(-\frac{1}{2} \frac{\eta^2}{\hat{\sigma}_2^2 + \hat{\tau}^2}\right)}{\frac{1 - \hat{p}}{\sqrt{\hat{\sigma}_1^2 + \hat{\tau}^2}} \exp\left(-\frac{1}{2} \frac{\eta^2}{\hat{\sigma}_1^2 + \hat{\tau}^2}\right) + \frac{\hat{p}}{\sqrt{\hat{\sigma}_2^2 + \hat{\tau}^2}} \exp\left(-\frac{1}{2} \frac{\eta^2}{\hat{\sigma}_2^2 + \hat{\tau}^2}\right)}. \end{aligned}$$

Repeat steps 1 and 2  $n_m$  times:  $a = 1, \dots, n_m$ .

Let  $p_m^{(b)} = n_m^{-1} \sum_{a=1}^{n_m} p^{(a)}$  be the average of  $n_m$  simulated values of  $p$ .

Repeat steps 1-4 a large number of times:  $b = 1, \dots, B$  (say,  $B = 500$ ).

Using the simulated values  $p_m^{(b)}$ ,  $b=1,\dots,B$ , estimate a “theoretical value”  $c_m^\alpha$  such that  $P\{p_m > c_m^\alpha\}$  is smaller than some predetermined level  $\alpha$ . This value depends on the number of units in area  $m$ .

If the actual value, obtained as (3.5.5), is higher than  $c_m^\alpha$ , then the area  $m$  has more outliers than would be in a “regular” area under the random occurrence assumption; thus, it can be regarded as an outlying area, and the bias correction (3.5.4) is applied; otherwise, the bias correction is not applied. In our simulations, for application of the bias adjustment, we required that an area had at least four sample units and  $\hat{p}_m > c_m^\alpha$ , where  $\alpha = 0.005$ :

$$\hat{u}_m^{BC2} = \begin{cases} \hat{\mu}_m^{BC1}, & \text{if } \hat{p}_m > c_m^\alpha \text{ and } n_m \geq 4 \\ \hat{u}_m^{N2}, & \text{otherwise} \end{cases} \quad (3.5.6)$$

The BC2 estimator is

$$\hat{Y}_{mr}^{N2+BC2} = \bar{\mathbf{x}}_{mr}^T \hat{\boldsymbol{\beta}} + \hat{u}_m^{BC2}, \quad (3.5.7)$$

*Remark 1.* The data consists of the individual measurements  $y_{mj}$ , with the corresponding area labels, while the area-level effects  $u_m$  are not observable. It is not obvious what is meant by “outlyingness” of an unobserved quantity in the REBLUP approaches. The mixture model formulation, on the other hand, allows the description of the outlying areas in terms of the observable quantities, i.e., as individual outliers clustered in certain areas.

*Remark 2.* Once an area is identified as an outlying area, one may ponder on the meaning of “good” prediction for such area. One can imagine a situation where

“borrowing strength” across areas for prediction in an outlying area is a wrong strategy. Since the area does not fit the model, it is possible that the best course of action is to recognize that using shrinkage estimator for such area would be rather misleading, remove it from the model and perhaps use the direct estimator as prediction for such area.

*Overall Bias Correction, (OBC).* By using (3.5.6), we correct biases in specific outlying areas. Still, it is possible that the assumption that outliers are distributed symmetrically around a common mean may not hold. Failure of this assumption would result in an overall bias. In the simulation study reported in this paper, we correct the initial estimate by adding a robust estimate of the overall mean of residuals to each small area prediction  $\hat{Y}_{mr}^{N2+BC2}$ . (Alternatively, the overall bias may be corrected by benchmarking the small area estimates to a more reliable aggregate level estimate. We did not pursue this approach here.) The data from all areas are involved in estimation of the overall bias. Thus, the OBC estimation is not a problem of small area estimation, and the assumptions may be considerably relaxed.

Denote residuals  $e_{mj}^{N2+BC2} = y_{mj} - \mathbf{x}_{mj}^T \hat{\boldsymbol{\beta}} - \hat{u}_m^{BC2}$ . The overall bias corrected estimator is

$$\hat{Y}_{mr}^{N2+OBC} = \hat{Y}_{mr}^{N2+BC2} + n^{-1} \sum_{m=1}^M \sum_{j=1}^{n_m} e_{mj}^*, \quad (3.5.8)$$

where  $e_{mj}^* = s \cdot \min(c_\alpha, \max(r_{mj}^{N2+BC2}, -c_\alpha))$ ,  $r_{mj}^{N2+BC2} = e_{mj}^{N2+BC2} / s$ ,  $s$  is a robust estimate of scale for the set of residuals  $\{e_{mj}^{N2+BC2}; j = 1, \dots, n_m, m = 1, \dots, M\}$  (e.g.,  $s = \text{med} |e_{mj}^{N2+BC2} - \text{med}(e_{mj}^{N2+BC2})| / 0.6745$ ),  $c_\alpha$  is a tuning parameter (e.g.,  $c_\alpha = 5$ ).

*Remark 3.* We could have slightly modified the initial mixture model assumption and allow the outlying part to have a different mean. This, in our view, would contradict the definition of outlier, which is an unusual observation for a given model: In the absence of additional information in the initial model, we opt for the assumption of symmetry.

The REBLUP and MQ estimators also can be corrected using the overall bias correction; however, the OBC alone would not correct the bias in particular outlying areas. For example, the following OBC for the REBLUP (SR or Fellner's versions) estimator can be considered.

Let  $e_{mj}^{REBLUP} = y_{mj} - \hat{Y}_{mr}^{REBLUP}$ , then the overall bias corrected REBLUP is

$$\hat{Y}_{mr}^{REBLUP+OBC} = \hat{Y}_{mr}^{REBLUP} + n^{-1} s^{REBLUP} \sum_{m=1}^M \sum_{j=1}^{n_m} \phi_b \left( \frac{e_{mj}^{REBLUP}}{s^{REBLUP}} \right), \quad (3.5.9)$$

where  $s^{REBLUP}$  is a robust measure of scale for the set of residuals

$\{e_{mj}^{REBLUP}; j = 1, \dots, n_m, m = 1, \dots, M\}$ , e.g.,  $s^{REBLUP} = \text{med} |e_{mj}^{REBLUP} - \text{med}(e_{mj}^{REBLUP})| / 0.6745$

and  $\phi_b$  is a bounded Huber's function  $\phi_b(x) = \min(b, \max(-b, x))$  with the tuning parameter  $b = 5$ .

### 3.6 Simulation study

The purpose of the simulation study is to compare the performances of different methods under different scenarios. For the first four scenarios, we use a similar setup as described in Chambers *et al.* (2009) with the only difference that we consider the unbalanced case. These scenarios explore cases where there is (1) no contamination in the random terms; (2) contamination in the random effect term only (describing outliers at the area level); (3) contamination in the random error term (describing individual outliers); (4) contamination in the random effect and random error terms (describing area-level and individual outliers). Outliers in scenarios 1-4 have different mean and a larger variance than the bulk of the data, thus we impose asymmetry on the distribution of the random terms. In the fifth scenario, we modify the setup to include a larger fraction of observations with large variance. Finally, in scenarios 6-8, the data is generated from models having the t-distribution with 2 degrees of freedom in random errors, random effects, or in both random terms. We now describe the details of the setup.

There are 40 areas. The sample sizes of the areas are  $n_1, n_3 = 1$ ,  $n_2, n_5 = 2$ ,  $n_3, n_6 = 3$ ,  $n_{7 \leq m \leq 11} = 7$ ,  $n_{12 \leq m \leq 16} = 9$ ,  $n_{17} = 10$ ,  $n_{18} = 50$ ,  $n_{19 \leq m \leq 38} = 5$ ,  $n_{39} = 10$ ,  $n_{40} = 30$ . The population sizes are  $N_m = 20n_m$ . From each area, a sample is selected using simple random sampling without replacement. The auxiliary variable  $x_{mj}$  is generated from the lognormal distribution with mean 1.004077 and standard deviation of 0.5 and the

population values  $y_{mj}$  are generated as  $y_{mj} = 100 + 5x_{mj} + u_m + \varepsilon_{mj}$ . The scenarios for distribution of  $u_m$  and  $\varepsilon_{mj}$  are described below.

- (1) No contamination scenario, **[0,0]**:  $u_m \sim N(0,3)$ ,  $\varepsilon_{mj} \sim N(0,6)$ ;
- (2) Outlying areas, **[0,u]**: for the first 36 areas,  $u_m \sim N(0,3)$ ; for the last four areas,  $u_m \sim N(9,20)$ ;  $\varepsilon_{mj} \sim N(0,6)$  for all observations;
- (3) Individual outliers, **[e,0]**:  $u_m \sim N(0,3)$  for all areas;  $\varepsilon_{mj} \sim N(0,6)$  with probability 0.97 and  $\varepsilon_{mj} \sim N(20,150)$  with probability 0.03;
- (4) Individual outliers and outlying areas, **[e,u]**: for the first 36 areas,  $u_m \sim N(0,3)$ ; for the last four areas,  $u_m \sim N(9,20)$ ;  $\varepsilon_{mj} \sim N(0,6)$  with probability 0.97 and  $\varepsilon_{mj} \sim N(20,150)$  with probability 0.03;
- (5) Individual outliers only, a high-peaked center of the distribution and very long tails, **[70/30]**:  $\varepsilon_{mj} \sim N(0,9)$  with probability 0.7 and  $\varepsilon_{mj} \sim N(0,900)$  with probability 0.3; random effects are  $u_m \sim N(0,9)$ ;
- (6) **[et,0]**: the  $t$  distribution with 2 degrees of freedom for the random error term  $\varepsilon_{mj} \sim t_2(0,9)$ ; random effects are  $u_m \sim N(0,9)$ ;
- (7) **[0,ut]**: the  $t$  distribution with 2 degrees of freedom for the random effect term  $u_m \sim t_2(0,9)$ ; random errors are  $\varepsilon_{mj} \sim N(0,9)$ ;
- (8) **[et,ut]**: the  $t$  distribution with 2 degrees of freedom for the random error and random effect terms,  $\varepsilon_{mj} \sim t_2(0,9)$ ,  $u_m \sim t_2(0,9)$ .

The tuning parameters in the bounded Huber's function for REBLUP are set to  $b=1.345$ ; for the bias-correction of REBLUP (Fellner and SR) and MQ, the tuning parameters are set to  $b=3$ . The tuning parameter for the overall bias correction is  $b=5$ . We used 250 simulation runs for each of the above scenarios and compared the estimates with the corresponding population area means.

To assess the quality of the estimators, we used the median value of the relative bias,

$$RB = 100 \cdot \text{med}_m \left\{ 250^{-1} \sum_{s=1}^{250} (\hat{Y}_{ms} - \bar{Y}_{ms}) / 250^{-1} \sum_{s=1}^{250} \bar{Y}_{ms} \right\},$$

and the median of the relative root mean squared error,

$$RRMSE = 100 \cdot \text{med}_m \left\{ \sqrt{250^{-1} \sum_{s=1}^{250} (\hat{Y}_{ms} - \bar{Y}_{ms})^2} / 250^{-1} \sum_{s=1}^{250} \bar{Y}_{ms} \right\},$$

index  $s = 1, \dots, 250$  denotes the simulation run.

The results of the simulation are presented in Table 11-Table 14 and plotted in Figure 4 and Figure 5. The meaning of the labels used in the tables is listed below:

- "Direct" is the direct sample estimate;
- "EBLUP" is the estimate based on the nested-error regression model;
- "REBLUP(F)" is REBLUP using Fellner's method, "F+BC" is its bias-corrected version;
- "REBLUP(SR)" is REBLUP using the Sinha-Rao method, "SR+BC" is its bias-corrected version;
- "MQ" is the M-quantile based estimate, "MQ+BC" is its bias-corrected version;

- “N2(1)”, “N2(2)”, and “N2(3)” are estimates based on the mixture model using, respectively, Methods 1, 2 or 3 of the EM algorithm (as described in Section 3.3.1);
- ”N2(1)+BC1”, ”N2(2)+BC1”, ”N2(3)+BC1” are the BC1-corrected versions of N2(1), N2(2), and N2(3), respectively;
- “N2(1)+BC2”, “N2(2)+BC2”, “N2(3)+BC2” are the BC2-corrected versions of N2(1), N2(2), and N2(3), respectively;
- “N2(1)+OBC”, “N2(2)+OBC”, “N2(3)+OBC” are the overall bias corrections after the individual area corrections N2(1)+BC2, N2(2)+BC2, N2(3)+BC2, respectively;
- “N2(1)+OBC\*”, “N2(2)+OBC\*”, “N2(3)+OBC\*” are the overall bias corrections of N2(1), N2(2), N2(3) without making the area-level corrections first;
- “F+OBC” is the overall bias corrections for Fellner’s REBLUP.

First, consider scenarios (1)-(4) (see Table 11 and Table 12).

In the no-outliers situations (the [0,0] and [0,u]/1-36 columns), the N2 estimators work similar to the regular EBLUP. The BC2 and OBC versions of N2 did not lose much efficiency compared to the uncorrected N2.

If there are only individual outliers (the [e,0] and [e,u]/1-36 columns), all robust estimators work similarly and significantly better than EBLUP. Bias correction reduces the efficiency somewhat, although the BC2 versions of N2 work better than the versions that do not use the random occurrence test.



In the outlying areas only case (the [o,u]/37-40 column), N2 estimator performs similar to EBLUP or REBLUP, while the MQ estimator has a larger bias. The BC estimators reduce the biases of the respective estimators and the random occurrence test in the N2 case verifies that the areas are outliers and the corrections are necessary.

The N2 estimator has a large bias when both the individual and area outliers are present (the [e,u]/37-40 column). This bias is corrected in the N2+BC versions. The efficiency of the N2+BC versions in the four outlying areas is better than that of EBLUP but it is worse than the efficiency of REBLUP.

Overall, N2+OBC\* estimators work well, except for the outlying areas; N2+OBC versions work similar to N2+OBC\* when there is no area-level outliers and are better in the outlying areas. As noted earlier, we may consider testing on area outlyingness using the proposed test, then estimating the outlying areas outside the model.

Plots in Figure 4 show relative errors for each area in scenarios (1)-(4). The areas on the plots are sorted in ascending order of their sample sizes.

For scenario (5) (see Table 13), all N2 versions are better than the other estimators. The bias correction after the random occurrence test works much better than the other versions of the bias corrected estimators, although there is still certain loss in efficiency. If a similar situation happens in the CES data, then a version of the N2 estimators may be preferred. Relative errors for areas in scenario (5) are shown in Figure 5 (panel 5).

Results for scenarios 6-8 are presented in Table 14. In the [et,0] scenario, where the random errors are generated from the t-distribution, all robust estimators have similar performance and are more efficient than EBLUP or the direct estimator; the BC versions that are applied without the test are significantly less efficient than the original estimators. In the [0,ut] scenario, where the random effects are generated from the t-distribution, the N2 estimator may be biased for some areas. The bias corrected versions repair this deficiency. After the correction, N2 performs similar to REBLUP or EBLUP, there is no gain in efficiency compared to the non-robust version of EBLUP. In scenario [et,ut], where both random terms are generated from the t-distribution, REBLUP versions perform better than the other estimators. The BC2 versions of N2 correct for the bias in the outlying areas and are more efficient than EBLUP or MQ but they are less efficient than the REBLUP versions. Plots for the t-distribution scenarios are shown in Figure 5 (panels 6-8). It is evident from the simulations that the random occurrence test and subsequent bias correction is necessary for the N2 versions in scenarios (7) and (8).

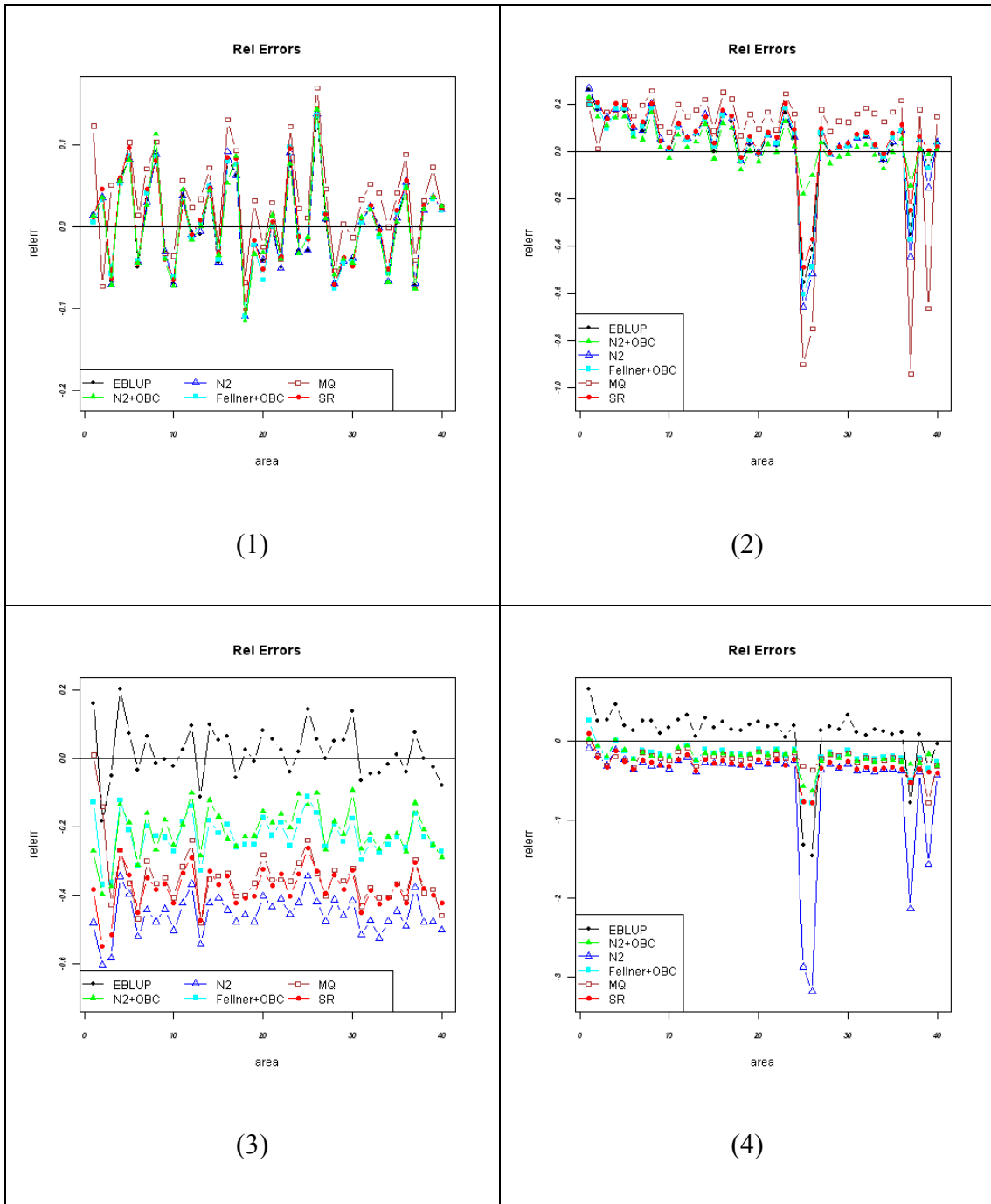


Figure 4. Relative errors for scenarios 1-4, areas are sorted in ascending order of the sample size: (1)  $[0,0]$  scenario; (2)  $[0,u]$  scenario; (3)  $[e,0]$  scenario; (4)  $[e,u]$  scenario.

Estimator / Scenario	<i>No outliers</i>		<i>Individual outliers only</i>		<i>Area outliers</i>	<i>Individual and area outliers</i>
	[0,0]	[0,u]/1-36	[e,0]	[e,u]/1-36	[0,u]/37-40	[e,u]/37-40
<b>Direct</b>	-0.05	-0.05	0.03	0.01	-0.04	0.11
<b>EBLUP</b>	0.00	0.06	0.01	0.18	-0.39	-1.06
<b>REBLUP (F)</b>	0.01	0.07	-0.38	-0.31	-0.43	-0.77
<b>REBLUP(SR)</b>	0.01	0.08	-0.38	-0.29	-0.31	-0.65
<b>MQ</b>	0.03	0.16	-0.36	-0.21	-0.83	-0.45
<b>N2(1)</b>	0.00	0.06	-0.46	-0.30	-0.48	-2.51
<b>N2(2)</b>	0.01	0.05	-0.45	-0.34	-0.35	-1.85
<b>N2(3)</b>	0.00	0.06	-0.45	-0.27	-0.40	-2.11
<b>F+BC</b>	0.00	0.01	-0.30	-0.28	-0.01	-0.23
<b>SR+BC</b>	0.00	0.02	-0.30	-0.27	-0.01	-0.21
<b>MQ+BC</b>	0.01	0.02	-0.28	-0.26	-0.10	-0.19
<b>N2(1)+BC1</b>	0.00	0.01	0.00	-0.01	0.01	0.06
<b>N2(1)+BC2</b>	0.01	0.06	-0.43	-0.29	-0.09	-0.54
<b>N2(2)+BC1</b>	0.00	0.01	0.00	0.00	0.01	0.06
<b>N2(2)+BC2</b>	0.01	0.05	-0.43	-0.33	-0.06	-0.35
<b>N2(3)+BC1</b>	0.00	0.01	0.00	-0.01	0.01	0.06
<b>N2(3)+BC2</b>	0.01	0.05	-0.43	-0.26	-0.01	-0.47
<b>N2(1)+OBC</b>	0.01	0.02	-0.22	-0.17	0.01	-0.43
<b>N2(1)+OBC*</b>	0.00	0.08	-0.23	0.11	-0.47	-2.12
<b>N2(2)+OBC</b>	0.01	0.02	-0.22	-0.21	-0.09	-0.24
<b>N2(2)+OBC*</b>	0.01	0.06	-0.22	0.01	-0.35	-1.53
<b>N2(3)+OBC</b>	0.01	0.02	-0.22	-0.17	0.05	-0.38
<b>N2(3)+OBC*</b>	0.00	0.06	-0.22	0.05	-0.40	-1.82
<b>F+OBC</b>	0.00	0.07	-0.23	-0.16	-0.43	-0.63

Table 11. Simulation results for scenarios 1-4 (250 runs) Median values of relative biases, expressed as a percentage.

<b>Estimator / Scenario</b>	<i>No outliers</i>		<i>Individual outliers only</i>		<i>Area outliers</i>	<i>Individual and area outliers</i>
	<b>[0,0]</b>	<b>[0,u]/1-36</b>	<b>[e,0]</b>	<b>[e,u]/1-36</b>	<b>[0,u]/37-40</b>	<b>[e,u]/37-40</b>
<b>Direct</b>	3.13	3.14	3.41	3.45	2.49	2.77
<b>EBLUP</b>	0.79	0.85	1.21	1.41	0.81	1.70
<b>REBLUP(F)</b>	0.82	0.82	0.97	0.97	0.84	1.10
<b>REBLUP(SR)</b>	0.83	0.84	0.98	0.97	0.82	1.02
<b>MQ</b>	0.83	0.85	0.99	0.98	1.44	1.21
<b>N2(1)</b>	0.80	0.85	0.98	0.95	0.92	3.56
<b>N2(2)</b>	0.81	0.87	1.00	1.00	0.81	2.81
<b>N2(3)</b>	0.79	0.85	0.98	0.96	0.81	3.09
<b>F+BC</b>	0.89	0.91	1.19	1.19	0.72	0.93
<b>SR+BC</b>	0.89	0.91	1.18	1.19	0.72	0.94
<b>MQ+BC</b>	0.89	0.90	1.19	1.20	0.77	1.02
<b>N2(1)+BC1</b>	0.89	0.91	1.64	1.62	0.72	1.33
<b>N2(1)+BC2</b>	0.82	0.85	1.11	0.96	0.75	1.54
<b>N2(2)+BC1</b>	0.89	0.91	1.64	1.64	0.72	1.33
<b>N2(2)+BC2</b>	0.84	0.87	1.10	1.02	0.75	1.36
<b>N2(3)+BC1</b>	0.89	0.91	1.63	1.65	0.72	1.33
<b>N2(3)+BC2</b>	0.83	0.86	1.10	0.97	0.73	1.47
<b>N2(1)+OBC</b>	0.82	0.85	1.05	0.93	0.76	1.49
<b>N2(1)+OBC*</b>	0.80	0.86	0.91	0.93	0.91	3.26
<b>N2(2)+OBC</b>	0.84	0.87	1.04	0.99	0.75	1.32
<b>N2(2)+OBC*</b>	0.81	0.87	0.93	0.96	0.81	2.58
<b>N2(3)+OBC</b>	0.83	0.86	1.04	0.95	0.74	1.44
<b>N2(3)+OBC*</b>	0.79	0.85	0.91	0.94	0.81	2.84
<b>F+OBC</b>	0.82	0.82	0.93	0.93	0.84	1.00

Table 12. Simulation results for scenarios (1)-(4). Median values of relative root mean squared errors, expressed as a percentage.

<i>Estimator</i>	<b>Med Rel Bias, %</b>	<b>Med Rel Root MSE, %</b>
<b>Direct</b>	0.02	6.75
<b>EBLUP</b>	0.05	2.92
<b>REBLUP (F)</b>	0.00	2.36
<b>REBLUP (SR)</b>	0.02	2.52
<b>MQ</b>	0.08	2.43
<b>N2(1)</b>	0.00	2.07
<b>N2(2)</b>	0.00	2.05
<b>N2(3)</b>	-0.01	2.09
<b>F+BC</b>	0.03	3.96
<b>SR+BC</b>	0.03	3.88
<b>MQ+BC</b>	0.03	4.14
<b>N2(1)+BC1</b>	0.01	5.72
<b>N2(1)+BC2</b>	0.00	2.20
<b>N2(2)+BC1</b>	0.05	5.73
<b>N2(2)+BC2</b>	0.00	2.18
<b>N2(3)+BC1</b>	0.02	5.72
<b>N2(3)+BC2</b>	-0.01	2.21
<b>N2(1)+OBC</b>	0.02	2.26
<b>N2(1)+OBC*</b>	0.03	2.15
<b>N2(2)+OBC</b>	0.03	2.22
<b>N2(2)+OBC*</b>	0.03	2.12
<b>N2(3)+OBC</b>	0.02	2.27
<b>N2(3)+OBC*</b>	0.02	2.18
<b>F+OBC</b>	0.02	2.40

Table 13. Simulation results for scenario 5, [70/30]. Median values of relative biases and relative root mean squared errors, expressed as a percentage.

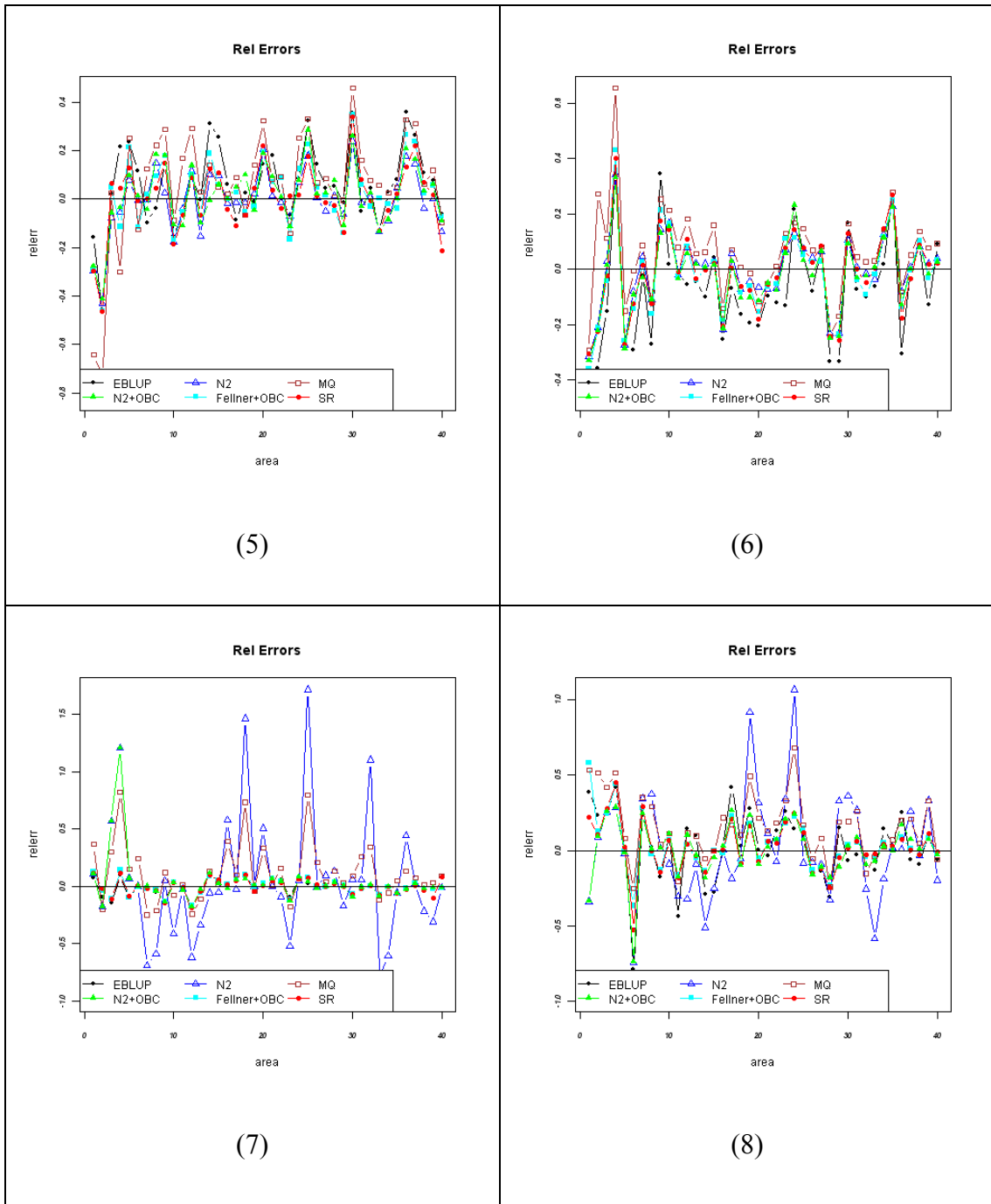


Figure 5. Relative errors for scenarios 5-8, areas are sorted in ascending order of the sample size: (5) [70/30] scenario (see Table 13); (6) [et,0] scenario; (7) [0,ut] scenario; (8) [et,ut] scenario.

<i>Estimator</i>	<b>Med Rel Bias, %</b>			<b>Med Rel Root MSE, %</b>		
	<b>[et,0]</b>	<b>[0, ut]</b>	<b>[et, ut]</b>	<b>[et,0]</b>	<b>[0, ut]</b>	<b>[et, ut]</b>
<b>Direct</b>	-0.03	-0.04	0.04	4.17	3.22	4.14
<b>EBLUP</b>	-0.07	-0.01	0.03	2.15	1.11	2.66
<b>REBLUP (F)</b>	-0.02	0.00	0.02	1.61	1.11	1.75
<b>REBLUP (SR)</b>	0.00	0.00	0.02	1.62	1.13	1.77
<b>MQ</b>	0.07	0.08	0.15	1.64	2.43	2.39
<b>N2(1)</b>	0.02	0.00	-0.03	1.65	4.32	3.87
<b>N2(2)</b>	0.01	-0.01	0.03	1.64	1.12	2.44
<b>N2(3)</b>	0.02	0.00	0.01	1.65	1.11	2.87
<b>F+BC</b>	-0.04	0.00	0.01	2.07	1.14	2.13
<b>SR+BC</b>	-0.03	0.00	0.01	2.08	1.14	2.13
<b>MQ+BC</b>	-0.01	0.02	0.03	2.06	1.51	2.34
<b>N2(1)+BC1</b>	-0.09	0.00	0.00	2.77	1.14	2.87
<b>N2(1)+BC2</b>	0.01	0.00	0.02	1.70	1.11	1.96
<b>N2(2)+BC1</b>	-0.09	-0.01	0.01	2.77	1.14	2.87
<b>N2(2)+BC2</b>	0.00	-0.01	0.00	1.66	1.11	2.06
<b>N2(3)+BC1</b>	-0.09	-0.01	0.01	2.77	1.14	2.87
<b>N2(3)+BC2</b>	0.00	-0.01	0.02	1.68	1.11	2.09
<b>N2(1)+OBC</b>	-0.01	0.00	0.02	1.71	1.11	1.96
<b>N2(1)+OBC*</b>	0.01	0.02	-0.02	1.65	4.31	3.87
<b>N2(2)+OBC</b>	-0.01	-0.01	-0.01	1.66	1.11	2.06
<b>N2(2)+OBC*</b>	-0.01	-0.01	0.03	1.63	1.12	2.44
<b>N2(3)+OBC</b>	-0.01	-0.01	0.02	1.69	1.11	2.09
<b>N2(3)+OBC*</b>	0.00	0.00	0.01	1.65	1.11	2.86
<b>F+OBC</b>	-0.02	0.00	0.02	1.62	1.11	1.76

Table 14. Simulation results for scenarios 6-8. Median values of relative biases and relative root mean squared errors, expressed as a percentage.



### 3.7 Properties of the parameter estimates when the number of areas increases

The simulation study of this subsection is designed to explore how the parameter estimates change with the increased number of areas. Consistent estimators would tend to true parameter values for a given model.

Consider three cases where the number of areas is  $M=20, 40, \text{ or } 60$ . The number of sample units in each area is  $n_m = 5$ . Similar to the simulation setup in the previous section, the auxiliary variable  $x_{mj}$  is generated from the lognormal distribution with mean 1.004077 and standard deviation of 0.5 and the sample values  $y_{mj}$  are generated as  $y_{mj} = 100 + 5x_{mj} + u_m + \varepsilon_{mj}$ . Random effects are  $u_m \sim N(0, 9)$ . Random errors are distributed as  $\varepsilon_{mj} \sim (1 - p)N(0, 9) + pN(0, 900)$ . Consider two scenarios for the portion of observations with larger variance (1)  $p = 0.03$  or (2)  $p = 0.30$ .

We considered the N2 estimators based on the three versions of the EM algorithm described in Section 3.3.1. Table 15 and Table 16 report mean estimates and the simulation standard errors (in parentheses) of the estimators of the parameters based on 1000 Monte Carlo iterations.

When  $p = 0.03$ , the general conclusion is that, as the number of areas increases, the standard error decreases and the estimates tend to the values of the parameters. For the larger fraction,  $p = 0.30$ , there is a considerable bias in the estimate of  $\tau^2$  and we cannot claim that it reduces when the number of areas increases. The bias is smaller for the Method 2 of N2.

	$\beta_1$	$\beta_2$	$\sigma_1^2$	$\sigma_2^2$	$\tau^2$	$p$
<i>True values</i>	100	5	9	900	9	0.03
<b>M=20</b>						
<b>N2 (1)</b>	100.00 (1.01)	5.01 (0.22)	9.21 (1.71)	835.05 (947.23)	7.93 (3.20)	0.07 (0.07)
<b>N2 (2)</b>	100.00 (1.10)	5.00 (0.22)	8.36 (1.65)	836.82 (956.08)	9.08 (3.35)	0.07 (0.07)
<b>N2(3)</b>	99.99 (1.01)	5.01 (0.22)	9.39 (1.75)	797.43 (937.97)	7.88 (3.24)	0.08 (0.08)
<b>M=40</b>						
<b>N2 (1)</b>	99.99 (0.70)	5.00 (0.15)	9.41 (1.23)	865.26 (674.41)	8.01 (2.40)	0.05 (0.04)
<b>N2 (2)</b>	100.01 (0.75)	5.00 (0.16)	8.52 (1.16)	864.31 (673.42)	9.05 (2.48)	0.04 (0.04)
<b>N2(3)</b>	99.99 (0.70)	5.00 (0.15)	9.49 (1.24)	809.44 (662.05)	7.93 (2.43)	0.05 (0.05)
<b>M=60</b>						
<b>N2 (1)</b>	100.02 (0.57)	5.00 (0.12)	9.56 (1.06)	925.69 (598.49)	8.14 (1.86)	0.04 (0.03)
<b>N2 (2)</b>	100.01 (0.61)	5.00 (0.12)	8.70 (1.00)	924.37 (598.76)	9.07 (1.91)	0.04 (0.03)
<b>N2 (3)</b>	100.02 (0.57)	5.00 (0.12)	9.58 (1.04)	856.85 (590.87)	8.04 (1.89)	0.04 (0.04)

Table 15. Mean estimates and the simulation standard errors (in parentheses) for scenario with  $p = 0.03$  (a method used in the EM algorithm is indicated in parentheses next to N2)

	$\beta_1$	$\beta_2$	$\sigma_1^2$	$\sigma_2^2$	$\tau^2$	$p$
<b>True values</b>	100	5	9	900	9	0.30
<b>M=20</b>						
<b>N2 (1)</b>	100.00 (1.35)	5.00 (0.34)	12.64 (3.92)	917.03 (268.32)	5.36 (2.83)	0.30 (0.06)
<b>N2 (2)</b>	100.03 (1.38)	5.00 (0.34)	10.92 (3.34)	916.76 (268.04)	7.47 (3.34)	0.30 (0.06)
<b>N2 (3)</b>	100.00 (1.36)	5.00 (0.35)	13.58 (4.21)	904.33 (266.21)	4.25 (2.86)	0.30 (0.06)
<b>M=40</b>						
<b>N2 (1)</b>	100.00 (0.87)	4.99 (0.23)	12.52 (2.51)	902.09 (190.51)	5.50 (2.10)	0.30 (0.04)
<b>N2 (2)</b>	100.01 (0.90)	4.99 (0.23)	10.79 (2.15)	902.07 (190.40)	7.46 (2.41)	0.30 (0.04)
<b>N2 (3)</b>	100.00 (0.88)	4.99 (0.23)	13.52 (2.73)	892.00 (189.15)	4.36 (2.12)	0.30 (0.04)
<b>M=60</b>						
<b>N2(1)</b>	99.99 (0.73)	5.01 (0.18)	12.54 (2.07)	916.64 (153.37)	5.54 (1.62)	0.30 (0.04)
<b>N2(2)</b>	99.99 (0.74)	5.00 (0.18)	10.80 (1.73)	915.74 (152.97)	7.43 (1.89)	0.30 (0.04)
<b>N2 (3)</b>	99.99 (0.74)	5.01 (0.18)	13.55 (2.23)	905.57 (152.08)	4.42 (1.63)	0.30 (0.04)

Table 16. Mean estimates and the simulation standard errors (in parentheses) for scenario with  $p = 0.30$  (a method used in the EM algorithm is indicated in parentheses next to N2)

### 3.8 Simulations for prediction confidence intervals using the parametric bootstrap

The simulation setup is similar to the one described in previous subsections. There are  $M = 40$  areas. The auxiliary variable  $x_{mj}$  is generated from the lognormal distribution with mean 1.004077 and standard deviation of 0.5 and the population values  $y_{mj}$  are generated as  $y_{mj} = 100 + 5x_{mj} + u_m + \varepsilon_{mj}$ . We used several patterns for the random terms:

- [0,0] pattern (no outliers): individual errors are  $\varepsilon_{mj} \sim N(0,6)$ ; random effects are  $u_m \sim N(0,3)$ ;
- [e0,0] pattern (individual outliers, symmetrical distribution): individual errors are  $\varepsilon_{mj} \sim N(0,6)$  with probability 0.97 and  $\varepsilon_{mj} \sim N(0,150)$  with probability 0.03; random effects are  $u_m \sim N(0,3)$ ;
- [70/30] pattern (individual outliers, symmetrical distribution, large fraction of the part 2 mixture units): individual errors are  $\varepsilon_{mj} \sim N(0,9)$  with probability 0.70 and  $\varepsilon_{mj} \sim N(0,900)$  with probability 0.30; random effects are  $u_m \sim N(0,9)$ .

Each area contains 1000 population units from which 5 units are selected using simple random sampling without replacement.

We used 100 simulated populations and corresponding samples. For each simulation run, we obtained 100 bootstrap estimates for each area. The 95% confidence intervals were constructed from the bootstrap pivots in all 40 areas.

Two alternative models were used: a nested error regression model (denoted EBLUP) and the N2 mixture model without bias correction. The results are summarized in Table 17.

<i>Scenario</i>	<i>EBLUP</i>	<i>N2</i>
<b>[0,0]</b>	95.0 (3.9)	94.8 (3.7)
<b>[e<sub>0</sub>,0]</b>	94.7 (5.1)	95.9 (4.1)
<b>[70/30]</b>	59.5 (10.7)	96.2 (7.7)

Table 17. Average coverage and median length of confidence intervals (nominal coverage 95%) using the NER model and the N2 mixture model, for different population patterns

Both models work well for the [0,0] (no outliers) scenario. In the other two scenarios, the confidence intervals based on the N2 model give approximately the nominal coverage. We encountered problems with estimation of parameters for EBLUP: in large percentage of the simulation runs, the NER model produced zeros for the variance of the random effects term. To avoid the appearance of zeros, we replaced zeros in variance by 0.0001. The length of the bootstrap intervals for EBLUP version is very unstable, and the result depends on the value we chose to replace the zero variances.

### 3.9 Linearization of a finite population target in small area estimation, with application to the CES survey

In order to apply a unit level model, when a target has a predefined form, we need to linearize the target population quantity, similar to the way discussed in Section 2.1 of Chapter 2. In this section, we first obtain linearization in the case of small areas, in general terms. Then we apply the method to estimation of the relative change in employment for small areas in CES.

There are two related purposes in linearizing a target quantity in small area context. First, it provides a way of formulating a small area model at the unit level. Second,

the form of a target finite population quantity, by the means of its influence function, dictates what observations are to be considered influential. Thus, the structure of the unit-level data is determined by the form of the target population parameter of interest. The role of the model is to provide a useful description of this structure.

To estimate a pre-defined target using a sample of a limited size, it is possible to use an area-level SAE model. To do this, one would first derive an estimate using the sample and then stabilize this direct sample estimate by applying an area-level method. In many situations, however, it is preferable to formulate a model at the unit level. If the unit-level auxiliary information is available, modeling incorporating such information can be especially beneficial. However, there are reasons to consider a unit-level modeling even in the absence of such auxiliary data. The direct sample-based estimates can be affected by influential observations. In such a case, using a model that is robust to the unit level outliers may be beneficial.

In the area level Fay-Herriot model, variances of the direct sample based estimates are considered to be known. In practice, some sort of a generalized variance function is used to supply the variances of the direct estimates. However, these smoothed variances do not always properly reflect the possibility that a particular realized sample contains extreme observations. If this happens, the harmful effect that such units have on the direct sample estimate carries over onto the resulting area-level model estimates.

Assume a vector of population measurements  $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_{N_m})$  in area  $m$  is a realization from a superpopulation distribution (each  $\mathbf{y}_j$  can be a vector of

measurements on a unit  $j$ ). Denote by  $F_{N_m}$  the empirical distribution function (edf) of the finite population in area  $m$ . The finite population quantity of interest is some  $T(F_{N_m})$ , it is assumed to be sufficiently regular to be linearized near  $F_m$ , the ideal distribution in area  $m$ , using a Taylor expansion. Similar to (2.1.1), we write

$$T(F_{N_m}) = T(F_m) + N_m^{-1} \sum_{j=1}^{N_m} IF_{F_m, T}(\mathbf{y}_j) + R_{N_m} \quad (3.9.1)$$

where  $T(F_m)$  is a superpopulation parameter and  $IF_{F_m, T}(\mathbf{y}_j)$  is the influence function of the functional  $T$ . As in Section 2.1, let us drop the remainder term of (3.9.1) and redefine the finite population target as

$$\tilde{T}(F_{N_m}) = T(F_m) + N_m^{-1} \sum_{j=1}^{N_m} IF_{F_m, T}(\mathbf{y}_j). \quad (3.9.2)$$

Given the *population* size  $N_m$  in area  $m$  is large, the remainder term is negligible, and this quantity is different from the ideal target by a small value.

Of course,  $T(F_m)$  in (3.9.2) is not known. If the sample is large enough, one could simply use a sample based estimate in its place.

In small domains, however, the direct sample estimator is not reliable. It is usual in small area estimation to make assumptions about proximity of the area levels to the aggregation of areas. Let  $F$  denote the distribution function of population measurements in the aggregation of areas and let us assume that  $T(F_{N_m})$  can be expanded in the neighborhood of  $F$  as

$$T(F_{N_m}) = T(F) + c_m^{-1} N_m^{-1} \sum_{j=1}^{N_m} IF_{F,T}(\mathbf{y}_j) + \tilde{R}_{N_m}, \quad (3.9.3)$$

for some  $c_m$  such that  $\sum_{m=1}^M p_m c_m = 1$ ,  $p_m = N_m/N$ ,  $N = \sum_{m=1}^M N_m$ ;  $\tilde{R}_{N_m}$  is a remainder term.

In general, we can make a supposition about the closeness of  $F_m$  to  $F$  by assuming that the remainder term is small. Then, similar to (3.9.2), we can redefine the target population parameter by dropping the remainder term:

$$\tilde{T}(F_{N_m}) = T(F) + c_m^{-1} N_m^{-1} \sum_{j=1}^{N_m} IF_{F,T}(\mathbf{y}_{mj}) \quad (3.9.4)$$

In what follows, we consider a particular case by setting  $c_m = 1$ .

Since  $T(F)$  is defined on aggregation of areas, it can be estimated from the sample with satisfactory precision. Let  $\hat{T}(F_N)$  denote an estimate of  $T(F)$ . The estimator of  $\tilde{T}(F_{N_m})$  takes the form

$$\hat{\tilde{T}}(F_{N_m}) = \hat{T}(F_N) + f_m \frac{1}{n_m} \sum_{j=1}^n \hat{u}_j + (1-f_m) \frac{E_S[(w_j - 1)\hat{u}_j | j \in S_m]}{E_S[w_j - 1 | j \in S_m]}, \quad (3.9.5)$$

where  $\hat{u}_j$  is an estimate of  $IF_{F,T}(\mathbf{y}_i)$ , it depends on the estimate  $\hat{T}(F_N)$ .

We next consider the application to CES.

In CES, the goal is to estimate the relative over-the-month change in employment at a given month  $t$  in areas  $m=1, \dots, M$ , where the areas are formed by cross-classifying



industries and metropolitan statistical areas (MSA). For area  $m$ , the target finite population quantity at month  $t$  is

$$R_{m,t} = \frac{\sum_{j \in P_{m,t}} y_{mj,t}}{\sum_{j \in P_{m,t}} y_{mj,t-1}}, \quad (3.9.6)$$

where  $P_{m,t}$  is a set of the area  $m$  population establishments having non-zero employment in both previous and current months, i.e.,  $y_{mj,t-1} > 0$  and  $y_{mj,t} > 0$ . The direct sample estimate is

$$\hat{R}_{m,t} = \frac{\sum_{j \in S_{m,t}} w_{mj} y_{mj,t}}{\sum_{j \in S_{m,t}} w_{mj} y_{mj,t-1}}, \quad (3.9.7)$$

where  $S_{m,t}$  is a set of the area  $m$  sample establishments having  $y_{mj,t-1} > 0$  and  $y_{mj,t} > 0$ ;  $w_{mj}$  is the sample weight for unit  $mj$ .

Assume the set of finite population observations at month  $t$

$\left\{ (y_{j,t-1}, y_{j,t}) \mid j \in P_t = \bigcup_{m=1}^M P_{m,t} \right\}$  to be independent realizations of a random vector

$(Y_{t-1}, Y_t)$  having a probability distribution  $F$ ; let  $(\theta_{t-1}, \theta_t)$  be a vector of

superpopulation means of  $(Y_{t-1}, Y_t)$ . The population measurements in area  $m$ ,

$\left\{ (y_{mj,t-1}, y_{mj,t}) \mid j \in P_{m,t} \right\}$  are independent realizations of a random vector  $(Y_{m,t-1}, Y_{m,t})$

with the probability distribution  $F_m$ . The superpopulation parameter of interest is a

function of the superpopulation means  $(\theta_{m,t-1}, \theta_{m,t})$ :

$$T(F_m) = T(\theta_{m,t-1}, \theta_{m,t}; F_m) = \frac{\theta_{m,t}}{\theta_{m,t-1}}$$

For  $\hat{T}(F_N)$  involved in formula (3.9.5), we use the survey weighted estimator

$$\hat{R}_t = \frac{\hat{\theta}_t}{\hat{\theta}_{t-1}} = \frac{\sum_{j \in S_t} w_j y_{j:t}}{\sum_{j \in S_{t-1}} w_j y_{j:t-1}} \quad (3.9.8)$$

based on the aggregation from all areas. The number of population units having nonzero employment in two consecutive months is not known and is estimated as

$\hat{N}_m = \sum_{j \in S_{m,t}} w_j$ , the sampling fraction is estimated as

$$\hat{f}_m = \frac{n_m}{\hat{N}_m} \quad (3.9.9)$$

Applying formula (3.9.5), we derive the following variable

$$y_{mj,t}^* = (1 - \hat{f}_m) \frac{(w_{mj} - 1) \hat{v}_{mj,t}}{\hat{w}_m - 1} + \hat{R}_t + \hat{f}_m \hat{v}_{m,t}, \quad (3.9.10)$$

where  $\hat{R}_t$  is the estimated ratio of employment at a statewide level;

$\hat{v}_{mj,t} = \hat{Y}_{t-1}^{-1}(y_{mj,t} - \hat{R}_t y_{mj,t-1})$  is the estimated influence function for the ratio;  $\hat{Y}_{t-1}$  is an

estimate of the previous month mean statewide employment;  $\hat{w}_m = n_m^{-1} \sum_{j \in S_{m,t}} w_{mj}$  is an

area  $m$  average weight;  $\hat{v}_{m,t} = n_m^{-1} \sum_{j \in S_{m,t}} \hat{v}_{mj,t}$ .

In this study, historical administrative data from the Quarterly Census of Employment and Wages (QCEW) program of the U.S. Bureau of Labor Statistics played the role of

the “real” data. (In real time production, the estimates are based on the data collected by CES.)

We compared performances of several estimators: one estimator is based on the area-level Fay-Herriot model and the other estimators are based on different unit-level models. We used the single slope, without intercept linear models, with the past year’s population trend  $R_{m,t-12}$  playing the role of an auxiliary variable (i.e., area-level auxiliary information for all observations in area  $m$ ).

We made estimates of the relative employment change in September 2006 for four States (Alabama, California, Florida, and Pennsylvania); the sample was drawn from the 2005 sampling frame, which mimics the production timeline. We fit the models separately for each State’s industrial supersector: a set of MSAs within States’ industrial supersectors defined the set of small areas. The resulting estimates were compared to the corresponding true population values  $R_{m,t}$  available from QCEW.

Performances of the estimators were measured using the 75th percentile of the absolute error  $E_{m,t} = 100 \left| \hat{R}_{m,t} - R_{m,t} \right|$  and the empirical root mean squared error

$$ERMSE_t = \left[ M^{-1} \sum_{m=1}^M E_{m,t}^2 \right]^{\frac{1}{2}}.$$

Summaries of results for each State are reported in Table 18 - Table 25.

The meanings of the column labels are as follows:

- “Dir” is the direct sample estimate;
- “FH” is the Fay-Herriot model based estimate;

- “NER” is the estimate based on the nested-error regression model;
- “F” is REBLUP using Fellner’s method, “FBC” is its bias-corrected version;
- “MQ” is the M-quantile based estimate, “MQBC” is its bias-corrected version;
- ”N2BC1” is the BC1-corrected N2, “N2BC2” is the BC2-corrected N2;
- “N2OBC\*” is the overall bias correction of N2 without making the area-level corrections first;
- “N2OBC” is the overall bias correction after the individual area corrections.

We used Method 1 of the EM algorithm (see Section 3.3.1) for estimation in N2.

The direct estimator does not perform well in comparison with the other estimators. So the use of a model is well warranted. In all states except Pennsylvania, the robust estimators outperform the FH or the NER-based EBLUP. Overall, the performance of N2 is close to the Fellner’s version of REBLUP. In Alabama and Florida, the N2 estimator is more efficient than the other estimators both in terms of ERMSE and the 75<sup>th</sup> percentile. In California, ERMSEs of REBLUP and MQ are smaller than of N2 but, in terms of the 75th percentile, these estimators are very close. In Pennsylvania, in several industries, the N2 estimator had a large error due to the overall bias, but the OBC versions helped reduce the bias.

Ind	Dir	FH	NER	F	MQ	N2	FBC	MQBC	N2BC1	N2BC2	N2OBC*	N2OBC
20	6.74	1.72	2.68	1.95	1.69	1.26	2.44	2.34	4.61	1.26	1.25	1.25
31	1.09	1.06	0.91	0.73	0.68	0.78	1.16	1.14	1.25	1.21	0.78	1.14
32	1.15	1.30	1.10	0.71	0.82	1.60	1.01	1.03	1.52	1.06	1.60	0.99
41	3.57	1.57	1.04	1.35	2.05	1.18	1.18	1.87	1.93	1.18	1.18	1.18
42	1.19	1.36	1.41	0.77	0.95	1.00	0.74	0.77	1.19	1.00	0.97	0.97
43	1.88	1.65	1.76	1.72	1.64	1.71	1.75	1.68	1.91	1.71	1.67	1.67
50	2.28	1.55	1.14	1.57	2.20	1.38	1.65	2.34	1.30	1.38	1.30	1.30
55	2.20	1.97	2.23	1.29	1.21	1.24	1.25	1.22	2.03	1.24	1.19	1.19
60	1.59	2.16	2.16	0.97	1.30	0.88	0.93	0.97	1.48	0.88	0.88	0.88
65	1.41	1.17	1.04	0.60	0.69	0.71	0.60	0.59	1.37	0.71	0.68	0.68
70	4.31	1.02	1.02	0.94	0.90	0.93	1.39	1.40	3.49	0.93	0.93	0.93
80	9.53	3.97	5.72	5.17	5.29	4.84	5.34	5.44	6.86	4.84	4.84	4.84
<b>Overall</b>	3.98	1.87	2.26	1.90	2.02	1.80	2.03	2.13	2.93	1.79	1.79	1.77

Table 18. Alabama, by Industry, Empirical Root Mean Squared Error

Ind	Dir	FH	NER	F	MQ	N2	FBC	MQBC	N2BC1	N2BC2	N2OBC*	N2OBC
20	3.72	1.69	1.96	2.03	2.04	1.22	1.89	2.03	2.98	1.22	1.25	1.25
31	1.24	1.27	1.12	0.75	0.75	0.94	1.10	1.10	1.10	0.98	0.95	0.88
32	1.30	1.01	0.70	0.73	0.84	0.82	0.61	0.52	0.98	0.82	0.95	0.92
41	2.68	1.66	1.14	1.56	2.21	1.46	1.50	1.80	1.92	1.46	1.47	1.47
42	1.09	1.36	1.54	0.67	0.94	1.26	0.53	0.61	1.07	1.26	1.27	1.27
43	1.83	2.01	2.18	2.08	1.84	2.09	2.08	1.93	2.40	2.09	2.01	2.01
50	2.42	1.68	1.29	1.89	2.01	1.67	1.93	2.33	1.47	1.67	1.68	1.68
55	2.17	2.26	2.57	1.73	1.61	1.57	1.70	1.70	2.51	1.57	1.50	1.50
60	1.29	2.46	2.32	0.85	1.29	1.17	0.82	1.10	1.51	1.17	1.16	1.16
65	1.26	1.57	1.42	0.46	0.72	0.76	0.36	0.33	1.09	0.76	0.74	0.74
70	2.19	1.09	1.09	1.06	0.93	0.94	1.56	1.57	2.30	0.94	0.97	0.97
80	9.44	3.13	5.35	3.71	3.13	2.85	4.14	3.90	9.07	2.85	2.84	2.84
<b>Overall</b>	2.14	1.84	1.70	1.38	1.46	1.35	1.67	1.69	1.99	1.39	1.30	1.31

Table 19. Alabama, by Industry, 75th Percentile Absolute Error

Ind	Dir	FH	NER	F	MQ	N2	FBC	MQBC	N2BC1	N2BC2	N2OBC*	N2OBC
20	3.99	1.90	1.86	1.64	1.45	1.69	1.94	1.99	3.41	1.68	1.66	1.66
31	3.09	1.70	1.76	1.51	1.82	1.72	1.99	2.23	2.22	1.72	1.72	1.72
32	7.74	4.95	4.80	4.03	4.12	5.62	3.98	3.91	4.88	5.06	5.60	5.07
41	3.64	2.01	2.66	1.42	1.03	1.36	1.46	1.01	4.86	1.36	1.35	1.35
42	2.06	0.99	1.05	0.55	0.65	0.64	0.74	0.81	1.59	0.64	0.64	0.64
43	8.26	4.85	4.05	3.21	3.66	2.62	4.16	4.46	5.03	2.57	2.61	2.57
50	2.65	1.52	1.38	1.14	0.95	1.25	1.18	1.01	1.36	1.25	1.24	1.24
55	3.11	2.11	1.20	0.88	0.84	0.89	0.90	0.78	4.38	0.89	0.88	0.88
60	2.22	1.55	1.53	1.24	1.25	1.58	1.69	1.68	2.95	2.52	1.57	2.52
65	2.41	1.69	1.24	0.92	0.80	0.90	0.89	0.86	2.23	0.90	0.90	0.90
70	2.26	1.23	1.38	1.25	1.19	1.15	1.63	1.62	2.22	1.15	1.16	1.16
80	5.53	1.65	1.68	3.69	2.84	1.62	4.12	3.87	5.24	4.07	1.62	4.07
<b>Overall</b>	4.40	2.50	2.34	2.10	2.04	2.15	2.39	2.38	3.64	2.36	2.14	2.36

Table 20. California, by Industry, Empirical Root Mean Squared Error, %

Ind	Dir	FH	NER	F	MQ	N2	FBC	MQBC	N2BC1	N2BC2	N2OBC*	N2OBC
20	4.31	1.47	1.57	1.53	1.65	1.58	1.88	1.87	3.17	1.58	1.41	1.45
31	2.90	1.23	1.42	1.19	0.70	1.49	1.39	1.37	1.41	1.49	1.49	1.49
32	1.87	2.71	2.57	3.14	3.14	3.50	2.80	2.80	2.80	3.30	3.50	3.32
41	3.23	1.95	1.94	1.21	1.42	1.08	1.16	1.12	3.78	1.08	1.08	1.08
42	1.56	0.75	1.00	0.68	0.70	0.79	0.63	0.66	1.15	0.79	0.74	0.74
43	3.26	2.54	2.12	1.17	1.96	1.17	1.23	1.23	2.17	1.17	1.17	1.18
50	1.94	1.42	1.48	1.23	1.04	1.41	1.27	1.28	1.42	1.41	1.39	1.39
55	1.55	1.26	1.29	0.82	0.93	0.85	0.75	0.79	1.71	0.85	0.85	0.85
60	2.26	0.98	1.04	1.14	1.08	1.10	1.13	1.12	2.56	1.10	1.08	1.08
65	1.58	1.52	0.71	0.74	0.64	0.67	0.85	0.88	1.45	0.67	0.71	0.71
70	2.84	1.31	1.78	1.47	1.43	1.28	2.10	2.06	2.59	1.28	1.28	1.28
80	6.22	1.48	1.84	1.44	1.29	1.35	1.48	1.38	5.00	1.31	1.35	1.31
<b>Overall</b>	3.01	1.57	1.61	1.23	1.23	1.20	1.35	1.28	2.61	1.19	1.21	1.19

Table 21. California, by Industry, 75th Percentile Absolute Error

Ind	Dir	FH	NER	F	MQ	N2	FBC	MQBC	N2BC1	N2BC2	N2OBC*	N2OBC
20	2.81	1.33	1.87	1.25	1.13	0.86	1.48	1.45	2.87	0.86	0.87	0.87
31	2.85	1.44	1.30	1.14	2.09	1.14	1.15	2.10	1.45	1.14	1.14	1.14
32	3.91	2.10	9.14	7.11	9.77	1.99	7.10	9.76	1.77	1.99	1.98	1.98
41	4.98	3.70	1.65	1.16	1.05	1.09	1.23	1.10	6.58	1.09	1.09	1.09
42	0.71	0.59	0.59	0.43	0.42	0.27	0.35	0.36	0.60	0.29	0.27	0.29
43	4.27	1.96	1.51	1.27	1.76	1.36	1.91	2.18	1.75	1.36	1.36	1.36
50	10.61	10.16	2.94	5.75	4.09	1.52	5.74	4.11	1.59	1.52	1.54	1.54
55	2.45	1.04	0.97	0.75	1.13	0.77	0.80	1.07	2.09	0.77	0.77	0.77
60	2.25	0.87	0.93	0.74	0.71	0.67	0.74	0.70	1.14	0.67	0.68	0.70
65	1.84	0.77	0.67	0.49	0.46	0.56	0.83	0.82	1.78	0.56	0.52	0.52
70	3.65	1.78	0.93	0.86	1.02	0.91	1.67	1.76	3.99	1.28	0.89	1.28
80	8.21	3.61	1.16	3.70	8.09	1.04	4.01	8.23	7.04	1.04	1.05	1.05
<b>Overall</b>	4.79	3.42	2.71	2.77	3.77	1.08	2.89	3.85	3.42	1.11	1.08	1.11

Table 22. Florida, by Industry, Empirical Root Mean Squared Error

Ind	Dir	FH	NER	F	MQ	N2	FBC	MQBC	N2BC1	N2BC2	N2OBC*	N2OBC
20	2.87	1.25	2.01	1.15	0.91	0.74	1.26	1.27	2.95	0.74	0.80	0.80
31	2.79	1.70	1.42	1.53	1.56	1.40	1.53	1.55	1.57	1.40	1.41	1.41
32	3.16	2.61	5.27	4.18	5.67	2.34	4.17	5.41	1.94	2.34	2.34	2.34
41	2.51	2.53	1.51	0.86	1.13	0.85	0.84	1.01	3.48	0.85	0.86	0.86
42	0.85	0.53	0.54	0.37	0.33	0.22	0.28	0.31	0.74	0.22	0.22	0.22
43	3.47	1.33	1.75	1.14	1.44	1.11	1.51	1.57	1.72	1.11	1.11	1.11
50	2.81	2.68	2.64	2.52	2.04	1.28	2.52	2.72	1.60	1.28	1.31	1.31
55	2.15	1.26	1.11	0.66	0.74	0.68	0.68	0.66	1.65	0.68	0.67	0.67
60	1.59	1.02	1.12	0.92	0.85	0.80	0.81	0.83	1.46	0.80	0.82	0.85
65	1.42	0.75	0.60	0.55	0.51	0.47	0.66	0.59	1.43	0.47	0.59	0.59
70	2.54	1.77	1.15	0.85	0.92	1.06	1.64	1.49	2.07	1.09	0.87	1.02
80	4.89	2.45	1.43	1.47	1.64	1.14	1.54	1.79	2.80	1.14	1.17	1.17
<b>Overall</b>	2.54	1.59	1.32	1.07	1.13	1.02	1.30	1.34	1.81	1.02	0.91	0.94

Table 23. Florida, by Industry, 75th Percentile Absolute Error

Ind	Dir	FH	NER	F	MQ	N2	FBC	MQBC	N2BC1	N2BC2	N2OBC*	N2OBC
20	4.95	1.29	1.18	1.48	1.83	1.99	2.45	2.51	4.34	1.99	1.24	1.24
31	2.20	0.77	0.78	0.74	0.91	0.74	1.76	1.88	2.41	0.74	0.71	0.71
32	2.31	1.07	2.46	1.06	1.19	0.95	1.31	1.47	1.47	0.95	0.95	0.95
41	2.42	0.66	0.63	0.84	0.71	0.77	1.03	0.98	1.73	0.77	0.73	0.73
42	1.73	0.62	0.42	0.74	0.54	0.52	0.97	0.94	1.33	0.86	0.52	0.85
43	5.25	1.52	1.59	4.34	3.80	4.12	4.53	4.53	4.90	4.12	3.38	3.38
50	1.85	1.21	1.13	1.03	1.52	1.11	1.03	1.53	1.29	1.11	1.08	1.08
55	4.15	2.90	0.94	0.83	0.91	1.02	0.87	1.02	2.46	1.02	0.89	0.89
60	2.59	1.16	0.99	0.88	0.98	0.97	1.23	1.25	2.38	0.97	0.92	0.92
65	1.28	0.49	0.54	0.64	0.67	0.63	0.79	0.80	1.27	0.62	0.58	0.60
70	3.29	1.54	1.43	2.71	2.29	2.00	3.44	3.31	3.43	2.00	2.07	2.07
80	5.66	1.91	2.19	1.56	1.56	1.52	1.76	1.77	5.97	1.52	1.52	1.52
<b>Overall</b>	3.46	1.42	1.32	1.75	1.66	1.67	2.09	2.13	3.15	1.68	1.44	1.46

Table 24. Pennsylvania, by Industry, Empirical Root Mean Squared Error

Ind	Dir	FH	NER	F	MQ	N2	FBC	MQBC	N2BC1	N2BC2	N2OBC*	N2OBC
20	4.68	1.47	1.23	1.53	1.38	2.44	3.19	3.20	4.27	2.44	1.48	1.48
31	1.35	0.69	0.72	0.57	0.91	0.51	1.06	1.49	1.02	0.51	0.60	0.60
32	1.86	0.79	1.96	1.32	1.31	0.81	1.01	1.75	1.77	0.81	0.69	0.69
41	3.19	0.56	0.53	0.74	0.74	0.90	0.89	0.91	1.24	0.90	0.85	0.85
42	1.19	0.59	0.60	0.47	0.35	0.50	0.72	0.69	1.03	0.50	0.50	0.50
43	5.72	1.66	1.73	4.62	4.26	5.26	5.11	4.88	5.69	5.26	4.41	4.41
50	2.20	1.60	1.28	1.17	1.32	1.35	1.21	1.40	1.48	1.35	1.28	1.28
55	2.97	2.46	0.90	0.96	1.03	1.03	0.72	1.18	2.13	1.03	0.85	0.85
60	2.63	1.15	1.20	0.96	1.22	0.90	1.45	1.40	2.51	0.90	1.06	1.06
65	1.47	0.53	0.60	0.73	0.83	0.83	0.90	0.91	1.28	0.83	0.77	0.81
70	3.87	1.62	1.09	2.90	2.54	2.24	3.82	3.76	3.82	2.24	1.78	1.78
80	6.01	1.23	1.83	1.27	1.43	0.99	2.38	2.26	5.81	0.99	0.95	0.95
<b>Overall</b>	3.39	1.17	1.19	1.28	1.31	1.39	1.67	1.73	2.49	1.40	1.11	1.12

Table 25. Pennsylvania, by Industry, 75th Percentile Absolute Error



Examples of the distribution of errors across areas are given in the plots below (see Figure 6 and Figure 7).

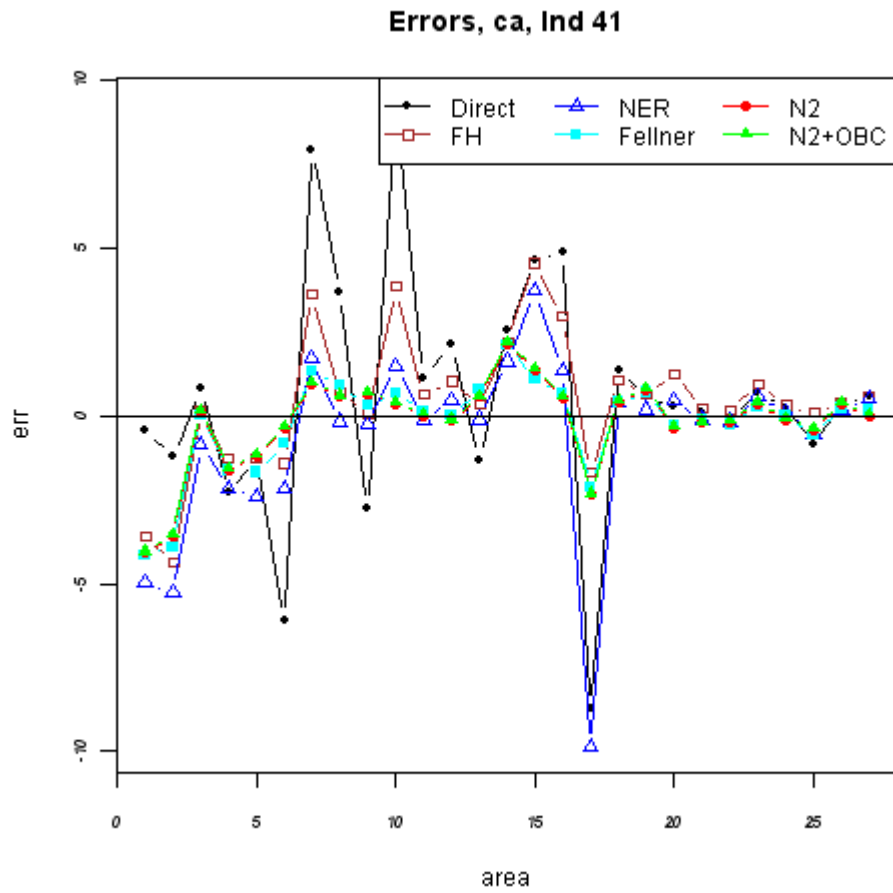


Figure 6. California, Wholesale Trade (industry 41) deviations from true population values (in hundreds) of the relative employment change estimates, by areas.

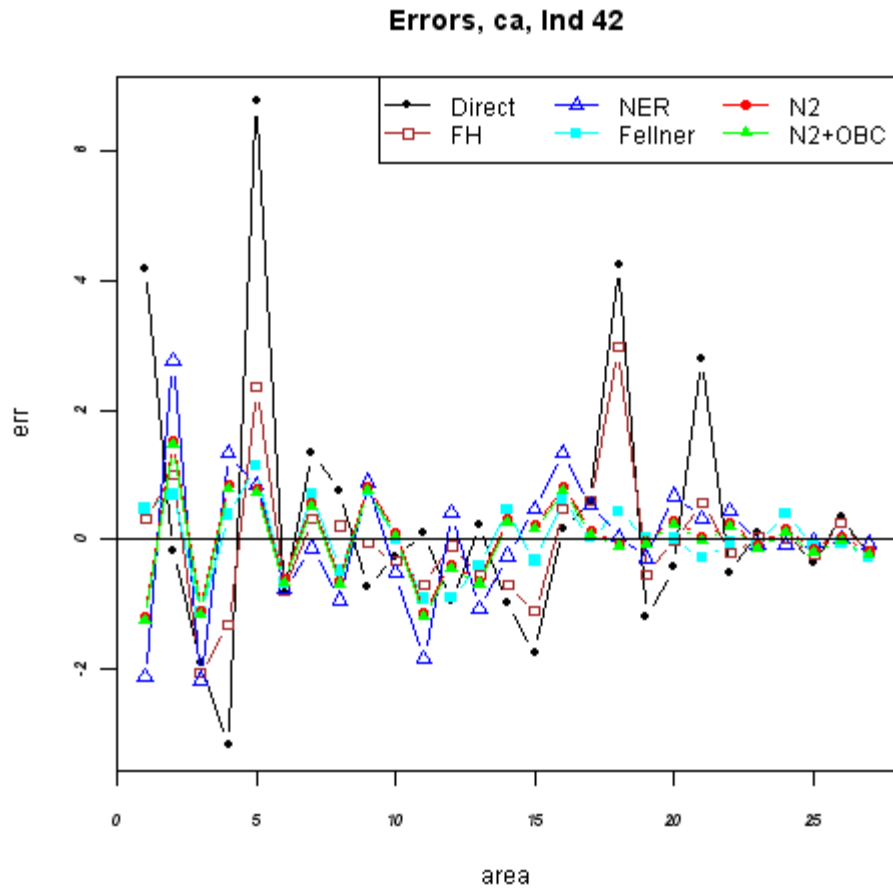


Figure 7. California, Retail Trade (industry 42) deviations from true population values (in hundreds) of the relative employment change estimates, by areas.

Deviations from true population values for areas in California Wholesale Trade (industry 41) and Retail Trade (industry 42) are shown for the direct estimator and estimates based on the nested-error regression (NER), Fay-Herriot (FH), Fellner, and N2 model. Areas on the plots are sorted in the ascending order of the number of sampled units. There were 27 areas in each industry. The number of sampled units range from 1 to 510 in Wholesale Trade and from 6 to 543 in Retail Trade. It can be seen that the direct estimator (black dots) is very inefficient. Errors of the Fay-Herriot estimator often mimic the errors of the direct estimator. This happens because the

variances of the direct estimators do not take into account the outliers that occur in the sample. Hence, in the weighted average, more weight is given to the direct estimator than to the synthetic part. The NER estimator also often has a larger error than the robust estimators. Performances of the robust estimators, Fellner, MQ, and N2, are for the most part similar.

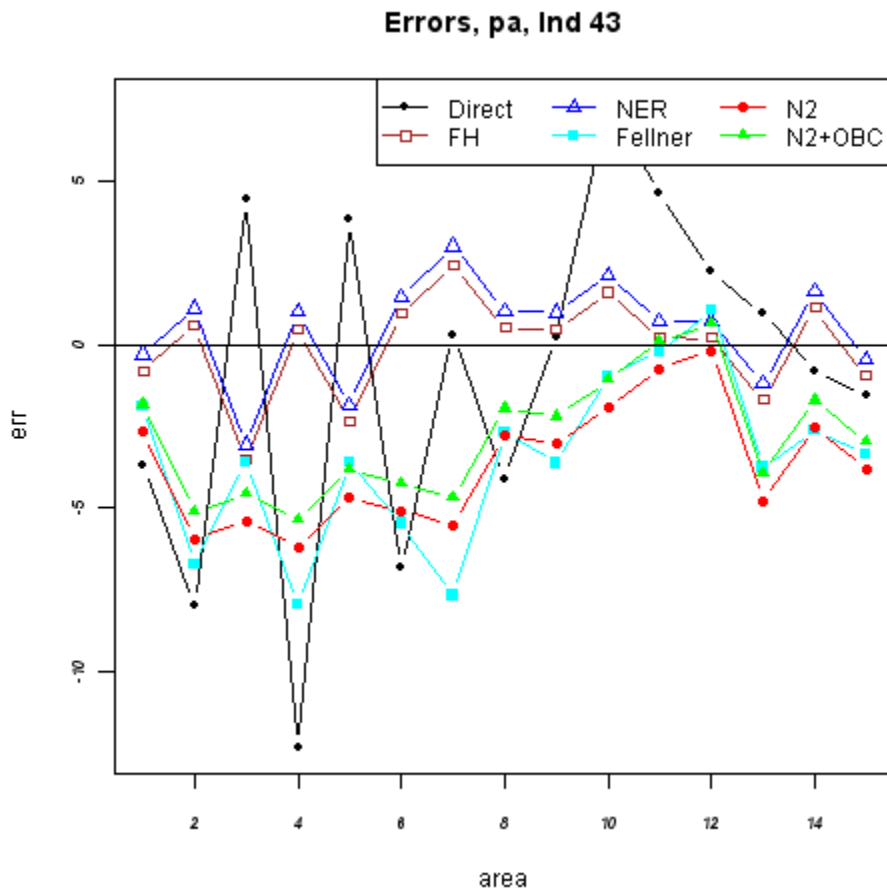


Figure 8. Pennsylvania, Transportation and Utilities (industry 43) deviations from true population values (in hundreds) of the relative employment change estimates, by areas.

An interesting case is shown in Figure 8 (Pennsylvania, industry 43). There are 15 areas included in the model, the smallest area has 2 units and the largest area has 60 units in the sample. Here, robust estimators perform worse than the Fay-Herriot or NER based estimators. The distribution of the residuals in this industry is asymmetric. The right tail units have a higher probability of being in part 2 of the sample. (This is an indication that, perhaps, the alternative models are also misspecified.) The bias incurred because observations tended to be downweighted more on the right tail of the distribution. The bias was somewhat corrected in the N2+OBC estimator.

## Summary

In this Chapter, we proposed a model that assumes that observations are generated from a mixture of two normal distributions with a common mean and different error variances. This model can be viewed as an extension of the nested error regression model, as it relaxes the assumption that the random error variance is constant.

When the fraction of the larger variance observations is small, the estimates from the model perform similar to the robust methods of Fellner (1986) and Sinha and Rao (2008) that are based on the Huber function. The model has potential to be especially useful when the fraction of the part 2 observations increases.

Another feature of the proposed method is that it estimates the conditional probabilities for observations to fall in each part of the mixture. This can serve as the basis for a formal test, such as the “random occurrence test” described in Section 3.5. The random occurrence test essentially is a check of validity of the model. If the test indicates that certain areas are outliers, such areas can be removed from the model

and the model can be re-fitted without the outlying areas. For the outlying areas, a separate set of assumptions has to be used. Depending on the context of a survey, subsequent treatment may include adding a bias correction term to the area estimate or excluding the area from the model and using the direct estimator for such an area.

We considered several scenarios for evaluation of the bootstrap procedure for the case of the mixture model. Bootstrap prediction confidence intervals provided approximately nominal coverage under each of these scenarios.

When the finite population target is not in a linear form, it can be linearized in order to apply a model at the unit level. The unit level modeling may be especially useful when outliers in the data affect the direct survey estimates. A study using CES sample data confirms this point.

## Chapter 4: Concluding Remarks and Future Research

We include in this Chapter a list of topics that we feel need to be explored in the future.

### 1. Asymptotic properties of the estimated parameters

In Section 3.7 of Chapter 3 we used a simulation study to investigate the performance of the parameter estimates when the number of areas increases. Although somewhat inconclusive due to large variance in some of the estimates, the results of the study suggest that the estimators of the parameters tend to the true values. It would be desirable to prove consistency of the estimators analytically. In particular, the consistency property is a necessary condition for the proper approximation of the true distribution of the pivot by the parametric bootstrap of Section 3.4.

### 2. Theoretical properties of the prediction confidence intervals obtained using the parametric bootstrap of Section 3.4

The goal is to prove theoretically that the distribution of the bootstrap pivot approximates the distribution of the corresponding quantity based on the original data and to derive the order of the approximation.

### 3. Improvements in the Monte Carlo part of the EM algorithm

The Monte Carlo part of the EM algorithm described in Section 3.3.1 works reasonably well when the probability of being in part 2 of the mixture is small. There is room for improvement of the algorithm. One problem is that, when the Monte Carlo error is large, the log-likelihood function does not necessarily increase at every

step of the EM algorithm and, as a result, the algorithm may not converge properly and the maximum will not be reached.

Booth and Hobert (1999) proposed several methods that help to control the performance of the EM algorithm. The methods include dynamic increase in the number of the Monte Carlo iterations depending on the error of the Monte Carlo estimates computed after each EM step. However, the error may be so large that it would call for an unrealistically large number of iterations. Therefore, the first goal would be to find an improvement in terms of efficiency of the Monte Carlo step.

## Appendix A. The proof of Result 1 from Section 2.3

First, note that it is always true that  $K \geq 0$  and  $L \leq 0$ . This follows immediately from conditions (2.3.2) and (2.3.3) and the fact that, for any  $j$ ,  $(K - u_j)J_j \leq 0$  and  $(L - u_j)I_j \geq 0$ .

Next, write the mean squared error as

$$MSE[\bar{u}(K, L)] = Var[\bar{u}(K, L)] + (Bias[\bar{u}(K, L)])^2, \quad (A.1.1)$$

where the bias is

$$\begin{aligned} Bias[\bar{u}(K, L)] &= E[\bar{u}(K, L) - \bar{u}] \\ &= E[\bar{u}(K, L)] - \bar{u} = n^{-1} E \sum_{j=1}^n [(K - u_j)J_j + (L - u_j)I_j] \end{aligned} \quad (A.1.2)$$

Consider the variance term:

$$\begin{aligned} Var[\bar{u}(K, L)] &= Var\left[\bar{u} + n^{-1} \sum_{j=1}^n \{(K - u_j)J_j + (L - u_j)I_j\}\right] = \\ &= Var(\bar{u}) + n^{-2} \sum_{j=1}^n Var\{(K - u_j)J_j + (L - u_j)I_j\} + 2n^{-2} E\left[\sum_{j=1}^n u_j \{(K - u_j)J_j + (L - u_j)I_j\}\right] \end{aligned}$$

(from the independence of  $u$ 's and that  $E(u_j) = 0$ )

$$\begin{aligned} &= Var(\bar{u}) + n^{-2} \sum_{j=1}^n \left( E[(K - u_j)J_j + (L - u_j)I_j]^2 - (E[(K - u_j)J_j + (L - u_j)I_j])^2 \right) \\ &\quad + 2E[u_j \{(K - u_j)J_j + (L - u_j)I_j\}] \\ &= Var(\bar{u}) + n^{-2} \sum_{j=1}^n \left( E[(K^2 - u_j^2)J_j + (L^2 - u_j^2)I_j] - (E[(K - u_j)J_j + (L - u_j)I_j])^2 \right). \end{aligned}$$

Note that whenever  $K \geq 0$  and  $L \leq 0$ , the variance of the Winsorized mean  $\bar{u}(K, L)$

does not exceed the variance of the sample mean  $\bar{u}$ ,  $Var(\bar{u})$ , since for all  $j$ ,



$(K^2 - u_j^2)J_j < 0$  and  $(L^2 - u_j^2)I_j \leq 0$ . When conditions (2.3.2) and (2.3.3) hold, the analogous result holds not only for the variance but also for MSE, as shown below.

The MSE is

$$\begin{aligned}
MSE[\bar{u}(K, L)] &= Var(\bar{u}) + n^{-2} \sum_{j=1}^n \left( E[(K^2 - u_j^2)J_j + (L^2 - u_j^2)I_j] - \left( E[(K - u_j)J_j + (L - u_j)I_j] \right)^2 \right) \\
&+ \left( Bias[\bar{u}(K, L)] \right)^2 \\
&= Var(\bar{u}) + n^{-2} \sum_{j=1}^n \left( E[K(K - u_j)J_j + L(L - u_j)I_j] + E[u_j(K - u_j)J_j + u_j(L - u_j)I_j] \right) \\
&- n^{-2} \sum_{j=1}^n \left( E[(K - u_j)J_j + (L - u_j)I_j] \right)^2 \\
&+ \left( Bias[\bar{u}(K, L)] \right)^2 \\
&= Var(\bar{u}) + n^{-2} E \left( K \sum_{j=1}^n (K - u_j)J_j + L \sum_{j=1}^n (L - u_j)I_j \right) \\
&+ n^{-2} \sum_{j=1}^n E[u_j(K - u_j)J_j + u_j(L - u_j)I_j] \\
&- n^{-2} \sum_{j=1}^n \left( E[(K - u_j)J_j + (L - u_j)I_j] \right)^2 \\
&+ \left( Bias[\bar{u}(K, L)] \right)^2
\end{aligned}$$

Since  $K$  and  $L$  satisfy, respectively, the equations  $K + \sum_{j=1}^n (K - u_j)J_j = 0$  and

$L + \sum_{j=1}^n (L - u_j)I_j = 0$ , the bias is

$$Bias[\bar{u}(K, L)] = -n^{-1}E(K + L).$$

The mean squared error is

$$\begin{aligned}
MSE[\bar{u}(K, L)] &= Var(\bar{u}) - n^{-2}(EK^2 + EL^2) \\
&+ n^{-2} \sum_{j=1}^n E[u_j(K - u_j)J_j + u_j(L - u_j)I_j] - n^{-2} \sum_{j=1}^n \left( E[(K - u_j)J_j + (L - u_j)I_j] \right)^2 \\
&+ n^{-2}(EK + EL)^2
\end{aligned}$$

$$\begin{aligned}
&= \text{Var}(\bar{u}) - n^{-2}\text{Var}(K) - n^{-2}\text{Var}(L) \\
&+ n^{-2} \sum_{j=1}^n E \left[ u_j (K - u_j) J_j + u_j (L - u_j) I_j \right] - n^{-2} \sum_{j=1}^n \left( E \left[ (K - u_j) J_j + (L - u_j) I_j \right] \right)^2 \\
&+ 2n^{-2} E(KL) \\
&\leq \text{Var}(\bar{u})
\end{aligned}$$

(the last inequality follows from noting that, for any  $j$ ,  $u_j J_j \geq 0$  and  $(K - u_j) J_j \leq 0$ .)

Thus,  $u_j (K - u_j) J_j \leq 0$ . Similarly, for any  $j$ ,  $u_j I_j \leq 0$  and  $(L - u_j) I_j \geq 0$ ; thus,

$$u_j (L - u_j) I_j \leq 0. \text{ Therefore, } n^{-2} \sum_{j=1}^n E \left[ u_j (K - u_j) J_j + u_j (L - u_j) I_j \right] \leq 0.$$

The term involving  $E(KL)$  also never exceeds zero because  $K \geq 0$  and  $L \leq 0$ .)

## Appendix B. R code for the Winsorization example of Section

### 2.3.1.

```
#####
#   KL function for finding cutoffs                                     #
#   -----                                                         #
# Input:                                                               #
#   X       sort and center the sample, then                         #
#           to find K: take values on the right from zero,         #
#           to find L: take absolute values on the left from zero  #
#           #                                                         #
#   S       length, defines the number of nodes,                   #
#           e.g., 10*length(X)                                       #
#           #                                                         #
# Output:                                                               #
#   KL      cutoff point (K or L)                                     #
#####
KL<-function(X,S){
  maxKL=max(X)
  i=0
  z=1
  while (z>0 && i<S){
    i=i+1                                # count nodes
    KL=(S-i)*maxKL/S                    # interpolation step
    P_K=length(X[X>KL])                 # (tail probability)*length(X)
    M_K=sum(X[X>KL])                    # (tail mean)*length(X)
    z=(KL+(KL*P_K-M_K))
  }
return(KL)
}

example_demo<-function(N, Sim, p, mu1, mu2, guess,seed){

  # N           sample size
  # Sim         number of simulation runs
  # p           contamination fraction (can be 0)
  # mu1        true mean for "good" units
  # mu2        true mean for contamination
  # guess      for initial guess: 1 - use truth; 2 - use mean

set.seed(seed)
est1<-matrix(0,Sim,1)
est2<-matrix(0,Sim,1)
K<-matrix(0,Sim,1)
L<-matrix(0,Sim,1)
n<-matrix(0,Sim,1)

for (sim in 1:Sim){
  n[sim]=sum(rbinom(N,1,p))
  x0=c(rnorm((N-n[sim]),mu1,1),rnorm(n[sim],mu2,sqrt(10)))
  truth=(1-p)*mu1+p*mu2

  ## lognormal
```

```

#x0=rlnorm(N,0,1)
#truth=exp(0.5)

if (guess==1) {mu0=truth} else
if (guess==2) {mu0=mean(x0)}

x=x0-mu0

rightx<--sort(-x[x>0])
K[sim]<-KL(rightx,10*length(rightx))
leftx<--sort(x[x<0])
L[sim]<--KL(leftx,10*length(leftx))

# Winsorized values
x_w<-
mu0+x*(x>L[sim])*(x<K[sim])+L[sim]*(x<=L[sim])+K[sim]*(x>=K[sim])

  est1[sim]=mean(x0)
  est2[sim]=mean(x_w)
}
### Summary
K<-mean(K)
L<-mean(L)

rmse1<-100*sqrt(mean((est1-truth)^2))
rmse2<-100*sqrt(mean((est2-truth)^2))

bias1=100*mean(est1-truth)
bias2=100*mean(est2-truth)

out<-
as.data.frame(cbind(Sim,N,p,mu1,mu2,guess,bias1,bias2,rmse1,rmse2,10
0*rmse2/rmse1,L,K))
names(out)<-
c("Sim","N","P","Mu1","Mu2","Guess","Bias1","Bias2","RMSE1","RMSE2",
"RMSE2/RMSE1","L","K")
print(out)
return(out)
}

### example call:

out<-example_demo(50,5000,0.03,0,0,1,2717)
write.table(out, "example_demo.csv",sep="," , append=1)

```

## Appendix C: EM algorithm for the scale mixture-fixed effects model WN2F from Section 2.4

(The algorithm is a slightly more general case of the scale mixture of K Normal distributions, strata means are modeled as fixed effects)

The Model:

$$y_{mj} \mid z_{mjk} = 1, \mu_m, \sigma_k^2 \stackrel{ind}{\sim} N(\mu_m, \sigma_k^2),$$

where  $k = 1, \dots, K; j = 1, \dots, n_m; m = 1, \dots, M; \sum_{m=1}^M n_m = n;$

$z_{mjk}$  is a mixture class indicator for an observation  $mj$  and class  $k$ ;

$\sigma_k^2$  is a variance parameter of the  $k^{th}$  component of the mixture.

Denote the observation vector by  $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_M^T)^T$ , where  $\mathbf{y}_m = (y_{m1}, \dots, y_{mn_m})^T$ .

The goal is to estimate the set of parameters  $\boldsymbol{\theta} = (\mu_1, \dots, \mu_M, \pi_1, \dots, \pi_K, \sigma_1^2, \dots, \sigma_K^2)$ .

The indicator  $z_{mjk}$  takes the value 1 if the observation  $(mj)$  belongs to class  $k$  and is 0 otherwise.

The complete data log likelihood is

$$\begin{aligned} l(\mathbf{y}, \mathbf{z} \mid \boldsymbol{\theta}) &= \sum_{m=1}^M \sum_{k=1}^K \sum_{j=1}^{n_m} z_{mjk} \log f(y_{mj} \mid z_{mjk}, \boldsymbol{\theta}) \\ &= -\frac{1}{2} \left( n \log 2\pi + \sum_{k=1}^K \log \sigma_k^2 \sum_{m=1}^M \sum_{j=1}^{n_m} z_{mjk} + \sum_{k=1}^K \sigma_k^{-2} \sum_{m=1}^M \sum_{j=1}^{n_m} z_{mjk} (y_{mj} - \mu_m)^2 \right) \end{aligned}$$

**The algorithm:**

Assign sets of initial values:

$$\mu_m^{(0)} \text{ for } \mu_m, m = 1, \dots, M,$$

$$\pi_k^{(0)} \text{ for prior probabilities to belong to the mixture part } k,$$

$$z_{mjk}^{(0)} = \pi_k^{(0)}, j = 1, \dots, n,$$

$$\sigma_k^{(0)} \text{ for } \sigma_k, k = 1, \dots, K$$

$p$  -- iteration

Run the loop as specified below.

1. Compute the log likelihood function for the current values of the parameters:

$$l(\theta^{(p)}; \mathbf{y}) = \sum_{m=1}^M \sum_{j=1}^{n_m} \log \left( \sum_{k=1}^K \frac{\pi_k^{(p)}}{\sqrt{2\pi}\sigma_k^{(p)}} \exp \left\{ -\frac{1}{2} \frac{(y_{mj} - \mu_m^{(p)})^2}{\sigma_k^{2(p)}} \right\} \right)$$

If  $z_{mjk}$  were observable, then the complete data log likelihood would look as follows:

$$l_C(\theta^{(p)}; \mathbf{y}) = -\frac{1}{2} \sum_{m=1}^M \left( n_m \log 2\pi + \sum_{k=1}^K \log \sigma_k^{2(p)} \sum_{j=1}^{n_m} z_{mjk}^{(p)} + \sum_{k=1}^K \sum_{j=1}^{n_m} z_{mjk}^{(p)} \frac{(y_{mj} - \mu_m^{(p)})^2}{\sigma_k^{2(p)}} \right)$$

2. (E-step) Mixture indicators  $z_{mjk}$  are replaced by their current conditional expectations

$$z_{mjk}^{(p+1)} = E[z_{mjk} | \mathbf{y}, \theta^{(p)}] = P\{z_{mjk} = 1 | \mathbf{y}, \theta^{(p)}\}.$$

At step  $p+1$ , we “impute” the posterior probabilities, using the Bayes formula,

$$z_{mjk}^{(p+1)} = \pi_k^{(p)} \frac{1}{\sigma_k^{(p)}} \exp \left\{ -\frac{1}{2} \frac{(y_{mj} - \mu_m^{(p)})^2}{\sigma_k^{2(p)}} \right\} \bigg/ \sum_{k=1}^K \left( \pi_k^{(p)} \frac{1}{\sigma_k^{(p)}} \exp \left\{ -\frac{1}{2} \frac{(y_{mj} - \mu_m^{(p)})^2}{\sigma_k^{2(p)}} \right\} \right)$$

for  $k = 1, \dots, K$

3. (M-step) Find MLE of the parameters:

$$\pi_k^{(p+1)} = \frac{1}{n} \sum_{m=1}^M \sum_{j=1}^{n_m} z_{mjk}^{(p+1)}$$

$$\mu_m^{(p+1)} = \frac{\sum_{k=1}^K \sigma_k^{-2(p)} \sum_{j=1}^{n_m} z_{mjk}^{(p+1)} y_{mj}}{\sum_{k=1}^K \sigma_k^{-2(p)} \sum_{j=1}^{n_m} z_{mjk}^{(p+1)}}$$

$$\sigma_k^{2(p+1)} = \frac{\sum_{m=1}^M \sum_{j=1}^{n_m} z_{mjk}^{(p+1)} \left( y_{mj} - \mu_m^{(p)} \right)^2}{\sum_{m=1}^M \sum_{j=1}^{n_m} \left[ z_{mjk}^{(p+1)} \right]}$$

4. Recompute the log likelihood using the new values of the parameters.

5. Check the convergence criteria:

$$\left| \Delta^{(p+1)} \right| < \varepsilon, \text{ where } \Delta^{(p+1)} = l(\theta^{(p+1)}; \mathbf{y}) - l(\theta^{(p)}; \mathbf{y})$$

## Appendix D: On the maximum likelihood estimator of $\boldsymbol{\beta}$ .

The derivative, with respect to  $\boldsymbol{\beta}$ , of the log-likelihood function of the mixture distribution given in the form (3.1.4) is

$$\begin{aligned} \frac{\partial \log h(\mathbf{y} | \boldsymbol{\theta})}{\partial \boldsymbol{\beta}} &= \sum_{k=1}^{2^n} \frac{\lambda_k f_k(\mathbf{y} | \boldsymbol{\theta})}{h(\mathbf{y} | \boldsymbol{\theta})} \frac{\partial L_k(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \\ &= \sum_{k=1}^{2^n} P\{i_k | \mathbf{y}, \boldsymbol{\theta}\} \frac{\partial L_k(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \\ &= E \left[ \frac{\partial L_k(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \middle| \mathbf{y}, \boldsymbol{\theta} \right], \end{aligned}$$

where  $L_k(\boldsymbol{\theta}) = \log f_k(\mathbf{y} | \boldsymbol{\theta})$  is the log-likelihood function of the mixed model corresponding to some  $k$ -th combination of the mixture indicators. The derivative with respect to  $\boldsymbol{\beta}$  is

$$\frac{\partial L_k(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = \mathbf{X}^T \boldsymbol{\Sigma}_k^{-1} \mathbf{y} - \mathbf{X}^T \boldsymbol{\Sigma}_k^{-1} \mathbf{X} \boldsymbol{\beta}.$$

Thus,

$$\begin{aligned} \frac{\partial \log h(\mathbf{y} | \boldsymbol{\theta})}{\partial \boldsymbol{\beta}} &= E \left[ \mathbf{X}^T \boldsymbol{\Sigma}_k^{-1} \mathbf{y} - \mathbf{X}^T \boldsymbol{\Sigma}_k^{-1} \mathbf{X} \boldsymbol{\beta} \middle| \mathbf{y}, \boldsymbol{\theta} \right] \\ &= \mathbf{X}^T E \left[ \boldsymbol{\Sigma}_k^{-1} \middle| \mathbf{y}, \boldsymbol{\theta} \right] \mathbf{y} - \mathbf{X}^T E \left[ \boldsymbol{\Sigma}_k^{-1} \middle| \mathbf{y}, \boldsymbol{\theta} \right] \mathbf{X} \boldsymbol{\beta} \end{aligned}$$

Thus, if all the other parameters are known, MLE of  $\boldsymbol{\beta}$  is a solution of the estimating

equations  $\frac{\partial \log h(\mathbf{y} | \boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = 0$ , i.e.,

$$\hat{\boldsymbol{\beta}} = \left( \mathbf{X}^T E \left[ \boldsymbol{\Sigma}_k^{-1} \middle| \mathbf{y}, \boldsymbol{\theta} \right] \mathbf{X} \right)^{-1} \mathbf{X}^T E \left[ \boldsymbol{\Sigma}_k^{-1} \middle| \mathbf{y}, \boldsymbol{\theta} \right] \mathbf{y}.$$



## References

- Arora, V., Lahiri, P. and Mukherjee, K. (1997), Empirical Bayes estimation of finite population means from complex surveys, *Journal of the American Statistical Association*, 92, 1555-1562.
- Balk, B.M. (1995). Axiomatic Price Index Theory: A Survey. *International Statistical Review*, 63, 69–93.
- Battese, G. E., Harter, R. M. and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data, *Journal of the American Statistical Association*, 83, 28-36
- Beaumont, J.-F. (2008), “A new approach to weighting and inference in sample surveys,” *Biometrika* (2008), 95, 3, pp. 539–553
- Booth, J. G., and Hobert, J. P. (1999), “Maximizing Generalized Linear Mixed Model Likelihoods with an Automated Monte Carlo EM Algorithm,” *Journal of the Royal Statistical Society, Ser. B*, 61, 265–285
- Bureau of Labor Statistics (2004), Chapter 2, “Employment, hours, and earnings from the Establishment survey,” *BLS Handbook of Methods*. Washington, DC: U.S. Department of Labor. <http://www.bls.gov/opub/hom/pdf/homch2.pdf>
- Chambers, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.
- Chambers, R. L. (2005) What If... ? Robust Prediction Intervals for Unbalanced Samples. Southampton, UK, Southampton Statistical Sciences Research Institute, 21pp. (S3RI Methodology Working Papers, M05/05)  
<http://eprints.soton.ac.uk/14075/>

- Chambers, R. and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika* 93,255-268.
- Chambers, R., Chandra, H., Salvati, N. and Tzavidis, N. (2009). Outlier Robust Small Area Estimation. Invited Presentation, ISI 2009, South Africa
- Chatterjee, A., Lahiri, P., and Li, H. (2008), Parametric bootstrap approximation to the distribution of EBLUP, and related prediction intervals in linear mixed models. *Annals of Statistics*. 36, 1221-1245
- Clements, K.W., Izan, H.Y., and Selvanathan, E.A. (2006). Stochastic Index Numbers: A Review. *International Statistical Review* , 74, 2, 235–270
- Cox, D.R., Hinkley, D.V. (1974) *Theoretical Statistics*, Chapman & Hall. [ISBN 0-412-12420-3](#)
- Datta, G. S. and Lahiri, P. (1995) “Robust Hierarchical Bayes Estimation of Small Area Characteristics in the Presence of Covariates and Outliers,” *Journal of Multivariate Analysis*, 54, 310-328.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), “Maximum Likelihood from Incomplete Data via the EM Algorithm,” *Journal of the Royal Statistical Society, Ser. B*, 39, 1–38.
- Diewert, W.E. (1981). *The Economic Theory of Index Numbers*. In *Essays in the Theory and Measurement of Consumer Behaviour (in Honour of Richard Stone)*, Ed. A. Deaton, pp. 163–208. New York: Cambridge University Press.
- Elliott, M. R. and Little, R. J. A. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, 16, 191-209.

- Fay, R.E. and Herriot, (1979). Estimates of Income for Small Places: an Application of James-Stein Procedure to Census Data, *Journal of American Statistical Association*, 74, 269-277
- Fellner, W. H. (1986), Robust Estimation of Variance Components," *Technometrics*, 28, 51-60.
- Fuller, W. A. (1991). Simple estimators for the mean of skewed populations. *Statistica Sinica*, 1, pp. 137-158.
- Gershunskaya, J. (2010). Robust Small Area Estimation Using a Mixture Model. *Proceedings of the Section on Survey Research Methods, American Statistical Association*
- Gershunskaya, J. and Huff, L., (2004). Outlier Detection and Treatment in the Current Employment Statistics Survey, *Proceedings of the Section on Survey Research Methods, American Statistical Association*
- Gershunskaya, J. and Lahiri, P., (2008). Robust Estimation of Monthly Employment Growth Rates for Small Areas in the Current Employment Statistics Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*
- Ghosh, M. (2009). Bayesian Developments in Survey Sampling, *Handbook of Statistics, Sample Surveys: Theory, Methods and Inference*, Eds. D. Pfeiffermann and C.R. Rao, Amsterdam:Elsevier BV. Vol. 29, Chapter 29
- Ghosh, M. and Lahiri, P. (1987), "Robust empirical Bayes estimation of means from stratified samples," *Journal of the American Statistical Association*, 82, 1153-1162.

- Ghosh, M., Maiti, T., Roy, A. (2008), "Influence functions and robust Bayes and empirical Bayes small area estimation," *Biometrika* (2008), 95, 3, pp. 573–585.
- Ghosh, M., and Meeden, G. (1986), Empirical Bayes Estimation in Finite Population Sampling, *Journal of the American Statistical Association*, 81, 1058-1062.
- Gross, S. (1980). Median estimation in sample surveys. Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, 181–184.
- Gwet, J.-P. and Rivest, L.-P. (1992). Outlier resistant alternatives to the ratio estimator. *J. Am. Statist. Assoc.* 87, 1174–82.
- Hampel, F. (1968). Contribution to the theory of robust estimation. Unpublished dissertation, Univ. of California, Berkley
- Hampel, F. (1974). The influence curve and its role in robust estimation, *Journal of American Statistical Association*, 69, 383–393
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986). *Robust Statistics: the Approach Based on Influence Functions*. New-York, John Wiley & Sns, Inc.
- Hansen, M. H., Madow, W. G. and Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys (with discussion), *Journal of the American Statistical Association*, 78, 776-793
- Huang, E.T. and Bell, W. R. (2006), "Using the t-Distribution in Small Area Estimation: An Application to SAIPE State Poverty Models," 2006 Proceedings of the American Statistical Association, Survey Research Methods Section [CD-ROM], Alexandria, VA: American Statistical Association

- Huber, P. J. (1964). Robust estimation of a location parameter, *Ann. Math. Statist.* 35: 73–101.
- Huber, P. J. (1972). Robust statistics: A review, *Ann. Math. Statist.* 43: 1041–1067.
- Huber, P.J. (1981, 2004). *Robust Statistics*. New-York, John Wiley & Sons, Inc.
- Hulliger, B. (1995). Outlier robust Horvitz-Thompson estimators. *Survey Methodology*, 21, 79-87.
- Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96
- Kokic, P. N., and Bell, P. A. (1994), "Optimal Winsorizing Cutoffs for a Stratified Finite Population Estimator," *Journal of Official Statistics*, 10, 419-435.
- Lee, H. (1995) Outliers in Business Surveys. In *Business Survey Methods* edited by B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. Colledge, and P. S. Kott, p. 503-526
- Li, Y, and Lahiri, P. (2007), "Robust model-based and model-assisted predictors of the finite population total," *Journal of the American Statistical Association*, 102, 664-673.
- Little, R. J. A. (1983). Comments on "An evaluation of model-dependent and probability-sampling inferences in sample surveys", *Journal of the American Statistical Association*, 78, 797-799
- Miller, J. J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *Ann. Statist.* 5 746-762.
- Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, 8, 343-366.

- Quenouille, M. (1956). Notes on bias in estimation,. *Biometrika* 43, 353–360
- Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.
- Pfeffermann, D., Krieger, A. M. and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8, 1087-1114.
- Pfeffermann, D., and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya*, 61, 166-186.
- Pfeffermann, D. and Sverchkov, M. (2003). Fitting generalized linear models under informative probability sampling. In: *Analysis of Survey Data*, eds. R. L. Chambers and C. J. Skinner, New York: Wiley, pp. 175-195.
- Pfeffermann, D. and Sverchkov, M. (2007). Small area estimation under informative probability sampling of areas and within the selected areas, *Journal of American Statistical Association*, 102, 1427-1439
- Pfeffermann, D. and Sverchkov, M. (2009). Inference under Informative Sampling, *Handbook of Statistics, Sample Surveys: Theory, Methods and Inference*, Eds. D. Pfeffermann and C.R. Rao, Amsterdam:Elsevier BV. Vol. 29, Chapter 39, to appear.
- Potter, F. (1988). Survey of procedures to control extreme sampling weights. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 453-458.

- Potter, F. (1990). A study of procedures to identify and trim extreme sampling weights. In Proceedings of the Section on Survey Research Methods, American Statistical Association, 225-230.
- Rao, J.N.K. (2003). Small Area Estimation, New-York, John Wiley & Sons, Inc.
- Rivest, L.-P. (1999). Stratum jumpers: Can we avoid them? Proceedings of the Survey Research Methods of the American Statistical Association, pp. 64-72.
- Särndal, C.-E., Swensson, B. and Wretman, J.H. (1992). Model Assisted Survey Sampling. New-York, Springer-Verlag.
- Searls, D.T. (1966), "An Estimator for a Population Mean Which Reduces the Effect of Large Observations," Journal of the American Statistical Association, 61, 1200-1204.
- Serfling, R. J. (1980). Approximation theorems of mathematical statistics. New York, Wiley.
- Sinha, S.K. and Rao, J.N.K. (2008). Robust methods for small area estimation. Proceedings of the American Statistical Association, Survey Research Methods Section, Alexandria, VA: American Statistical Association, 27-38
- Sverchkov, M., and Pfeffermann, D. (2004), "Prediction of Finite Population Totals Based on the Sample Distribution," Survey Methodology, 30, 79-92.
- Tukey, J. W. (1958). Bias and confidence in not quite large samples. Annals of Mathematical Statistics 29, 614.
- Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling

- (I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow and H. B. Mann, eds.) 448-485. Stanford Univ. Press.
- Tukey, J. W. (1970) *Exploratory Data Analysis (Limited Preliminary Edition)*, Reading, Massachusetts: Addison-Wesley.
- Tukey, J. W. and McLaughlin, D. H. (1963), “Less Vulnerable Confidence and Significance Procedures for Location Based on a Single Sample: Trimming/Winsorization 1,” *Sankhya A*, 25, 331–352.
- Valliant, R., Dorfman, A. and Royall, R.M. (2000). *Finite population sampling: a prediction approach*. New-York, John Wiley & Sons, Inc.
- Weiss, L. (1971). Asymptotic properties of maximum likelihood estimators in some nonstandard cases. *J. Amer. Statist. Assoc.* 66 345-350.
- Weiss, L. (1973). Asymptotic properties of maximum likelihood estimators in some nonstandard cases II. *J. Amer. Statist. Assoc.* 68 428-430.
- Welsh, A.H., and Ronchetti, E. (1998). Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society, Series B*, 60, 413-428.
- Xie, D., Raghunathan, T. E., and Lepkowski, J. M. (2005), “Estimation of Prevalence of Overweight in Small Areas – A Robust Extension of Fay-Herriot Model,” 2005 Proceedings of the American Statistical Association, Survey Research Methods Section [CD-ROM], Alexandria, VA: American Statistical Association, 3701-3712.
- Yakowitz, S. J. and Spragins, J. D. (1968) On the identifiability of finite mixtures, *Ann. Math. Statist.*, 39, pp. 209-214.



Zaslavsky, A.M., Schenker, N. and Belin, T.R. (2001). Downweighting influential clusters in surveys: application to the 1990 post enumeration survey. *Journal of the American Statistical Association*, 96, 858-869.