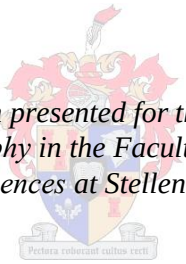
Statistical Inference of the Multiple Regression Analysis of Complex Survey Data

by

Retha Luus



Dissertation presented for the degree of
Doctor of Philosophy in the Faculty of Economic and Management Sciences at Stellenbosch University

Supervisor: Prof. Tertius de Wet
Co-supervisor: Dr. Ariane Neethling

December 2016

Declaration

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

.....
R. Luus

December 2016

Abstract

The quality of the inferences and results put forward from any statistical analysis is directly dependent on the correct method used at the analysis stage. Most survey data analyzed in practice originate from stratified multistage cluster samples or complex samples. In developed countries the statistical analysis, for example linear modeling, of complex sampling (CS) data, otherwise known as survey-weighted least squares (SWLS) regression, has received some attention over time. In developing countries such as South Africa and the rest of Africa, SWLS regression is often confused with weighted least squares (WLS) regression or, in some extreme cases, the CS design is ignored and an ordinary least squares (OLS) model is fitted to the data. This is in contrast to what is found in the developed countries. Furthermore, especially in the developing countries, inference concerning the linear modeling of a continuous response is not as well documented as is the case for the inference of a categorical response, specifically in terms of a dichotomous response. Hence, the decision was made to research the linear modeling of a continuous response under CS with the objective of illustrating how the results could differ if the statistician ignores the complex design of the data or naïvely applies WLS in comparison to the correct SWLS regression.

The complex sampling design leads to observations having unequal inclusion probabilities, the inverse of which is known as the design weight of an observation. Once adjusted for unit non-response and differential non-response, the sampling weights can have large variability that could have an adverse effect on the estimation precision. Weight trimming is cautiously recommended as a remedy for this, but could also increase the bias of an estimator which then affects the estimation precision once more. The effect of weight trimming on estimation precision is also investigated in this research.

Two important parts of regression analysis are researched here, namely the evaluation of the fitted model and the inference concerning the model parameters. The model evaluation part includes the adjustment of well-known prediction error estimation methods, viz. leave-one-out cross-validation, bootstrap estimation and .632 bootstrap estimation, for application to CS data. It also considers a number of outlier detection diagnostics such as the leverages and Cook's distance. The model parameter inference includes bootstrap variance estimation as well as the construction of bootstrap confidence intervals, viz. the percentile, bootstrap-t, and BCa confidence intervals.

Two simulation studies are conducted in this thesis. For the first simulation study a model was developed and then used to simulate a hierarchical population such that stratified two-stage cluster

samples can be selected from this population. The second simulation study makes use of stratified two-stage cluster samples that are sampled from real-world data, i.e. the Income and Expenditure Survey of 2005/2006 conducted by Statistics South Africa. Similar conclusions are made from both simulation studies. These conclusions include that the incorrect linear model applied to CS data could lead to wrong conclusions, that weight trimming, when conducted with care, further improves estimation precision, and that linear modeling based on resampling methods such as the bootstrap, could outperform standard linear modeling methods, especially when applied to real-world data.

Uittreksel

Die gehalte van die inferensie en resultate wat deur enige statistiese analise voortgebring word, is afhanklik daarvan dat die korrekte analise metode gebruik word. In praktyk is dit meestal so dat die data wat geanaliseer word, ingesamel is volgens ‘n gestratifiseerde meerstadium trossteekproef, wat ook bekendstaan as ‘n komplekse steekproef (KS). Die statistiese analise, byvoorbeeld lineêre modelering, van komplekse steekproewe, het in ontwikkelde lande reeds heelwat aandag ontvang. Veral in ontwikkelende lande, soos Suid-Afrika, is daar gevind dat navorsers dikwels hierdie tipe lineêre modelering verwar met geweegde kleinste kwadrate regressie of selfs sover gaan as om die komplekse ontwerp van die steekproef te ignoreer en ‘n gewone kleinste kwadrate model te pas. Daar is ook gevind dat inferensie oor die lineêre modelering van ‘n kontinue afhanklike veranderlike nie so goed gedokumenteer is in vergelyking met die literatuur wat bestaan vir die inferensie rondom ‘n kategoriese afhanklike veranderlike nie. Dus is ‘n besluit geneem om te illustreer hoe die afvoer van gewone en geweegde kleinste kwadrate modelle kan verskil van die korrekte lineêre model wanneer ‘n kontinue afhanklike veranderlike gemodeleer word.

Komplekse steekproefneming het gewoonlik ongelyke insluitingswaarskynlikhede tot gevolg. Die inverse van hierdie insluitingswaarskynlikhede staan bekend as die ontwerpgewig van ‘n waarneming. Die ontwerpgewigte word aangepas ten opsigte van eenheid nie-respons en differensiële nie-respons waarna hulle bekend staan as steekproefnemingsgewigte. Hierdie gewigte kan groot variasie toon wat ‘n negatiewe invloed op die gehalte van die beraming kan hê. ‘n Moontlike oplossing hiervoor is om die gewigte versigtig te snoei en sodanig die variasie te verminder, maar hierdie aanpassing mag tot ‘n toename in beramingsydigheid lei wat ook nie na wense is nie. Die effek van gewigsnoeiing op die gehalte van die inferensie word ook hier ondersoek.

Twee belangrike dele in regressie word hier oor navorsing gedoen, naamlik die evaluering van die gepaste model asook inferensie met betrekking tot die modelparameters. Die model evaluering gedeelte sluit onder andere die uitbreiding van bekende voorspellingsfoutberamingsmetodes, naamlik los-een-uit kruisgeldigheidsbepaling, bootstrap beraming en .632 bootstrap beraming, vir die toepassing in KS in. ’n Aantal uitskieter opsporings diagnostiese toetse soos die hefboom en Cook se afstand is ook beskou. Skoenlus variansieberaming en die berekening van vertrouensintervalle, naamlik die persentiel, bootstrap-t en BCa intervalle, vorm deel van die model parameter inferensie.

Daar is twee simulasie studies onderneem in hierdie tesis. Vir die eerste simulasie studie is ’n simulasie model ontwikkel en daarna gebruik vir die simulasie van ’n hiërargiese populasie waaruit

gestratifiseerde tweestadium trossteekproewe geneem kan word. Die tweede simulasie studie maak gebruik van gestratifiseerde tweestadium trossteekproewe wat geneem is vanuit werklike data, naamlik die Inkomste en Uitgawe Opname van 2005/2006, 'n opname gedoen deur Statistiek Suid-Afrika. Beide simulasie studies het soortgelyke gevolgtrekkings getoon. Hierdie gevolgtrekkings sluit onder andere in dat verkeerde gevolgtrekkings gemaak kan word indien die verkeerde lineêre model op komplekse steekproefdata gepas word, dat die gewigsnoeiing, indien dit versigtig toegepas word, die beramingsgehalte kan verbeter en dat hersteekproefnemingsmetodes goed werk, veral as dit op werklike data toegepas word.

To my parents, Errol and the late Emma Kirchoff

Acknowledgements

I would like to express my sincere gratitude to the following persons and institutes:

- My father, Errol, and late mother, Emma. It has been 26 years since I received my first academic award for "overall performance and neatness". From that moment, you encouraged me, without fail, to never be satisfied with anything less. My gratitude to you for this is immeasurable.

- My husband, André. Your unfailing support, motivation, patience and understanding were invaluable during this time. Thank you seems so insignificant for all the technical support you provided me with while I developed the various R functions for this research. You truly are a remarkable person.

- Prof. Tertius de Wet. Our paths crossed for the first time in 2005. I was an uncertain student whose potential was spotted immediately by a department head who did not know me at all. The confidence you showed in my abilities encouraged me to work hard and to never deliver anything other than my best work. It was not until later that I learned what an accomplished scholar and great man of God you are. I feel very fortunate to have had the opportunity to do research with and learn from you. Thank you for everything.

- Dr. Ariane Neethling. You are an incredibly strong, passionate and hard working woman and I feel very privileged to have been able to learn from you. Thank you for your guidance and support during this time.

- My friends and colleagues. Thank you for your support, encouragement, suggestions and friendliness. I appreciate each one of you.

- I would like to acknowledge the National Research Foundation of South Africa that supported, in part, the research on which this work is based (UNIQUE GRANT NO: 93148).

- Computations were performed using the University of Stellenbosch's Rhasatsha HPC: http://www.sun.ac.za/hpc. I would like to thank their support staff for all their help.

Soli Deo Gloria!

Contents

List of Figures

List of Tables

Notation

Herewith a list, in alphabetical order, of the regularly used notation:

- $\underline{\beta}$: $(p+1) \times 1$ vector of unknown regression model parameters.
  - $\underline{\hat{\beta}}_{OLS}$: OLS estimator of the unknown parameters.
  - $\underline{\hat{\beta}}_{SWLS}$: SWLS estimator of the unknown parameters.
  - $\underline{\hat{\beta}}_{WLS}$: WLS estimator of the unknown parameters.
- $bias\left(\hat{\beta}_j\right)$: "true" bias of the estimator of the jth parameter, $j = 1, ..., (p+1)$.
- $\widehat{bias}_B\left(\hat{\beta}_{r_j}\right)$: bootstrap estimated bias of the estimator of the jth parameter from the rth replicate sample, $j = 1, ..., (p+1)$ and $r = 1, ..., R$.
  - $\widehat{bias}_B\left(\hat{\beta}_j\right)$: overall bootstrap estimated bias of the estimator of the jth parameter, $j = 1, ..., (p+1)$.
- d: the design/base weight of a sampled unit.
  - d_{CS}: theoretical design weights.
  - d_{SRS}: alternative design weights.
- $Dev_{bias}\left(\hat{\beta}_j\right)$: difference between the bootstrap estimated bias and the "true" bias of the jth parameter, $j = 1, ..., (p+1)$.
- $Dev_{MAD}\left(\hat{\beta}_j\right)$: difference between the bootstrap estimated MAD and the "true" MAD of the jth parameter, $j = 1, ..., (p+1)$.
- $Dev_{MSE}\left(\hat{\beta}_j\right)$: difference between the bootstrap estimated MSE and the "true" MSE of the jth parameter, $j = 1, ..., (p+1)$.
- $MAD\left(\hat{\beta}_j\right)$: "true" median absolute deviation of the estimator of the jth parameter, $j = 1, ..., (p+1)$.
- $\widehat{MAD}_B\left(\hat{\beta}_{r_j}\right)$: bootstrap estimated MAD of the estimator of the jth parameter from the rth replicate sample, $j = 1, ..., (p+1)$ and $r = 1, ..., R$.

  - $\widehat{MAD}_B\left(\hat{\beta}_j\right)$: overall bootstrap estimated MAD of the estimator of the jth parameter, $j = 1, ..., (p+1)$.

- $MSE\left(\hat{\beta}_j\right)$: "true" mean squared error of the estimator of the jth parameter, $j = 1, ..., (p+1)$.

- $\widehat{MSE}_B\left(\hat{\beta}_{r_j}\right)$: bootstrap estimated MSE of the estimator of the jth parameter from the rth replicate sample, $j = 1, ..., (p+1)$ and $r = 1, ..., R$.

  - $\widehat{MSE}_B\left(\hat{\beta}_j\right)$: overall bootstrap estimated MSE of the estimator of the jth parameter, $j = 1, ..., (p+1)$.

- N: population size.

  - In cluster sampling this is the number of clusters (psu's) in the population.

- n: sample size.

  - In cluster sampling this is the number of clusters (psu's) to be selected from the population.

- N_h: population number of PSU's in stratum h, $h = 1, ..., H$.

  - n_h: number of PSU's sampled from stratum h, $h = 1, ..., H$.

- N_{hj}: population number of SSU's in the jth PSU in the hth stratum, $j = 1, ..., N_h$ and $h = 1, ..., H$.

  - n_{hj}: number of SSU's sampled from the jth sampled PSU in the hth stratum, $j = 1, ..., n_h$ and $h = 1, ..., H$.

- $\tilde{PE}$: the Luus approach to the "true" prediction error.

- $\tilde{PE}_r$: the Molinaro approach to the "true" prediction error.

- $\hat{PE}^{Apparent}$: the apparent prediction error.

- $\hat{PE}^{Apparent}_{SWLS}$: the apparent prediction error under complex sampling.

- $\hat{PE}^{LOOCV}$: leave-one-out cross-validation (LOOCV) estimated prediction error.

- $\hat{PE}^{LOOCV}_{SWLS}$: leave-one-out cross-validation (LOOCV) estimated prediction error under complex sampling.

- $\hat{PE}^{.632}$: .632 bootstrap estimated prediction error.

- $\hat{PE}^{.632}_{SWLS}$: .632 bootstrap estimated prediction error under complex sampling.

- $\hat{PE}^{BS}$: bootstrap estimated prediction error.
- $\hat{PE}^{BS}_{SWLS}$: bootstrap estimated prediction error under complex sampling.
- π: inclusion probability of a sampling unit.
- R: number of samples selected from the simulated population or the surrogate population.
- R^2_{OLS}: OLS coefficient of multiple determination.
- R^2_{SWLS}: WLS coefficient of multiple determination
- R^2_{WLS}: SWLS coefficient of multiple determination
- θ: population parameter of interest.
- $\hat{\theta}$: estimator of the parameter of interest.
- U: finite population.
- $\hat{V}_B\left(\hat{\beta}_j\right)$: bootstrap estimated variance of the estimator of the jth parameter, $j = 1, ..., (p+1)$.
- $\hat{V}_{JK}\left(\hat{\beta}_j\right)$: jackknife estimated variance of the estimator of the jth parameter, $j = 1, ..., (p+1)$.
- $\hat{V}_M\left(\hat{\beta}_j\right)$: model estimated variance of the estimator of the jth parameter, $j = 1, ..., (p+1)$.
- w: the final sampling weight of a sampled unit.
  - $w^{pp_1}_{CS}$: d_{CS} design weight benchmarked to person-level auxiliary variables based on the linear distance function.
  - $w^{pp_1}_{SRS}$: d_{SRS} design weight benchmarked to person-level auxiliary variables based on the linear distance function.
  - $w^{pp_2}_{CS}$: d_{CS} design weight benchmarked to person-level auxiliary variables based on the exponential distance function.
  - $w^{pp_2}_{SRS}$: d_{SRS} design weight benchmarked to person-level auxiliary variables based on the exponential distance function.
  - $w^{ph_1}_{CS}$: d_{CS} design weight benchmarked to person- and household-level auxiliary variables based on the linear distance function.
  - $w^{ph_1}_{SRS}$: d_{SRS} design weight benchmarked to person- and household-level auxiliary variables based on the linear distance function.
  - $w^{ph_2}_{CS}$: d_{CS} design weight benchmarked to person- and household-level auxiliary variables based on the exponential distance function.

  - $w_{SRS}^{ph_2}$: d_{SRS} design weight benchmarked to person- and household-level auxiliary variables based on the exponential distance function.

- $\boldsymbol{X}$: $n \times p$ matrix of predictors.

- $\boldsymbol{y}$: $n \times 1$ vector of responses.

Abbreviations

Herewith a list, in alphabetical order, of the regularly used abbreviations:

- CS: complex sample.
- CV: cross-validation.
- DU: dwelling unit.
- EA: enumberated area.
- EC: Eastern Cape province.
- ECKZN: population simulated based on the characteristics of the EC and KZN derived from the IES.
- EPSEM: equal probability sampling.
- EVI: extreme value index.
- 4Avg: weight trimming method with threshold of 4 times the weight average.
- 5Avg: weight trimming method with threshold of 5 times the weight average.
- 5IQR: weight trimming method with threshold of the median weight plus 5 times the weight interquartile range.
- Hill: weight trimming method with threshold determined using the newly derived approach to the Hill estimator of the EVI and Hill plot, Retha's Hill method.
- HPC: high-performance computing.
- HPC1: Rhasatsha Linux cluster at Stellenbosch University.
- IES: Income and Expenditure Survey 2005/2006
- KZN: Kwa-Zulu Natal province.
- LOOCV: leave-one-out cross-validation.

- MOS: measure of size.
- M_0: weight trimming method with threshold determined using Berning's M_0 estimator of the EVI.
- M_3: weight trimming method with threshold determined using Berning's M_3 estimator of the EVI.
- NCP: total non-coverage probability of a confidence interval.
- 1.5IQR: weight trimming method with threshold of the third quartile weight plus 1.5 times the weight interquartile range.
- OLS: ordinary least squares.
- PE: prediction error.
- PPS: probability proportionate to size.
- PSU: primary sampling unit.
- 6IQR: weight trimming method with threshold of the median weight plus 6 times the weight interquartile range.
- SRS: simple random sampling without replacement.
- SS: systematic sampling.
- SSU: secondary sampling unit.
- SWLS: survey-weighted least squares.
- 3.5Med: weight trimming method with threshold of 3.5 times the median weight.
- USU: ultimate sampling unit.
- WC: Western Cape province.
- WCEC: population simulated based on the characteristics of the WC and the EC derived from the IES.
- WLS: weighted least squares.

Chapter 1

Introduction

1.1 Problem Statement

The quality of the inferences and results put forward from any statistical analysis are directly dependent on the correct method used at the analysis stage. General statistical textbooks usually only discuss statistical analysis methodology applied to data obtained from simple random samples (SRS) while most survey data analyzed in practice originate from non-SRS designs. These designs typically combine different sampling methods, such as stratified and cluster sampling, called complex sampling (CS), a technique employed to ensure that the sample collected represents the target population as closely as possible.

The statistical analysis of CS data has received some attention over time, especially in developed countries such as Europe, USA and the United Kingdom. A comment once made by a statistics professor, namely that "every statistician needs to have regression in their toolbox", was the catalyst for this research topic. While researching regression and CS data, it became apparent that, in contrast to the developed countries, some researchers in developing countries, such as South Africa and the rest of Africa, confuse linear regression using CS data, i.e. survey-weighted least squares (SWLS) regression, with weighted least squares (WLS) regression. It was also found that inference of linear models based on CS data, specifically for a continuous response, was not as well documented as for the linear modeling of a dichotomous response. Hence, the decision was made to research the linear modeling of a continuous response under CS with the objective of illustrating how the results could differ if the researcher ignores the complex design of the data, i.e. applies ordinary least squares (OLS) regression, or naïvely applies WLS in comparison to the correct SWLS regression.

The multistage designs by which CS data are sampled lead to the sample observations usually having unequal inclusion probabilities. The stratification is used for the reduction of variances as well as to be able to obtain seperate estimates for different groups (or domains) of interest. The cluster sampling is used to reduce survey costs and is often the only feasible method to

use due to the information available in the samplingframe. Therefore, if CS data are analysed under the assumption of being independent and identically distributed, the estimated standard errors, confidence intervals and hypothesis tests will be incorrect. In obtaining estimates and their standard errors from CS data, the sample design, coverage errors and non-response must be taken into account since the estimates may otherwise be biased and have inaccurate variances. This is achieved by assigning a weight to each sample unit, a quantity that denotes the number of population units represented by the associated sample unit. The weight is developed in three stages. Firstly, the sample unit is assigned a design weight which is calculated as the inverse of the inclusion probability of the unit. When any sample is selected it is possible that some units omit parts of the information required from them or refuse to participate in the survey altogether. Hence, the second stage sees the adjustment of the design weights to compensate for non-response. It also frequently occurs that the achieved sample does not represent the target population as closely as intended in terms of certain subgroups being under-represented in the collected sample. Thus, the final stage of the sampling weight developement sees the correction of the non-response adjusted design weights through the use of auxiliary information such that the weighted estimates of some population totals conform to the actual known population totals of such variables. The final sampling weights contain all the information required for the calculation of point estimates for the population parameters of interest.

When conducting inference using CS data the sampling weights are incorporated in the inference, but the weight development process described here could result in extreme sampling weights that inflate the variability within the sampling weight distribution. As such the precision of the results reported from the analysis, when the variability is quite large, could be adversely affected. It has been proposed that the sampling weights be trimmed or smoothed to reduce this variability, but although the weight adjustment will decrease the variability in the weight distribution, it could also increase the bias to such an extent that the mean squared error of estimation increases as well. This is not a desirable consequence. Some approaches to weight trimming have been proposed in literature and some new approaches are introduced in this thesis.

Linear models have underlying assumptions that have to be met in order to assure a good linear model is fitted. However, real-world data rarely meet these assumptions. As such it was decided to investigate the use of resampling methods as an alternative to the standard methods, especially for the estimation of the variances of the estimators of the model parameters and the construction of confidence intervals.

The issues outlined in the above discussion, led to the following as the main objectives of this research:

- the effect of ignoring the survey design, i.e. using standard statistical methods to analyze CS data;

- the importance of calculating design weights and then benchmarking to final sampling weights;
- the effect of trimming the sampling weights on the inference precision; and
- using resampling methods such as the bootstrap and jackknife to estimate the variances of the estimators of the unknown parameters versus using the model estimated variances.

To achieve these objectives an extensive simulation study was undertaken. For the first simulation study a model was developed and used to simulate a hierarchical population from which complex samples could be selected. Simulated CS data were used to ensure that the assumptions underlying the linear modeling are met, and that any differences observed between the output of the different linear models are attributable to the type of linear model. The same simulation study was then repeated using real-world data in the form of the Income and Expenditure survey (IES), a survey conducted by Statistics South Africa in 2005. The IES was selected since it is a large data set available to researchers to use and its structure meets the design requirements of the samples that had to be selected for the thesis. Furthermore, it was important to find a data set that contained a possible continuous dependent variable and a number of appropriate covariates to include in the model. The IES met these requirements as well.

The method of simulating data makes it possible for the researcher to know what the true values of the parameters of interest are. The large real-world data set, however, was considered as a surrogate population such that "true" parameter values could be, at the very least, determined approximately. The known values made it possible to do comparisons in terms of effectiveness and accuracy of the developed models used for the estimation and/or prediction. From the simulated populations as well as the surrogate population a number of replicate samples were drawn. These samples each followed a stratified two-stage cluster design which, in the case of the IES samples, was similar to the surrogate population from which they were selected.

The simulation study has two parts, namely the evaluation of the linear models as well as inference concerning the model parameters. The linear models were evaluated based on their prediction errors through the comparison of the respective estimated prediction errors to the "true" prediction error. Within each replicate sample the model parameter inference, based on no weighting (OLS) versus weighting (incorrect WLS and correct SWLS), was investigated at two levels. The first level considered the comparison of the OLS, WLS and SWLS estimators to the "truth" by evaluating diagnostics such as the bias, mean squared error and median absolute deviation. The next level considered the estimation of the parameters and their estimated variances by making use of resampling methods for the estimation of mean squared errors, biases, etc. The results were averaged and then compared to the known population values. Also included under the model parameter inference are the various parameteric and non-parametric confidence intervals for the model parameters which were evaluated based on their non-coverage probabilities of the "true"

model parameters as well as their lengths and standardized lengths. The relationship between the main objectives and the simulation study is illustrated in figure 6.1.3. This diagram is repeated for the inference concerning the model parameters, i.e. point and interval estimation, as well as the evaluation of the model fit. Finally the summarized results were presented in tabular and graphic form and conclusions were made in line with the research objectives.

This section discussed the problem statement on which the research is based. The next section presents the scope and contribution of this research and the chapter is concluded with an outline of the thesis.

1.2 Scope and Contribution of the Thesis

This section presents the scope of the thesis as well as a list of the contributions of this research. The scope of the research is summarized in the list, given below, of the main points that will be addressed in this thesis:

1. The development of the sampling weights, that are associated with the units in a complex sample, is considered. The sampling weights are integral to any inference conducted on CS data and an important part of the research is to observe the effect of the large variation that often occurs in sampling weights, on the inference precision. For this reason various weight trimming (or winsorizing) methods exist that are used to adjust outlier weights and a selection of the commonly implemented methods is included in the study. Along with these weight trimming methods two new trimming methods were developed as part of this research.

2. The theory and methodology of multiple linear regression, viz. ordinary least squares regression, weighted least squares regression and survey-weighted least squares regression, are considered since the comparison of the results obtained from these linear models, is of importance in this research. The estimated model parameters have estimated variances and this is where one of the main differences between OLS, WLS and SWLS, resides. The parametric methodology is extended to the non-parametric linear model. The non-parametric bootstrap resampling method is widely used for the estimation of standard errors, the construction of confidence intervals, etc., and this has been extended to complex sampling to some degree. However, using the bootstrap for inference concerning the regression coefficients under complex sampling, i.e. under the SWLS model, is not well documented and has been included in this research, specifically for the estimation of the standard errors of the estimated coefficients and their confidence intervals.

3. In linear modeling it is important to be able to assess how well the estimated model will predict a future response. One way of doing this is through the assessment of the model's

prediction error, a measure that can be estimated using, for example, cross-validation or the bootstrap resampling method. Although well-known and widely used in the simple random sampling case, these methods were not developed for use under complex sampling and are thus adjusted for use in the SWLS context in this thesis.

4. The development of new techniques brings about a desire to evaluate these techniques under controlled conditions. Researchers are known to simulate data such that the data meet the assumptions underlying the technique under evaluation. This makes it possible to determine whether any differences observed between the results from the methods being compared, are attributable to the specific method applied and not due to the assumptions being violated. Simulating simple random sampling data is well-known and widely used, but the simulation of complex sampling data is not. The building of such a model is necessarily a complex matter, one that has received special attention in this thesis.

Given this scope of the research as background, the main contributions of the thesis, in the area of complex sampling, can be summarized as follows:

1. In linear modeling there are a number of important assumptions that need to be met in order to ensure the quality of the linear model being fitted. Real data typically do not meet any or all of the theoretical assumptions. The research conducted in this thesis requires CS data that meet the assumptions underlying linear modeling, but a literature review on the simulation of CS data presented a very sparse collection of such information. In this thesis a multilevel model is developed and it is shown how the model can be used for the simulation of a hierarchical population. More importantly, it is also shown how multilevel modeling can be used to ensure that the variability within the clusters remains fairly constant while the between-cluster variation is quite large, a common situation in practice. An important part of this contribution is the development of an R function that can be used for the simulation of this population. The program allows the user to specify the parameters required by the probability distributions used to simulate the data. It also allows for a hierarchical relationship between two covariates, to be simulated. Furthermore, the user can specify the number of strata he/she wishes to simulate as well as the number of clusters within a stratum and a variable cluster size. A diagram of this function is presented in figure 5.3.5. This function is available to interested researchers who wish to use the function as is, or to adjust it to their research requirements.

2. The estimation of the prediction error of a survey-weighted linear model, whether using cross-validation or bootstrap prediction error estimation methods, has not received sufficient attention in the literature and is considered one of the main contributions made by this research. The leave-one-out cross-validation method, the bootstrap estimator of prediction error and the .632 bootstrap estimator of prediction error are defined for SRS data and then

extended to CS data. A major challenge with this development, was the programming of the functions. Integrating the layouts, presented in figures 6.2.2 to 6.2.13, into the scope of the simulation study (figure 6.1.3), especially under complex sampling, was a complex procedure. Correctly accounting for the clustering in the samples and knowing when and how to adjust the sampling weights, were challenging. Finally three R functions, viz. loope, bootpe, and bootpe.63, were programmed for this thesis. These functions are available to interested parties who wish to make use of them or to adjust/improve them as part of their research.

3. Another important component of the model parameter inference are the confidence intervals for the parameters. Most statistical software report the standard (asymptotic) interval based on the variance estimated by the software, but this interval assumes that the distributional assumptions are met. The non-parametric bootstrap confidence intervals, viz. the percentile interval, the bootstrap$-t$ interval and the BCa interval, are thus introduced and defined as alternatives to the parametric standard confidence interval. These intervals are newly presented for the linear model parameter inference under CS. To obtain these non-parametric confidence intervals for the coefficients in an SWLS model, an R function was programmed. This function also had to be aligned with figure 6.1.3 such that the confidence intervals could be compared and interpreted within the scope of this thesis. Nevertheless, the function is available to other researchers and could be adjusted to align with the objectives of their work.

4. A final challenge of this research was the computer power required to carry out these complex simulations. Special "run" functions had to be programmed such that these individual functions could be submitted, as job arrays, to a cloud computer. Hence, the challenge was converted to a contribution since these functions are now available to researchers who wish to make use of cloud computing for their own computer simulations.

Apart from the above, the following are smaller contributions of this research:

1. Summary functions that were programmed in R for the calculation of various diagnostic measures, viz. bias, MSE, etc., and for the automatic production and storage of graphs and tables based on these diagnostics.

2. The novel analyses of the IES data serve as guidelines for future analyses by Statistics South Africa and other research institutions for similar data sets.

3. Of interest in this research are the extreme weights that are located in the tail of the sampling weight distribution. Two new extreme weight thresholds are introduced in this thesis, one based on Tukey's outlier detection rule, called the 1.5IQR trimming method, and the other developed from extreme value theory (EVT), an area of statistics that is concerned with

extreme deviations that reside within the tails of probability distributions. Of interest in EVT is the estimation of the extreme value index (EVI) and the Hill estimator is a well-known estimator thereof. A new trimming method was proposed called Retha's Hill method. This method determines a possible percentage of sampling weights in the tail of the weight distribution by automatically locating the biggest change in the slope of the cumulative distribution. This percentage is then used to construct the Hill plot. Finally the threshold for extreme weights is automatically identified from the point associated with the first change in the slope of the Hill plot. Both the 1.5IQR and Retha's Hill method were found to perform well.

4. It has been found, through private survey sampling consultation conducted by Dr Ariane Neethling, that some survey researchers in South Africa do not follow the sampling weight development process outlined previously. Instead of assigning a design weight to each element as the inverse of the inclusion probability of that element and then benchmarking the design weights, these researchers benchmark the raw data. Essentially, they let the design weight equal the inverse of the SRS inclusion probability of an element and then benchmark these "SRS" design weights. The investigation of this alternative approach to the calculation of sampling weights versus the theoretical approach and the difference in the inference results based on the different sets of sampling weights, is an important part of all the analyses conducted for this research. A comparison of the various diagnostics for the estimators based on the two sets of sampling weights shows that differences do exist between the same analyses conducted using both sampling weight sets.

In the next section a chapter outline is given that indicates where each of these topics are discussed.

1.3 Outline of the Thesis

The thesis consists of nine chapters. In chapter 2 sampling is discussed in general with specific focus on complex sampling, since the data sets considered in the simulation study are based on a complex sample, and the development of the final sampling weights, an intricate part of complex sampling that leads to wrong conclusions when not carried out correctly. The weight development process may lead to extreme sampling weights that could have an adverse effect on the precision of the inference carried out using these sampling weights. Chapter 3 presents weight trimming as an approach to reducing the variability within the sampling weight distribution.

Chapter 4 presents the methodology and computation of linear models and in particular the specification of the model, the estimation of model parameters and their associated variances, as well as the evaluation of the model. This chapter also contains the extension of non-parametric linear models to CS data, non-parametric confidence intervals, and the estimation of prediction errors under complex sampling.

The ability to test models under "controllable" circumstances by ensuring that the data meet the assumptions underlying the models, is the great advantage of being able to simulate data. However, simulating data that exhibit a hierarchical structure such that complex samples can be selected, is uncommon. Chapter 5 discusses the development of a multilevel model with a hierarchical structure which can be used to generate the required CS data. Chapter 6 contains the design of the simulation study based on the simulated data. Only a small subset of the results are presented to limit the length of the thesis document, but the summaries and conclusions in sections 6.2.4 and 6.3.3 are based on all of the results that are grouped into folders on a CD. Also included on the CD is a document with instructions on how to find the results in the different folders.

The simulation study outlined in chapter 6 is based on simulated data, but it is also important to investigate the main objectives based on real-world data. It has already been said that the data set used for this purpose is the IES and chapter 7 then describes the data, the methods that were used to collect the data, how non-response was dealt with, the design of the survey, and the calculation of the sampling weights. Chapter 8 considers the simulation study, outlined in chapter 6, applied to the IES. It should be noted that chapter 8 mostly contains summaries of the findings with very few results being presented. Again, this is to restrict the length of the thesis document, but all results are available and are included on the CD for the reader to peruse.

The thesis is concluded with chapter 9 which presents overall summaries of the findings of this research. The chapter concludes with a list of topics that have been identified for further research.

Chapter 2

Review of Probability Sampling Techniques

Consider a finite population U of size N and a parameter of interest, θ, to be measured. Ideally one would want to use the entire population to determine this θ, but the population is usually too large to measure, which increases cost, or too complex to collect each population unit's information necessary to calculate θ. Hence, one collects a sample of size n from the population which provides the information with which θ can be estimated. Let this estimator be denoted by $\hat{\theta}$. The quality in terms of the precision of $\hat{\theta}$ as an estimator of θ relies inter alia on how well the sample represents the population of interest. A "perfect" sample is one that mirrors every characteristic of the population, but the best chance of achieving such a "scaled-down" version of the population is to measure the entire population. Instead one aims for a good sample by ensuring that the characteristic of interest in the population, θ, can be estimated from the sample by $\hat{\theta}$, and that the accuracy of the estimation can be determined (Lohr, 2010).

When selecting a sample one can decide between using a probability sampling technique or a non-probability sampling technique. The methodology underlying non-probability techniques such as convenience or purposive sampling, automatically excludes certain population units from the sampled population since these techniques select sample units through subjective evaluation. This type of sample selection, in general, causes the estimate, $\hat{\theta}$, to be biased and in the absence of any probability techniques in the selection process, the extent of the bias is unknown. Furthermore, the non-random sample selection makes the estimation of sampling errors impossible. To conclude, any inference inferred from non-probability samples will be subject to an unknown amount of bias (Lohr, 2010).

Probability sampling techniques ensure that each possible sample of size n collected from the finite population has a known probability of being the selected subset. The employment of a random mechanism to determine which population units to select for the sample decreases the possibility of changing a pre-selected unit for a different unit based on personal judgement. Hence, through the application of a probability sampling technique, each population unit has a known positive chance of appearing in the sample. The probabilities underlying all possible samples of

size n collected using a probability sampling technique make it possible to establish the sampling distribution of $\hat{\theta}$, the estimator of θ. This makes it possible to conduct inference using $\hat{\theta}$ and also to determine the quality of the inference through the evaluation of standard errors, biases, etc. of the estimators (Lohr, 2010).

The purpose of this chapter is to revise well-known probability sampling methods such as

- simple random sampling;
- systematic sampling;
- stratified sampling; and
- cluster sampling,

and also to discuss each method's advantages and disadvantages. This will be followed by a discussion of complex sampling, a stratified multistage cluster sampling method used to improve the representativeness of the collected sample. Finally, sampling weights will be defined and discussed as an integral part of complex sampling.

2.1 Simple Random Sampling

A simple random sample of size n is the most well-known and widely applied probability sampling method in which every set of n elements has the same probability of being selected as the sample (Lohr, 2010). Simple random sampling can be conducted in two ways.

1. Simple random sampling with replacement (SRSWR)

   When using SRSWR, a sample of size n is selected from a population of size N by replacing a selected element after it's been selected. This method can be thought of as taking n independent samples of size 1. The first sample element is selected with probability $\frac{1}{N}$ after which it is placed back into the sample. Since the population size has not changed, the second element is also selected with probability $\frac{1}{N}$. Hence, if $\mathbf{y} = y_1, ..., y_n$ denotes the sample then the probability of the ith unit appearing in the sample is $\frac{1}{N}$. This procedure is repeated until the desired sample size is achieved. Note that the achieved sample may contain duplicates due to the replacement of previously selected elements (Lohr, 2010).

2. Simple random sampling without replacement (SRS)

   When the same element appears more than once in the sample there is a loss of information and this can be rectified by selecting a sample without replacement. The aim with this type of sampling is that each distinct sample of n elements has the same probability of being selected. There are $\binom{N}{n}$ possible subsets of size n that can be selected from a population of

size N with probability $\frac{1}{\binom{N}{n}}$ of being the selected sample. Under this sampling method it follows that element i appears in the sample with probability $\frac{n}{N}$ (Lohr, 2010).

The main advantage of SRS is the simplicity of the application of this sampling method. Also, it is the only assumption free sampling method. However, selecting a sample using SRS does not guarantee a sample representative of the target population and it requires a complete and up-to-date list, i.e. sampling frame, of sampling units in the population from which to select the sample (Lohr, 2010).

2.2 Systematic Sampling

Systematic sampling (SS) can be used, instead of SRS, when there is no list of the population units, e.g. in a production line, or when the population has been ordered according to some ordering scheme (Lohr, 2010).

Consider a population of size N from which a sample of size n has to be selected. The SS method requires a selection interval as well as a random starting point to commence the sample selection. Firstly, calculate a value

$$k = \frac{N}{n}.$$

If k is not an integer, one possible approach is to choose the next integer after k as the selection interval length to use in the application of the sampling method. The sampling method proceeds by selecting a random integer between 1 and k, say L. This integer represents the first observation of the sample and then every kth observation will be added to the sample until the desired sample size is reached. This results in a sample of the form

$$S = \{L,\ L + k,\ L + 2k,\ ...,\ L + (n-1)\,k\}\,.$$

SS forms part of the probability sampling methods as long as it makes use of a random starting point. In contrast to SRS, all subsets of size n do not have the same probability of being the selected sample since, given a selection interval length of k, the probability of selecting two consecutive observations is highly unlikely. However, if the population is in random order it will be much like an SRS. The importance of this statement lies in the fact that the achieved sample can be compared to an SRS and SRS methods can be applied in the inference from the systematic sample (Lohr, 2010).

Systematic sampling is a fast and convenient method to use and it could result in a more representative sample than an SRS, especially if the units are in a specific order. On the other hand, if some form of periodicity exists in the population, then SS does not necessarily produce a representative sample. Especially if the period is a multiple or factor of the selection interval

length, the sample will be unrepresentative of the population and consequently less accurate than an SRS. For example, if sales data is sorted according to seasons and the interval length is 4, one could end up with a sample of just summer sales figures which will introduce bias into any results obtained from the data. Also, if a sampling frame is used, it should be complete and up-to-date (Lohr, 2010).

2.3 Stratified Random Sampling

In stratified random sampling the population of size N is divided into subgroups called strata such that each population unit belongs to only one stratum. The aim is to divide the population in such a way that population units in a stratum are similar, homogeneous, which ensures that the within-stratum variation is minimized. Also, the subgroups have to be set up such that the between-strata variation is maximized, i.e. that heterogeneity is achieved. When the maximum between-strata variation and minimum within-stratum variation is achieved it follows that a stratified random sample provides estimates with smaller standard error, i.e. better precision, when compared to estimates obtained from an SRS (Lohr, 2010).

Along with improved precision, another advantage of stratified random sampling in comparison to SRS is the improved chance of obtaining a representative sample of the population. The subdivision of the population into non-overlapping subgroups according to some characteristic implies that the subgroups can be treated as independent sub-populations and a sample can thus be selected independently from each subgroup. This ensures that each subgroup is properly represented in the final sample without increasing the selection bias. Furthermore, since independent samples can be collected from each stratum it follows that inferences about the individual subgroups can be made, information which could be lost in more general random sampling (Lohr, 2010).

A further advantage of stratified random sampling in comparison to SRS includes the application of different sampling methods within strata to collect the samples from the respective strata. Once again the independence of the strata makes this possible, a valuable characteristic if it should happen that some sampling methods are more appropriate in certain strata than in others. Finally, these samples tend to be easy to administer and could also decrease the survey cost (Lohr, 2010).

2.4 Cluster Sampling

Cluster sampling divides the population into subgroups called clusters and a sample of clusters is then selected of which all or some of the units in each cluster can be included in the sample. When all the units in a sampled cluster are included in the sample the method is referred to as one-stage cluster sampling and the clusters are called primary sampling units (PSU's). When the units in a selected PSU are sub-sampled the units are called secondary sampling units (SSU's) and

the sampling method is called two-stage cluster sampling (Lohr, 2010).

A sample of clusters can be drawn in two ways (Lohr, 2010):

1. With equal probability irrespective of the number of population units in each cluster (EPSEM); or

2. With probability proportionate to some measure of the size of the clusters (PPS).

As an illustration of the difference between the two selection methods, suppose the number of grade 12 learners in the Western Cape that take mathematics as a subject must be estimated from a cluster sample of grade 12 learners. The schools in the Western Cape are used to divide the population into clusters from which a number of schools need to be selected. Suppose there are N clusters in the population from which n must be selected. If the clusters are selected with equal probability irrespective of the number of grade 12 learners in a school, the inclusion probability of the jth school, $j = 1, ..., N$, is given by

$$\pi_j = P\left(School\ j\ selected\right) = \frac{n}{N}.$$

However, suppose the clusters will be selected according to the number of grade 12 learners in a school. Here the number of grade 12 learners will be considered as the measure of size (MOS). Let A_j denote the number of grade 12 learners in the jth cluster, $j = 1, ..., N$, such that $\sum_{j=1}^{N} A_j$ is the total MOS of the population. Now the inclusion probability of the jth school becomes

$$\pi_j = P\left(School\ j\ selected\right) = n \cdot \frac{A_j}{\sum A_j},$$

where n is the number of cluster to be selected for the sample.

The benefits of cluster sampling is that this is often the only feasible method of sampling in the absence of a complete list of population observation units from which a sample must be selected. Also, it is considered a cost efficient form of sampling. Bare in mind that cluster sampling does not necessarily guarantee a representative sample. Furthermore, where stratification generally improves estimation precision, clustering decreases precision. This occurs since units within the same cluster tend to be more similar, i.e. homogeneous, and thus less information is gained by sampling units within the same cluster than when randomly sampling units from a population (Lohr, 2010).

2.5 Complex Sampling

A complex sample (CS) is defined as a stratified multistage cluster sample. The procedure for selecting a CS starts by dividing the population into non-overlapping subgroups called strata. Recall from the earlier discussion of stratified random sampling, the stratification succeeds at ensuring

that all strata are represented in the final sample and by extension improves the representativeness of the sample. Each stratum is then divided into meaningful clusters from which a predetermined number is selected. These first stage clusters are called the primary sampling units (PSU's). It is important to ensure that at least two PSU's are selected per stratum to enable variance estimation. Each of the selected PSU's is then once again divided into smaller clusters from which a predetermined number is selected. These second stage clusters are called the secondary sampling units (SSU's). Note that the PSU's may be stratified before the SSU's are formed and selected. One continues in this way until the population units of interest are reached and thus selected for the sampling. The final stage units are called the ultimate sampling units (USU's) (Lohr, 2010).

Figure 2.5.1 below illustrates a general stratified two-stage cluster sample where the population has been stratified into H strata. Consider stratum h and suppose the stratum contains N_h PSU's of which a sample of n_h PSU's is selected. Let the jth selected PSU contain M_{hj} SSU's and suppose a sample of m_{hj} SSU's is selected from the jth PSU in the hth stratum, $j = 1, ..., n_h$ and $h = 1, ..., H$.

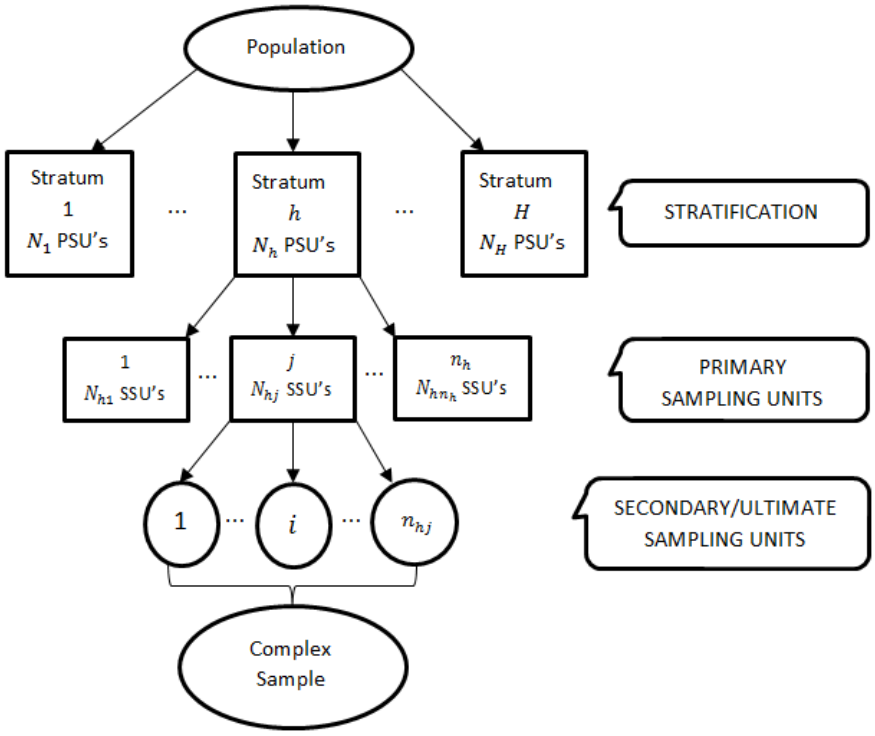


Figure 2.5.1: Stratified Two-stage Cluster Design

Complex sampling makes the step-by-step design of a sample possible and by making use of a combination of meaningful stratification and clustering a representative sample could be designed. Furthermore, if a complete list of the desired population observation units does not exist, through making use of a customized complex sample design a sample of these observation units can be obtained without a complete list of these units. Hence, CS does not require a complete sampling

frame of observation units.

Under the discussion of, respectively, stratified and cluster sampling it was explained how stratification improves estimation precision while cluster sampling tends to deliver estimates with lower precision in comparison to an SRS. Since CS consists of a combination of stratification and clustering it is possible to obtain estimates from a CS that has lower precision than those obtained from an SRS. Hence, a larger sample may be required to achieve the same estimation precision as under an SRS, but using CS is still more convenient and has lower cost per unit than an SRS. This means that it is possible to obtain the same precision through CS as through SRS at a lower cost even if a larger sample is required (Lohr, 2010).

2.6 Weighting

The aim in sampling theory is to make inferences about population parameters of interest and thus the conclusions made from the sample need to be generalisable to the population. Herein lies the importance of using sampling weights in inference. The weights are used to correct for imperfections such as unequal probabilities, population groups not adequately represented in the achieved sample as well as non-response. To summarize, sample weights are developed in different stages to

1. compensate for unequal inclusion probabilities;
2. compensate for non-response; and
3. adjust the achieved sample to represent the target population more closely.

The first stage assigns a design weight, also called a base weight, to each sample unit to adjust for the unequal inclusion probabilities. This is followed by an adjustment of the design weights of the respondents such that the respondents represent the non-respondents as well. Final weight adjustments are necessary since the achieved sample does not necessarily coincide with the target population due to over-/ under-representation of certain population subgroups. This adjustment is made through the use of auxiliary information of the population available through resources such as censuses (Neethling and Galpin, 2006; Lohr, 2010).

This section will discuss these different stages involved in the development of the final sampling weight to be used in the estimation of population parameters of interest.

2.6.1 Calculating the Design Weight

The use of probability sampling techniques make it possible to determine the inclusion probability of a population unit in the achieved sample. Let the inclusion probability of the ith population

unit be defined as π_i and let d_i denote the design weight. The design weight is defined as the inverse of the inclusion probability of a population unit to be selected for the sample,

$$d_i = \frac{1}{\pi_i},\ i = 1, ..., n,$$

where n is the size of the selected sample, and is interpreted as the number of population units represented by the ith sampled unit. Consequently $N = \sum_i d_i$, the size of the population from which the sample is selected.

Consider a sample $\{y_i\}$, $i = 1, ..., n$ of size n selected from a population of size N using an equal probability sampling (EPSEM) method, namely simple random sampling. In practice the populations from which samples are selected are typically large enough for the samples to be selected without replacement. When SRS is used the inclusion probability of the ith observation is defined as $\pi_i = \frac{n}{N}$. From the definition of a design weight it thus follows that the design weight of the ith observation in this scenario is given by

$$d_i = \frac{1}{\pi_i} = \frac{N}{n},$$

where d_i represents the design weight of observation i. Let $\{d_i\}$, $i = 1, ..., n$ denote the design weights of all n sample observations. When the sum of the design weights is considered,

$$\sum_i d_i = \sum_i \frac{N}{n} = N,$$

it is seen that this sum indeed represents the population size from which the sample was selected (Lohr, 2010).

When a sample is selected in such a way that the inclusion probabilities of all units in the sample are equal, namely $\pi_i = \pi$, $i = 1, ..., n$, then the sample is self-weighting. Since the design weight of a sampled unit is equal to the inverse of its inclusion probability, this phenomenon thus implies that the design weights of all units in a self-weighting sample are also equal. Hence, each observed unit represents the same number of unobserved units in the population (Lohr, 2010). In the SRS case described above the design weights of the observations remain the same which means that SRS always results in self-weighting samples. This, however, is not always the case for all designs, as discussed below.

In stratified sampling an independent sample is selected from each stratum which can result in different inclusion probabilities in the various strata leading to unequal design weights if disproportional allocation is applied. Suppose the population is divided into H non-overlapping strata and each stratum contains N_h, $h = 1, ..., H$ population elements from which a sample of size n_h, $h = 1, ..., H$ is to be selected per stratum. These samples can be selected proportionately or disproportionately to the number of population units in each stratum.

When proportional allocation is used it implies that the number of sampled units in each

stratum is proportional to the population size of the stratum. Hence, the inclusion probability of the ith observation from the hth stratum is equal to

$$\pi_{hi} = \frac{n_h}{N_h} = \frac{n}{N},$$

and remains the same for all strata. Finally, the design weight associated with the ith observation in the hth stratum is

$$d_{hi} = \frac{N_h}{n_h},$$

where the sum over all design weights over all the strata equals the population total (Lohr, 2010). Since the inclusion probabilities and, by extension, the design weights remain the same, it follows, from the definition of a self-weighting sample, that stratified sampling with proportional allocation results in self-weighting samples.

Disproportionate and optimal allocation can be used when there is a desire to apply different sampling rates to the different strata. The need for this arises when, for example, some of the strata are quite small, but contain information of great importance to the estimation of the parameter of interest. Applying different sampling rates across the strata will ensure that the smaller strata are adequately represented by oversampling from them (Lohr, 2010).

Cluster sampling, where the population is divided into N non-overlapping clusters from which a sample of n must be selected, is another example of how the design weights can differ between observations. Let M_j denote the number of population units in the jth selected cluster and let m_j denote the number of observations to be sampled from each of the selected clusters for $j = 1, ..., n$.

Consider the case of one-stage cluster sampling where a number of clusters is selected from the population without further sampling from the selected clusters. This type of cluster sampling can be used when the cost of sampling from the selected clusters is negligible compared to the cost of sampling the clusters. In this case it follows that $M_j = m_j$.

As explained before, the clusters can be sampled with equal probability (EPSEM) or with PPS. If the clusters are sampled with equal probability, then the probability of selecting the jth cluster is given by

$$\pi_j = \frac{n}{N},$$

where N is the number of clusters in the population and n is the number of sampled clusters. Since all units within the selected cluster are included in the sample, it follows that the probability of selecting an observation given that the jth cluster has been selected, is one. From this it follows that the inclusion probability of an observation under one-stage cluster sampling will be equal to

$$\pi_{ji} = \left(\frac{n}{N}\right) \cdot (1) = \frac{n}{N} = \pi_j,\ j = 1, ..., n,$$

and thus the design weight under one-stage cluster sampling with equal probability is equal to

$$d_{ji} = d_j = \frac{N}{n}.$$

It should be noted that one-stage EPSEM cluster sampling thus also results in self-weighting samples.

If the units within a cluster are very similar, measuring all the units in the cluster is unnecessary and does not contribute information to the sample. Since the variability within a cluster is usually smaller than the variability between clusters, it will be more meaningful to draw more clusters and then take a sample of units from each sampled cluster for a given sample size. This approach is called two-stage cluster sampling by which a sample of clusters is selected at the first stage after which a sample of units from each sampled cluster is selected at the second stage. Recall that there are N clusters in the population from which n are selected. Within each cluster there is M_j units from which a sample of m_j is selected. Suppose the ith observation within the jth sampled cluster is selected under EPSEM. The inclusion probability of this unit consists of two parts:

1. The selection probability of the jth cluster; and

2. The selection probability of the ith observation given that the jth cluster is selected.

As before, under EPSEM the probability of selecting the jth cluster is equal to $\pi_j = \frac{n}{N}$. However, under two-stage sampling the probability of selecting the ith observation given that the jth cluster has been selected is equal to

$$\pi_{i|j} = \frac{m_j}{M_j}.$$

Now the inclusion probability of the ith observation within the jth sampled cluster is given by

$$\pi_{ji} = \pi_j \cdot \pi_{i|j} = \frac{n}{N} \cdot \frac{m_j}{M_j},$$

where $i = 1, ..., m_j$ and $j = 1, ..., n$ (Lohr, 2010). Finally, the design weight under two-stage equal probability cluster sampling is given by

$$d_{ji} = \frac{N}{n} \cdot \frac{M_j}{m_j}.$$

By selecting the SSU's with the same proportion, two-stage EPSEM cluster sampling also results in self-weighting samples (Lohr, 2010).

Under PPS, define a meaningful measure of size (MOS) and let A_j denote the MOS of the jth cluster such that $\sum_j A_j$ is the total MOS of the population. Consider the section on cluster sampling for an example of this. For one-stage PPS cluster sampling it follows that the probability of selecting the jth cluster is equal to

$$\pi_j = n \cdot \frac{A_j}{\sum_j A_j},$$

where n is the number of clusters to be selected for the sample. Recall that in one-stage cluster sampling all units within a selected cluster are included in the sample giving these units a selection probability of 1. Hence, the inclusion probability of the ith observation in the jth cluster is equal to

$$\pi_{ji} = \left(n \cdot \frac{A_j}{\sum_j A_j} \right) \cdot (1),$$

from which it follows that the design weight of the ith observation under one-stage PPS cluster sampling is equal to

$$d_{ji} = d_j = \frac{1}{n} \cdot \frac{\sum_j A_j}{A_j}.$$

However, as explained previously, under two-stage cluster sampling the selection probability of the ith observation differs from one. Following the same reasoning as before, but using PPS cluster sampling, the inclusion probability of the ith observation is given by

$$\pi_{ji} = \pi_j \cdot \pi_{i|j} = \left(n \cdot \frac{A_j}{\sum_j A_j} \right) \cdot \frac{m_j}{M_j},$$

where m_j and M_j are, respectively, the number of units sampled from cluster j and the total number of units in cluster j. Note that in two-stage PPS cluster sampling where the MOS is chosen as the number of units in the cluster, M_j, the above inclusion probability simplifies to

$$\pi_{ji} = \pi_j \cdot \pi_{i|j} = \left(n \cdot \frac{M_j}{\sum_j A_j} \right) \cdot \frac{m_j}{M_j} = n \cdot \frac{m_j}{\sum_j A_j}.$$

Finally, the design weight of this observation under general MOS is calculated as

$$d_{ji} = \frac{1}{n} \cdot \frac{\sum_j A_j}{A_j} \cdot \frac{M_j}{m_j},$$

and where $A_j = M_j$,

$$d_{ji} = \frac{1}{n} \cdot \frac{\sum_j A_j}{m_j},$$

Lohr (2010).

Now, let the number of SSU's in a PSU be the same as the MOS used to select the PSU's. Thus, $M_j = A_j$. To achieve a self-weighting sample, let the same number of SSU's now be selected from each PSU, i.e. $m_j = c$. Then the inclusion probability of the ith SSU becomes

$$\pi_{i|j} = \frac{m_j}{M_j} = \frac{c}{A_j}.$$

Finally, the inclusion probability of the ith SSU in the jth PSU is equal to

$$\pi_{ji} = \left(\frac{nA_j}{\sum_j A_j}\right) \cdot \left(\frac{c}{A_j}\right) = \frac{nc}{\sum_j A_j},$$

which remains constant. Thus the design weight in this case will also remain constant resulting in a self-weighting two-stage cluster sample (Lohr, 2010).

Now consider a stratified two-stage cluster sample design, an example of a CS. The sample is selected from a population that has firstly been stratified into H strata. Suppose stratum h has been divided into N_h PSU's of which n_h has been sampled, $h = 1, ..., H$ with equal probability. It follows that the selection probability of the jth PSU in the hth stratum, π_{hj}, is given by

$$\pi_{hj} = \frac{n_h}{N_h}.$$

Let the jth sampled PSU be clustered into M_{hj} SSU's of which m_{hj} are sampled with equal probability, $j = 1, ..., n_h$. The selection probability of the ith SSU given that the jth PSU in the hth stratum has been selected, $\pi_{i|hj}$, is defined as

$$\pi_{i|hj} = \frac{m_{hj}}{M_{hj}}.$$

Finally, the inclusion probability of the ith SSU in the jth PSU of the hth stratum is calculated as

$$\pi_{hji} = \pi_{hj} \times \pi_{i|hj} = \left(\frac{n_h}{N_h}\right) \cdot \left(\frac{m_{hj}}{M_{hj}}\right),\ h = 1, ..., H,\ j = 1, ..., n_h,\ i = 1, ..., m_{hj},$$

and thus the design weight, when PSU's and SSU's are sampled with equal probability, is given by

$$d_{hji} = \left(\frac{N_h}{n_h}\right) \cdot \left(\frac{M_{hj}}{m_{hj}}\right).$$

Similar reasoning is applied when the PSU's and SSU's are sampled with other sampling mechanisms.

Provided that proportional allocation was applied to the strata, it is possible to design self-weighting complex samples, which might result at last in unequal weights. In household surveys, for example, the PSU's (e.g. enumerated areas) as well as the SSU's (e.g. households) could be selected in such a way as to yield a self-weighting sample on household level. Since one person is

often selected per household to be interviewed and household sizes differ, the sample at person level is no longer self-weighting. Also, samples often have unit non-response as well as certain levels of differential non-response (under/over coverage) of certain groups, that need to be corrected in ways that will change the weight of the sample units. Furthermore, disproportional allocation to strata is usually used especially when precise estimates are required per stratum and consequently smaller strata need to be over-sampled to obtain sufficient information for this purpose. Disproportional allocation will lead to unequal inclusion probabilities and thus a non-self-weighting sample (Ajayi et al., 2005).

One of the advantages of a self-weighting sample is that standard statistical methods may be used to obtain point estimates from the sample. However, in complex samples where this property is rarely achieved the use of these standard statistical methods will lead to wrong standard errors, confidence intervals and hypothesis test results and biased point estimates (Lohr, 2010).

Now, let the design weight under a stratified two-stage CS be classified as *approach one* and denoted as $d_{CS_{hji}}$, $h = 1, ..., H$, $j = 1, ..., n_h$ and $i = 1, ..., m_{hj}$, or in short d_{CS}. In contrast to this approach, some survey practitioners do not assign such a design weight to each observation before commencing with benchmarking, the final stage in the development of the sampling weights that will be discussed later. Instead, the design weight becomes

$$d_i = \frac{N}{n},\ i = 1, ..., n,$$

which corresponds to the design weight under equal probability SRS. Here the subscripts "h" and "j" are omitted to emphasize that the sample design has been ignored. Let this be known as *approach two* and let this design weight be denoted as d_{SRS_i}. This second approach to the calculation of the design weights of a CS is often observed in practice and does not follow the general theory of the calculation of design weights. Part of the analyses conducted for this thesis will be to investigate the effect of using d_{SRS} and the benchmarked d_{SRS} weights as apposed to the d_{CS} and benchmarked d_{CS} weights.

It can be seen here that the number of PSU's, SSU's and USU's selected is directly used in the calculation of the correct design weights, d_{CS}. Recall that the sum of the design weights should equal the population size from which the sample was selected. When the survey is carried out and the collected information is considered it is usually found that some information, even entire records, are missing. If the sum of the design weights are now calculated it will not equal the population size due to these non-responses. The next section discusses the types of non-response that typically occur and explains how this phenomenon can be handled.

2.6.2 Adjustment of Sample Weights for Non-response

When some population units selected for the sample do not respond it might have an effect on the survey design since non-respondents often differ quite substantially from respondents. Non-

response occurs in two ways, namely:

1. Item non-response

   This non-response occurs where the sampled units omit answers to certain questions in a questionnaire. This is mainly observed due to the sampled unit refusing to make certain information known.

2. Unit non-response

   Here the entire sampled unit's information is missing. This type of non-response can occur due to the fieldworker not being able to make contact with the sampled unit, the sampled unit being unable to take part in the survey or the sampled unit refusing to be part of the survey.

Lohr (2010) states that the best way to deal with non-response is to prevent it. However, it is rarely the case that the desired information is obtained from all the sampled units. Furthermore, since respondents and non-respondents typically differ from each other, and if there is a non-negligible non-response in the survey, the estimates of the parameters of interest based only on the respondents' information will be biased (Lohr, 2010). Thus, it is important to take care of non-response. There are different approaches to follow:

- Prevention by designing the survey in such a way that the response rate is high. This is the preferred method;
- Taking a representative sample of non-respondents and using it to make inferences about the other non-respondents;
- Using a model to predict the values of the non-respondents; or
- Ignoring the non-response, but this is not recommended.

Even after carefully designing a survey in such way as to minimize the non-response there is always some non-response that occurs and this needs to be dealt with. Define an indicator variable R such that when a sample unit responds the variable takes on the value one, and when the unit does not respond it takes on the value zero. Furthermore, let y_i denote the response of interest and let $\mathbf{x}_i$ be a vector of information that is known about the ith unit in the sample and used in the survey design. Then the probability that unit i, which is selected for the sample, responds is given by

$$\phi_i = P\left(R_i = 1\right), \tag{2.6.1}$$

which is an unknown probability (Lohr, 2010).

In terms of unit non-response there are three types to consider, such as

1. responses missing completely at random (MCAR);

2. responses missing at random (MAR); and

3. responses not missing at random (NMAR).

MCAR non-response is observed when the unit's chance of responding, ϕ_i, is independent of the response required, y_i, as well as the information known about the unit, $\mathbf{x}_i$. In this case the respondents are representative of the non-respondents and the estimator of the parameter of interest obtained only from the respondents' information, will be unbiased. When the information is MAR the chance of responding depends only on $\mathbf{x}_i$, the known information. This type of non-response is often referred to as ignorable in the sense that the non-response can be ignored once a model has explained the non-response mechanism. Finally, when the chance of non-response cannot be explained by the observed variables, then the non-response is NMAR. Models can help in this case, but cannot completely adjust for the non-response (Lohr, 2010). Lohr (2010) is of the opinion that the non-response in surveys is expected to be MAR.

When the non-response is MAR the design weights of the respondents can be adjusted such that the achieved sample is closer to the intended sample. Suppose information exists that is known for all sampled units and that makes it possible to divide the sample into C weighting classes by cross-classifying the categories of these variables. It is assumed that the respondents and non-respondents in the same weighting class are similar. Recall that each sampled unit has a design weight, discussed in section 2.6.1, denoted by $d_i = \frac{1}{\pi_i}$, $i = 1, ..., n$. In each weighting class the response probability ϕ of that class is estimated as

$$\hat{\phi}_c = \frac{sum\ of\ respondent\ weights\ in\ class\ c}{sum\ of\ all\ weights\ of\ units\ in\ class\ c},\ c = 1, ..., C.$$

Define an indicator variable such that $x_{ci} = 1$ if the ith sampled unit is in weighting class c. The weight of the ith unit in class c, which also responded, is adjusted to

$$\tilde{d}_i = d_i \sum_c \frac{x_{ci}}{\hat{\phi}_c},$$

such that the adjusted design weight of a respondent in weighting class c is inflated to $\tilde{d}_i = \frac{d_i}{\hat{\phi}_c}$ while the non-respondent's design weight becomes zero (Lohr, 2010; Ajayi et al., 2005). By doing this the respondents become representative of the non-respondents and the sum of the design weights should again equal the population size from which the sample was selected.

Now that the design weights of the respondents have been adjusted to account for the non-respondents, some sample characteristics, such as gender and race, can be verified to determine whether the achieved sample represents the target population as closely as intended. Skewness in the achieved sample, such as the over- and/or under-representation of certain population sub-groups, could occur due to the random sampling mechanism, the sampling frame, non-response,

etc. Correcting for this unintended skewness through the benchmarking of the non-response adjusted design weights, is discussed in the next section.

2.6.3 Adjustment of Sample Weights for Differential Non-response

As mentioned before, it is often the case that the achieved sample does not represent the target population as closely as intended in terms of certain subgroups being under-/ over-represented. This occurrence (such as too few young males or small households) is quite common in practice and could lead to biased results if ignored. Therefore it should be identified and controlled. Some of the approaches to handling coverage errors due to under-/ over-representation, are:

1. Improved field procedures, and/or
2. Compensating for over-coverage and/or under-representation through the adjustment of design weights.

The final stage of weight construction is then where this design weight adjustment occurs. This stage makes use of auxiliary information, obtained from census data or other population data sources, to adjust the non-response adjusted weights of the sampled units such that the weighted estimates of the population totals conform to the actual known population totals of such variables (Neethling and Galpin, 2006). The following weight adjustment methods exist under this approach and will be considered here:

1. Post-Stratification,
2. Cell-Weighting,
3. Calibration Weighting, and
4. Integrated Weighting.

2.6.3.1 Post-Stratification and Cell-Weighting

An adjustment made by means of post-stratification consists of dividing the sample elements into subgroups called post-strata. After this has been done, an adjustment is made to the weight of each element in a given subgroup by using the known population counts. The adjustment is made to correct the effects of differential non-response in the post-strata, or to reduce the variances of estimators involving variables correlated with characteristics used to partition the population into post-strata. If the fixed total for each post-stratum is equal to the expected value of the sample estimate of that total, then the procedure introduces no bias (Rust et al., 1996). Post-stratification makes use of a ratio estimator within each subgroup to adjust by the true population count. Let

$$x_{ai} = \begin{cases} 1, & if\ i\ is\ a\ respondent\ in\ post-stratum\ a \\ 0, & otherwise \end{cases}.$$

Then let

$$w_i^* = \sum_{a=1}^{A} d_i x_{ai} \cdot \frac{N_a}{N_{aR}}, \tag{2.6.2}$$

where A is the number of post-strata, N_a is the population total in post-stratum a, N_{aR} is the population total in post-stratum a based only on respondents and d_i is the design weight (or $\tilde{d}_i$) of the ith sampling unit. The weight defined in (2.6.2), w_i^*, is called the post-stratum weight (Lohr, 2010).

When using the post-stratification adjusted weights the estimates will be approximately unbiased within each post-stratum under the following circumstances (Lohr, 2010):

1. when every unit has the same ϕ_i as defined in (2.6.1); and
2. when the non-response can be classified as MAR.

These are strong assumptions, but Lohr (2010) states that in practice the researchers often use many post-strata to make sure the assumptions are met. Note that this could be problematic since using a large number of post-strata can lead to empty or too few respondents per post-stratum which leads to unstable estimates. To avoid this problem be sure that there are at least 20 observations per post-stratum or that the response rate is above 50% (Lohr, 2010).

Cell-weighting and post-stratification work well where population numbers in the interlaced cells are known and the sample is large enough, but population information is often available only at certain levels. It is also ineffective when cells that are too small or empty appear in the sample. This is where calibration and integrated weighting can be used (Neethling, 2004). Calibration weighting is discussed in section 2.6.3.2 and integrated weighting is discussed in section 2.6.3.3.

2.6.3.2 Calibration Weighting

The calibration technique was introduced by Deville and Särndal (1992) and by Deville et al. (1993). It is a widely used procedure for obtaining improved estimates in sampling surveys by making use of auxiliary information in the form of known population totals to produce a new adjusted set of weights, called calibration weights. Here, suppose a two-stage cluster sample with PSU's sampled at the first level and SSU's sampled at the second level.

The following notation should be introduced (Neethling and Galpin, 2006):

- A sample, S, of n PSU's with a total of m units is drawn from a finite population, U, of N PSU's with a total of M units. Weighting cells are formed by using categorical variables that

are known for all units in the sample and subgroups (cells) are formed by cross-classifying the categories of these variables. It is assumed that respondents and non-respondents in the same cell are similar (Lohr, 2010). The weights of the respondents are then adjusted so that the achieved sample represents the intended sample, and hence the population (Neethling, 2004). Let

- m_j, the number of SSU's in PSU j, $j = 1, ..., n$.
- m, the number of SSU's sampled, $\sum_{j=1}^{n} m_j = m$.
- π_i, the inclusion probability of the ith population element.
- $\Pi = diag(\pi_i)$, the $N \times N$ diagonal matrix of inclusion probabilities.
- $d_i = \frac{1}{\pi_i}$, the design weight of $i \epsilon U$.
- y_i, the study variable.
- $\mathbf{Y} = (y_1, ..., y_N)^{'}$, the N-vector of values of the study variable.
- $x_1, ..., x_Q$, the Q auxiliary variables.
- $\mathbf{x}_i = (x_{i1}, ..., x_{iQ})^{'}$, the Q-vector for each $i \epsilon U$.
- $\mathbf{X}_T = \sum_{i \epsilon U} \mathbf{x}_i$, the Q-vector with known population totals.
- $\hat{\mathbf{X}}_\pi = \sum_{i \epsilon S} d_i \mathbf{x}_i$, the Horvitz-Thompson estimator of the auxiliary variables.

Vectors and matrices for the sample will be denoted by the subscript S (Neethling, 2004).

The auxiliary information can be obtained from external sources such as census data. The calibration estimator is given by

$$\hat{Y}_{cal} = \sum_{i \epsilon S} w_i y_i, \tag{2.6.3}$$

where w_i are the calibration weights that are as close as possible to the design weights, d_i (Neethling and Galpin, 2006). The calibration weights are subject to a set of constraints, namely

$$\sum_{i \epsilon S} w_i \mathbf{x}_i = \mathbf{X}_T, \tag{2.6.4}$$

where the vector $\mathbf{X}_T$ contains the known population totals and $\mathbf{x}_i$ is a vector containing the values of the different auxiliary variables for each element in the population. Equation (2.6.4) ensures that the sample sums of the weighted auxiliary variables equal the known population totals for those variables (Neethling, 2004).

Consider a general distance function

$$G\left(w_i, d_i\right) = d_i v_i G\left(\frac{w_i}{d_i}\right),$$

that measures the distance between the original design weight d_i and the new weight w_i, where v_i is a known positive weight unrelated to d_i (Deville and Särndal, 1992).

Now, new weights w_i, $i \epsilon S$, have to be found that minimize the average distance for the whole sample,

$$min_{w_i} \sum_{i \epsilon S} G\left(w_i, d_i\right),$$

subject to the constraint in (2.6.4). From this it follows that the calibration weights are given by

$$w_i = d_i F\left(\frac{\mathbf{x}_i^{'} \lambda_c}{v_i}\right),$$

where $\lambda_c = \left(\lambda_1, ..., \lambda_J\right)^{'}$ is the Lagrange multiplier vector and F is the inverse function of $\frac{dG(\psi)}{d\psi}$ for $\psi = \frac{w_i}{d_i}$ (Neethling and Galpin, 2006). Thus, the calibration estimator in (2.6.3) is now given by

$$\hat{Y}_{cal} = \sum_{i \epsilon S} d_i F\left(\frac{\mathbf{x}_i^{'} \lambda_c}{v_i}\right) y_i.$$

Several distance functions have been suggested in the literature, inter alia the linear, exponential (or the so called raking ratio), logit (truncated exponential) and truncated linear methods. In the case of the linear method the calibration weights are given by

$$w_i = d_i \left(1 + \mathbf{x}_i^{'} \lambda_c / v_i\right),$$

where λ_c is determined by the solution to the system

$$\left(\sum_{i \epsilon S} d_i \mathbf{x}_i \mathbf{x}_i^{'} / v_i\right) \lambda_c = \mathbf{X}_T - \hat{\mathbf{X}}_\pi,$$

and v_i is usually set equal to one (Neethling, 2004).

The efficiency of the estimator $\hat{Y}_{cal}$ depends on how well the auxiliary variables explain the variability of the variable of interest. Thus, the weights perform well given that there exists a strong correlation between auxiliary variables and study variables (Neethling, 2004).

One of the disadvantages of this method is that it may produce weights that are either negative, resulting from an over-constrained system, or large and positive, leading to an increase in the standard error of the estimator. Also, the shortcoming of using a calibration technique for adjusting USU weights, is that the weights will usually differ from USU to USU within the same SSU

(e.g. among household members within the same household). Hence, it does not produce a representative SSU (e.g. household) weight which could be used to estimate SSU variables of interest. Furthermore, the calibration estimators do not take the SSU as a cluster into account (Neethling and Galpin, 2006).

2.6.3.3 Integrated Weighting

In the past, surveys generally used separate weighting procedures for estimating USU and SSU characteristics. As a result different sets of weights were obtained. Since calibration weighting produces weights that differ between SSU elements, it does not produce a representative SSU weight either. This can introduce some uncertainty in estimating SSU variables. Integrated linear weighting was developed to achieve a single set of weights that can be used for both USU and SSU estimation (Neethling, 2004).

Integrated Weighting: SSU Level Let us assume the finite population U contains N PSU's with a total of M SSU's. A sample of n PSU's has been drawn with a total of m SSU's. Let L be an $N \times M$ matrix that links SSU and PSU data by (Neethling, 2004)

$$L_{ji} = \begin{cases} 1, & i \epsilon j \\ 0, & otherwise \end{cases}.$$

Here j refers to the PSU to which the ith SSU belongs. A method proposed by Lemaître and Dufour (1987) replaces X_S with Z_{pp}, where $\{X_S\}_{i_q}$ is the (i_q)th entry of the $n \times Q$ matrix, indicating the value of auxiliary variable q for SSU i in the sample and $\{Z_{pp}\}_{ji_q}$ is the proportion of SSU's in the jth chosen PSU with auxiliary characteristic q. The subscript "pp" denotes the use of SSU-based auxiliary variables only (Neethling and Galpin, 2006). The elements of this matrix are given by (Neethling, 2004)

$$z_{ji} = \frac{a_{ji}}{m_j},$$

and are defined for SSU i of PSU j with m_j members. Note that

$$a_{ji} = \sum_{i \epsilon j} x_{i_q},$$

is the total for characteristic q in PSU j. Thus, the matrix Z_{pp} at SSU level is defined as

$$Z_{pp} = L_S K_{HS}^{-1} A_{HS},$$

where K_{HS} is a $n \times n$ diagonal matrix containing the PSU sizes m_j, $j = 1, ..., n$, and A_{HS} is a $m \times Q$ matrix given by

$$A_{HS} = L_S^{'} X_S,$$

that includes the auxiliary variables through the $n \times Q$ matrix X_S, aggregated per PSU (Neethling, 2004).

When both SSU and PSU auxiliary variables exist, the matrix that already contains the SSU variable information, can now be extended by adding columns for each category of the PSU auxiliary variable under consideration. The entry is then simply the inverse of the PSU size for the category in which the PSU falls and zero for all other categories (Neethling, 2004). The Z matrix that includes both SSU and PSU auxiliary variables will be denoted by Z_{ph}.

The $n \times 1$ SSU level vector of weights is

$$\mathbf{W}_S = \Pi_S^{-1} \mathbf{1}_n + \Pi_S^{-1} Z_S \left(Z_S^{'} \Pi_S^{-1} Z_S \right)^{-1} \left(X_T - \hat{X}_\pi \right), \tag{2.6.5}$$

where

$$\Pi_S = diag\left(\pi_i\right)$$

is a diagonal matrix containing the inclusion probability of the ith sampled element and Z_S denotes Z_{pp} or Z_{ph}. These weights satisfy a set of constraints (Neethling, 2004),

$$Z_S^{'} \mathbf{W}_S = \mathbf{X}_T. \tag{2.6.6}$$

Integrated Weighting: PSU Level The above integrated weights, calculated on SSU level, can also be calculated on a PSU-based data set. The method proposed for the PSU auxiliary variable case replaces matrices X_S and Z_{pp} by A_{HS}, the matrix of aggregates of the auxiliary characteristics of PSU members. Furthermore, if reliable population counts are also available for PSU's, this information can be added in the form of additional columns to the matrix A_{HS} such that dummy variables denote whether a PSU belongs to a certain category or not (Neethling, 2004).

PSU weights are defined as

$$\mathbf{W}_{HS} = \Pi_{HS}^{-1} \mathbf{1}_m + \Pi_{HS}^{-1} V_{HS}^{-1} A_{HS} \left(A_{HS}^{'} \Pi_{HS}^{-1} V_{HS}^{-1} A_{HS} \right)^{-1} \left(X_T - \hat{X}_\pi \right), \tag{2.6.7}$$

and are subjected to the same set of constraints (2.6.6) as the SSU level weights.

Now all elements of a PSU retain the same weight and when the weights are multiplied by the number of SSU's in each category of a SSU-level auxiliary variable, the weighting estimates agree with the marginal population totals of that variable at SSU level (Neethling, 2004).

Finally, the link between the SSU-based weights in (2.6.5) and the SSU-based weights in (2.6.7) is given by either

$$\mathbf{W}_{HS} = K_{HS}^{-1} L_S^{'} \mathbf{W}_S,$$

or

$$\mathbf{W}_S = L_S K_{HS} \mathbf{W}_{HS},$$

where K_{HS} is a $n \times n$ diagonal matrix containing the PSU sizes. It has been shown that the integrated weighting technique based on SSU level data yields the same final weights than the technique based on PSU level data. Thus, the decision of which data to use relies on the current situation, the auxiliary information available as well as the desired estimators (Neethling and Galpin, 2006).

Adjusting the non-response adjusted design weights through calibration and integrated weighting, also sometimes called benchmarking, completes the development of the final sampling weights to be used in the estimation of population parameters of interest. It has been shown in previous work by Kirchoff (2010) and Luus et al. (2012) that the use of sampling weights in the estimation of parameters of interests improves the precision of the estimators. However, once the design weights have been benchmarked it could happen that some of the sampling weights become excessively large, increasing the variability within the distribution of the sampling weights. The increased variability could reverse the positive effect of the sampling weights on the estimation precision by increasing the mean squared error and bias of the estimators. The extreme weights can be adjusted through the use of weight trimming methods, but the weight trimming itself could affect the estimation precision. The next chapter discusses a selection of weight trimming methods already used in some software packages and also introduces a few new methods. The effect of the trimming of weights will be investigated as part of the analyses in this thesis.

Chapter 3

Weight Trimming Methods

One of the advantages of making use of complex sampling data is the improved estimation of quantities through the use of sampling weights. Unfortunately, these sampling weights may increase the sampling variance of the estimators due to large variation in their values and consequently result in lower inference precision (Ajayi et al., 2005). Such variation could be attributed to the specific sampling procedure used, errors in the sampling frame, non-response adjustments or various other sources (Potter, 1990). The trimming or truncation of weights identified to be extreme could assist in the reduction of the large weight variation which in turn would improve the precision of the estimation (Chowdhury et al., 2007).

Although the weight trimming results in a reduction of the sampling variance, it also introduces some bias into the estimators constructed using the trimmed weights. The increase in bias could possibly increase the overall mean squared error, a measure of estimation precision, of the estimator resulting in a reduction in the precision of the estimator which is not desirable. Hence, the aim of these weight trimming procedures is to reduce the sampling variance by a large enough amount such that the possible increase in estimation bias will be compensated for and that there will be an overall gain in terms of the mean squared error (Potter, 1990).

This chapter will review some of the currently used trimming procedures as well as some other alternatives. Each weight trimming method will include a procedure to identify the weights that are flagged as possibly extreme and also a method for the redistribution of the trimmed portion of the weights among the untrimmed weights to ensure that the weights still sum to the correct population total. These procedures will be compared in a later simulation study and the results will be discussed.

It should be noted that although these methods are called trimming methods, they are in fact all examples of Winsorizing. In trimming a value is determined above (or below) which other observations are completely removed from consideration. Winsorizing, on the other hand, determines a cut-off value and all observations above (or below) this cut-off are set equal to the cut-off. Hence, with trimming one decreases the number of observations while with Winsorizing one

does not lose observations, one merely adjusts a proportion of the observations. For simplicity and to remain in line with the literature summarized here, the adjustment of the weight distribution will be referred to as weight trimming.

Furthermore, the weight trimming methods will only consider the upper tail of the weight distribution for the identification of extreme weights since this was the approach followed in the literature reviewed for this chapter.

3.1 Some Commonly used Weight Trimming Methods

There are no strict rules nor formal procedures to either define what an extreme weight is or how to appropriately trim such weights. The method used differs from survey to survey, but some commonly used procedures are discussed briefly below (Izrael et al., 2009).

3.1.1 4Avg Trimming Method

The 4Avg trimming method gets its name from the cut-off used in this procedure for identifying outlier weights. Let w_i denote the final sampling weight of the ith observation and let $4\bar{w}$ denote the 4Avg cut-off where $\bar{w} = \frac{\sum_i w_i}{n}$. For simplicity only subscript i will be used here. The method proceeds by identifying all weights greater than this cut-off as outliers. All of the outlying weights are then set equal to this cut-off after which the trimmed and untrimmed weights are readjusted to ensure that the weights still sum to the correct population total (Izrael et al., 2009). The trimming procedure as described in Valliant et al. (2013) is employed here:

1. Set an upper bound for the weights at $4\bar{w}$.
2. Let all weights greater than $4\bar{w}$ be equal to this bound and let $\left\{w_i^{4Avg}\right\}_{i\epsilon S}$ denote the set of trimmed weights.
3. Determine the net amount of weight lost due to the trimming: $K = \sum_{i\epsilon S}\left|w_i - w_i^{4Avg}\right|$.
4. Divide K equally among all units whose weights were not trimmed.
5. Repeat steps 2-4 until no weights are in excess of the bound set in step 1.

The final set of weights $\left\{w_i^{4Avg}\right\}$, $i = 1, ..., n$ that requires no further trimming will then be used in the estimation of the parameters of interest.

3.1.2 5Avg Trimming Method

The 5Avg trimming method is applied similarly to the 4Avg method discussed in the previous section, but here the cut-off for outlying weights is set equal to $5\bar{w}$ where $\bar{w} = \frac{\sum_i w_i}{n}$.

3.1.3 5IQR Trimming Method

The 5IQR method is based on the inter-quartile range (IQR_w) of the sampling weights. Let w_i denote the sampling weight of the ith observation and let the cut-off for outlier weights be $Q_{2w} + 5IQR_w$ where Q_{2w} is the median of the n sampling weights and $IQR_w = Q_{3w} - Q_{1w}$, where Q_{1w} and Q_{3w} are, respectively, the first and third quartiles of the sampling weights (Izrael et al., 2009). All weights in excess of this limit are flagged as outliers and hence set equal to the limit. Once again the method in Valliant et al. (2013) is used iteratively until the final set of sampling weights is obtained that requires no further adjustment. Let this set be denoted by $\left\{w_i^{5IQR}\right\}$, $i = 1, ..., n$, the weights to be used in the estimation of the parameters of interest.

3.1.4 6IQR Trimming Method

The 6IQR method is similar to the 5IQR method but instead sets the limit for outlying weights at $Q_{2w} + 6IQR_w$ and all weights above this limit are set equal to the limit (Izrael et al., 2009). Furthermore the method is applied in the same manner as the 5IQR method.

3.1.5 3.5Med Trimming Method

This cut-off is discussed in Valliant et al. (2013) and is based on the median of the sampling weights. Let w_i denote the sampling weight of the ith observation and let $3.5 \times Q_{2w}$, where Q_{2w} is defined as before, be the cut-off for outlying weights. The sampling weights greater than this cut-off are set equal to this value after which the trimming procedure discussed before is applied here too. Let $\left\{w_i^{3.5Med}\right\}$, $i = 1, ..., n$ be the final set of weights that needs no further adjustment. These are the weights that will be used in the estimation of the parameters of interest.

3.2 Newly Introduced Trimming Methods

The first two sections of this chapter reviewed a selection of commonly used as well as other developed weight trimming methods. This section discusses two new weight trimming methods that have been developed for this thesis.

3.2.1 1.5IQR

The 1.5IQR method comes from Tukey's outlier detection rule (Tukey, 1977). Let Q_1 denote the first quartile of a data set, Q_3 the third quartile and let $IQR = Q_3 - Q_1$ denote the inter-quartile range of the data set. The rule states that if a data value is either less than $Q_1 - 1.5IQR$ or greater than $Q_3 + 1.5IQR$ it is considered an outlier.

Analogous to this, let Q_{1w}, Q_{3w} and IQR_w denote the first quartile, third quartile and inter-quartile range of the sampling weight distribution. Then an outlying weight is identified when it is either less than $Q_{1w} - 1.5IQR_w$ or greater than $Q_{3w} + 1.5IQR_w$. To remain in line with the trimming methods discussed previously only the upper bound will be considered as a cut-off for outlying weights. The trimming procedure as described in Valliant et al. (2013) is also employed here:

1. Set an upper bound for the weights at $Q_{3w} + 1.5IQR_w$.
2. Let all weights greater than the bound set in step 1 be equal to this bound and let $\left\{w_i^{1.5IQR}\right\}_{i\epsilon S}$ denote the set of trimmed weights.
3. Determine the net amount of weight lost due to the trimming: $K = \sum_{i\epsilon S}\left|w_i - w_i^{1.5IQR}\right|$.
4. Divide K equally among all units whose weights were not trimmed.
5. Repeat steps 2-4 until no weights are in excess of the bound set in step 1.

The final set of weights $\left\{w_i^{1.5IQR}\right\}$, $i = 1, ..., n$ that requires no further trimming will then be used in the estimation of the parameters of interest.

This is a new trimming method and results obtained when using 1.5IQR trimmed weights will be compared to the results obtained from the other trimming methods discussed.

3.2.2 Hill Estimator and Hill Plot

Many high quality data sets require heavy tailed distributions for appropriate modeling. Drees et al. (2000) states that a heavy tailed distribution is one that satisfies

$$1 - F(x) \sim x^{-\alpha} L(x),$$

with $x \to \infty$, $\alpha > 0$ and $L(x)$ a slow varying function that satisfies $\lim_{t\to\infty} \frac{L(tx)}{L(t)} = 1$ for all $x > 0$.

This distribution function requires the estimation of the shape parameter, α, from a sample from a stationary sequence. The Hill estimator is a popular estimator of the so-called extreme value index (EVI), γ, which is assigned the value $\frac{1}{\alpha}$. Let $X_1, X_2, ..., X_n$ be a sample from a stationary process $(X_n)_{n\epsilon\mathbb{N}}$. Denote the ordered sample by $X_{1,n} < X_{2,n} < ... < X_{n,n}$. For a choice of k, the Hill estimator is given by

$$H_{k,n} := \frac{1}{k}\sum_{j=1}^{k}\left[\log\left(X_{n-j+1,n}\right) - \log\left(X_{n-k,n}\right)\right], \tag{3.2.1}$$

the average log distance between the jth upper order statistics and $(k+1)$th upper order statistics. The Hill plot is obtained by plotting $H_{k,n}$ for $1 \leq k \leq n-1$ and the "best" estimate of γ is inferred from a stable region in the graph (Drees, De Haan, et al., 2000).

Typically in extreme value theory the stable region is selected such that the tail of the distribution contains as many observations as possible. This is done to ensure accurate estimation of quantities from these observations. In the application of this technique to survey weights the idea is to retain as many of the original weights such that only the truly extreme weights be adjusted and also that the bias introduced due to the weight trimming be kept to a minimum. Hence, here the stable region will be identified such that the minimum number of weights lie in the tail of the distribution.

Consider a plot of the stratum, i.e. province, sampling weights distributions of the Income and Expenditure Survey (IES) of 2005, one of the data sets to be used in the later analyses:

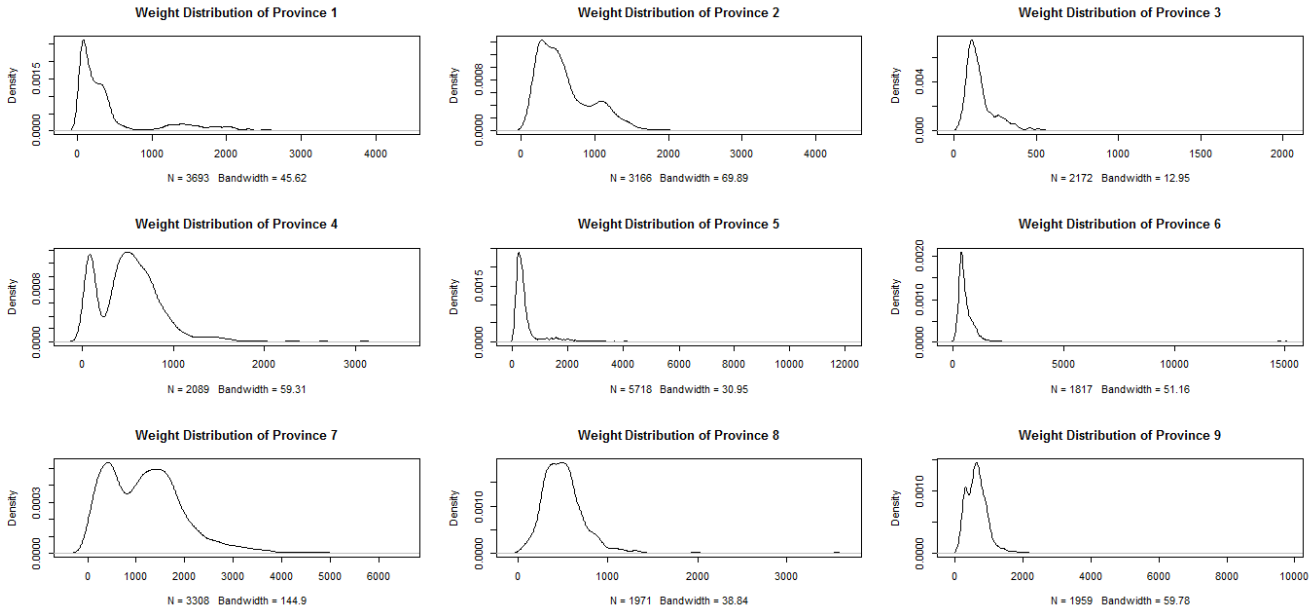


Figure 3.2.1: IES2005 Stratum Weight Distributions

The reason for considering the stratum weight distributions is since sampling weights are calculated by stratum and thus the assumption was made that the weight trimming methods should be applied per stratum.

It is clear from this graph that each of the distributions is fairly heavy tailed. Hence, the Hill estimator could be used to determine an appropriate cut-off above which a weight is considered to be in the tail of the weight distribution, i.e. an outlying weight.

Let w_i denote the sampling weight of the ith sampled unit, $i = 1, ..., n$ and let $w_{1,n} < w_{2,n} < ... < w_{n,n}$ denote the ordered arrangement of the sampling weights. Then the Hill estimator of the EVI of the weight distribution, analogous to (3.2.1) is given by

$$H_w(k) = \frac{1}{k}\sum_{j=1}^{k}\left[\log\left(w_{n-j+1,n}\right) - \log\left(w_{n-k,n}\right)\right]. \tag{3.2.2}$$

The Hill plot is obtained by plotting $H_w(k)$ for $1 \leq k \leq n-1$ and the "best" estimate of γ is inferred from a stable region in the graph. The weight that corresponds to the Hill estimate, $w_{n-k,n}$, is then used as the cut-off for outlying weights. All weights above this cut-off are set equal to the cut-off $w_{n-k,n}$ and then the approach of Valliant et al. (2013), discussed previously, is used to determine the final set of sampling weights, $\left\{w_i^{Hill}\right\}$, $i = 1, ..., n$. This Hill trimming method will be applied and compared to the results obtained from the other trimming methods discussed above.

Drees et al. (2000) mentions that the performance of the Hill estimator is very dependent on the choice of k, the number of observations in the tail of the distribution, and consequently performs poorly in terms of accuracy of estimation (Berning, 2015). The use of the Hill plot to infer a stable region signifying the estimator of γ, is very subjective and due to the extensive research that has been done on the Hill estimator, several threshold selection methods have been proposed in literature:

1. Guillou and Hall (2001);
2. Beirlant et al. (2004); and
3. Drees and Kaufmann (1998).

However, research has shown that the Hill estimator still performs unsatisfactorily even when these methods are used for threshold selection.

A possible approach to the threshold selection, developed for this thesis, considers the proportion of weights that lie within each interval of a histogram of the weight distribution as an indication of the proportion of weights that form the tail of the distribution. The assumption is that the sum of the interval proportions above a point, where a significant change in interval proportions is observed, can be used as an indication of the length of the tail of the distribution. The manual examination of histograms, however, can become quite tedious and hence a way was devised to automate this process. A computer program was developed to identify the point where the maximum absolute gradient of the histogram occurs and to calculate the sum of the interval proportions above this point. The proportion of weights above this threshold, i.e. the output from the program, represents the tail proportion of the weight distribution. Let this proportion be denoted by $k\%$. The $k\%$ is then used for the calculation of the Hill estimators, using (3.2.2), and the construction of the Hill plot. The Hill plot is then used to identify the weight above which other weights are considered outliers. Here too it was necessary to automate the manual investigation of the Hill plots and thus a computer program was developed for this purpose. In this program it is important to find the first possible stable region from the Hill plot to ensure

that the minimum number of weights is trimmed. Hence, selection is based on the point where the minimum gradient of the Hill plot is observed. This point identifies the optimal number of weights in the tail of the weight distribution, say k, for which the Hill plot stabilizes and $w_{n-k,n}$, the weight in the kth position of the ordered weight distribution, is used as the bound for outlier weights. After this bound is determined the approach of Valliant et al. (2013) is used to determine the final set of sampling weights, $\left\{w_i^{Hill}\right\}$, $i = 1, ..., n$, to use in the inference.

Another approach, proposed by Berning (2010; 2015), is to use a bias-reduced estimator obtained by fitting a perturbed Pareto distribution (PPD) to the excesses in which case a larger range of values of k is obtained for which the bias is quite small. Furthermore, a measure of instability of the estimates of γ over the range of values of k is developed and the stable region, a region within which the estimates do not vary excessively, is then defined as the region for which the instability measure is the lowest (Berning, 2015).

For the instability measure, let $y_1, y_2, ..., y_m$, $m \geq 2$, denote the observed values of y with corresponding chosen values $x_1, x_2, ..., x_m$ of x. It follows that the instability of y with respect to x, θ^2, is given by

$$\theta^2 = \sigma^2 + b^2, \tag{3.2.3}$$

where σ^2 is the sample variance of the values of y (Hill estimates) and b is the slope of the line of the simple least squares regression of y on x (the corresponding thresholds)(Berning, 2015).

Now, let $y_1, ..., y_m$ denote the set of estimated EVI's, i.e. let $\{y_i\} = \{\hat{\gamma}_i\}$, $i = 1, ..., m$, and let $x_1, ..., x_m$ denote the values corresponding to the respective choices of k, i.e. let $\{x_i\} = \left\{X_{(k),i}\right\}$ $i = 1, ..., m$. For the purpose of weight trimming, $x_i = w_i$, $i = 1, ..., n$, the sampling weights that correspond to the respective choices of k, i.e. $\{w_i\} = \left\{W_{(k),i}\right\}$ $i = 1, ..., m$. Note that it is assumed that the x values are equally spaced and that the x and y values are scaled to ensure location and scale invariance. Also, for simplicity, the xy-notation will be used throughout the discussion. The x values are normalized by calculating $x_i^* = \frac{(x_i - x_1)}{(x_m - x_1)}$, where x_1 is the smallest and x_m is largest x value, and the y values are normalized by calculating $y_i^* = \frac{y_i}{\bar{y}}$, where $\bar{y}$ is the average of the y values (Berning, 2015).

Before applying the methods there are two algorithms developed that need to be applied to the data. The first algorithm rounds the values to the nearest 5% of their mean while the second algorithm is applied to remove from consideration the region of estimates where the bias becomes significant (Berning, 2015):

1. Round $y_1, ..., y_m$ to the nearest 5% of their mean (2 decimals); and

2. Remove from consideration the region of estimates where the bias becomes significant:

   (a) $y_{m+1} \leftarrow 2 \times y_m$;

   (b) Set $m^* \leftarrow m$;

(c) while $(y_{m*} \geq y_{m*-1})$ and $(y_{m^*-1} \geq y_{m^*-2})$ and $(m > 2)$, then $m^* \leftarrow m - 1$.

(d) while $(y_{m^*} = y_{m^*-1})$ and m^* is less than its initial value, then $m^* \leftarrow m + 1$.

(e) while $(y_{m^*} = y_{m^*+1})$ and m^* is less than its initial value, then $m^* \leftarrow m + 1$.

After this procedure is applied the values $y_1, ..., y_m$ are reduced to $y_1, ..., y_{m^*}$ where $m^* \leq m$.

Once these procedures have been applied the methods for optimal stable region selection can be applied. A summary of the methods given in (Berning, 2015) is supplied here.

Method 0

This method simply applies the two procedures described above and regards the remaining values as the optimal stable region (Berning, 2015).

Method 1

Here the region length is fixed beforehand, say k, and hence the instability measure is calculated for regions $\{y_1, ..., y_k\}, \{y_2, ..., y_{k+1}\}, ..., \{y_{m^*-k+1}..., y_{m^*}\}$. The region that results in the smallest instability measure is chosen as the optimal region (Berning, 2015).

Method 2

The region is trimmed systematically until a "stable" instability measure is obtained. Firstly, the instability measure is calculated over the entire region $y_1, ..., y_{m^*}$ after which either the first or the last value within the region is deleted. Deciding whether to delete the first or last value depends on which omission results in the largest reduction of the instability measure. This procedure is repeated until neither the deletion of the first nor the last value decreases the instability measure (Berning, 2015).

Method 3

The final method fixes the upper limit of the region and proceeds by trimming values from the left. For example, if the upper limit is set at 5, then the instability measure will be calculated using $\{y_1, ..., y_5\}, \{y_2, ..., y_5\}, ..., \{y_4, y_5\}$. The region resulting in the lowest instability measure is retained as the optimal stable region (Berning, 2015).

Although Berning (2015) presents 4 methods for determining the optimal stable region, he showed through simulation that the best results were obtained by method 0 (M0) and method 3 (M3). Methods M0 and M3 will be employed for threshold selection after which the approach of Valliant et al. (2013) will be used to obtain the final set of sampling weights, $\left\{w_i^{M0}\right\}$, $i = 1, ..., n$ and $\left\{w_i^{M3}\right\}$, $i = 1, ..., n$ respectively, that will be used in the estimation of the parameters of interest.

3.3 Other suggested Weight Trimming Methods

Other more formal weight trimming methods that have been proposed in literature, include:

- the estimated MSE trimming procedure where the estimated MSE is evaluated at various trimming levels to determine the optimal trimming level (Potter, 1990);
- The NAEP (National Assessment of Educational Progress) procedure is another more formal procedure that uses the comparison of the contribution of each weight to the sampling variance of an estimator by systematically comparing all weights to a value computed from the sum of the squared weights for the sample. Any weight above this computed value is assigned this value and the other weights are adjusted such that they sum to the original weight total. This procedure is repeated until all adjusted weights are below or equal to the value based on the sum of the adjusted squared weights (Potter, 1990);
- the Taylor series procedure that makes use of the estimated MSE as well as the estimated relative bias computed for each data item at possible trimming levels. The optimal trimming level is the one that corresponds to a minimum combined score for both the estimated MSE and relative bias (Potter, 1990);
- the weight distribution procedure where a sampling weight distribution is assumed. It has been shown that this distribution is essentially the inverse of a beta distribution. The parameters of the distribution are estimated using the sampling weights and a trimming level is selected with a prespecified probability of occurrence. Any sampling weights greater than this trimming level are set at this level and the excess is distributed among the untrimmed weights. The procedure is repeated and the distribution parameters are now estimated using the trimmed weights, a new trimming level is specified and any weights in excess of this level are set to the adjusted trimming level and the excess distributed among the untrimmed weights. The process is repeated ten times (Potter, 1990);
- Chowdhury et al. (2007) follows a similar approach to the weight distribution procedure, but instead assumes that the tail weights follow an exponential distribution with parameter $\lambda = \frac{1}{\mu}$ where μ is the mean of the tail weights. This approach is called the alternative weight distribution procedure; and
- Elliot (2007) refers to the trimming discussed thus far in this section as direct weight trimming methods. As an alternative to these he has developed weight trimming methods that utilize Bayes methodology (Elliott, 2007; Elliott, 2008; Elliott, 2009). For example, the unequal inclusion probabilities are accounted for by letting the sampling weights be stratification variables within strata, defined by the inclusion probabilities. After this the standard weighted estimates are obtained by treating the weighted stratum means as fixed effects and then the weights are trimmed by treating the weighted stratum means as random effects.

The trimming procedures listed above will not be included in this research, but will be considered as part of further work.

To conclude, recall that this chapter considered the various weight trimming that can be used to reduce the variability within the sampling weight distribution such that the precision of an estimator is not adversely affected by the potentially large variability in the sampling weights. A selection of more generally used procedures was presented and discussed in section 3.1 followed by the introduction and discussion of alternative weight trimming methods not employed by statistical software such as SAS, R, etc. All of the procedures discussed in these sections will be used in the analyses and the precision of the estimators, obtained using untrimmed weights versus the various trimmed weights, will be compared.

Chapter 4

Regression Methodology and Computation

4.1 Introduction to Linear Regression

Regression analysis is a widely applied area in Statistics. In a nutshell, it is the modeling of a response (or various responses), also known as the dependent variable, based on its relationship with one or more explanatory variables or independent variables. Three different cases of regression can be distinguished based on the number of these dependent and independent variables (Rencher, 2002):

1. Simple linear regression, the modeling of a single dependent variable based on its relationship with a single independent variable.

2. Multiple linear regression, where a single dependent variable is related to several independent variables. This is also known as univariate multiple regression to emphasize the single dependent variable.

3. Multivariate multiple linear regression, where several dependent variables are modeled based on their relationship with several independent variables and also referred to as simply multivariate regression.

The main objective, in either of the three regressions listed above, is to establish a linear relationship between the response and the explanatory variable(s) for the purpose of prediction. Suppose this relationship is defined by the model

$$\mathbf{y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}, \tag{4.1.1}$$

where $\mathbf{y}$ represents the response variable of interest, $\mathbf{X}$ represents the explanatory variable(s), $\underline{\beta}$ represents the vector of unknown regression parameters and $\underline{\varepsilon}$ the random error terms or residuals. The existence of the error lies therein that the explanatory variables included in the model do not

necessarily enclose all possible explanatory variables of importance to the prediction of the response of interest. It is assumed that the random errors meet the following assumptions (Rencher, 2002):

1. The residuals $\underline{\varepsilon}$ are normally distributed;
2. $E(\underline{\varepsilon}) = \mathbf{0}$ such that $E(\mathbf{y}) = \mathbf{X}\underline{\beta}$;
3. $V(\underline{\varepsilon}) = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix, such that $V(\mathbf{y}) = \sigma^2\mathbf{I}$ (homoscedasticity);
4. The residuals are independent.

It is of importance, for the validity and quality of further inference as well as prediction to be carried out, that these assumptions be verified and met. Upon investigating the effects of violating the assumptions underlying regression modeling the following conclusions are made (Lumley et al., 2002; Osborne et al., 2002):

- Linearity

  The intercept, estimated coefficients and predicted responses will be biased, the slope and intercept won't be meaningful and the predicted responses will also be wrong, especially when the model is applied to out-of-sample data. The true relationship will be underestimated and the significance of some covariates may be overestimated.

- Normality

  This assumption must be in place for any inference conducted from the model. Significance tests of the coefficients, predictions and confidence intervals all become problematic. Also, if normality is not present it is considered a red flag that some other assumptions are also not met.

- Homoscedasticity

  This makes it difficult to gauge the true standard deviation of the prediction errors, usually resulting in confidence intervals that are too wide or too narrow. A serious violation of homoscedasticity can distort findings and weaken the analysis.

To ensure model quality and prediction accuracy, regression is typically carried out in four different stages (Heeringa et al., 2010):

1. Model specification;
2. Estimation;
3. Evaluation; and
4. Inference.

Each of these will be addressed to some extent in this chapter, however step 3 will mostly be considered under further research.

The application of linear regression is quite common for independent and identically distributed data, but it has been found that when investigators perform regression analysis on survey data that resulted from a complex sampling design (CS), they sometimes choose to ignore the design used for the collection of the observation units, run the data through standard software packages and report the output obtained without further consideration of the design. The following can happen in complex surveys (Lohr, 2010):

- The observations may have different inclusion probabilities as is often the case; and
- Non-respondents can change the relationship between the response and the predictor variable.

Although the estimators of the regression parameters are approximately design unbiased, the standard errors given by non-survey software packages will likely be wrong if the design involves clustering (Lohr, 2010).

The most important feature in CS is the design weights that are developed in such a way that they take care of the effects of stratification and clustering on estimates (Lohr, 2010). When these weights are weakly related to the variables of interest and have large variation, the estimation of quantities from a CS may be inefficient (Beaumont, 2008).

The purpose of this chapter is to examine the theory of linear regression analysis. It considers different types of linear regression applied to data from both simple random sampling (SRS) as well as CS. It should be mentioned that although the first section considers simple linear regression while the focus of the thesis is on multiple linear regression, the section is included as an introduction to notation. The estimation of variances of estimated regression coefficients, a discussion of some diagnostic measures used to ascertain the quality of the regression and finally inference regarding the regression coefficients and response, will be included.

4.2 Model Specification and Parameter Estimation

4.2.1 Multiple Regression

Consider a finite population of size N and define an N-vector of responses, $\mathbf{y}$. Furthermore, suppose p predictor variables, $\mathbf{x}_1, ..., \mathbf{x}_p$, exist and define a p-vector $\mathbf{x}_i = (x_{i1}, ...x_{ij}, ..., x_{ip})$, where x_{ij} represents the value of the jth predictor variable for the ith observation. Now, let the population model that defines the relationship between the response and the predictors be given by

$$\mathbf{y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}, \tag{4.2.1}$$

where $\underline{\beta}$ is a p-dimensional vector of unknown regression coefficients (Rencher, 2002).

Consider an SRS of size n where the response variable is denoted by $\mathbf{y} = [y_1, y_2, ..., y_n]^{'}$ and the predictor variables by the $n \times p$ matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_p]$ where $\mathbf{x}_j = [x_{1j}, x_{2j}, ..., x_{nj}]^{'}$ is the jth explanatory variable, $j = 1, ..., p$. The objective is to estimate the unknown regression coefficients and consequently $E(\mathbf{y})$. Let $\underline{\hat{\beta}} = \left[\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p\right]^{'}$ denote the estimator of $\underline{\beta}$ and let $\widehat{E(\mathbf{y})} = \hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, ..., \hat{y}_n]^{'}$ be the estimator of $E(\mathbf{y})$. The method of least squares is well known and finds $\underline{\hat{\beta}}$ as the solution that minimizes the total squared deviations between the observed $\mathbf{y}$ and their corresponding model predicted values, $\hat{\mathbf{y}}$. Thus, find $\underline{\hat{\beta}}$ that minimizes (Rencher, 2002)

$$SSE_{OLS} = \sum_{i \epsilon s} \left(y_i - \mathbf{x}_i^{'} \underline{\hat{\beta}}\right)^2, \tag{4.2.2}$$

where SSE_{OLS} is the sum of squared errors and $\mathbf{x}_i^{'} = [x_{i1}, x_{i2}, ..., x_{ip}]$ denotes the ith row of the predictor matrix $\mathbf{X}$. Since $y_i - \mathbf{x}_i^{'}\underline{\hat{\beta}}$ is the ith element of the vector $\mathbf{y} - \mathbf{X}\underline{\hat{\beta}}$ it follows that, in matrix notation,

$$SSE_{OLS} = \left(\mathbf{y} - \mathbf{X}\underline{\hat{\beta}}\right)^{'} \left(\mathbf{y} - \mathbf{X}\underline{\hat{\beta}}\right).$$

From the minimization of the SSE_{OLS} with regards to $\underline{\hat{\beta}}$ the BLUE estimator of the unknown regression coefficients is given by

$$\underline{\hat{\beta}} = \left(\mathbf{X}^{'}\mathbf{X}\right)^{-1} \mathbf{X}^{'}\mathbf{y}, \tag{4.2.3}$$

with variance

$$V\left(\underline{\hat{\beta}}\right) = \sigma^2 \left(\mathbf{X}^{'}\mathbf{X}\right)^{-1}, \tag{4.2.4}$$

where σ^2 is the unknown variance of the residuals. It is assumed that $\mathbf{X}^{'}\mathbf{X}$ is nonsingular and that none of the independent variables can be expressed as a linear combination of the other independent variables (Rencher, 2002).

It can be shown that

$$E(SSE_{OLS}) = \sigma^2 (n - p - 1),$$

where n is the number of observations and p the number of dependent variables. Using this result an unbiased estimator of the residual variance, σ^2, is given by

$$\hat{\sigma}^2 = S^2 = \frac{SSE_{OLS}}{n - p - 1}.$$

4.2.2 Weighted Least Squares Regression

In the analysis of real-world data the assumptions underlying the linear regression model are easily violated (Heeringa et al., 2010). One of these assumptions is that of homoscedasticity of the residual variance. In the absence of constant residual variance while the other assumptions are met, one will still obtain unbiased and consistent estimators of the regression coefficients, but the estimators will no longer have minimum variance. To obtain the minimum variance characteristic of the estimators one should account for the difference in reliability of the response observations, since observations with larger variances provide less reliable information about the regression function. In this case it is suggested to apply weighted least squares regression (WLS). The difference between OLS and WLS lies therein that WLS assigns smaller weights to observations with large error variations as opposed to OLS where each observation receives a constant weight. Recall the least squares criterion given in equation (4.2.2). Under WLS the same minimizing criterion is defined, but now it incorporates a weight to account for the heteroscedasticity,

$$SSE_{WLS} = \sum_{i \epsilon s} w_i \left(y_i - \mathbf{x}_i' \underline{\hat{\beta}} \right)^2, \tag{4.2.5}$$

where w_i is the weight associated with the ith observation (Kutner et al., 2005; Neter et al., 1983). The weights of all the observations are then combined into a diagonal weight matrix, $\mathbf{W}$, and finally the weighted least squares estimator of $\underline{\beta}$ is given by

$$\underline{\hat{\beta}}_{WLS} = \left(\mathbf{X}'\mathbf{W}\mathbf{X} \right)^{-1} \mathbf{X}'\mathbf{W}\mathbf{y}. \tag{4.2.6}$$

Furthermore it follows that, under WLS, the variance of the unknown regression parameters is given by

$$V\left(\underline{\hat{\beta}}_{WLS} \right) = \sigma^2 \left(\mathbf{X}'\mathbf{W}\mathbf{X} \right)^{-1}, \tag{4.2.7}$$

where σ^2 is the unknown variance of the model residuals which is estimated by

$$\hat{\sigma}^2 = \frac{\sum_i w_i \left(y_i - \hat{y}_i \right)^2}{n - p},$$

(Kutner et al., 2005; Neter et al., 1983).

4.2.3 Survey-weighted Least Squares Regression

The data from a CS is not independently and indentically distributed, as is the case with the data from an SRS that are used in OLS and WLS. The variation in sample selection and inclusion probabilities when making use of CS necessitates the inclusion of sampling weights, discussed in section 2.6, when developing unbiased estimators of general unknown parameters. Here the intent

is to estimate regression parameters by finding the estimator of $\underline{\beta}$ that minimizes

$$SSE_{SWLS} = \sum_{i \epsilon s} w_i \left(y_i - \mathbf{x}_i' \hat{\underline{\beta}} \right)^2. \tag{4.2.8}$$

Notice the similarity between equation 4.2.8 and equation 4.2.5. Due to the similarity of the minimization criterions of WLS and SWLS it follows that, even when analysts naïvely use the sampling weights under WLS, both WLS and SWLS will give the same unbiased estimators of the regression parameter. Thus, the estimator of $\underline{\beta}$ under SWLS is then, similarly to WLS, given by

$$\hat{\underline{\beta}}_{SWLS} = \left(\mathbf{X}' \mathbf{W} \mathbf{X} \right)^{-1} \mathbf{X}' \mathbf{W} \mathbf{y}, \tag{4.2.9}$$

where $\mathbf{W} = diag\left[w_1, w_2, ..., w_n\right]$ is the $n \times n$ matrix of sampling weights. Since WLS and SWLS estimators of the regression parameters are the same when the sampling weights are specified it could give the impression that one can also use WLS estimated standard errors of these estimated regression parameters for further inference. However, naïvely doing so will lead to severely biased estimated standard errors which will affect further inference using the WLS results (Heeringa et al., 2010).

Firstly it should be said that the estimation of the variances of most parameters of interest under CS will not be simple linear functions, i.e. the independence in the responses are disturbed due to the complex design, and thus finding closed-form solutions for these variances, is scarce. It is thus necessary to consider alternative variance estimation methods when making use of CS. The most commonly used variance estimation methods under CS, are the Taylor series linearization (TSL) approach and resampling procedures such as the jackknife, balanced repeated replication or the bootstrap. These methods are popular since they provide robust, non-parametric approaches to variance estimation (Heeringa et al., 2010).

Using the TLS it was found that the estimated variance of the estimator under CS is given by

$$\hat{V}\left(\hat{\underline{\beta}}_{SWLS} \right) = \left(\sum_{i \epsilon s} w_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \hat{V} \left(\sum_{i \epsilon s} w_i \mathbf{x}_i \left(y_i - \mathbf{x}_i' \hat{\underline{\beta}}_{SWLS} \right) \right) \left(\sum_{i \epsilon s} w_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1}. \tag{4.2.10}$$

See Lohr (2010). This is in contrast to the well known estimated variance obtained in the case of WLS, given by

$$\hat{V}\left(\hat{\underline{\beta}}_{WLS} \right) = \hat{\sigma}^2 \left(\mathbf{X}' \mathbf{W} \mathbf{X} \right)^{-1},$$

and which can also be written as

$$\hat{V}\left(\hat{\underline{\beta}}_{WLS} \right) = \hat{\sigma}^2 \left(\sum_{i \epsilon s} w_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1},$$

where $\hat{\sigma}^2$ is an estimator of σ^2. Hence, naïvely using WLS variances under SWLS will lead to biased, and in fact wrong, conclusions.

In this thesis, use will mostly be made of the jackknife and the bootstrap methods for variance estimation.

4.3 Jackknife and Bootstrap Variance Estimation

Two well known and popular non-parametric methods of inference is the jackknife and the bootstrap. Their application to SWLS inference will be discussed in the following sections.

4.3.1 A Jackknife approach to Regression Analysis

The jackknife method predates the bootstrap method in the estimation of bias and standard errors of an estimator $\hat{\theta}$. Its name was used by Tukey in 1958 as a way of conveying the broad usefulness of this technique (Knight, 2000). A jackknife is synonymous to a penknife or a switchblade, which is a multipurpose knife that can perform the functions of a number of more specialized knives. Thus, the jackknife can be used as a substitute for a variety of more specialized techniques. Here the jackknife is going to be used to estimate the variance of the regression parameter estimator.

Consider the ith data pair, $(y_i, \mathbf{x}_i)$. By the jackknife method the ith data pair is deleted, $(1, ..., i-1, i+1, ..., n)$, with $(n-1)$ pairs of data remaining in the ith jackknife sample, $i = 1, ..., n$. Suppose an OLS is fitted to the ith jackknife sample and let $\underline{\hat{\beta}}_{(i)_{OLS}}$ be the jackknife replicate of the estimator of the regression parameters,

$$\underline{\hat{\beta}}_{(i)_{OLS}} = \left(\mathbf{X}'_{(i)}\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}'_{(i)}\mathbf{y}_{(i)}, \tag{4.3.1}$$

where $\mathbf{X}_{(i)}$ is the matrix of independent variables with the ith row deleted and $\mathbf{y}_{(i)}$ is the associated response vector with the ith response deleted (Sahinler et al., 2007).

The same process is repeated for each of the data pairs, each time removing one data pair and using the remaining $(n-1)$ pairs to fit a linear model and calculate a jackknife replicate. This results in n jackknife replicates of the estimated OLS regression parameters, $\left\{\underline{\hat{\beta}}_{(i)_{OLS}}\right\}$, $i = 1, ..., n$. Consider the jth regression parameter, $\hat{\beta}_{j_{OLS}}$. The jackknife estimated variance of this parameter under OLS is given by

$$\hat{V}_{JK}\left(\hat{\beta}_{j_{OLS}}\right) = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_{j_{(i)_{OLS}}} - \tilde{\beta}_{j_{OLS}}\right)^2, \tag{4.3.2}$$

where

$$\tilde{\beta}_{j_{OLS}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_{j_{(i)_{OLS}}},$$

the average of the n jackknife replicates of the jth regression parameter estimates and thus a jackknife estimate of the jth regression parameter (Sahinler et al., 2007). Notice the factor $\frac{n-1}{n}$ instead of $\frac{1}{n}$ or $\frac{1}{n-1}$, which one would expect in the calculation of a variance, is used (Rust et al., 1996). This agrees with the discussion in Efron and Tibshirani (1998) where it is explained that this "inflation factor" is necessary since the jackknife deviations $\left(\hat{\theta}_{(hj)} - \hat{\theta}\right)^2$, in this jackknife application $\left(\hat{\beta}_{j_{(i)OLS}} - \tilde{\beta}_{j_{OLS}}\right)^2$, are much smaller than the deviations for other resampling techniques such as the bootstrap. This is the case due to the jackknife sample being more similar to the original sample than a typical bootstrap sample.

The above application of the jackknife method to OLS modeling, where it is assumed that the residual variance is constant, is called the balanced data case according to Miller (Miller, 1974). When using WLS it is assumed that the error variance is not constant. Wu (1986)), Hinkley (1977), and Miller (1974), to name a few, refer to this as unbalanced regression data. Three shortcomings of the ordinary jackknife method, when applied to unbalanced data, are pointed out in Hinkley (1977):

1. $\tilde{\beta}_{j_{OLS}}$ is an unbiased estimator of $\hat{\beta}_{j_{OLS}}$, but generally has a bigger variance than $\hat{\beta}_{j_{OLS}}$;
2. In general, $\hat{V}_{JK}\left(\hat{\beta}_{j_{OLS}}\right)$ is a biased estimator of $V\left(\hat{\beta}_{j_{OLS}}\right)$ and $V\left(\tilde{\beta}_{j_{OLS}}\right)$; and
3. Depending on how balanced the predictor matrix $\mathbf{X}$ is, the bias of $\tilde{\beta}_{j_{OLS}}$ is of the order n^{-1} or n^{-2}.

Due to this, Hinkley (1977) proposed a weighted jackknife method, but not to be confused with the weights specified under WLS or even the sampling weights. The weights here serve the purpose of distance measures, namely the distance between a single design point and the center of the design,

$$\mathbf{x}_i^{'}\left(\mathbf{X}^{'}\mathbf{X}\right)^{-1}\mathbf{x}_i.$$

The objective in this thesis is to illustrate the effect of naïvely using WLS on CS data and thus the weighted jackknife method will not be discussed further.

Now consider the application of jackknife in a stratified multistage cluster sample (CS) with H strata and n_h PSU's in each stratum. Let $\underline{\hat{\beta}}_{(hj)_{SWLS}}$ be the estimator of $\underline{\beta}$, obtained from fitting an SWLS model to the data after the jth PSU in the hth stratum has been deleted, $j = 1, ..., n_h$ and $h = 1, ..., H$, and the weights of all other units from the hth stratum have been inflated by a factor of $\frac{n_h}{n_h-1}$ (Rust et al., 1996). Thus

$$w_{i(hj)} = \begin{cases} w_i,\ i \notin h \\ w_i \cdot \frac{n_h}{(n_h-1)},\ i \in h,\ i \notin j \\ 0,\ i \in (h,j) \end{cases}, \tag{4.3.3}$$

where w_i is the sampling weight of the ith unit. $w_{i(hj)}$ is the jackknife-adjusted sampling weight, defined in such a way that only the weights of the units in the stratum to which the deleted PSU belongs, are adjusted. Notice how the CS jackknife method deletes a PSU at a time instead of a data row at a time. The reason for this lies therein that the PSU structure needs to be preserved. Also, the sampling weights are adjusted to ensure that the overall sum of the weights still equals the population total.

These jackknife weights are then used when fitting an SWLS model to the sample without the (hj)th PSU. The jackknife replicate of the estimator of the regression parameters is thus calculated as

$$\underline{\hat{\beta}}_{(hj)_{SWLS}} = \left(\mathbf{X}'_{(hj)}\mathbf{W}_{(hj)}\mathbf{X}_{(hj)}\right)^{-1}\mathbf{X}'_{(hj)}\mathbf{W}_{(hj)}\mathbf{y}_{(hj)}, \tag{4.3.4}$$

where $\mathbf{W}_{(hj)}$ is the notation used to denote the diagonal matrix of jackknife sampling weights. Consider the jth regression estimator, $\hat{\beta}_{j_{SWLS}}$. The jackknife estimator of the variance of the jth regression estimator is given by

$$\hat{V}_{JK}(\hat{\beta}_{j_{SWLS}}) = \sum_{h=1}^{H}\left(\frac{n_h - 1}{n_h}\right)\sum_{j=1}^{n_h}\left(\hat{\beta}_{(hj)_{j_{SWLS}}} - \tilde{\beta}_{j_{SWLS}}\right)^2, \tag{4.3.5}$$

where $\tilde{\beta}_{j_{SWLS}}$is the average of the jackknife replicates of the jth regression estimator estimates and thus a jackknife estimator of the jth regression parameter.

Advantages of the jackknife that have to be mentioned are that the same procedure is used to estimate the variance of every statistic for which the jackknife can be used and it provides a consistent estimator of the variance when θ, a general parameter of interest, is a smooth function of population totals. On the other hand, it performs badly if the statistic is not smooth (Lohr, 2010). Results obtained when applied in unequal probability sampling designs where sampling is done without replacement should not be trusted since little is known about the performance of the jackknife method under these circumstances (Lohr, 2010).

4.3.2 A Bootstrap approach to Regression Analysis

The bootstrap resampling method was introduced as a computer intensive method for estimating the variance of an estimator. A pleasing property of this resampling method is that there is no need to derive theoretical variances and the bootstrap estimate is available regardless of how mathematically complicated the estimator may be (Efron et al., 1998).

Bootstrap resampling methodology has long been used in inference for variance estimation, confidence intervals, etc., but here its application in linear regression analysis is considered. Two approaches to bootstrap regression are described in Efron and Tibshirani (1998), namely bootstrapping residuals and bootstrapping pairs and this section will briefly discuss both. The bootstrap

resampling method forms an important part of the analysis in this thesis, specifically pertaining to the inference concerning the regression parameters in terms of the estimation of diagnostics such as bias and mean squared error and confidence interval estimation. It will also form part of the model evaluation diagnostics.

4.3.2.1 Bootstrapping Residuals

The probability model for linear regression consists of two components, $P = \left(\underline{\beta}, F\right)$ where $\underline{\beta}$ is the parameter vector of regression coefficients and F is the probability distribution of the error terms. Firstly, suppose $\underline{\beta}$ is estimated by using the OLS method to obtain $\underline{\hat{\beta}}_{OLS}$. The question one is faced with is how to estimate the error distribution (Efron et al., 1998).

The error terms in the linear model can be written as

$$\varepsilon_i = y_i - \mathbf{x}_i\underline{\beta},$$

where $\mathbf{x}_i$ is the ith row of the matrix of predictors, $\mathbf{X}$. One can then calculate approximate errors (residuals) using the estimated regression coefficients $\underline{\hat{\beta}}_{OLS}$,

$$\hat{\varepsilon}_{i_{OLS}} = y_i - \mathbf{x}_i\underline{\hat{\beta}}_{OLS},\ i = 1, ..., n.$$

The bootstrap method in this case consists of resampling the residuals with equal probability of $\frac{1}{n}$ (Efron et al., 1998).

Denote a generated bootstrap sample by $\{\hat{\varepsilon}_i^*\}$, $i = 1, ..., n$. These bootstrap residuals are then used to generate a set of bootstrap responses,

$$y_i^* = \mathbf{x}_i^{'}\underline{\hat{\beta}}_{OLS} + \hat{\varepsilon}_i^*,$$

where $\underline{\hat{\beta}}_{OLS}$ is still the OLS regression parameters estimated from the orginal sample. To obtain the bootstrap least squares estimator of $\underline{\hat{\beta}}$, find the estimator that minimizes the residual squared error of the bootstrap data,

$$\sum_{i=1}^{n}\left(y_i^* - \mathbf{x}_i^{'}\underline{\hat{\beta}}^*\right)^2 = \min_{\mathbf{b}}\sum_{i=1}^{n}\left(y_i^* - \mathbf{x}_i^{'}\mathbf{b}\right)^2.$$

Hence,

$$\underline{\hat{\beta}}^* = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}^*, \tag{4.3.6}$$

where $\underline{\hat{\beta}}^*$ is the bootstrap estimated regression parameters and $\mathbf{X}$ is the original sample matrix of predictors. To summarize:

1. Calculate $\underline{\hat{\beta}}_{OLS} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$.

2. Calculate the approximate residuals, $\hat{\varepsilon}_{i_{OLS}} = y_i - \mathbf{x}_i' \underline{\hat{\beta}}_{OLS}$.

3. Obtain the bootstrap sample of error terms by sampling, with replacement, from $\{\hat{\varepsilon}_{i_{OLS}}\}$. This yields $\{\hat{\varepsilon}_i^*\}$.

4. Calculate the bootstrap response variable, $y_i^* = \mathbf{x}_i' \underline{\hat{\beta}}_{OLS} + \hat{\varepsilon}_i^*$.

5. Calculate the bootstrap estimate of the least squares regression coefficients,

$$\underline{\hat{\beta}}^* = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{y}^*.$$

Repeat this process B times to obtain B estimates of the unknown regression coefficients $\underline{\hat{\beta}}_1^*, ..., \underline{\hat{\beta}}_B^*$. Consider the jth regression estimator and denote its bth bootstrap replicate by $\hat{\beta}_{j_b}^*$. The bootstrap estimated variance of the jth estimated regression parameter is now calculated as

$$\hat{V}_B\left(\hat{\beta}_j\right) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\beta}_{j_b}^* - \hat{\bar{\beta}}_j^*\right)^2, \tag{4.3.7}$$

where $\hat{\bar{\beta}}_j^* = \frac{1}{B}\sum_b \hat{\beta}_{j_b}^*$, the average of the B bootstrap replicates of the estimator of the jth regression parameter.

Next, consider a design where the population is stratified into H strata. Within stratum h a number of subgroups is formed, called primary sampling units (PSU's), and a sample of PSU's, n_h, is selected of which all or some of the elements can be included in the final sample. Let the sampling weight for this design be denoted by w_{hji}, $h = 1, ..., H$, $j = 1, ..., n_h$, $i = 1, ..., n_{hj}$. This method begins by obtaining $\underline{\hat{\beta}}_{SWLS}$, the estimator of the regression parameters under complex sampling, and using the estimator to find the residuals, $\underline{\hat{\varepsilon}}_{SWLS}$, from which the resampling will be done.

Let the residuals of the hth stratum be denoted by $\underline{\hat{\varepsilon}}_{h_{SWLS}} = \left\{\hat{\varepsilon}_{hj_{SWLS}}\right\}$, $j = 1, ..., n_h$ $h = 1, ..., H$. Independently within each of the H strata, select a with replacement sample of $(n_h - 1)$ PSU's and then extracting the residuals that belong to these PSU's. Note that due to the with replacement sampling, the weights have to be adjusted to compensate for some PSU's being over-sampled, undersampled or not sampled at all. Define m_{hj}^* as the number of times the jth PSU is sampled. The bootstrap weights are then calculated as

$$w_{hji}^* = w_{hji}\left[\left(\frac{n_h}{n_h - 1}\right) \cdot m_{hj}^*\right], \tag{4.3.8}$$

where w_{hji} is the original sampling weight (Rust et al., 1996). Now the bootstrap sample of residuals, $\underline{\hat{\varepsilon}}^* = \left\{\underline{\hat{\varepsilon}}_h^*\right\}$, $h = 1, ..., H$, is used to calculate the bootstrap response,

$$\mathbf{y}^* = \underline{\hat{\varepsilon}}^* + \mathbf{X}_{hj}\underline{\hat{\beta}}_{SWLS},$$

where $\mathbf{X}_{hj}$ is the original independent variables measured for the jth PSU in the hth stratum. Using the new bootstrap response variable $\mathbf{y}^*$ together with the original $\mathbf{X}_{hj}$ and the bootstrap weights $\mathbf{w}^*_{hj}$, the bootstrap estimator of the estimator of the regression coefficients is given by

$$\underline{\hat{\beta}}^* = \left(\mathbf{X}'\mathbf{W}^*\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^*\mathbf{y}^*,$$

where $\mathbf{W}^*$ is the diagonal matrix of bootstrap weights. This procedure is repeated a large number of times, B, resulting in B bootstrap estimators of the estimator of the regression parameters, $\hat{\mathbf{B}} = \left\{\underline{\hat{\beta}}^*_b\right\}$, $b = 1, ..., B$. Consider the jth estimator and associated bth bootstrap replicate of this estimator. The bootstrap estimated variance of the jth estimated regression parameter is calculated as

$$\hat{V}_B\left(\hat{\beta}_j\right) = \frac{1}{B-1}\sum_b \left(\hat{\beta}^*_{j_b} - \overline{\hat{\beta}^*_j}\right)^2.$$

4.3.2.2 Bootstrapping Pairs Method

Consider first the SRS case with data $(y_i, \mathbf{x}_i)$, $i = 1, ..., n$. Resample from these with equal probabilities, giving a bootstrap sample of $(y^*_i, \mathbf{x}^*_i)$, $i = 1, ..., n$. Use the bootstrap sample to estimate $\underline{\beta}$, giving

$$\underline{\hat{\beta}}^* = \left(\mathbf{X}^{*'}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*'}\mathbf{y}^*.$$

This procedure is repeated a large number of times, B, and finally the bootstrap estimated variance of the jth estimator of the regression parameters is calculated as before, namely

$$\hat{V}_B\left(\hat{\beta}_j\right) = \frac{1}{B-1}\sum_{b=1}^{B} \left(\hat{\beta}^*_{j_b} - \bar{\hat{\beta}}^*_j\right)^2. \tag{4.3.9}$$

Consider now the CS case. Assume the same design as described before. Within each stratum, select a with replacement sample of $(n_h - 1)$ PSU's and let these form the bootstrap sample. Let $\mathbf{y}^* = \left\{\mathbf{y}^*_{hj}\right\}$, $h = 1, ..., H$ denote the responses and $\mathbf{X}^* = \left\{\mathbf{X}^*_h\right\}$, $h = 1, ..., H$ denote the predictors corresponding to the PSU's in the bootstrap sample. As explained before the sampling weights have to be adjusted to compensate for the with replacement sampling. The bootstrap estimator of the estimator of the regression parameters is now given by

$$\underline{\hat{\beta}}^* = \left(\mathbf{X}^{*'}\mathbf{W}^*\mathbf{X}^*\right)^{-1}\mathbf{X}^{*'}\mathbf{W}^*\mathbf{y}^*.$$

This procedure is repeated a large number of times, B, resulting in B bootstrap estimators of the estimator of the regression parameters. Finally, the bootstrap estimated variance of the jth estimator of the regression parameters is calculated as explained under the SRS application,

$$\hat{V}_B\left(\hat{\beta}_j\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}^*_{j_b} - \bar{\hat{\beta}}^*_j\right)^2. \tag{4.3.10}$$

Two bootstrap regression approaches have been presented here and a natural question is which approach is considered superior. According to Efron and Tibshirani (1998) it mostly depends on whether the linear model can be trusted. The discussion of the two approaches makes it clear that the bootstrapping pairs approach should be less sensitive to the linear model assumptions. In fact, it had been found that the bootstrapping pairs estimated standard error is reasonable even when the model is not exactly linear or the distributional assumptions of the residuals are not met (Efron et al., 1998).

In this thesis only the bootstrapping pairs approach will be considered. Some reasons for this are:

- since the bootstrapping residuals method assumes a linear model and that resampling be carried out on the residuals of this model, the model assumptions underlying the residuals are assumed, and
- although computer power has significantly improved over the years, the simulation study of this thesis is substantial and thus the time was not available to include both bootstrap approaches.

Once the model has been specified and the regression parameters have been estimated along with their variances, the next step is to evaluate the model.

4.4 Model Evaluation

This section considers some diagnostics that are used for evaluating the fitted model. Here too the diagnostics will be discussed for OLS and WLS and then illustrate how the same diagnostics under SWLS differ from the OLS and WLS diagnostics. The section begins with the very well known and commonly used coefficient of multiple determination, R^2, as well as the adjusted coefficient, R^2_{adj}, followed by the estimation of the fitted model's prediction error (PE). Finally a selection of outlier detection diagnostics will be discussed.

4.4.1 Coefficient of Multiple Determination

The coefficient of multiple determination, denoted by R^2, is a diagnostic that measures the proportion of the total variation in the responses that can be attributed to the regression on the independent variables, i.e.

$$R^2 = \frac{Regression\, sum\, of\, squares}{Total\, sum\, of\, squares},$$

and is bound between 0 and 1.

Consider OLS. The regression sum of squares (SSR) is defined as

$$SSR_{OLS} = \hat{\beta}'_{OLS}\mathbf{X}'\mathbf{y} - n\bar{y}^2,$$

while the total sum of squares (SST) is defined as (Rencher, 2002)

$$SST_{OLS} = \mathbf{y}'\mathbf{y} - n\bar{y}^2.$$

Thus,

$$R^2_{OLS} = \frac{SSR_{OLS}}{SST_{OLS}},$$

or

$$R^2_{OLS} = 1 - \frac{SSE_{OLS}}{SST_{OLS}},$$

where SSE_{OLS} is defined as before. Rule of thumb is that the inclusion of additional independent variables will move R^2 closer to its upper limit, but never decrease it since SSE_{OLS} will only become smaller as additional variables are added to the model. Because of this property, a modified R^2 was introduced which adjusts for the number of variables in the model. This measure is called the adjusted coefficient of multiple determination and is defined as

$$R^2_{OLS_{adj}} = 1 - \frac{\frac{SSE_{OLS}}{n-p}}{\frac{SST_{OLS}}{n-1}},$$

where n is the sample size, p is the number of variables in the fitted model, $(n-p)$ is the degrees of freedom of SSE_{OLS} and $(n-1)$ is the degrees of freedom of SST_{OLS} (Kutner et al., 2005).

The adjusted R^2 might decrease with the addition of another variable to the model since the decrease in SSE_{OLS} might be offset by the loss of a degree of freedom, $(n-p)$ (Kutner et al., 2005).

The same definition applies when making use of WLS. Thus,

$$R^2_{WLS} = 1 - \frac{SSE_{WLS}}{SST_{WLS}},$$

where SSE_{WLS} is as defined previously and

$$SST_{WLS} = \sum_i w_i \left(y_i - \bar{y}\right)^2,$$

where w_i is the weight associated with the ith observation. Furthermore, if the weights specified

under WLS are the sampling weights, then

$$R^2_{SWLS} = R^2_{WLS},$$

and

$$R^2_{SWLS_{adj}=}R^2_{WLS_{adj}}.$$

It is mentioned in Heeringa et. al (2010) that, although analysts have been trained through textbook examples to expect R^2 values in excess of 0.8, the experience in practice is much different. Specifically social scientists are mentioned to not be discouraged when R^2 values of between 0.2 and 0.4 are observed. This should be expected (Heeringa et al., 2010).

A final word on R^2 is that one should not necessarily feel confident when a fitted model achieves a large R^2 value. It could be the case that the variation within the independent variables is small due to only a few levels being observed and thus the model might not predict well outside these levels. Alternatively, collinearity among the independent variables is also known to inflate R^2 and thus give a distorted impression of the quality of the fitted model (Kutner et al., 2005).

4.4.2 Model Prediction Error Estimation

Consider a sample of n observations where each observation is associated with a p-vector of measured covariates, $\mathbf{x}$, and a continuous response, y, with an unknown distribution, P. One of the aims of modeling is the construction of a rule which implements the information from $\mathbf{x}$ in order to predict y such that a future unobserved outcome, say y_0, can be predicted based on its associated measures in $\mathbf{x}_0$. Let this rule, or predictor, be defined as ψ such that

$$\hat{y} = \psi(\mathbf{x}),$$

where $\hat{y}$ is the predicted outcome associated with the observed $\mathbf{x}$. In the case of a continous response these predictors can be built via regression modeling (Molinaro et al., 2005).

Let ψ be written as $\psi(\cdot|P_n)$ such that P_n, the empirical distribution of the data, emphasizes the prediction rule's dependence on the observed data. To evaluate the performance of a prediction rule one can make use of loss functions, and most commonly the squared error loss, $L(y, \psi)$, given by

$$L(y, \psi) = (y - \psi(\mathbf{x}))^2.$$

For the purpose of evaluating the prediction rule, define an expected loss,

$$\hat{\theta} = R(\psi, P) = \int L(y, \psi(x))\, \partial P(x, y).$$

Since the distribution P is usually unknown, the prediction rule has an expected loss, or generalization error, given by

$$\tilde{\theta}_n = R\left(\psi\left(\cdot|P_n\right), P\right) = \int L\left(y, \psi\left(x|P_n\right)\right) \partial P\left(x, y\right),$$

which is also called the test error (Molinaro et al., 2005).

Keep in mind that, when evaluating the prediction rule, there are two seperate goals that might be of interest. The first is the selection of the "best" model through the evaluation of the performances of different models. In this case the aim is to find the model that minimizes the generalization error out of a collection of potential models. Once the final model has been chosen the goal is to determine how well the model predicts an out-of-sample response. Hence, one is interested in estimating the generalization error or prediction error (PE) (Molinaro et al., 2005; Hastie et al., 2009).

In an ideal world an independent dataset will be available for the purpose of model selection and to estimate the PE, but in reality the observed data is all one has available. Estimating the PE using the observed data gives the apparent error,

$$\hat{\theta}_n = R\left(\psi\left(\cdot|P_n\right), P_n\right) = \int L\left(y, \psi\left(x|P_n\right)\right) \partial P_n\left(x, y\right).$$

When a dataset is used to construct a prediction rule, the fitting method used to construct the rule adapts to the data to which it is fitted. Hence, using the same data to construct the rule and evaluate its performance, i.e. using the apparent error to estimate the generalization error, will lead to an estimated PE that is too optimistic (Molinaro et al., 2005; Hastie et al., 2009).

To address the problem of a biased PE estimate, techniques such as cross-validation (CV) and resampling methods such as the jackknife and bootstrap have been utilized to construct artificial extra-samples to be used as "new" observations to be predicted by the constructed prediction rule. This being said, these PE estimation methods may be well known in the SRS case, but not necessarily in the CS case. Each PE estimation method will thus be described for the SRS case and then developed for the CS case.

4.4.2.1 Cross-Validation

Considered to be the simplest and most widely used method for estimating PE, cross-validation (CV) splits the data into a set on which the model is fitted and a set on which the fitted model is tested (Hastie et al., 2009). At a minimum the data is split once into two parts, called split-sample CV, but this is only satisfactory for large data sets (Efron et al., 1998).

In general the data is split into K parts of roughly equal size. $K-1$ parts are used to fit the model while the remaining part is used for testing the fitted model. This is called K-fold cross-validation (KCV). Suppose the kth part of the K parts is retained as test sample and the

remaining $K-1$ parts are used as a learning sample on which the linear model is fitted. Suppose the kth part contains $n_k = \frac{n}{K}$ observations. Let the predicted value of the ith observation, y_i, be defined as $\hat{y}_i^{-k(i)}$to stress that this predicted value has been obtained using a model fitted to the data with the kth part removed. The estimated prediction error in this case is calculated as

$$\hat{PE}^k = \frac{1}{n_k} \sum_i \left(y_i - \hat{y}_i^{-k(i)} \right)^2,$$

where $\hat{PE}^k$ denotes the estimated prediction error of the kth test sample. This procedure is repeated for all K parts resulting in K estimates of prediction error. The final K-fold cross-validation estimated prediction error is taken as the average of the K estimated PE's (Efron et al., 1998),

$$\hat{PE}^{KCV} = \frac{1}{K} \sum_k \hat{PE}^k.$$

In K-fold CV both the proportion of observations in the test set and the number of estimates to average can affect the error estimate. When increasing K the proportion of observations in the test set decreases while the proportion in the learning set increases. This will cause a decrease in bias. Furthermore, a large number of estimates to average may also decrease the bias (Molinaro et al., 2005).

Leave-one-out Cross-Validation

Leave-one-out cross-validation (LOOCV) is a special case of K-fold CV where the sample is split into $K = n$ parts, one part for each observation. Let the part containing the first observation be the test sample and let the remaining $n-1$ parts be used to fit the linear model. Let the predicted value of the first observation be $\hat{y}_1^{-(1)}$where the superscript $-(1)$ is used to emphasize that the first observation was excluded from the data used to fit the linear model. The error in predicting the first observation is calculated as

$$\hat{PE}_1 = \left(y_1 - \hat{y}_1^{-(1)} \right)^2.$$

This is repeated for all n observations resulting in n estimates of PE, $\left\{ \hat{PE}_i \right\}$, for $i = 1, ..., n$ and finally the LOOCV estimated PE is calculated as (Efron et al., 1998)

$$\hat{PE}^{LOOCV} = \frac{1}{n} \sum_i \hat{PE}_i. \tag{4.4.1}$$

The LOOCV method is summarized in figure 4.4.1 below.

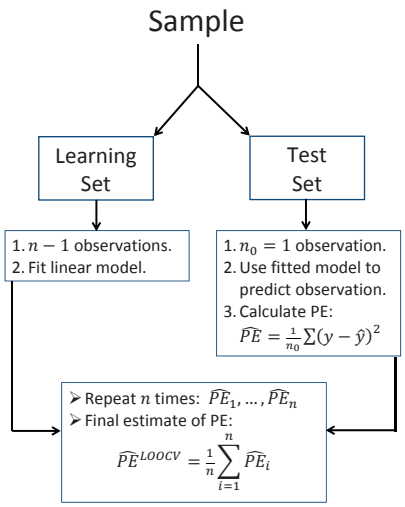


Figure 4.4.1: Diagram of the Leave-one-out Cross-Validation Method under SRS

Now consider the application of LOOCV to a CS design as in figure 2.5.1. In CS the cross-validation will be carried out in each stratum since strata are considered to be independent non-overlapping subgroups into which the entire population has been divided. Furthermore, the units within each stratum that are to be divided into a learning set and a test set will be the PSU's, the first level of sampling within each stratum. The reason for this is to ensure that the structure within the PSU's remains preserved.

The application of the LOOCV is quite similar to the jackknife resampling method where each of the units is omitted one at a time while the remaining units are used to estimate the parameter of interest. In this case the "parameter of interest" is the linear model and the remaining units comprise the learning set to which the linear model is fitted. The omitted unit forms the test set and the model fitted to the learning set is used to predict this unit in the test set. Recall that the CS design in figure 2.5.1 comprises of H strata and if the hth stratum is considered, recall that it contains n_h PSU's. Suppose the jth PSU in the hth stratum is assigned to the test set with the remaining $n_h - 1$ PSU's comprising the learning set. The sampling weights associated with the units within the learning set have to be adjusted to compensate for the missing PSU such that the sum of the sampling weights still equals the correct population total. Thus, from Rust and Rao (1996),

$$w_{i(hj)} = \begin{cases} w_i,\ i \notin h \\ w_i \cdot \frac{n_h}{(n_h-1)},\ i \in h,\ i \notin j \\ 0,\ i \in (h,j) \end{cases}, \tag{4.4.2}$$

where $w_{i(hj)}$ is the adjusted sampling weight of the ith unit after the (hj)-th PSU has been removed to form the test sample. These new weights are then incorporated when fitting a survey-weighted least squares regression model to the learning set. The fitted model is then used to predict

the test set, $\hat{y}_1^{-(hj)}, ..., \hat{y}_{n_{hj}}^{-(hj)}$, where n_{hj} is the number of SSU's in the test set which contains the jth PSU of the hth stratum. The estimated PE will then be calculated as

$$\hat{PE}_{(hj)} = \frac{1}{\sum w_{hji}} \sum_{i=1}^{n_{hj}} w_{hji} \left(y_i^{-(hj)} - \hat{y}_i^{-(hj)} \right)^2,$$

where w_{hji} is the original sampling weight associated with the ith SSU in the jth PSU that is now considered the test set. This is repeated for each $j \in h$ and $h = 1, ..., H$.

Once each PSU in stratum h has had a turn to be the test set, n_h PE's will have been calculated. Thus, the overall LOOCV estimated PE under SWLS will be similar to a jackknife estimated average and thus calculated as

$$\hat{PE}_{SWLS}^{LOOCV} = \sum_{h=1}^{H} \frac{1}{n_h} \sum_{j=1}^{n_h} \hat{PE}_{(hj)}.$$

LOOCV, with its small proportion in the test set, represents the best example of a bias-variance trade-off since it achieves a small bias, but with increased variances due to the number of estimated PE's included in the final LOOCV estimated PE. Although, traditionally, LOOCV was considered to be computationally (too) expensive, modern computing power has made it an attractive method to use. This is especially true where there is no clear guide lines for the choice of K. In the applications discussed in later chapters, LOOCV was employed.

4.4.2.2 Bootstrap Methods

This section considers two bootstrap methods of PE estimation as alternatives to the well-known cross-validation estimation. The sections begins with a discussion of the bootstrap estimator of prediction error under SRS, a method that determines an optimism with which the apparent prediction error is adjusted, after which it is expanded for application to CS data. This is followed by a discussion of the .632 bootstrap estimator of prediction error under SRS. This more recent approach sees the optimism calculated as a weighted average of the difference between an out-of-sample error rate and the apparent error. The SRS .632 bootstrap discussion is then also followed by an extension of the method to CS data.

Bootstrap Estimator of Prediction Error

Consider a sample of size n and let the responses of the sample be denoted by $y_1, ..., y_n$. The sample is used to fit a linear model which is evaluated by estimating the response of the sample from which the model was obtained and calculating its PE. The prediction error calculated in this regard is called the apparent error rate,

$$\hat{PE}^{Apparent} = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2 , \tag{4.4.3}$$

where y_i is the observed and $\hat{y}_i$ the estimated response of the ith observation.

To obtain the bootstrap estimate of PE, generate a with-replacement bootstrap sample of size n from the observed sample and fit a linear model to the bootstrap sample. Firstly the model, from which the bootstrap estimator of the estimator of the regression parameters is obtained, is used to predict the response of the observed sample,

$$\hat{\mathbf{y}} = \mathbf{X}\underline{\hat{\beta}}^*, \tag{4.4.4}$$

where $\mathbf{X}$ is the matrix of predictor variables of the original sample and $\underline{\hat{\beta}}^*$ is the bootstrap estimator. These predicted responses are then used to obtain a PE estimate,

$$\hat{PE}_1^{B_1} = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2 ,$$

where the superscript B_1 is used to label the above PE as the PE calculated from the predicted responses of the observed sample obtained from the bootstrap linear model. Next the fitted model is used to estimate the responses of the bootstrap sample to which the model has been fitted,

$$\hat{\mathbf{y}}^* = \mathbf{X}^*\underline{\hat{\beta}}^*, \tag{4.4.5}$$

where $\mathbf{X}^*$ is the bootstrap matrix of predictor variables and $\hat{\mathbf{y}}^*$ is the vector of estimated responses of the bootstrap sample. These estimated responses are used to obtain a second PE,

$$\hat{PE}_1^{B_2} = \frac{1}{n}\sum_i (y_i^* - \hat{y}_i^*)^2 ,$$

with superscript B_2 to emphasize that the PE is calculated using the estimated bootstrap responses. Finally, the difference between the two estimated PE's is calculated,

$$\widehat{Diff}_1 = \hat{PE}_1^{B_1} - \hat{PE}_1^{B_2}. \tag{4.4.6}$$

The process is repeated for each bootstrap sample resulting in B differences, $\left\{\widehat{Diff}_b\right\}$, $b = 1, ..., B$. The differences are used to calculate the optimism,

$$optimism = \frac{1}{B}\sum_b \widehat{Diff}_b, \tag{4.4.7}$$

a number which represents the amount by which the apparent error rate underestimates the true PE (Efron et al., 1998). Finally, the bootstrap estimator of prediction error is obtained as the sum of the apparent prediction error and the optimism,

$$\hat{PE}^{BS} = \hat{PE}^{Apparent} + Optimism. \tag{4.4.8}$$

A diagram of the bootstrap estimation of PE is given in figure 4.4.2 as a summary of the method.

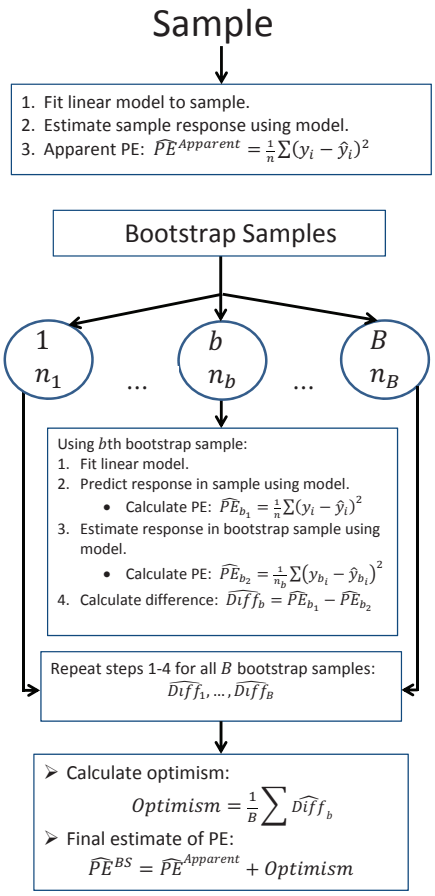


Figure 4.4.2: Diagram of the Bootstrap PE Estimation Method under SRS

Now, suppose a complex sample (CS) of H strata, with N_h PSU's. In the hth stratum n_h PSU's have been sampled and from the jth PSU a sample of n_{hj} SSU's has been selected, $j = 1, ..., n_h$ and $h = 1, ..., H$. An SWLS is fitted to the CS and the model is used to estimate the response of the sample which is used to calculate the apparent PE which, under CS, is given by

$$\hat{PE}^{Apparent}_{SWLS} = \sum_h \frac{N_h}{N} \sum_j \frac{1}{n_{hj}} \sum_i (y_{hji} - \hat{y}_{hji})^2, \tag{4.4.9}$$

where y_{hji} and $\hat{y}_{hji}$ are, respectively, the observed and estimated response of the ith SSU in

the jth PSU of the hth stratum. Notice how the sampling weights have been employed in the SWLS model, but not in the calculation of 4.4.9. The reasoning behind this lies therein that the objective with the PE is to gauge how well the model predicts the sample and not a population. Also, sampling weights are mostly employed to improve the precision of the estimators of unknown parameters under CS and in this regard the sampling weights have fulfilled their purpose in the estimation of $\underline{\beta}$. However, it is still important to respect the clustering and the stratification within the data. Thus, in 4.4.9,

$$\frac{1}{n_{hj}} \sum_{i} \left(y_{hji} - \hat{y}_{hji}\right)^2$$

represents the calculation of the PE for the jth cluster in the jth stratum, and

$$\sum_{h} \frac{N_h}{N} \sum_{j} (\cdot)$$

represents the overall apparent PE calculated as a weighted average of PSU PE's.

The bootstrap resampling technique is applied independently within each stratum by sampling as discussed in section 4.3.2. The bootstrap weights, defined in 4.3.8, are used in the SWLS model fitted to the bootstrap sample and the model, i.e. $\underline{\hat{\beta}}^*_{SWLS}$, is then used in the same two ways as for OLS, namely to predict the response of the original sample, as given in 4.4.4, and to estimate the response of the bootstrap sample, as given in 4.4.5. The simple prediction error is calculated as

$$\hat{PE}^{simple}_{SWLS} = \sum_{h} \frac{N_h}{N} \sum_{j} \frac{1}{n_{hj}} \sum_{i} \left(y_{hji} - \hat{y}_{hji}\right)^2,$$

similarly to the apparent PE, but recall that the responses have been predicted as in 4.4.4. Next the estimate of improved prediction error is obtained as

$$\hat{PE}^{improved}_{SWLS} = \sum_{h} \frac{N_h}{N} \sum_{j} \frac{1}{n^*_{hj}} \sum_{i} \left(y^*_{hji} - \hat{y}^*_{hji}\right)^2,$$

where y^*_{hji} and $\hat{y}^*_{hji}$ are, respectively, the observed and estimated response of the ith observation in the jth PSU in the bootstrap sample from stratum h. This is followed by calculating the difference between the simple and improved bootstrap estimated PE's defined previously,

$$\widehat{Diff} = \hat{PE}^{simple}_{SWLS} - \hat{PE}^{improved}_{SWLS}.$$

This process is repeated for all B bootstrap samples resulting in $\left\{\widehat{Diff}_b\right\}$, $b = 1, ..., B$, and then the optimism is calculated as the average of these differences. Now the bootstrap estimate of prediction error under CS is given by,

$$\hat{PE}^{BS}_{SWLS} = \hat{PE}^{Apparent}_{SWLS} + Optimism_{SWLS}.$$

The .632 Bootstrap Estimator of Prediction Error

The application of this method for OLS regression begins by generating B bootstrap samples. Then, for each $i = 1, ..., n$, divide the bootstrap samples into those that contain the ith observation and those that don't. Since the prediction error for the ith observation will be larger for a bootstrap sample that does not contain the observation, it is proposed to use the prediction error from only these cases to adjust the optimism in the apparent error rate (Efron et al., 1998).

As before, consider a sample of size n to which a linear model is fitted from which the apparent error rate as defined in (4.4.3), is obtained. Commence by generating B bootstrap samples of size n, with-replacement, from the observed sample. Consider the ith observation of the observed sample and determine which of the bootstrap samples do not contain this observation. Let this number be denoted by B_i.

Consider sample b_i of these B_i bootstrap samples. Using this sample,

1. fit a linear model to the bootstrap sample,
2. use the model to predict the ith observation, and
3. calculate $\hat{PE}_{b_i} = (y_i - \hat{y}_i)^2$.

The above steps are repeated for each of the bootstrap samples identified to not contain the ith observation, i.e. for $b_i = 1, ..., B_i$. Each of the $\left\{\hat{PE}_{b_i}\right\}$ is used to calculate the overall prediction error of the ith observation,

$$\hat{PE}_i = \frac{1}{B_i} \sum_{b_i} \hat{PE}_{b_i}.$$

The above procedure is repeated for each of the observations in the observed sample resulting in n estimated PE's $\hat{PE}_1, ..., \hat{PE}_n$ which are used to calculated the average estimated error rate,

$$\hat{\varepsilon}_0 = \frac{1}{n} \sum_{i} \hat{PE}_i.$$

From this, the .632 estimate of optimism is given by

$$optimism^{.632} = 0.632 \left[\hat{\varepsilon}_0 - \hat{PE}^{Apparent}\right],$$

where .632 represents from the probability that a given observation is in a bootstrap sample of size n. This factor adjustment in optimism is said to make the .632 prediction error estimator approximately unbiased (Efron et al., 1998).

Finally, the .632 bootstrap estimated prediction error is calculated as

$$\hat{PE}^{.632} = \hat{PE}^{Apparent} + optimism^{.632}.$$

A summary of the .632 bootstrap estimator of PE is given in the figure 4.4.3.

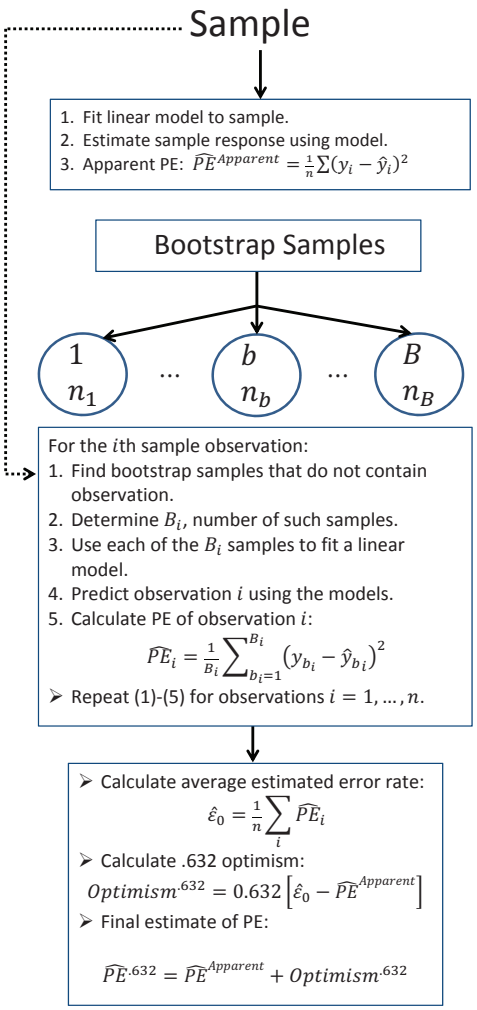


Figure 4.4.3: Diagram of the .632 Bootstrap Estimation of PE under SRS

Next, consider a CS of H strata and in the hth stratum there is n_h PSU's, $h = 1, ..., H$. Firstly, fit an SWLS model to the data and determine the apparent error rate, $\hat{PE}^{Apparent}_{SWLS}$, as given in (4.4.9).

Consider stratum h with n_h PSU's from which B bootstap samples are selected as explained in section 4.3.2. Now consider the jth PSU in stratum h in the original sample. Determine which of the B bootstrap samples do not contain this PSU and let this number be denoted by B_{hj}, $j = 1, ..., n_h$, and consider sample b_{hj} of these samples. Using SLWS regression a model is fitted to this bootstrap sample after which the model is used to predict the jth PSU. Recalling that the

jth PSU contains n_{hj} SSU's, the PE is then calculated as

$$\hat{PE}_{b_{hj}} = \frac{1}{n_{hj}} \sum_{i=1}^{n_{hj}} \left(y_{hji} - \hat{y}_{hji}\right)^2 .$$

This is repeated for all B_{hj} bootstrap samples that do not contain the jth PSU resulting in PE's $\hat{PE}_1, ..., \hat{PE}_{B_{hj}}$ after which the PE for the jth PSU in the hth stratum is calculated as

$$\hat{PE}_{hj} = \frac{1}{B_{hj}} \sum_{b_{hj}=1}^{B_{hj}} \hat{PE}_{b_{hj}}.$$

The procedure is repeated for all n_h PSU's in stratum h and the average estimated error rate is then calculated as

$$\hat{\varepsilon}_{h_0} = \frac{1}{n_h} \sum_{j=1}^{n_h} \hat{PE}_{hj}.$$

The overall estimated error rate is then calculated as the weighted average of the PE's estimated for the PSU's,

$$\hat{\varepsilon}_0^{SWLS} = \sum_h \frac{N_h}{N} \hat{\varepsilon}_{h_0},$$

where N_h is the number of PSU's in stratum h and N is the total number of PSU's in the CS.

Here, as with the bootstrap estimator of PE, the sampling weights are used under SWLS regression, but not in the calculation of the PE's and error rates. The reasoning behind not using the sampling weights past the SWLS model was explained as part of the bootstrap estimator of PE discussion. An argument can also be made for incorporating the sampling weights further, especially if the desire is to estimate how well a model will predict a population response. However, this will be considered under further research.

Now, the .632 estimate of optimism is given by

$$Optimism_{SWLS}^{.632} = 0.632 \left[\hat{\varepsilon}_0^{SWLS} - \hat{PE}_{SWLS}^{Apparent}\right],$$

and then the .632 bootstrap estimated prediction error is calculated as

$$\hat{PE}_{SWLS}^{.632} = \hat{PE}_{SWLS}^{Apparent} + Optimism_{SWLS}^{.632}.$$

The implementation of the methods discussed here forms part of the outline of the analyses presented at a later stage in the thesis.

4.4.3 Outlier Detection Diagnostics

Once a model has been fitted it is important to measure the quality of the model due to the presence of phenomena such as collinearity and extreme points which could influence any inference with regard to the model (Liao and Valliant, 2012; Liao, 2010). Extreme points can exist due to factors such as

- outliers in the predictors, dependent variables or both;
- large weights when working with survey data; as well as
- interaction between the weights and variables.

Although much research has been done on regression diagnostics for non-survey data, this is not the case for survey data. Liao and Valliant (2012) are of the opinion that the work done in this area over the past decade was mostly focused on assessing the quality of the regression on survey data through the identification of influential points (outliers that greatly affect the slope of the regression line) and influential groups through abnormal data values or weights. The following work was cited (Liao and Valliant, 2012):

- Li and Valliant (2011; 2009)
  - Adaption and extension of traditional diagnostic techniques to regression on CS data, mainly on the identification of influential observations and influential groups.
  - Other topics covered include:
    * Residuals;
    * Leverages;
    * DFBetas;
    * DFfits;
    * Cook's distance; and
    * Forward search.

Li and Valliant (2015) have recently extended the calculation of diagnostics for influential observations to, specifically, stratified multistage cluster samples (CS). For this purpose, consider a population divided into H strata where stratum h contains N_h PSU's, $h = 1, ..., H$. Furthermore, the (hj)-th PSU contains N_{hj} SSU's. Suppose a stratified two-stage cluster sample is selected from the population. This CS is also made up of H strata and n_h PSU's are selected from stratum h.

It is mentioned in (Li et al., 2015) that the PSU's are selected with replacement, but in practice this sample selection will be carried out without replacement. The reason specified for using with replacement sampling is that it provides simpler design-based variance formulae which was of importance for that article. Finally, n_{hj} SSU's are selected from the (hj)-th selected PSU. Let $\mathbf{x}_{hji}$ be a p-dimensional vector of independent variables for the ith observation in the jth PSU in the hth stratum. A response variable, Y_{hji}, collected in the CS follows a linear model,

$$Y_{hji} = \mathbf{x}_{hji}^{'}\underline{\beta} + \varepsilon_{hji},$$

with the variance-covariance of the (hji)-th residual defined as

$$Cov_M\left(\varepsilon_{hji}, \varepsilon_{hji}^{'}\right) = \begin{cases} \sigma^2, & h = h^{'}, j = j^{'}, i = i^{'} \\ \rho\sigma^2, & h = h^{'}, j = j^{'}, i \neq i^{'} \\ 0, & otherwise \end{cases},$$

where ρ is the intracluster correlation (ICC), a measure of how homogeneous the units within a PSU are (Li et al., 2015). This model thus suggests that the units have a common variance and that the ICC is the same for all PSU's. It follows that the SWLS estimator of $\underline{\beta}$ can be written as

$$\underline{\hat{\beta}}_{SWLS} = \sum_{h=1}^{H}\sum_{j=1}^{n_h} \mathbf{A}^{-1}\mathbf{X}_{hj}^{'}\mathbf{W}_{hj}\mathbf{y}_{hj},$$

where

- $\mathbf{X}_{hj}$ is the $n_{hj} \times p$ matrix of predictors for the units in the(hj)-th PSU;
- $\mathbf{W}_{hj}$ is the $n_{hj} \times n_{hj}$ diagonal matrix of sampling weights for the units in the (hj)-th PSU;
- $\mathbf{y}_{hj}$ is the $n_{hj} \times 1$ vector of responses for the units in the (hj)-th PSU; and
- $\mathbf{A} = \sum_h \sum_j \mathbf{X}_{hj}^{'}\mathbf{W}_{hj}\mathbf{X}_{hj}$.

According to Li and Valliant (2015) the model variance of $\underline{\hat{\beta}}_{SWLS}$ is given by

$$V_M\left(\underline{\hat{\beta}}_{SWLS}\right) = \sum_h \sum_j \mathbf{A}^{-1}\mathbf{X}_{hj}^{'}\mathbf{W}_{hj}V_M\left(\mathbf{y}_{hj}\right)\mathbf{W}_{hj}\mathbf{X}_{hj}\mathbf{A}^{-1},$$

which can be further simplified to

$$V_M\left(\underline{\hat{\beta}}_{SWLS}\right) = \sum_h \sum_j \mathbf{A}^{-1}\mathbf{X}_{hj}^{'}\mathbf{W}_{hj}\left[\left(1-\rho\right)\sigma^2\mathbf{I}_{n_{hj}} + \rho\sigma^2\mathbf{1}_{n_{hj}}\mathbf{1}_{n_{hj}}^{'}\right]\mathbf{W}_{hj}\mathbf{X}_{hj}\mathbf{A}^{-1},$$

where $\mathbf{I}_{n_{hj}}$ is an $n_{hj} \times n_{hj}$ identity matrix and $\mathbf{1}_{n_{hj}}$ is a vector of n_{hj} 1's.

The model-based variance of $\underline{\hat{\beta}}_{SWLS}$ requires the estimation of $\rho\sigma^2$. Now, using the residuals from an OLS regression, the following quantities can be defined (Li et al., 2015):

- $\hat{P} = \frac{1}{n}\sum_h \sum_j \frac{1}{n_{hj}-1}\sum_i \left(\hat{\varepsilon}_{hji} - \bar{\hat{\varepsilon}}_{hj}\right)^2$, where $\bar{\hat{\varepsilon}}_{hj}$ is the average OLS residual in the (hj)-th PSU;
- $\hat{Q} = \frac{\sum_h \sum_j n_{hj}\left(\bar{\hat{\varepsilon}}_{hj} - \bar{\hat{\varepsilon}}_h\right)^2}{n-1}$, where $\bar{\hat{\varepsilon}}_h$ is the average OLS residual in the hth stratum and n is the number of PSU's in the sample; and
- $\hat{D} = \frac{\left(m - \sum_h \sum_j \frac{n_{hj}^2}{m}\right)}{n-1}$, where m is the total number of observations in the sample.

These estimates can now be used to estimate the unknown component of the regression parameter model variance (Li et al., 2015), as

$$\widehat{\rho\sigma^2} = \frac{\hat{Q} - \hat{P}}{\hat{D}}.$$

To conclude the model-based variance estimator, it is mentioned that it is very sensitive to departures from the model and due to this non-robustness, replication estimators are preferred. Jackknife and bootstrap resampling methods have been discussed in section 4.3. The advantage of the model-based variance, however, lies in its usefulness to determine cut-offs for diagnostics (Li et al., 2015).

The variances of regression parameter estimators give a first impression of the accuracy of the estimated parameters, but should be used with caution since it is known that the variances of the regression estimators can be inflated by collinearity. As such, along with the variances, other diagnostics regarding the identification of spurious observations should be considered and thus the remainder of this section is devoted to such measures.

4.4.3.1 Hat Matrix and Leverages

The hat matrix along with its diagonal elements, termed leverages, are well known measures to use for the identification of outliers in the predictor variables. It is of importance to identify such cases since these may influence the model fitting. Work done by Li and Valliant (2009) showed that not only are the hat matrix and leverages useful for detecting predictor outliers, but also for detecting large sample weights that could be equally influential on the model fit.

Before embarking on the survey sampling case, consider the hat matrix as defined under OLS. As discussed previously, the OLS estimator of $\underline{\beta}$ is given by

$$\underline{\hat{\beta}}_{OLS} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}.$$

If $\underline{\hat{\beta}}_{OLS}$ is re-written as

$$\underline{\hat{\beta}}_{OLS} = \mathbf{A}^{-1}\mathbf{X}'\mathbf{y},$$

where $\mathbf{A} = \mathbf{X}'\mathbf{X}$ is a square and invertible matrix, then the fitted values of $\mathbf{y}$ can be defined as

$$\hat{\mathbf{y}}_{OLS} = \mathbf{X}\hat{\underline{\beta}}_{OLS} = \mathbf{X}\mathbf{A}^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}_{OLS}\mathbf{y}, \tag{4.4.10}$$

where $\mathbf{H}_{OLS} = \mathbf{X}\mathbf{A}^{-1}\mathbf{X}'$ is the hat matrix with ith diagonal element $h_{ii} = \mathbf{x}_i'\mathbf{A}^{-1}\mathbf{x}_i$, $i = 1, ..., n$, the leverage of the ith observation. Some special properties of $\mathbf{H}_{OLS}$ are:

1. it is symmetric;
2. it is idempotent, i.e. $\mathbf{H}_{OLS} = \mathbf{H}_{OLS}^2$;
3. $\mathbf{H}_{OLS}\mathbf{X} = \mathbf{X}$;
4. the leverages lie between 0 and 1, i.e. $0 \leq h_{ii} \leq 1$; and
5. $\sum_i h_{ii} = p$, where p is the number of independent variables and also the rank of $\mathbf{X}$.

It has also been shown that if the model contains an intercept, then the hat matrix has two additional properties, namely

1. $\sum_i h_{ii} = 1$, and
2. $h_{ii} = \frac{1}{n} + (\mathbf{x} - \bar{\mathbf{x}})' \mathbf{A}^{-1} (\mathbf{x} - \bar{\mathbf{x}})$,

where $\mathbf{A} = \mathbf{X}'\mathbf{X}$ and $\bar{\mathbf{x}}$ is the vector of means (Li et al., 2009). The leverage measures the impact of y_i on its associated fitted value, $\hat{y}_i$. An extreme leverage is one that twice exceeds the mean leverage, $\bar{h} = \frac{\sum h_{ii}}{n}$ (Li et al., 2009).

When using WLS to account for unequal variances, recall that the estimator of $\underline{\beta}$ becomes

$$\hat{\underline{\beta}}_{WLS} = \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}.$$

In this case the predicted values become

$$\hat{\mathbf{y}}_{WLS} = \mathbf{X}\hat{\underline{\beta}}_{WLS} = \mathbf{X}\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}\mathbf{y} = \mathbf{H}_{WLS}\mathbf{y}, \tag{4.4.11}$$

where $\mathbf{H}_{WLS} = \mathbf{X}\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W} = \mathbf{X}\mathbf{A}^{-1}\mathbf{X}'\mathbf{W}$ represents the hat matrix under WLS with the leverage of the ith observation under WLS defined as (Li et al., 2009)

$$h_{ii} = \mathbf{x}_i'\mathbf{A}^{-1}\mathbf{x}_i w_i. \tag{4.4.12}$$

Since these leverages are constructed from covariates and weights, they are not affected by variation in the responses, $\mathbf{y}$ (Valliant, 2010). Furthermore, since the leverages do not include the standard errors of the estimated regression estimators and the commonly followed approach

of setting $\mathbf{V} = \mathbf{I}$ is assumed, it follows that under SWLS the leverages will be the same as under WLS as long as the sampling weights are used under WLS. Hence,

$$\mathbf{H}_{SWLS} = \mathbf{H}_{WLS},$$

and, assuming a stratified two-stage cluster sample with H strata, n_h PSU's selected from the hth stratum and n_{hj} SSU's selected from the (hj)-th PSU, the leverage of the (hji)th observation is given by

$$h_{hji,i} = \mathbf{x}_{hji}^{'}\mathbf{A}^{-1}\mathbf{x}_{hji}w_{hji}.$$

By incorporating the sampling weights into the hat matrix it is no longer symmetric, but properties (2) - (6), assuming that $\mathbf{H}_{OLS}$ is replaced by $\mathbf{H}_{SWLS}$, still hold. Some additional properties that the hat matrix now possesses, are (Li et al., 2009)

1. $\mathbf{W}\mathbf{H}_{SWLS} = \mathbf{H}_{SWLS}^{'}\mathbf{W}$,
2. $\mathbf{X}^{'}\mathbf{H}_{SWLS}\left(\mathbf{I} - \mathbf{H}_{SWLS}\right) = \mathbf{0}$, and
3. $w_i h_{i'i} = w_i h_{ii'}$.

In Li and Valliant (2009) scatterplots are used to plot the response versus each independent variable. For the OLS leverages the points on the scatterplot that exceed the proposed cut-off are identified. As opposed to this, bubble plots are used to visualize the response by each independent variable and the relative size of the bubble are proportional to the sampling weight of the point. The OLS cut-off is also used for the SWLS leverages and these are also identified on the bubble plots.

Numerical studies conducted in both Li and Valliant (2009) and Valliant (2010) found that when comparing leverages from OLS and SWLS, SWLS resulted in more leverages and furthermore, the leverages identified under OLS differed from those identified under SWLS.

4.4.3.2 Standardized Residuals

The calculation of standardized residuals is another way of determining which observations could be outlying and influential on the model fit. In OLS, residuals are calculated as

$$e_i = y_i - \hat{y}_i,\ i = 1, ..., n,$$

with mean of approximately zero and estimated variance

$$\hat{\sigma}^2 = \frac{\sum_{i \epsilon s} e_i^2}{n - p}.$$

In the field of regression diagnostics it is useful to standardize these residuals to have variance of approximately 1. This can be achieved by using either the root mean squared error of the residuals, $\sqrt{MSE}$, or their estimated standard error (Valliant, 2010).

In the case of SWLS regression the residual for unit (hji) is calculated as

$$\hat{\varepsilon}_{hji} = y_{hji} - \mathbf{x}_{hji}^{'}\underline{\hat{\beta}}_{SWLS},$$

with model variance,

$$V_M\left(\hat{\varepsilon}_{hji}\right) = \sigma^2\left[\left(1 - h_{hji,hji}\right)^2 + \sum_{i' \neq i} h_{ii'}^2\right],$$

where $h_{hji,hji}$ is the leverage of the (hji)th observation. In Li and Valliant (2015) the model-based standard deviation of the residuals is shown to be

$$\hat{\sigma} = \sqrt{\hat{P} + \left(\hat{Q} - \hat{P}\right)\hat{D}^{-1}},$$

where $\hat{P}$, $\hat{Q}$ and $\hat{D}$ are the estimators defined earlier.

The standard deviation of the residuals is used for the standardization and the standardized residuals are compared to the percentiles of a standard normal distribution. In cases where the distribution of the e_i's departs from normality, cut-off values can be obtained from the Gauss inequality. Suppose a distribution has a single mode defined as μ_0. Then the Gauss inequality is given by

$$P\left\{|X - \mu_0| > \lambda\tau\right\} \leq \frac{4}{9\lambda^2}, \tag{4.4.13}$$

where $\tau^2 \equiv \sigma^2 + \left(\mu - \mu_0\right)^2$ (Valliant, 2010). Assuming that the residual follows a symmetric distribution with both its mean and mode at zero, (4.4.13) implies that

1. the absolute value of a residual should be less than twice its standard deviation with 90% probability, and

2. the absolute value of a residual should be less than thrice its standard deviation with 95% probability.

4.4.3.3 DFBetas

Along with the leverages and standardized residuals, the DFBeta is also a well-known outlier detection diagnostic. Under OLS regression with estimated regression parameters, $\underline{\hat{\beta}}_{OLS}$, and associated model variances, $V\left(\underline{\hat{\beta}}_{OLS}\right)$, the DFBeta of the ith observation is defined as

$$DFBETA_i = \hat{\underline{\beta}}_{OLS} - \hat{\underline{\beta}}_{OLS(i)},\ i = 1, ..., n,$$

where $\hat{\underline{\beta}}_{OLS(i)}$ is the estimated regression parameters with the ith observation removed.

In survey sampling, when taking the weights $\mathbf{W}$ into consideration the DFBeta of the hjith observation is calculated as

$$DFBETA_{hji} = \hat{\underline{\beta}}_{SWLS} - \hat{\underline{\beta}}_{SWLS(hji)},\ h = 1, ..., H, j = 1, ..., n_h, i = 1, ..., n_{hj}$$

where $\hat{\underline{\beta}}_{SWLS}$ is the SWLS estimate of $\underline{\beta}$ with all observations included and $\hat{\underline{\beta}}_{SWLS(hji)}$ is the SWLS estimate when the ith observation in the (hj)th PSU is deleted. It can be showed that this simplifies to

$$DFBETA_i = \frac{\mathbf{A}^{-1}\mathbf{x}_{hji}\hat{\varepsilon}_{hji}w_{hji}}{1 - h_{hji,hji}}, \tag{4.4.14}$$

where $h_{hji,hji}$ is the leverage of the (hji)th observation and $\mathbf{A} = \mathbf{X}'\mathbf{W}\mathbf{X}$ (Valliant, 2010; Li et al., 2015). Let the effect of the hjith unit on the kth coefficient be defined as

$$DFBETAS_{hji,k} = \frac{\frac{c_{hji,k}\hat{\varepsilon}_{hji}}{\left(1-h_{hji,hji}\right)}}{\sqrt{V_M\left(\hat{\underline{\beta}}_{SWLS_k}\right)}},\ h = 1, ..., H, j = 1, ..., n_h, i = 1, ..., n_{hj}\, k = 1, ..., p, \tag{4.4.15}$$

where $c_{hji,k} = \left(\mathbf{A}^{-1}\mathbf{x}_{hji}\hat{\varepsilon}_{hji}w_{hji}\right)_j$ and $V_M\left(\hat{\underline{\beta}}_{SWLS_k}\right)$ is the model-based variance of the kth survey weighted estimated regression coefficient that takes account of the unequal probabilities, stratification and other design complexities of a survey sample. The cutoff value for outliers is $\frac{z}{\sqrt{n}}$, with z equal to 2 or 3. Alternatively, use $\frac{t_{\frac{\alpha}{2};n-p}}{\sqrt{n}}$ as cut-off value where $t_{\frac{\alpha}{2};n-p}$ is the $\left(1-\frac{\alpha}{2}\right)$th percentile of the Student -t distribution with $n-p$ degrees of freedom (Valliant, 2010; Li et al., 2015).

4.4.3.4 DFFits

The DFFIT measure, a further outlier diagnostic, is obtained by multiplying the DFBETA, defined in (4.4.14) by $\mathbf{x}'_{hji}$ to give

$$DFFIT_{hji} = \mathbf{x}'_{hji}\left(\hat{\underline{\beta}}_{SWLS} - \hat{\underline{\beta}}_{SWLS(hji)}\right) = \frac{h_{hji,hji}\hat{\varepsilon}_{hji}}{1 - h_{hji,hji}},\ h = 1, ..., H, j = 1, ..., n_h, i = 1, ..., n_{hj}. \tag{4.4.16}$$

This measure is used to examine the change in the (hji)th fitted value when the (hji)th observation is deleted (Valliant, 2010).

In Li and Valliant (2015) the model-based variance of the (hji)th predicted value is defined as

$$V_M\left(\hat{y}_{hji}\right) = \mathbf{x}_{hji}^{'} V\left(\underline{\hat{\beta}}_{SWLS}\right) \mathbf{x}_{hji}.$$

Now, when scaling the measure by dividing (4.4.16) by the standard deviation of the (hji)th predicted value, the DFFITS diagnostic is obtained and given by

$$DFFITS_{hji} = \frac{\frac{h_{hji,hji}\hat{\varepsilon}_{hji}}{\left(1-h_{hji,hji}\right)}}{\sqrt{V_M\left(\hat{y}_{hji}\right)}}. \tag{4.4.17}$$

In this case the cutoff for extreme points is set at $z\sqrt{\frac{p}{n}}$, where z is set equal to 2 or 3 (Li et al., 2015).

4.4.3.5 Extended and Modified Cook's Distance

Another diagnostic often used to identify outliers, is Cook's distance. It is constructed as the distance between $\underline{\hat{\beta}}$, estimated using the full set of observations, and $\underline{\hat{\beta}}_{(i)}$, estimated with the ith observation deleted. It measures the effect of a single unit on the estimate $\underline{\hat{\beta}}$.

In survey weighted least squares the hjith Cook's Distance measure is calculated as

$$\begin{aligned} ED_{hji} &= \left(\underline{\hat{\beta}}_{SWLS} - \underline{\hat{\beta}}_{SWLS(hji)}\right)^{'} \left[V_M\left(\underline{\hat{\beta}}_{SWLS}\right)\right]^{-1} \left(\underline{\hat{\beta}}_{SWLS} - \underline{\hat{\beta}}_{SWLS(hji)}\right), \\ &\quad h = 1, ..., H, j = 1, ..., n_h, i = 1, ..., n_{hj}, \end{aligned} \tag{4.4.18}$$

where $\underline{\hat{\beta}}_{SWLS(hji)}$ is the SWLS estimated regression parameters with the (hji)th observation deleted and $\underline{\hat{\beta}}_{SWLS}$ is the estimate using the full set of observations (Li et al., 2015). For convenience, standardize the measure and then take its square root. This modification leads to the modified Cook's Distance,

$$MD_{hji} = \sqrt{\frac{\left\{n\bar{m}\left[1+\hat{\rho}\left(\bar{m}-1\right)\right]\right\} ED_{hji}}{p}},\ h = 1, ..., H, j = 1, ..., n_h, i = 1, ..., n_{hj},$$

where

- n is the total number of PSU's in the sample,
- $\bar{m}$ is the average number of observations in a PSU,
- $\hat{\rho}$ is an estimate of the ICC, and
- p is the number of independent variables.

The modified Cook's Distance diagnostic can be compared to a cut-off of 2 or 3 (Li et al., 2015). It was found in Valliant (2010) that different observations are identified as being influential when using OLS compared to SWLS.

4.5 Model Parameter Inference

The previous section outlined the estimation of the regression parameters under OLS, WLS and SWLS and also illustrated the difference between the variance of the estimated regression parameters under WLS and SWLS. This section now considers further inference concerning the regression parameters. Specifically point estimators, standard errors and confidence intervals are considered. Further parameter inference will be part of further research.

4.5.1 Survey-weighted Least Squares Inference

Since in this case

$$\underline{\hat{\beta}}_{SWLS} = \left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}\mathbf{y},$$

where $\mathbf{W}$ is an $n \times n$ diagonal matrix of the sampling weights associated with each observation in the complex sample, it follows immediately that

$$E\left(\underline{\hat{\beta}}_{SWLS}\right) = \underline{\beta}.$$

However, as pointed out previously, the standard errors of the estimated coefficients will be incorrect if not calculated in a way that takes the complex design of the survey into account. Conventional maximum likelihood variance estimators for SRS data cannot be used on data collected from a complex survey design and this is one of the first aspects of the inferential process that differs between OLS/WLS and SWLS. Non-parametric methods such as the Taylor series linearization method (TSL), balanced repeated replication (BRR) method, and resampling techniques such as the jackknife and bootstrap methods, need to be used when estimating the standard errors of the estimated regression parameters under SWLS (Heeringa et al., 2010).

The other aspect that differs is the degrees of freedom of the Student-t distribution that need to be adjusted such that the reduced degrees of freedom under CS standard error estimation is reflected. It is mentioned in (Heeringa et al., 2010) that it is difficult to determine the degrees of freedom for variance estimation under CS. To illustrate how the CS degrees of freedom is derived, consider the pivotal t-statistic for estimating the population mean under SRS,

$$t_{n-1,SRS} = \frac{(\bar{y} - \mu_0)}{\sqrt{\frac{s^2}{n}}},$$

where n is the sample size, $\bar{y}$ is the sample mean, μ_0 is the hypothesized population mean and s^2 is the sample variance. Substituting the calculation formula for s^2 into the t-statistic,

$$t_{n-1,SRS} = \frac{(\bar{y} - \mu_0)}{\sqrt{\frac{\frac{1}{n-1}\sum_i (y_i - \bar{y})^2}{n}}}.$$

Once the sample mean is known, only $n-1$ unique pieces of information remains for estimating the variance. Hence, the t-statistic under SRS follows a Student-t distribution with $n-1$ degrees of freedom (Heeringa et al., 2010).

When this same statistic is considered under CS, then

$$t_{df,CS} = \frac{(\bar{y}_w - \mu_0)}{\sqrt{V(\bar{y}_w)}},$$

where df refers to degrees of freedom, $\bar{y}_w$ is the sample mean calculated under CS and $V(\bar{y}_w)$ its associated variance. Suppose the CS design consists of H strata where stratum h contains N_h PSU's and within PSU j there are N_{hj} SSU's where $h = 1, ..., H$ and $j = 1, ..., N_h$. According to Heeringa et al. (2010),

$$V(\bar{y}_w) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(\frac{1}{N_h}\right) \left[\frac{1}{(N_h - 1)} \left\{\sum_{j=1}^{N_h} \frac{1}{N_{hj}^2} \left(y_{hj} - \frac{y_h}{N_h}\right)^2\right\}\right],$$

where N is the total number of PSU's. From this it is seen that each stratum contributes $(N_h - 1)$ pieces of information to the variance estimation. Thus, the t-statistic no longer follows a Student-t distribution with $(n-1)$ degrees of freedom, but rather the correct degrees of freedom under CS are

$$df_{CS} = \sum_{h=1}^{H} (N_h - 1) = N - H.$$

This is called the fixed degrees of freedom rule and is mostly used in computer software programs (Heeringa et al., 2010; Lohr, 2010).

Suppose TSL, BRR, jackknife or bootstrap has been used to estimate the variance of the estimated regression parameters. According to Lohr (2010) the degrees of freedom now become the difference between the number of PSU's sampled and the number of strata, i.e. $n - H$. Now the $100(1-\alpha)\%$ confidence interval for the jth model parameter is given by

$$\hat{\beta}_{SWLS_j} \pm t_{\frac{\alpha}{2}} \sqrt{\hat{V}\left(\hat{\beta}_{SWLS_j}\right)}, \tag{4.5.1}$$

where $t_{\frac{\alpha}{2}}$is a percentile from the Student-t distribution with $n - H$ degrees of freedom.

The confidence interval given in (4.5.1) is the standard (asymptotic) confidence interval. As alternatives to this confidence interval, section (4.5.2.1) will present discussions of some bootstrap

confidence intervals, viz.

- standard (asymptotic) confidence interval using the resampling variance estimator,
- bootstrap percentile confidence interval,
- bootstrap-t confidence interval, and
- BCa confidence interval.

4.5.2 Non-parametric Model Parameter Inference

The previous section considered model parameter inference under the assumption that the model parameter estimator meets certain distributional assumptions, i.e. parametric model parameter inference. In this section a non-parametric approach to model parameter inference is presented with specific reference to the non-parametric bootstrap resampling method.

Firstly consider a general parameter of interest, θ, and its associated estimator, $\hat{\theta}$, defined as $\hat{\theta} = \hat{\theta}(\mathbf{y})$, where $\mathbf{y}$ is an observed SRS. The bias of $\hat{\theta}$ is defined as

$$bias_F\left(\hat{\theta}\right) = E_F\left(\hat{\theta}\right) - \theta, \tag{4.5.2}$$

where the subscript F in denotes the probability distribution from which the sample, $\mathbf{y}$, was taken. The aim is to have a small bias. A plug-in estimator, such as $\hat{\theta}$, is not necessarily unbiased, but its bias tends to be small in comparison to its standard error which is one of the pleasing properties of plug-in estimators (Efron et al., 1998).

The bootstrap can be used to assess the bias of an estimator, $\hat{\theta}$, and is defined by making use of the plug-in principle and replacing F in (4.5.2) with $\hat{F}$, the empirical distribution function that places probability $1/n$ on each unit,

$$bias_{\hat{F}}\left(\hat{\theta}\right) = E_{\hat{F}}\left(\hat{\theta}\right) - \hat{\theta}.$$

The bootstrap method starts by generating B independent bootstrap samples, $\mathbf{y}_1^*, ..., \mathbf{y}_B^*$, where B is a large number. For each bootstrap sample the bootstrap replicate, $\hat{\theta}_b^* = \hat{\theta}(\mathbf{y}_b^*)$, $b = 1, .., B$, is calculated. The bootstrap approximation of $E_{\hat{F}}\left(\hat{\theta}\right)$ is given by

$$\tilde{\theta}^* = \frac{1}{B}\sum_{b_r=1}^{B}\hat{\theta}_{b_r}^*,$$

the average of the B bootstrap replicates (Efron et al., 1998). The bootstrap estimate of bias is then

$$\widehat{bias}_B\left(\hat{\theta}\right) = \tilde{\theta}^* - \hat{\theta}. \tag{4.5.3}$$

Now consider a stratified multistage cluster sample with H strata and n_h PSU's in stratum h. The bootstrap method as described for a simple random sample is applied independently in each stratum by selecting n_h PSU's with replacement and calculating the bootstrap weight, defined in equation (4.3.8). The bootstrap weights are then used to calculate the bootstrap replicates, $\left\{\hat{\theta}_b^*\right\}$. The bootstrap estimated bias of $\hat{\theta}$ is then given by

$$\widehat{bias}_B\left(\hat{\theta}\right) = \bar{\hat{\theta}}^* - \hat{\theta},$$

where

$$\bar{\hat{\theta}}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*.$$

Here the parameter of interest is β_j, the jth population regression model coefficient, estimated by $\hat{\beta}_{OLS_j}$, $\hat{\beta}_{WLS_j}$ or $\hat{\beta}_{SWLS_j}$. Recall that the bootstrapping pairs regression approach discussed in section 4.3.2. Letting $\hat{\theta} = \underline{\hat{\beta}}_j$ and following the same reasoning as above, the bootstrap estimated bias of $\hat{\beta}_j$ is calculated as

$$\widehat{bias}_B\left(\hat{\beta}_j\right) = \bar{\hat{\beta}}_j^* - \hat{\beta}_j,$$

where

$$\bar{\hat{\beta}}_j^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_{j_b}^*,$$

and $\hat{\beta}_{j_b}^*$ is the bth bootstrap replicate of $\hat{\beta}_j$ obtained from the bth bootstrap sample.

The next section will consider various bootstrap confidence intervals that can be employed to estimate the intervals of the parameter of interest, β_j.

4.5.2.1 Bootstrap Confidence Intervals for Regression Parameter Inference

In this section different bootstrap approaches to the construction of confidence intervals (CI) will be discussed. Firstly, a brief overview of the standard asymptotic interval is given followed by discussions on the percentile interval, the bootstrap-t interval as well as the bias-corrected and accelerated (BCa) interval. The section will be concluded with a short summary of the advantages and disadvantages of each technique.

Standard (asymptotic) Interval

Consider the jth unknown regression parameter, β_j, to be estimated by $\hat{\beta}_j$ and suppose that $\hat{\beta}_j$ is approximately normally distributed with expected value β and estimated variance $\hat{V}\left(\hat{\beta}_j\right)$. An approximate $100(1-\alpha)\%$ CI for β_j is then given by

$$\left[\hat{\beta}_j - t_{\frac{\alpha}{2}}\sqrt{\hat{V}\left(\hat{\beta}_j\right)}; \hat{\beta}_j + t_{\frac{\alpha}{2}}\sqrt{\hat{V}\left(\hat{\beta}_j\right)}\right], \tag{4.5.4}$$

with $t_{\frac{\alpha}{2}}$ the relevant Student-t quantile. An interval estimator, as given in (4.5.4), can be more useful than a point estimator $\hat{\beta}_j$ viewed alone. When the point estimator and the interval estimator are combined they give an indication of what the "best guess" for β may be as well as how far that "guess" may be from the actual value of the parameter of interest.

The CI in (4.5.4) holds whether β_j is estimated under OLS, WLS or SWLS although the answers will differ. Now suppose the variance of $\hat{\beta}_j$ is estimated using bootstrap resampling as discussed in section 4.3.2. Especially for CS it has been recommended to make use of, for example, the bootstrap resampling method to estimate the variance of estimators since quite a few of the estimators under CS do not have a closed-form variance formula (Lohr, 2010; Heeringa et al., 2010). This is also recommended under SWLS and thus it might be sensible to replace the model-based variance estimator of $\hat{\beta}_{SWLS_j}$ with its bootstrap estimated variance when calculating the standard CI for β_j. When this is done, the $100(1-\alpha)\%$ standard CI for β_j is given by

$$\left[\hat{\beta}_{SWLS_j} - t_{\frac{\alpha}{2}}\sqrt{\hat{V}_B\left(\hat{\beta}_{SWLS_j}\right)}; \hat{\beta}_{SWLS_j} + t_{\frac{\alpha}{2}}\sqrt{\hat{V}_B\left(\hat{\beta}_{SWLS_j}\right)}\right]. \tag{4.5.5}$$

The coverage performance of the standard CI in (4.5.4) can be compared to that of the standard CI in (4.5.5) to determine how well the bootstrap estimated variance performs compared to the model-based variance.

The Percentile Interval

The derivation of the percentile interval through the bootstrap procedure is quite simple. Generate B bootstrap samples from the estimated probability model $\hat{P}$, $\mathbf{y}_1^*, ..., \mathbf{y}_B^*$, where B is a large number. For each bootstrap sample a bootstrap replicate, $\hat{\theta}_b^*$, $b = 1, ..., B$, is calculated. Once the bootstrap replicates, $\left\{\hat{\theta}_b^*\right\}$, have been computed for each bootstrap sample, they are sorted in ascending order, $\left\{\hat{\theta}_{(b)}^*\right\}$. The $\frac{\alpha}{2}$th point of the percentile interval is the $B\cdot\frac{\alpha}{2}$th largest value of these sorted replicates. In the same way the $\left(1-\frac{\alpha}{2}\right)$th point of the percentile interval is the $B\cdot\left(1-\frac{\alpha}{2}\right)$th largest value. In cases where $B\cdot\frac{\alpha}{2}$ is not an integer let $k = \left[(B+1)\frac{\alpha}{2}\right]$, the largest integer less than or equal to $(B+1)\frac{\alpha}{2}$. Then the empirical $\frac{\alpha}{2}$ and $\left(1-\frac{\alpha}{2}\right)$ quantiles are the kth and the $(B+1-k)$th largest values of $\left\{\hat{\theta}_{(b)}^*\right\}$, respectively. The $100\,(1-\alpha)\,\%$ bootstrap percentile interval is then given by (Efron et al., 1998)

$$\left[\hat{\theta}_{lo}, \hat{\theta}_{up}\right] = \left[\hat{\theta}^*_{([B\frac{\alpha}{2}])}, \hat{\theta}^*_{([B(1-\frac{\alpha}{2})])}\right]. \tag{4.5.6}$$

The percentile interval under complex sampling would be exactly the same as outlined above.

The only difference occurs in the calculation of the bootstrap replicates for each sample where the bootstrap weights in (4.3.8), w^*_{hji}, are incorporated into the calculation.

Recall that the parameter of interest is β_j, the jth population regression model coefficient, estimated by $\hat{\beta}_{OLS_j}$, $\hat{\beta}_{WLS_j}$ or $\hat{\beta}_{SWLS_j}$. Also be reminded that the bootstrapping pairs regression approach, which has been discussed in section (4.3.2), is followed in this thesis. Following the same reasoning it can be deduced that the $100\,(1-\alpha)\,\%$ bootstrap percentile interval for β_j should be given by

$$\left[\hat{\beta}^*_{j([B\frac{\alpha}{2}])}, \hat{\beta}^*_{j([B(1-\frac{\alpha}{2})])}\right],$$

where $\hat{\beta}^*_{j(B\cdot\frac{\alpha}{2})}$ and $\hat{\beta}^*_{j([B(1-\frac{\alpha}{2})])}$ are, respectively, the $B\cdot\frac{\alpha}{2}$th and $B\cdot\left(1-\frac{\alpha}{2}\right)$th largest bootstrap replicates of $\hat{\beta}_j$, $\left\{\hat{\beta}^*_{j(b)}\right\}$, $b=1,...,B$.

The percentile interval would in general be preferable to the standard interval. The first objection to the use of the standard interval is the normal approximation that underlies it. If n is small this approximation may not be accurate. One way of improving the standard interval is through the use of an appropriate transformation and then mapping the endpoints of the interval back to the original scale. The problem with this approach is that you are required to know a different transformation, such as the log-transformation or the exponential-transformation, for each estimator, $\hat{\theta}$, of the parameter of interest, θ. The advantage of the percentile method is that it can be thought of as an algorithm that automatically incorporates these transformations and as a result it extends the effectiveness of the standard interval. In situations where the standard interval would be correct if the appropriate transformation was applied, the percentile method automatically incorporates the transformation and thus it is not necessary to know all the appropriate transformations of $\hat{\theta}$; you only need to assume they exist (Efron et al., 1998). The percentile interval does not work particularly well in general cases, but in certain cases it is better than the bootstrap-t interval that will be discussed later in section 4.5.2.1. The percentile method also works well for the estimation of quantiles (Kovar et al., 1988).

An advantage of the percentile method should be the improved coverage performance. Although it still tends to under cover, it is more balanced in both sides of the interval than the standard interval. This undercoverage occurs because of the non-parametric inference used. The percentile method has no knowledge of the underlying distribution and uses the empirical distribution instead (Efron et al., 1998).

A further advantage of this method is that it is transformation respecting. When the interval, obtained after the application of an appropriate transformation on the estimator, $\hat{\theta}$, of the parameter of interest, θ, is mapped back to the original scale, it results in the same interval as before the transformation. This is not the case with the standard interval (Efron et al., 1998). The transformation is used to improve the interval and once the endpoints of the interval are transformed back to the original $\hat{\theta}$ scale, it sometimes results in a shorter or longer interval than the interval based

on the untransformed estimator. This reflects what is meant by the transformation respecting property.

A third advantage of the percentile method is the range-preserving property. Some parameters are defined on a certain range of values, for example the correlation is defined from -1 to 1. The endpoints of the percentile interval are values of the bootstrap replicates themselves that automatically fall within the allowable range. Confidence procedures that are range-preserving tend to be more accurate and reliable (Efron et al., 1998).

The Bootstrap-t Interval

Consider a general parameter of interest, θ, estimated by $\hat{\theta}$. The bootstrap methodology makes it possible to obtain accurate intervals without making assumptions about approximate normality. The bootstrap-t method considers the "t-statistic"

$$T \equiv \frac{\hat{\theta} - \theta}{\widehat{se}},$$

and the approximate confidence interval

$$P\left(\underline{\delta} \leq T \leq \overline{\delta}\right) = 1 - \alpha,$$

where $\underline{\delta}$ and $\overline{\delta}$ represent, respectively, the lower quantile and upper quantile of the distribution of T and $\widehat{se}$ is the estimated standard error of $\hat{\theta}$. In the "ideal world" the confidence interval would be

$$P\left(\hat{\theta} - \overline{\delta} \cdot \widehat{se} \leq \theta \leq \hat{\theta} - \underline{\delta} \cdot \widehat{se}\right) = 1 - \alpha,$$

but since only a single sample is taken from the population, resampling methods need to be used to estimate the "ideal" situation. The bootstrap-t method uses resampling on the data to generate a bootstrap t-statistic

$$T^* \equiv \frac{\hat{\theta}^* - \hat{\theta}}{\widehat{se}^*}, \tag{4.5.7}$$

where $\hat{\theta}^*$ is the statistic calculated on the bootstrap sample, $\hat{\theta}$ is the statistic calculated on the original sample and $\widehat{se}^*$ is the bootstrap standard error of $\hat{\theta}^*$. The latter is calculated by resampling from the bootstrap sample, calculating the statistic for each resample and computing the standard error of those bootstrap statistics. Then the bootstrap T^* value is calculated for each bootstrap sample and ordered, from which the bootstrap quantiles are obtained (Efron et al., 1998).

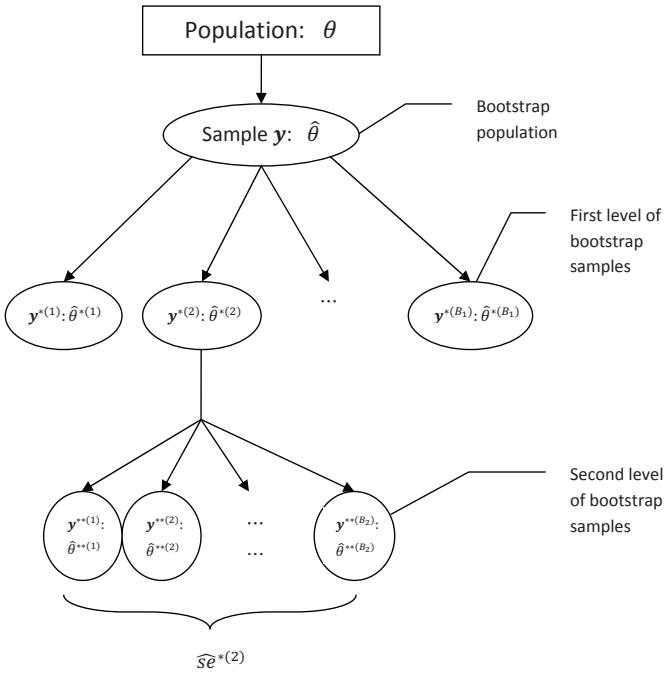


Figure 4.5.1: Bootstrap-t Confidence Interval

Consider figure 4.5.1 where $\mathbf{y}$ denotes the original sample and $\hat{\theta} = \hat{\theta}(\mathbf{y})$ is the estimator of the parameter θ. Firstly, generate B_1 bootstrap samples , $\mathbf{y}_1^*, ..., \mathbf{y}_{B_1}^*$, with replacement from the original sample and calculate the bootstrap replicate, $\hat{\theta}_b^* = \hat{\theta}(\mathbf{y}_b^*)$, $b = 1, ..., B_1$, for each bootstrap sample. B_1 is usually a fairly large number. For each bootstrap sample, calculate the bootstrap t-statistic

$$t_b^* = \frac{\hat{\theta}_b^* - \hat{\theta}}{\widehat{se}_b^*}, \tag{4.5.8}$$

where $\hat{\theta}$ is the parameter calculated on the original sample and $\widehat{se}_b^*$ is the estimated standard error of $\hat{\theta}_b^*$ for the bootstrap sample $\mathbf{y}_b^*$ (Efron et al., 1998). This estimated standard error is obtained by taking B_2 bootstrap samples from the current bootstrap sample, $\mathbf{y}_b^*$, calculating the replicates for each resample, $\left\{\hat{\theta}_b^{**}, b = 1, ..., B_2\right\}$, and then obtaining the standard error of those replicates. It is necessary here to distinguish between B_1 and B_2 to emphasize the use of a nested bootstrap. B_1 bootstrap samples are selected from the original sample and then B_2 samples are selected, with replacement, from each of the B_1 bootstrap samples . Both B_1 and B_2 are typically large, but need not be the same size. The $\frac{\alpha}{2}$th percentile of t_b^*, $b = 1, ..., B_1$, is estimated by the value $\hat{t}^{(\frac{\alpha}{2})}$ such that

$$\frac{\#\left\{t_b^* \leq \hat{t}^{(\frac{\alpha}{2})}\right\}}{B_1} = \frac{\alpha}{2}.$$

The $\{t_b^*\}$ values are sorted in ascending order and the $\frac{\alpha}{2}$th point is the $B_1 \cdot \frac{\alpha}{2}$th largest value of these $\left\{t_{(b)}^*\right\}$ values. In the same way the $\left(1-\frac{\alpha}{2}\right)$th point is the $B_1 \cdot \left(1-\frac{\alpha}{2}\right)$th largest value. In cases where $B_1 \cdot \frac{\alpha}{2}$ is not an integer the same procedure, as explained in section 4.5.2.1, is followed (Efron et al., 1998). The $100\,(1-\alpha)\,\%$ bootstrap-t interval for θ is then given by

$$\left(\hat{\theta} - \hat{t}^{(1-\frac{\alpha}{2})} \cdot \widehat{se}_{B_1}, \hat{\theta} - \hat{t}^{(\frac{\alpha}{2})} \cdot \widehat{se}_{B_1}\right), \tag{4.5.9}$$

where $\widehat{se}_{B_1}$ is the estimated standard error of $\hat{\theta}$ calculated as the standard error of $\left\{\hat{\theta}_b^*,\, b = 1, ..., B_1\right\}$. It should be noted that $B_1 = 100$ or 200 is not adequate for the construction of confidence intervals. Many more bootstrap samples are required to accurately estimate the parameter of interest, θ, according to the argument of Booth and Sarkar (1998), and then there is a second level of bootstrapping needed to estimate the standard error of each bootstrap replicate, $\hat{\theta}_b^*$.

This is a major computational difficulty with the use of the bootstrap-t interval. The standard error, $\widehat{se}_b^*$, has to be estimated for each bootstrap sample which is not a problem when the parameter of interest is the sample mean, because there exists a formula for its standard error. Unfortunately there exists very few standard error formulas which means that the standard error for other statistics will have to be estimated using resampling methods and this leads to a nested bootstrap. Thus, in a nested bootstrap where B_1 bootstrap samples are taken from the original sample and B_2 samples are taken from each of the B_1 bootstrap samples to estimate the standard error, $B_1 \cdot B_2$ bootstrap samples are required. This is a large number and hence computationally intensive. Given this difficulty with the computational demand of the bootstrap-t method, the jackknife variance estimation method has been proposed for the estimation of the bth bootstrap replicate's variance, i.e. $\hat{V}_{JK}\left(\hat{\theta}_b^*\right)$. However, previous work by Luus et al. (2012) made use of this proposition, but found the performance of the bootstrap-t intervals unsatisfactory irrespective of the large number of first-level bootstrap samples selected. Given this as well as the improvement in computing power, it has been decided to, along with the interval using the proposed second-level jackknife variance, use a second-level bootstrap to estimate the variance in question. Hence, the variance of the bth bootstrap replicate will in fact be computed as $\hat{V}_{B_2}\left(\hat{\theta}_b^*\right)$.

Now consider the application of the bootstrap-t interval in complex sampling using the jackknife variance estimator at the second level. In Rao and Wu (1987) these intervals were obtained for smooth functions, $\hat{\theta} = \hat{\theta}(\mathbf{y})$, by approximating the distribution of

$$T_{JK} = \frac{\left(\hat{\theta} - \theta\right)}{\sqrt{\hat{V}_{JK}\left(\hat{\theta}\right)}}, \tag{4.5.10}$$

through the use of the bootstrap method. Recall that $\hat{\theta}$ is the estimator of the population parameter θ and $\hat{V}_{JK}\left(\hat{\theta}\right)$ is the jackknife estimator of the variance of $\hat{\theta}$. The bootstrap counterpart is given by

$$T^*_{JK} = \frac{\left(\hat{\theta}^* - \hat{\theta}\right)}{\sqrt{\hat{V}_{JK}\left(\hat{\theta}^*\right)}}, \tag{4.5.11}$$

where $\hat{V}_{JK}\left(\hat{\theta}^*\right)$ is similar to the jackknife estimate of the variance of $\hat{\theta}$ under complex sampling in (4.3.5), but is the estimated variance of the jackknife replicates calculated from the second level of samples. The jackknife weights, defined in (4.3.3) will be recalculated on the second level and used in the calculation of the jackknife replicates. The two-sided $100\,(1-\alpha)\,\%$ bootstrap-t confidence interval under complex sampling is then given by

$$\left[\hat{\theta} - t^*_U \cdot \sqrt{\hat{V}_{B_1}\left(\hat{\theta}\right)}, \hat{\theta} - t^*_L \cdot \sqrt{\hat{V}_{B_1}\left(\hat{\theta}\right)}\right], \tag{4.5.12}$$

where t^*_L and t^*_U are the lower and upper $\frac{\alpha}{2}$-points obtained from sorted bootstrap replicates of the statistic

$$t^*_{JK_b} = \frac{\left(\hat{\theta}^*_b - \hat{\theta}\right)}{\sqrt{\hat{V}_{JK}\left(\hat{\theta}^*_b\right)}},\ b = 1, ..., B_1.$$

$\hat{V}_{B_1}\left(\hat{\theta}\right)$ is the bootstrap estimated standard error of $\hat{\theta}$, as defined in (4.3.10) and $\hat{V}_{JK}\left(\hat{\theta}^*_b\right)$ is the jackknife estimate of variance calculated from the second level of sampling (Rao et al., 1992).

In the case where the variance of the bth bootstrap replicate will be computed as $\hat{V}_{B_2}\left(\hat{\theta}^*_b\right)$, the bootstrap counterpart will be given by

$$T^*_{BS} = \frac{\left(\hat{\theta}^* - \hat{\theta}\right)}{\sqrt{\hat{V}_{B_2}\left(\hat{\theta}^*\right)}}, \tag{4.5.13}$$

where $\hat{V}_{B_2}\left(\hat{\theta}^*\right)$ is similar to the bootstrap estimate of the variance of $\hat{\theta}$ but is the estimated variance of the bootstrap replicates calculated from the second level of sampling. The bootstrap weights, defined in (4.3.8), will be recalculated on the second level and used in the calculation of the second level bootstrap replicates. Now the $100\,(1-\alpha)\,\%$ bootstrap-t confidence interval under complex sampling is given by

$$\left[\hat{\theta} - t^*_U \cdot \sqrt{\hat{V}_{B_1}\left(\hat{\theta}\right)}, \hat{\theta} - t^*_L \cdot \sqrt{\hat{V}_{B_1}\left(\hat{\theta}\right)}\right], \tag{4.5.14}$$

where t^*_L and t^*_U are the lower and upper $\frac{\alpha}{2}$-points obtained from sorted bootstrap replicates of the statistic

$$t^*_{BS_b} = \frac{\left(\hat{\theta}^*_b - \hat{\theta}\right)}{\sqrt{\hat{V}_{B_2}\left(\hat{\theta}^*_b\right)}},\ b = 1, ..., B_1.$$

and $\hat{V}_{B_1}\left(\hat{\theta}\right)$ is the bootstrap estimated standard error of $\hat{\theta}$, as defined in (4.3.10).

The lower and upper bootstrap percentiles, namely t^*_L and t^*_U, correspond to the $B_1 \cdot \frac{\alpha}{2}$th and the $B_1 \cdot \left(1 - \frac{\alpha}{2}\right)$th largest values of the sorted $\{t^*_b\}$ values. If $B_1 \cdot \frac{\alpha}{2}$ is not an integer, the same argument can be followed as given before.

Also in complex sampling, a variance stabilizing transformation can be used to correct uneven error rates, but the bootstrap provides an alternative when such transformations do not exist or are unknown (Rao et al., 1992).

Consider the application of the bootstrap-t interval to regression with the bootstrapping pairs approach. The parameter of interest is β_j, the jth population regression model coefficient, estimated by $\hat{\beta}_{OLS_j}$, $\hat{\beta}_{WLS_j}$ or $\hat{\beta}_{SWLS_j}$. If the same reasoning is applied, then the $100\,(1-\alpha)\,\%$ bootstrap-t interval for β_j, using the jackknife variance estimator at the second level, should be given by

$$\left[\hat{\beta}_j - t^*_U \cdot \sqrt{\hat{V}_{B_1}\left(\hat{\beta}_j\right)}, \hat{\beta}_j - t^*_L \cdot \sqrt{\hat{V}_{B_1}\left(\hat{\beta}_j\right)}\right],$$

where t^*_L and t^*_U are the lower and upper $\frac{\alpha}{2}$-points obtained from $t^*_{(1)}, ..., t^*_{(B_1)}$,

$$t^*_b = \frac{\left(\hat{\beta}^*_{j_b} - \hat{\beta}_j\right)}{\sqrt{\hat{V}_{JK}\left(\hat{\beta}^*_{j_b}\right)}},\ b = 1, ..., B_1,$$

$\hat{V}_{B_1}\left(\hat{\beta}_j\right)$ is the first-level bootstrap estimated variance of $\hat{\beta}_j$ and $\hat{V}_{JK}\left(\hat{\beta}^*_{j_b}\right)$ is the second-level jackknife estimated variance of $\hat{\beta}^*_{j_b}$, the bth first-level bootstrap replicate of $\hat{\beta}_j$.

As for the SRS case, and for the same reasons given before, the variance of the bth bootstrap replicate, $\hat{\beta}^*_{j_b}$, will be estimated using the jackknife as well as the bootstrap methods. When using a second-level bootstrap, the estimated variance of $\hat{\beta}^*_{j_b}$ will be given by

$$\hat{V}_{B_2}\left(\hat{\beta}^*_{j_b}\right) = \frac{1}{B_2 - 1}\sum_{b=1}^{B_2}\left(\hat{\beta}^{**}_{j_b} - \bar{\hat{\beta}}^{**}_{j_b}\right)^2,$$

where $\bar{\hat{\beta}}^{**}_{j_b} = \frac{1}{B_2}\sum_b \hat{\beta}^{**}_{j_b}$, and B_2 is the number of bootstrap samples taken at the second level. Then the $100\,(1-\alpha)\,\%$ bootstrap-t interval for β_j becomes

$$\left[\hat{\beta}_j - t^*_U \cdot \sqrt{\hat{V}_{B_1}\left(\hat{\beta}_j\right)}, \hat{\beta}_j - t^*_L \cdot \sqrt{\hat{V}_{B_1}\left(\hat{\beta}_j\right)}\right],$$

where t_L^* and t_U^* are the lower and upper $\frac{\alpha}{2}$-points obtained from $t_{(1)}^*, ..., t_{(B_1)}^*$,

$$t_b^* = \frac{\left(\hat{\beta}_{j_b}^* - \hat{\beta}_j\right)}{\sqrt{\hat{V}_{B_2}\left(\hat{\beta}_{j_b}^*\right)}},\ b = 1, ..., B_1.$$

It has been shown that the coverage of the bootstrap-t interval tends to be closer to the desired level than that of the standard interval. Unfortunately the gain in accuracy goes hand in hand with a loss in generality, since the bootstrap-t interval applies only to the given sample. The interval generated by the bootstrap-t method is not symmetric about zero. It is this asymmetry that plays an important part in the coverage improvement that is enjoyed by the bootstrap-t (Efron et al., 1998).

The bootstrap methodology provides a good measure for both smooth and non-smooth functions. It is the preferred method for one-sided intervals, but if suitable variance-stabilizing transformations can be found then other methods, such as the normal-theory one-sided interval, may be used and may perform better. As it is generally difficult to find these transformations, the bootstrap intervals will be used (Kovar et al., 1988).

Bias-Corrected and Accelerated Confidence Interval

In the case of the standard CI constructed from a general estimator $\hat{\theta}$ for some parameter of interest, θ, the assumption underlying the interval is that

$$\hat{\theta} \sim N\left(\theta, \sigma^2\right).$$

When the bootstrap estimated cumulative distribution function (cdf) is perfectly normal the percentile interval agrees with the standard method of CI construction, but when the bootstrap cdf is decidedly non-normal, the percentile CI is quite different from the standard method. The question to be asked is, which method should be used (Efron et al., 1986)?

Now suppose that some monotone transformation, $g(\cdot)$, exists such that

$$\hat{\phi} \sim N\left(\phi, \tau^2\right),$$

where $\hat{\phi} = g\left(\hat{\theta}\right)$, $\phi = g(\theta)$ and τ is a constant standard error of $\hat{\phi}$. It has been found that if the standard CI is used in this case its limits are very inaccurate whereas the percentile interval's limits, due to its transformation respecting property mentioned in section 4.5.2.1, are correct (Efron et al., 1986).

What if the estimator of θ is not unbiased as is assumed under the standard method? The bias-corrected (BC) interval was developed to adjust for this type of bias by defining a bias-correction,

$$z_0 \equiv \Phi^{-1}\left[\hat{G}\left(\hat{\theta}\right)\right],$$

where Φ^{-1} is the inverse function of the standard normal cdf and $\hat{G}(\cdot)$ is the bootstrap cdf of $\hat{\theta}$,

$$\hat{G}\left(\hat{\theta}\right) = P^*\left(\hat{\theta}^* < \hat{\theta}\right),$$

where the notation $P^*(\cdot)$ is used to indicate that the probability is computed according to the bootstrap distribution of $\hat{\theta}^*$ (Efron et al., 1986).

This bias-correction is quite important for equalizing the error probabilities at the two endpoints of the CI and if the accurate estimation of this bias-correction is possible, then the BC interval is preferred (Efron et al., 1986).

However, the BC method is not always successful. One example in Efron and Tibshirani (1986) illustrates how the BC interval is still an improvement on the standard CI, but the improvement is only approximately 50%. Thus, a further improvement was made to the BC interval in an attempt of generalize the standard interval further. Now it is assumed that for some monotone transformation, g, bias-correcting constant z_0 and acceleration constant, a, the transformation gives

$$\hat{\phi} \sim N\left(\phi - z_0\tau\sigma_\phi, (\tau\sigma_\phi)^2\right),$$

where $\sigma_\phi = 1 + a\phi$. From here it is quite easy to find the confidence interval for ϕ and then to transform the endpoints of the interval back to the original parameter's scale. This method is called the bias-corrected and accelerated (BCa) confidence interval and one of its advantages is that it automatically produces the interval on the original parameter's scale and does not require any knowledge of the transformation (Efron, 1987). Furthermore, it is also transformation respecting which implies that if the parameter of interest is transformed to some function of the parameter, the endpoints of the BCa interval transform correctly to the parameter scale (Efron et al., 1998).

At this point it is clear that, in order to construct the BCa interval, the calculation of two constants, i.e. the bias-correction and acceleration, is required. This can either be done parametrically or non-parametrically. Since the application of bootstrap methods in this thesis will follow the non-parametric application of the bootstrap, the parametric approach will not be discussed. See Efron (1987) and Efron and Tibshirani (1986; 1998) for more information on this.

Let the data be obtained from an SRS and let it consist of a single response variable, $\mathbf{y}$ and p predictors, $\mathbf{X}$. As discussed in section 4.3.2 and according to the bootstrapping pairs approach, a bootstrap sample is selected and used to calculate a bootstrap replicate of $\hat{\theta}$ denoted by $\hat{\theta}^*$. The process is repeated B times resulting in B replicates of $\hat{\theta}$, $\left\{\hat{\theta}_b^*\right\}$. The bias-correction constant is calculated directly from the proportion of bootstrap replicates that is less than $\hat{\theta}$,

$$\hat{z}_0 = \Phi^{-1}\left[\frac{\#\left\{\hat{\theta}_b^* < \hat{\theta}\right\}}{B}\right]. \tag{4.5.15}$$

The calculation of the acceleration is not as straightforward. In Efron (1987) an approximation for a is proposed which depends on

$$U_i = \lim_{\triangle \to 0} \frac{t\left[(1-\triangle)\hat{F} + \triangle\delta_i\right] - t\left(\hat{F}\right)}{\triangle},\ i = 1, ..., n,$$

the empirical influence function of $\hat{\theta} = t\left(\hat{F}\right)$ with δ_i a point mass at x_i and U_i the derivative of $\hat{\theta}$ with respect to the mass on x_i. Efron (1987) furthermore points out that Jaeckel's infinitesimal jackknife estimate of $\hat{\theta}$'s standard error is the square root of $\frac{1}{n}\sum_i U_i^2$ and shows that the acceleration can be approximated by

$$a \doteq \frac{\sum_i U_i^3}{6\left[\left(\sum_i U_i^2\right)^{3/2}\right]}.$$

Following this approximation, Efron and Tibshirani (1998) propose using the jackknife replicates of $\hat{\theta}$ to calculate the acceleration as

$$\hat{a} = \frac{\sum_i \left(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)}\right)^3}{6\left[\left(\sum_i \left(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)}\right)^2\right)^{3/2}\right]}, \tag{4.5.16}$$

where $\hat{\theta}_{(i)}$ is the jackknife replicate of $\hat{\theta}$ calculated with the ith observation removed and $\hat{\theta}_{(\cdot)}$ is the average of the jackknife replicates. This constant is called the acceleration since it continually changes the natural measurement units as one moves along the ϕ axis (Efron, 1987). This means that where the standard interval assumes, for every θ, that the standard error of $\hat{\theta}$ remains constant, the acceleration corrects for this often unrealistic assumption (Efron et al., 1998). Since it might not be clear why (4.5.16) provides an estimate of the acceleration, especially since the jackknife is used which in some opinions has been surpassed by the bootstrap, some clarity might be gained by considering Efron (1987).

The BCa endpoints are also based on the percentiles of the bootstrap distribution, just like the percentile interval, but the percentiles are adjusted to correct for the shortcomings of the percentile interval through the use of the two constants, (4.5.15) and (4.5.16). Specifically, the lower endpoint is based on the probability α_1,

$$\alpha_1 = \Phi\left[\hat{z}_0 + \frac{\hat{z}_0 + z_{\frac{\alpha}{2}}}{1 - \hat{a}\left(\hat{z}_0 + z_{\frac{\alpha}{2}}\right)}\right], \tag{4.5.17}$$

where $z_{\frac{\alpha}{2}}$ is the usual standard normal percentile, and the upper endpoint on α_2 (Efron et al., 1998),

$$\alpha_2 = \Phi\left[\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\frac{\alpha}{2}}}{1 - \hat{a}\left(\hat{z}_0 + z_{1-\frac{\alpha}{2}}\right)}\right]. \tag{4.5.18}$$

The procedure discussed up to this point is appropriate for OLS and WLS regression parameters, but for SWLS it is important that the jackknife and bootstrap methods be carried out as explained in section 4.3.1 and section 4.3.2.

Now, let the parameter of interest be β_j, the jth population regression model coefficient, to be estimated by $\hat{\beta}_{OLS_j}$, $\hat{\beta}_{WLS_j}$ or $\hat{\beta}_{SWLS_j}$. The $100\,(1-\alpha)\,\%$ BCa interval of β_j is given by

$$\left[\hat{\beta}^*_{j_{\left(B\cdot\frac{\alpha_1}{2}\right)}}, \hat{\beta}^*_{j_{\left(B\cdot\frac{\alpha_2}{2}\right)}}\right], \tag{4.5.19}$$

where $\hat{\beta}^*_{j_{\left(B\cdot\frac{\alpha_1}{2}\right)}}$ and $\hat{\beta}^*_{j_{\left(B\cdot\frac{\alpha_2}{2}\right)}}$ are, respectively, the $B\cdot\frac{\alpha_1}{2}$th and $B\cdot\frac{\alpha_2}{2}$th largest bootstrap replicates of $\hat{\beta}_j$, $\left\{\hat{\beta}^*_{j(b)}\right\}$, $b = 1, ..., B$.

A second advantage of the BCa interval is that it has been shown to be second-order accurate, i.e. its error in matching the true interval of θ approaches zero at a rate of $1/n$. A discussion about this is presented in Efron (1987). This is an improvement from the standard and percentile intervals that are only first-order accurate (Efron et al., 1998).

The objective of this chapter was to discuss OLS, WLS and SWLS linear modeling and to compare the three linear models at the different stages of regression analysis. OLS and WLS assume that the data comes from an SRS and are independent and identically distributed (iid) while SWLS is applied to CS data where the sample units are not iid. It has been observed in practice that, although there is an important difference between SRS and CS data, some practitioners would naïvely use OLS (or WLS) when the data has been collected according to a complex design. Thus, one of the important results presented in this chapter was the illustration that, although the estimated regression coefficients under WLS and SWLS agree when using the sampling weights, the WLS standard errors are wrong and thus any inference based on the WLS output will be affected.

An important stage in regression analysis is the evaluation of the fitted model. This section commenced with a short discussion of the coefficient of multiple determination which was followed by the estimation of the model's prediction error. The methods considered for this purpose, were the LOOCV, the bootstrap estimator of PE and the .632 bootstrap estimator of PE. All three methods are well-known under SRS data, but needed to be developed for application to CS data, which is considered one of the contributions of this chapter. The model evaluation part was concluded by a discussion of different outlier detection diagnostics and their extension from OLS to SWLS.

After the evaluation of the fitted model it was natural to continue with the model parameter inference and now the estimation of the variances of the regression estimators came into question. OLS and WLS estimators have closed-form variance formulas, but as with many parameter estimators under CS, the definition of an SWLS closed-form variance formula is complex and this shifted the attention to non-parametric variance estimation. Here, the jackknife and bootstrap methods were discussed with specific focus on their application to CS data. This was followed by a discussion of the standard confidence interval and various bootstrap confidence intervals.

The theory discussed in this chapter, viz.

- regression parameter estimation;
- variance estimation;
- model evaluation; and
- regression parameter inference,

will be applied in the simulation study of this thesis and the results will be compared for OLS, WLS and SWLS. The simulation study will be discussed in subsequent chapters.

Chapter 5

Simulation Model

In order to evaluate inferential practice under "controllable" conditions a need exists to simulate data that meet the requirements of inferential practices. The simulation of independent and identically distributed (SRS) data is quite common, but part of the inferential practices to evaluate requires stratified multistage cluster sampling (CS) data and the simulation thereof is not as common. A literature review was pursued to determine whether this was possible. Work by Asparouhov et al. (2005) suggested the possibility of being able to do this and Pfeffermann et al. (1998) and Rabe-Hasketh and Skrondal (2006) discussed the use of multilevel modeling to this avail. Multilevel, or hierarchical, data refers to data where the units are grouped at different levels. For example, learners that are grouped into schools in which case learners will be the level 1 units while schools will be the level 2 units. When these groupings occur, whether random or not, the groups become differentiated since members influence and become influenced by other group members. Overlooking this relationship may lead to missing the group effect when analyzing such data and even reaching conclusions that are completely wrong (Goldstein, 2003).

This is similar to the structure that exists in CS data whereby the units of a target population are grouped into subgroups at different levels. To use the same example, a target population can be grouped into schools (PSU's) with learners (SSU's) nested within the schools. This hierarchical structure implies interdependence among sampling units which cannot be ignored when simulating such data. In CS data it is essential to recognize the importance of the clustering in these designs. While statistical inference has been adapted to take account of this, it is the opinion of Goldstein (2003) that the population structure mirrored by the CS design is considered a nuisance factor and as such does not receive the regard it should. Multilevel modeling, by contrast, regards the structure as a potential interest in itself (Goldstein, 2003).

The objective here is not to apply the multilevel models to CS data as an alternative to the linear models discussed in 4, but to explore whether multilevel models can be used to simulate a hierarchical population from which a CS can be selected.

5.1 The Two-level Model

Consider a simple model,

$$y_i = \alpha + \beta x_i + \varepsilon_i,$$

where α, the intercept, β, the slope and ε_i, the random error follow the standard assumptions discussed in chapter 4. This model does not take any hierarchy within the data into account. For example, let x_i denote the test score of a learner at eight years old and y_i the score at eleven years old. The above model is set up to relate the eleven year score to the eight year score without taking account of the school to which each learner belongs. This is called a single-level relationship (Goldstein, 2003).

However, if one wishes to describe the relationship for several schools at the same time, then the model for school j is given by

$$y_{ji} = \alpha_j + \beta_j x_{ji} + \varepsilon_{ji},$$

where the subscript j refers to the level 2 unit, namely the school, and the subscript i to the level 1 unit, namely the learner. Currently this model is still a single-level model since it describes a separate relationship for each of the level 2 units (Goldstein, 2003).

For the above model to become a two-level model, let the parameters α_j and β_j be random variables. In particular, let $\alpha_j = \beta_{0_j}$ and let $\beta_j = \beta_{1_j}$ where

$$\beta_{0_j} = \beta_0 + u_{0_j},$$

and

$$\beta_{1_j} = \beta_1 + u_{1_j},$$

with $E\left(u_{0_j}\right) = E\left(u_{1_j}\right) = 0$, $V\left(u_{0_j}\right) = \sigma^2_{u_0}$, $V\left(u_{1_j}\right) = \sigma^2_{u_1}$ and $Cov\left(u_{0_j}, u_{1_j}\right) = \sigma_{u_{01}}$. Now the model becomes

$$\begin{aligned} y_{ji} &= \beta_{0_j} + \beta_{1_j} x_{ji} + \varepsilon_{ji} \\ &= \left(\beta_0 + u_{0_j}\right) + \left(\beta_1 + u_{1_j}\right) x_{ji} + \varepsilon_{ji} \\ &= \beta_0 + \beta_1 x_{ji} + \left(u_{0_j} + u_{1_j} x_{ji} + \varepsilon_{0_{ji}}\right), \end{aligned} \tag{5.1.1}$$

where $V\left(\varepsilon_{0_{ji}}\right) = \sigma^2_{\varepsilon_0}$. Note that the additional suffix in the level one residual, $\varepsilon_{0_{ji}}$, will become necessary at a later stage. The model is thus made up of a fixed part,

$$\beta_0 + \beta_1 x_{ji},$$

which can be written in matrix notation as

$$E(\mathbf{Y}) = \mathbf{X}\underline{\beta},$$

where $\mathbf{Y} = \{y_{ji}\}$ and $E(y_{ji}) = X_{ji}\underline{\beta} = \left(\mathbf{X}\underline{\beta}\right)_{ji}$ where $\mathbf{X} = \{X_{ji}\}$, and a random part, $u_{0_j} + u_{1_j}x_{ji} + \varepsilon_{0_{ji}}$. Note that $\{\}$ denotes a matrix, $\mathbf{X}$ denotes the design matrix of explanatory variables, X_{ji} is the ji-th row of $\mathbf{X}$ and for (5.1.1), $\mathbf{X} = \left\{ \begin{array}{cc} 1 & x_{ji} \end{array} \right\}$. The variables in the random part of (5.1.1) are called residuals and it is this presence of more than one residual term that distinguishes this model from the standard linear models (Goldstein, 2003).

5.2 The Two-level Model in Sample Surveys

Consider once again the CS design described before, namely a stratified two-stage cluster sample design whereby the population is grouped into H strata and the units in each stratum have been grouped into PSU's. A sample of PSU's is selected from each stratum and then the units in each selected PSU is grouped into SSU's. A sample of SSU's is taken from each of the sampled PSU's and these then form the USU's of the sample. Recall that the USU's are finally assigned a sampling weight defined as the inverse of the inclusion probability of each unit. Refer to section 2.6 for a discussion of how the sampling weights are developed. This sampling design is illustrated in the figure below.

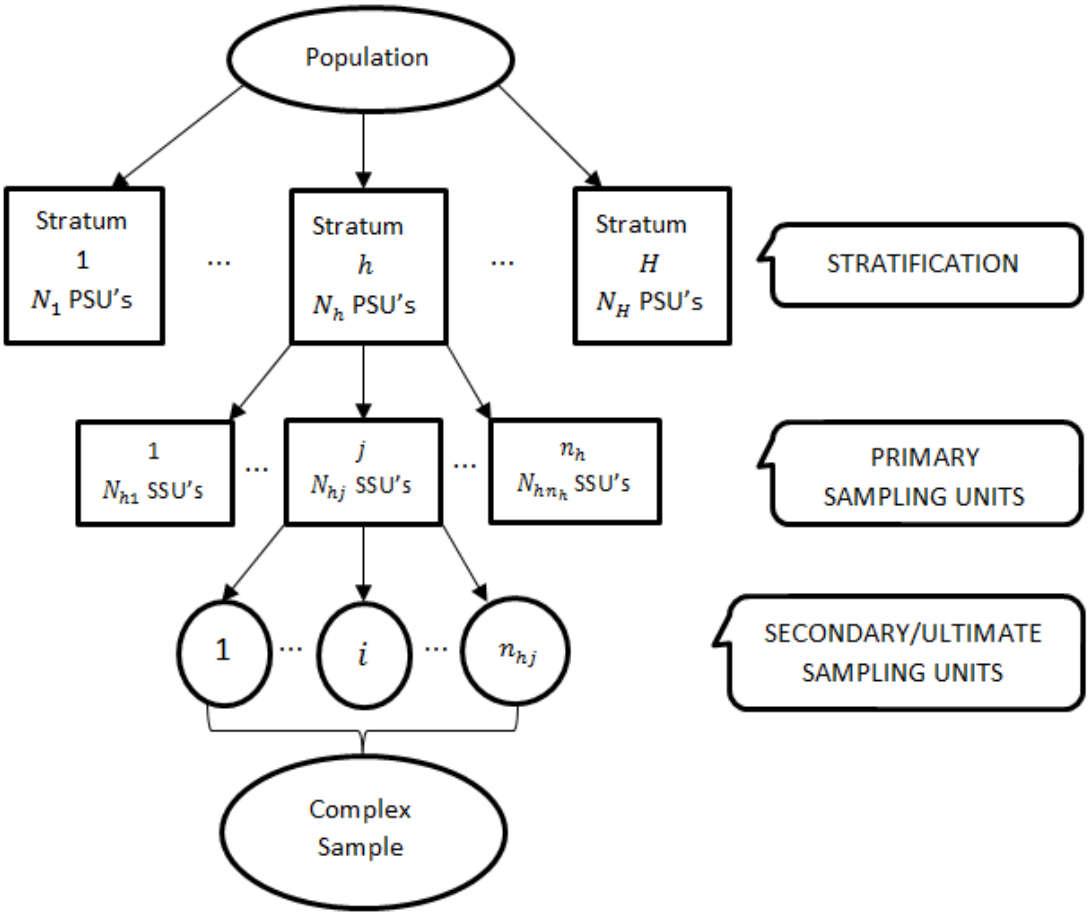


Figure 5.2.1: General Stratified Two-stage Cluster Sample Design

Now, relate this CS design to the hierarchical structure described in the previous section. by, firstly, considering each stratum as an independent population. This is a reasonable assumption to make since strata are defined as non-overlapping subgroups of the population as a whole. Consider stratum h, for which the units have been grouped into N_h PSU's and from which a sample of size n_h has been selected. The PSU level in a CS design relates to the level two units in the multilevel model. A further sampling level is defined by grouping the units in each PSU into SSU's. For the jth sampled PSU in stratum h, that contains N_{hj} SSU's, n_{hj} SSU's are selected. The SSU level in this design relates to the level one units defined in the multilevel model. Since no further sampling will take place beyond the SSU level, these units will be considered the ultimate sampling units, or USU's.

The simulation of this type of data is quite uncommon and thus limited examples were available to study. However, the few that were available tended to define sampling schemes in such a way that they delivered informative samples. Kim and Skinner (2013) defines informative sampling simply as a sampling scheme related to the response variable of a regression analysis, conditional on the independent variables. For a discussion on informative and non-informative sampling, the reader is referred to Kim and Skinner (2013).

Now consider two examples of using multilevel models to simulate CS data:

- Pfeffermann et al. (1998) defined the model,

$$y_{ij} = \beta + u_j + \nu_{ij},\ j = 1, ..., N,\ i = 1, ..., N_j,$$

where N is the number of level 2 (or PSU's) and N_j the number of level 1 (or SSU's) units, respectively, to simulate, $u_j \sim N(0, \omega^2)$, the level 2 random effect, and $\nu_{ij} \sim N(0, \sigma^2)$, the level 1 random effect. The authors chose $\beta = 1$, $\omega^2 = 0.2$, $\sigma^2 = 0.5$ and $M = 300$. The sizes of the level 2 units were calculated as

$$N_j = 75 \exp(\tilde{u}_j),$$

where $\tilde{u}_j$ was generated from $N(0, \omega^2)$ and then limited to lie within the interval $1.5\omega \leq \tilde{u}_j \leq 1.5\omega$. Thus, the size range of the level 2 units is from 38 to 147 when considering the parameter values specified (Pfeffermann et al., 1998). Next the authors defined 3 different sampling schemes for sampling from the simulated population:

1. Sample n level 2 units, PPS, with the measure of size (MOS), X_j, simulated using the level 2 random effect, $X_j = 75 \exp(u_j)$. From this it follows that the selection probability of the jth level 2 unit is calculated as $\pi_j = \frac{n \cdot X_j}{\sum_{j=1}^{N} X_j}$. Next, the level 1 units in each sampled level 2 unit were partitioned according to their associated random effects, hence forming 2 strata. Specifically, level 1 units with $\nu_{ij} > 0$ were assigned to a first stratum and the other level 1 units to the second stratum. SRS was used to sample level 1 units from each stratum, $0.25 \cdot n_j$ from stratum 1 and $0.75 \cdot n_j$ from stratum 2, and n_j was either a fixed quantity or proportional to N_j, the number of population level 1 units in the jth sampled level 2 unit (Pfeffermann et al., 1998). Suppose n_j is chosen as a fixed quantity, say n. It follows that the selection probability of the ith level 1 unit given the jth level 2 unit was selected, is $\pi_{i|j} = \frac{n}{N_j}$. This sampling scheme is used to ensure informative sampling at both levels, i.e. the inclusion probability of the ultimate sampling unit, conditional on the covariates, is related to the outcome of interest (Kim et al., 2013). The inclusion probability of a level 1 unit in this case is given by $\pi_{ij} = \left(\frac{n \cdot X_j}{\sum_{j=1}^{N} X_j}\right)\left(\frac{n_j}{N_j}\right)$.
2. The second sampling scheme defined by Pfeffermann et al. (1998) ensures that the sampling is informative only at level 2. It is the same as the sampling scheme defined in (1), but with SRS employed to sample the level 1 units within each sampled level 2 unit.
3. The final sampling scheme leads to a non-informative sample, i.e. completely random at both levels. This is similar to the sampling scheme described in (2), but here the

MOS, X_j, is set equal to the population size of the jth level 2 unit, i.e. $X_j = N_j$ (Pfeffermann et al., 1998). Now the inclusion probability of a level 1 unit is given by $\pi_{ij} = \left(\frac{n \cdot N_j}{\sum_{j=1}^{N} N_j}\right)\left(\frac{n_j}{N_j}\right)$, which simplifies to $\pi_{ij} = \frac{n \cdot n_j}{\sum_j N_j}$.

- Asparouhov et al. (2005) defined the model,

$$y_{ij} = \mu_j + \lambda_j \eta_i + \varepsilon_{ij},\ i = 1, ..., n,\ j = 1, ..., 5,$$

to simulate a population from which a CS, similar to the one described in figure 5.2.1, can be selected with the PSU's being selected, without replacement, with equal probabilities and the SSU's with or without replacement, also with equal probabilities. The data are to be used for factor analysis. In the model, μ_j is the intercept parameter, λ_j is the loading parameter, $\eta_i \sim N(0, \psi)$ is the factor variable, and $\varepsilon_{ij} \sim N(0, \theta_j)$ is the residual variable (Asparouhov et al., 2005). Hence, the parameters used for this model, are

$$\Theta = (\mu_1, ..., \mu_5, \lambda_1, ..., \lambda_5, \theta_1, ..., \theta_5, \psi).$$

The authors proceeded by generating a population of size 50000 with 5 outcomes, each distributed normally with mean and variance given by the model, using the parameter values specified in Θ. After doing so the authors grouped the simulated population observations in such a way as to resemble a two-level structure. The observations were, firstly, grouped into 140 PSU's by ordering the observations according to some function, f, which the authors chose as

$$f_i = \sum_j y_{ij},$$

to ensure informative sampling. The observations were then ranked according to their respective f-scores and then assigned to the PSU's. Of the 140 PSU's, the first 120 received 250 observations each and the remaining 20 received 1000 each. Finally, the two-stage sample was selected as described above (Asparouhov et al., 2005).

Note that in further discussion the level 2 units will be called PSU's and the level 1 units SSU's to remain in line with the terminology used in figure 5.2.1.

5.3 The Simulation of Complex Sampling Data

Multilevel modeling presents the possibility to simulate a population with a hierarchical structure from which a CS can be selected. The general two-level model was discussed in section 5.1 and

this section will describe how a population will be simulated from which a CS will be selected that will be used in the analyses of this thesis.

The ideal is to keep the simulated population as close to reality as possible, but still controlling the parameters and distributions from which the data is simulated. Firstly a real-world survey that acts as a surrogate population, whose characteristics can be borrowed and adjusted for the simulated population, needs to be identified. The South African Income and Expenditure Survey (IES) of 2005 was identified for this purpose, mostly due to familiarity of the author with it. The survey will be used in two ways:

1. to identify variables to mimic in the simulated population; and
2. to obtain descriptive measures and graphical displays as an indication of the distributions and parameters required for the simulation model.

The original IES has been reduced to meet the requirements of building a model which predicts personal income based on a selection of independent variables. Initially all persons at least 20 years old were considered irrespective of whether a positive income was captured for this person or not. Since then it has been decided to limit the age range to all persons at least 21 years of age and not older 65 years for which a positive income was captured. The IES data will be discussed in a later chapter.

Since the ideal is to simulate a population that is representative of real-world data, one continuous, three categorical variables and a dependent variable have been identified, from the IES, whose characteristics will be portrayed in the simulated population. Furthermore, it was decided to simulate two populations, one based on two strata from the IES that differ in terms of the characteristics of the dependent variable and the other based on strata that are more alike. The IES is stratified by the provinces of South Africa and the following provinces have been identified for each population:

1. Western Cape (WC) versus Eastern Cape (EC); and
2. Eastern Cape (EC) versus Kwa-Zulu Natal (KZN) .

Let the first population be denoted by WCEC and the second population by ECKZN. The next section will explain the simulation of the first population, WCEC, followed by the section that describes the simulation of the second population, ECKZN.

5.3.1 WCEC Simulation Process

5.3.1.1 Identifying the Variable Characteristics

This section provides some graphs and descriptive measures that aid in the visualization of the characteristics of the WC and EC variables that have been identified for use in the simulation study.

The continuous variable identified, is age. Consider the WC and EC probability density functions of the continuous variable.

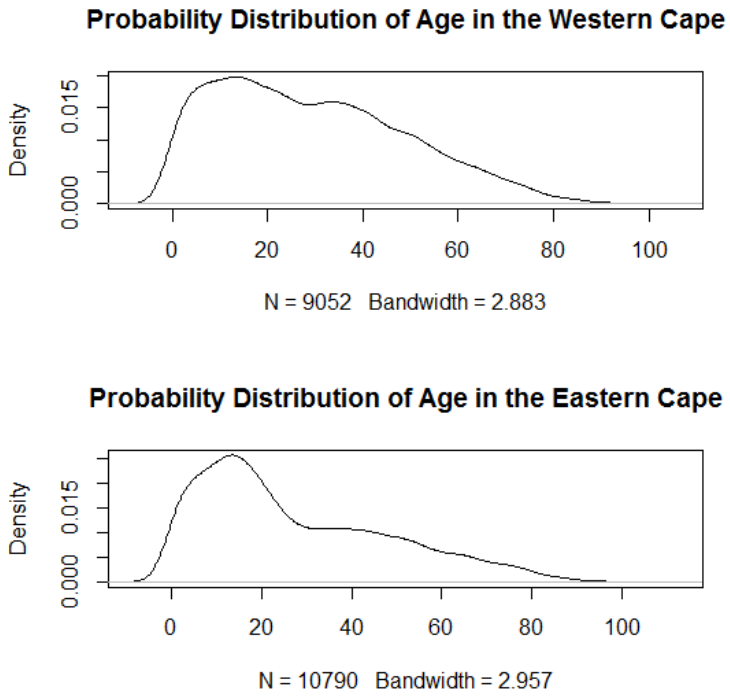


Figure 5.3.1: Age Probability Density Functions

The shapes of the distributions in both the WC and the EC, are positively skewed. Now consider summary statistics of this variable in each of the provinces to further assess the shape of each distribution. The results are given in the table below and have been measured in years.

	WC	EC
Min	0	0
Mean	29.39	27.62
Max	98	104
Std. Dev.	19.81	21.05
Skew	0.52	0.78
Kurt	2.501	2.705

Table 5.3.1: Summary Statistics of Age (in years)

The minimum age in both provinces, is zero while the respective maximum ages differ as well as the mean ages and the respective standard deviations. The slightly larger age standard deviation in the EC agrees with the larger age range in the province. The skewness statistics are both positive, indicative of positively skewed distributions, but the EC skewness is further from zero than the WC skewness. This is also seen in figure 5.3.1. The EC kurtosis statistic is larger than the WC's

equivalent value, also indicative of the EC distribution being slightly more positively skewed than that of the WC.

Finally, consider the QQ plots of each of the age distributions.

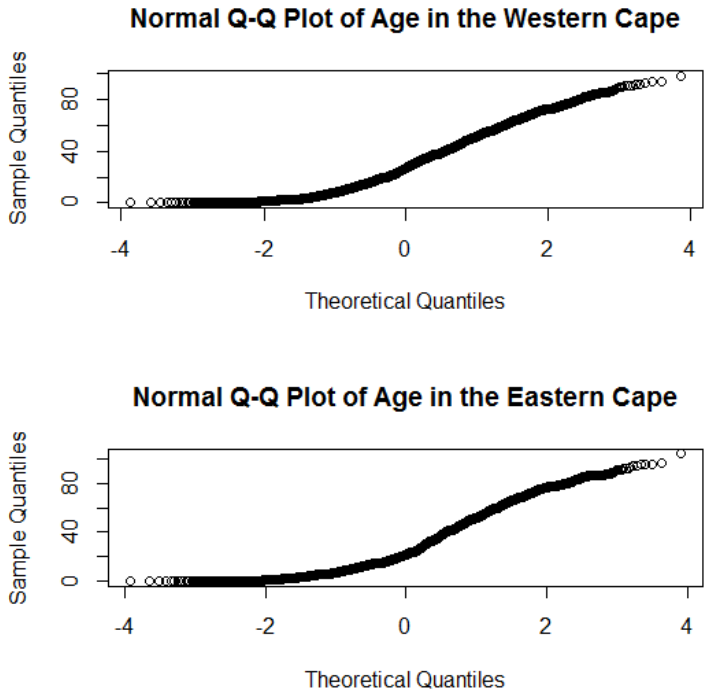


Figure 5.3.2: Normal Q-Q Plots of Age

The Q-Q plots agree with the density functions in figure 5.3.1, namely that the distributions are clearly non-normal. This is also supported by the positive skewness statistics as well as the respective kurtosis statistics.

Let the probability distribution from which the continuous variable, X_1, will be simulated, be denoted by F_1. Two continuous distributions considered for this purpose are:

1. the gamma distribution

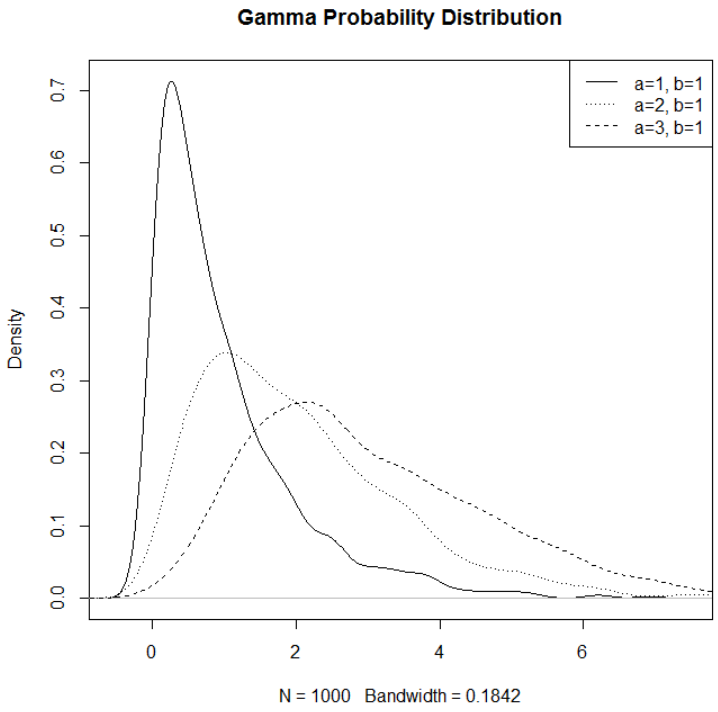


Figure 5.3.3: Examples of the Gamma Distribution at different parameter values

From the age density functions in figure 5.3.1 it was clear that the continuous variable has a positively skewed distribution. The gamma distribution was considered due to its positively skewed shape, as illustrated in the above figure. The distribution requires two parameters, namely a shape and a scale parameter. Let the parameters be defined as the vector θ_1,

$$\theta_1 = \begin{pmatrix} \alpha & \beta \end{pmatrix},$$

where α represents the shape parameter and β the scale parameter. The distribution was plotted for different choices of α and β before a decision was made.

2. the Weibull distribution

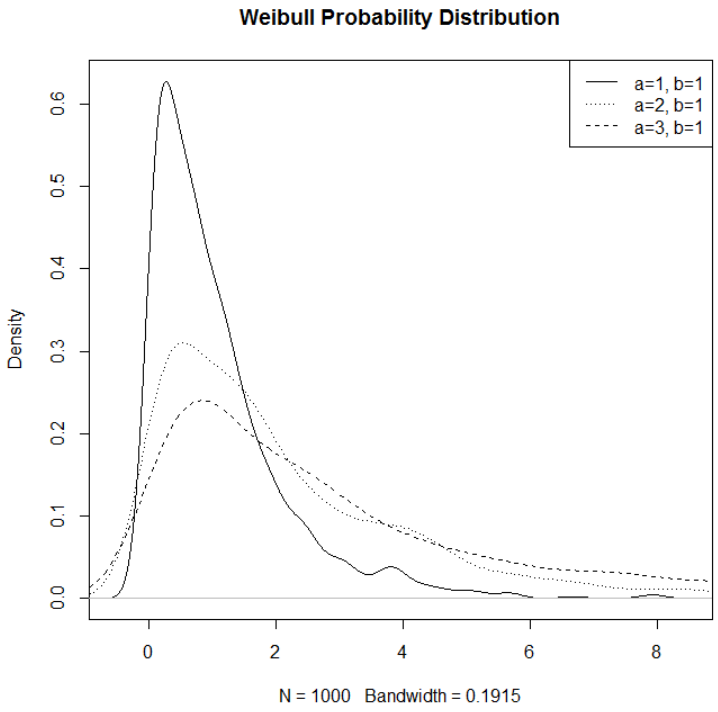


Figure 5.3.4: Examples of the Weibull Distribution at different parameter values

This distribution was considered as an alternative to the gamma distribution since it also has a positively skewed shape.

Finally, after plotting both distributions for difference parameter values, the decision was made to use the gamma distribution since it better suited the shape of the age distribution in this case.

Real-world surveys usually contain many categorical variables and thus it was decided to include 3 categorical variables, namely a two-, three-, and four-level categorical variable. The most common two-level categorical variable seems to be gender and this variable in the IES was identified to mimic in the simulation. The proportion of observations in each category, by WC/EC, are given in the table below.

GENDER		
	MALE	FEMALE
WC	50.99%	49.01%
EC	46.34%	53.66%

Table 5.3.2: Gender Relative Frequency

Gender proportions in the WC are approximately the same while in the EC there is a slightly higher proportion of females.

Let the probability distribution from which the two-level categorical variable, X_2, will be simulated, be denoted by F_2. The distribution identified for this purpose, is the Bernoulli distribution

which requires only one parameter, namely the probability of observing a successful outcome. Let this parameter be denoted by θ_2.

The four-level categorical variable chosen to mimic, is race, and the category proportions by WC/EC are presented in the table below.

RACE				
	BLACK	COLOURED	INDIAN	WHITE
WC	18.74%	64.80%	0.22%	16.25%
EC	76.44%	14.15%	0.25%	9.16%

Table 5.3.3: Race Relative Frequency

The WC is made up largely of coloured people and of approximately equal proportions of blacks and whites. The population in the EC contains mainly black people.

Since these variables, namely gender and race, would portray a hierarchical relationship in the IES it was decided to simulate these variables in a similar way. The table below presents the proportions of male/female within each race group in each province. Note that black = 1, coloured = 2, indian = 3, and white = 4.

WESTERN CAPE				
	RACE			
GENDER	1	2	3	4
MALE	10.78%	31.52%	0.14%	8.56%
FEMALE	7.96%	33.28%	0.08%	7.69%
EASTERN CAPE				
	RACE			
GENDER	1	2	3	4
MALE	33.54%	7.11%	0.22%	5.46%
FEMALE	42.89%	7.04%	0.03%	3.70%

Table 5.3.4: Gender by Race Relative Frequency

In both provinces there are very few people in the cross-section of gender and race 3. Thus, when deciding on the parameters to use for the distribution from which this relationship will be simulated, some adjustments will have to be made to ensure that the achieved population contains approriate numbers in each category.

Let the four-level categorical variable be denoted by X_3 and the distribution function from which it will be simulated, by F_3. The appropriate probability distribution from which to simulate this intended hierarchical relationship between X_2 and X_3, would be the multinomial distribution such that

- when $X_2 = 0$, then $\theta_3 = \begin{pmatrix} p_1^0 & p_2^0 & p_3^0 & p_4^0 \end{pmatrix}$, and

- when $X_2 = 1$, then $\theta_4 = \begin{pmatrix} p_1^1 & p_2^1 & p_3^1 & p_4^1 \end{pmatrix}$,

where θ_3 is the vector of category proportions to use for F_3 when $X_2 = 0$ and θ_4 the vector of proportions to use when $X_2 = 1$. To determine which of the four categories of X_3 should be assigned to an X_2 outcome, simulate a uniform number, U, between 0 and 1. If

- $0 < U \leq p_1$, then $X_{3|2}$ receives a I;
- $p_1 < U \leq p_1 + p_2$, then $X_{3|2}$ receives a II;
- $p_1 + p_2 < U \leq p_1 + p_2 + p_3$, then $X_{3|2}$ receives a III; and
- $p_1 + p_2 + p_3 < U \leq 1$, then $X_{3|2}$ receives a IV.

The final proportion choices will be presented under the simulation parameter information section.

The education level variable in the IES will be mimicked by the three-level categorical variable, but first the education levels had to be grouped into three categories for this purpose. The groupings of the 26 education levels are as follows:

1. no education;
2. primary (grades 1 - 7) and high school (grades 8 - 12) education; and
3. post grade 12 education.

The characteristics of the regrouped variable that will be used for the three-level categorical variable, are presented in the table below.

	EDUCATION LEVELS		
	1	2	3
WC	7%	80%	13%
EC	10%	77%	12%

Table 5.3.5: Level of Education Relative Frequency

The EC has a slightly higher proportion of people with no education and slightly fewer people that have completed school in comparison to the WC.

Let the three-level categorical variable be denoted by X_4 and let the probability distribution function from which it will be simulated, be denoted by F_4. Similar to the above situation, it would be appropriate to let F_4 be the multinomial probability distribution and then to let $\theta_5 = \begin{pmatrix} p_1 & p_2 & p_3 \end{pmatrix}$ be the vector of category proportions to be used for the distribution function. However, the same approach will be followed as the one described above.

5.3.1.2 Defining the Random Effects

In chapter 2 it was explained how strata are constructed such that the strata are heterogeneous within while being homogeneous between strata. Keeping this in mind, consider the within-stratum variation of the income variable in the IES, the variable on which the response in the simulation will be modeled. The standard deviation is also included.

	Variance	Std Dev
WC	13138231547.66	114622.125
EC	2785775180.60	52780.44316

Table 5.3.6: Within Stratum Variation of Income

It is clear that the income variation in the WC is between four and five times the variation in the EC according to the data captured in the IES. This implies that although the EC, on average, is considered a poorer province than the WC, the WC seems to portray a greater inequality in its income distribution.

Now, let the random effect of the hth stratum be denoted by e_h. By the error assumptions of the linear model it is assumed that the errors are normally distributed with a mean of zero and constant variance. In line with this assumption the stratum effect will be generated from a normal distribution with zero mean and constant variance,

$$e_h \sim N\left(0, \sigma_h^2\right),\ h = 1, ..., H$$

where σ_h^2 is the within stratum variation specified for stratum h such that it mimics the results presented in the above table. The stratum effect of the hth stratum will remain constant within that stratum.

Also explained in chapter 2 was that, theoretically, clusters are formed such that the units within a cluster are homogeneous while the clusters are heterogeneous between each other. The rate of homogeneity (roh, ρ) is a quantity used to examine the variability within PSU's (Lohr, 2010). It is defined as

$$\rho = \frac{deff - 1}{N_{hj} - 1},$$

where $deff$ denotes the design effect and N_{hj} is the population size of the jth cluster of the hth stratum. The design effect is defined as

$$deff = \frac{V_{CS}\left(\hat{\theta}\right)}{V_{SRS}\left(\hat{\theta}\right)},$$

where $V_{CS}\left(\hat{\theta}\right)$ denotes the variance of an estimator, $\hat{\theta}$, under CS relative to the variance of the estimator under SRS. The design effect measures the precision gained or lost by using a CS

design instead of an SRS. As explained before, stratification generally improves precision while clustering decreases it and hence $deff$ can be less than or greater than 1 depending on whether more precision is gained through stratification than lost through clustering. This implies that ρ will also be less than or greater than 1. When,

- $-1 < \rho < 0$, then the clusters are more heterogeneous than the strata,
- $\rho = 0$, then the variation in the clusters is the same as in the strata or population,
- $0 < \rho < 1$, then the clusters are more homogeneous than the strata or population, and
- $\rho = 1$, then there is complete homogeneity in the clusters.

Consider the jth cluster in the hth stratum and let this cluster contain N_{hj} population observations. Let the (hj)th within-cluster variation of the response, y, be defined as

$$\sigma^2_{hj} = \sum_{i=1}^{N_{hj}} \frac{\left(y_{hji} - \bar{y}_{U_{hj}}\right)^2}{N_{hj} - 1},$$

where $\bar{y}_{U_{hj}} = \frac{1}{N_{hj}} \sum_i y_{hji}$ (Lohr, 2010). The total within-cluster variation, $\sigma^2_{h_{within}}$, of the hth stratum is then obtained by simply adding the N_h variations together,

$$\sigma^2_{h_{within}} = \sum_j \sigma^2_{hj},\ j = 1, ..., N_h.$$

An alternative definition of ρ, as seen in Killip et al. (2004) and obtained through personal communication with Steve Heeringa, is

$$\rho = \frac{\sigma^2_{h_{between}}}{\sigma^2_{h_{between}} + \sigma^2_{h_{within}}},$$

where $\sigma^2_{h_{between}}$ is the total between-cluster variation of the hth stratum. Solving for $\sigma^2_{h_{between}}$ it follows that

$$\sigma^2_{h_{between}} = \frac{\rho \sigma^2_{h_{within}}}{1 - \rho}.$$

It has been determined from previous surveys that ρ for the IES should be approximately 0.05. The total within-cluster WC and EC variation was calculated and, letting $\rho = 0.05$, the total between-cluster variation was calculated and both are presented in the below table.

$\rho = 0.05$		
	Within Variation	Between Variation
WC	489540679310.03	25765298911
EC	124425645237.38	6548718170

Table 5.3.7: Within- and Between-cluster Variation

From table 5.3.7 it is seen that the WC between-cluster variation is approximately four times that of the EC variation.

Finally, let the cluster effect of the hth stratum be denoted by e_{N_h}. The standard error distributional assumption will be applied here too and thus,

$$e_{N_h} \sim N\left(0, \sigma^2_{h_{between}}\right),$$

where $\sigma^2_{h_{between}}$ is the betweem-cluster variation specified for stratum h such that it mimics the results presented in the above table. A between-cluster effect will be generated for each cluster j in stratum h, $j = 1, ..., N_h$ and $h = 1, ..., H$, and this value will remain constant for that cluster.

Recall the within-cluster variation presented in table 5.3.7. It is seen that the within-cluster variation in the WC is once again approximately four times that of the EC. Let the SSU effect be denoted by $e_{N_{hj}}$, a value that represents the variation between the observations (SSU's) in a PSU. Following the assumption that the errors in a linear model are normally distributed, the SSU effect within the jth PSU of the hth stratum will be generated as

$$e_{N_{hj}} \sim N\left(0, \sigma^2_{h_{within}}\right),$$

where $\sigma^2_{h_{within}}$ will be chosen to reflect the within-cluster variation observed from the IES. For each observation simulated in a PSU, a new $e_{N_{hj}}$ will be generated.

To summarize the random effects,

1. stratum effect, $e_h \sim N\left(0, \sigma^2_h\right)$, $h = 1, ..., H$, remains constant for all units within the stratum,
2. cluster/PSU effect, $e_{N_h} \sim N\left(0, \sigma^2_{h_{between}}\right)$, $\forall\, j \epsilon h$ and $h = 1, ..., H$, remains constant for all units within the PSU, and
3. SSU/USU effect, $e_{N_{hj}} \sim N\left(0, \sigma^2_{within}\right)$, $\forall\, i \epsilon\, (hj)$ where $j = 1, ..., N_h$ and $h = 1, ..., H$.

5.3.1.3 Simulation Parameter Information

The goal is to simulate a two stratum population, i.e. $H = 2$. Initially it was decided to simulate the PSU's to be of a fixed size, but since one of the objectives of this thesis is to consider the effect of weight trimming on estimation precision, the sampling weights require enough variation such that the trimming methods can be applied. The variability in the weight distribution after simulating same-size PSU's was not sufficient. Thus, a decision was made to simulate the PSU's with varying numbers of observations. The PSU simulation was carried out independently for each stratum. This simulation setup should result in a multilevel population which will enable the selection of a stratified two-stage cluster sample. The sampling scheme will be discussed in the next chapter.

The descriptive statistics and graphs perused in the previous section will guide the choice of the random effects as well as the probability distributions to be used in the simulation process. Recall that the stratum and cluster variations presented previously were very large. It was decided to scale these values by firstly taking the square root of the variation and then multiplying the standard deviation by a constant. The adjusted values are presented below.

	PARAMETERS	IES	$\sqrt{V}$	$^1/_{100000}$	FINAL (σ^2)
Stratum	σ_1^2	13138231548	114622.125	1.146221	1.314
	σ_2^2	2785775181	52780.44316	0.527804	0.279
PSU	$\sigma^2_{1_{between}}$	16481687714	128381.0255	1.28381	1.648
	$\sigma^2_{2_{between}}$	35631479947	188763.0259	1.88763	3.563
SSU	$\sigma^2_{1_{within}}$	489540679310	699671.8369	6.996718	48.954
	$\sigma^2_{2_{within}}$	124425645237	352740.1951	3.527402	12.443

Table 5.3.8: Adjusted Random Effects

However, the simulation model also requires regression coefficients and the table below presents the estimated coefficients after fitting a main effects and first-order interactions SWLS to the IES. The last column in the table is obtained by simply scaling the parameter values such that the simulated data is more user-friendly.

		PARAMETERS	IES	$^1/_{100000}$
MAIN EFFECTS	VARIABLE			
	Intercept	β_0	-31444.96362	-0.31445
	Stratum	β_1	-9781.83246	-0.09782
	Continuous	β_2	896.1250341	0.008961
	Two-level categorical	β_3	7358.766641	0.073588
	Four-level categorical	β_4	-152.2464525	-0.00152
		β_5	8512.915246	0.085129
		β_6	19400.99633	0.19401
	Three-level categorical	β_7	4578.617964	0.045786
		β_8	7170.598233	0.071706
FIRST-ORDER INTERACTIONS				
	Two-level by Four-level	β_9	2316.773548	0.023168
		β_{10}	12983.12525	0.129831
		β_{11}	62602.0838	0.626021
	Two-level by Three-level	β_{12}	2352.07548	0.023521
		β_{13}	5843.588432	0.058436
	Four-level by Three-level	β_{14}	-4011.371423	-0.04011
		β_{15}	-637.7533588	-0.00638
		β_{16}	-8900.966295	-0.08901
		β_{17}	-9637.126936	-0.09637
		β_{18}	24891.80185	0.248918
		β_{19}	-58902.72487	-0.58903

Table 5.3.9: Regression Parameter Values

Taking all of the above into consideration, the following distributions and initial parameter values will be used for stratum $h = 1$:

- e_1, simulated from $N(0, \sigma_1^2)$ with $\sigma_1^2 = 1.31$;

- N_1, the number of PSU's to simulate for stratum 1, $N_1 = 1800$;
- e_{N_1}, simulated from $N\left(0, \sigma^2_{1_{between}}\right)$ with $\sigma^2_{1_{between}} = 1.65$;
- p, the number of explanatory variables, $p = 4$ (1 continuous, 3 categorical);
- N_{1j}, the number of SSU's in the jth PSU, a number to be generated as $N_{1j} \sim U(7, 19)$ and then rounded to be between 7 and 19;
- $e_{N_{1j}}$, simulated from $N\left(0, \sigma^2_{1_{within}}\right)$ with $\sigma^2_{1_{within}} = 48.95$;
- B_1, the vector of regression parameters for stratum 1. Notice that the second coefficient is zero since stratum 1 is considered the reference category of the stratification variable,

$$B_1 = \begin{pmatrix} -0.31 & 0 & 0.01 & 0.07 & -0.01 & -0.09 & 0.19 & 0.05 & 0.07 & 0.02 \\ 0.13 & 0.63 & 0.02 & 0.06 & -0.04 & -0.01 & -0.09 & -0.10 & 0.25 & -0.59 \end{pmatrix};$$

- $\theta_1 = \begin{pmatrix} 2 & 1 \end{pmatrix}$, the vector of parameter values for distribution F_1 from which the continuous variable will be simulated,

$$F_1 \sim gamma\left(\theta_1\right);$$

- $\theta_2 = 0.45$, the parameter value for distribution F_2 from which the 2-level categorical variable will be simulated,

$$F_2 \sim bernoulli\left(\theta_2\right);$$

- $\theta_3 = \begin{pmatrix} 0.21 & 0.50 & 0.10 & 0.19 \end{pmatrix}$, the vector of parameter values for distribution F_3 given that $X_2 = 0$.
- $\theta_4 = \begin{pmatrix} 0.15 & 0.60 & 0.10 & 0.15 \end{pmatrix}$, the vector of parameter values for distribution F_3 given that $X_2 = 1$.
- $\theta_5 = \begin{pmatrix} 0.10 & 0.30 & 0.60 \end{pmatrix}$, the vector of parameter values for distribution F_4.

The following distributions and initial parameter values will be used for stratum $h = 2$:

- e_2, simulated from $N\left(0, \sigma^2_2\right)$ with $\sigma^2_2 = 0.28$;
- N_2, the number of PSU's to simulate for stratum 2, $N_2 = 1200$;
- e_{N_2}, simulated from $N\left(0, \sigma^2_{2_{between}}\right)$ with $\sigma^2_{2_{between}} = 3.56$;

- p, the number of explanatory variables, $p = 4$ (1 continuous, 3 categorical);
- N_{2j}, the number of SSU's in the jth PSU, a number to be generated as $N_{2j} \sim U(7, 19)$ and then rounded to be between 7 and 19;
- $e_{N_{2j}}$, simulated from $N\left(0, \sigma^2_{2_{within}}\right)$ with $\sigma^2_{2_{within}} = 12.44$;
- B_2, the vector of regression parameters for stratum 2,

$$B_2 = \begin{pmatrix} -0.31 & -0.10 & 0.01 & 0.07 & -0.01 & -0.09 & 0.19 & 0.05 & 0.07 & 0.02 \\ 0.13 & 0.63 & 0.02 & 0.06 & -0.04 & -0.01 & -0.09 & -0.10 & 0.25 & -0.59 \end{pmatrix};$$

- $\theta_1 = \begin{pmatrix} 2 & 1 \end{pmatrix}$, the vector of parameter values for distribution F_1 from which the continuous variable will be simulated,

$$F_1 \sim gamma\left(\theta_1\right),$$

  which is the same as for stratum 1;
- $\theta_2 = 0.55$, the parameter value for distribution F_2 from which the 2-level categorical variable will be simulated,

$$F_2 \sim bernoulli\left(\theta_2\right);$$

- $\theta_3 = \begin{pmatrix} 0.16 & 0.60 & 0.10 & 0.14 \end{pmatrix}$, the vector of parameter values for distribution F_3 given that $X_2 = 0$.
- $\theta_4 = \begin{pmatrix} 0.22 & 0.50 & 0.10 & 0.18 \end{pmatrix}$, the vector of parameter values for distribution F_3 given that $X_2 = 1$.
- $\theta_5 = \begin{pmatrix} 0.50 & 0.30 & 0.20 \end{pmatrix}$, the vector of parameter values for distribution F_4.

Finally, the linear model that will be used for the simulation of the ith SSU in the jth PSU of the hth stratum, is

$$\begin{aligned} Y_{hji} &= \beta_{h_0} + \beta_{h_1} I_{hji} + \beta_{h_2} X_{hji_1} + \beta_{h_3} X_{hji_2} + \beta_{h_4} X^{D_2}_{hji_3} + \beta_{h_5} X^{D_3}_{hji_3} + \beta_{h_6} X^{D_4}_{hji_3} + \\ &\quad \beta_{h_7} X^{D_2}_{hji_4} + \beta_{h_8} X^{D_3}_{hji_4} + \left(e_h + e_{hj} + e_{hji}\right) + \left(first - order\, interactions\right), \end{aligned} \quad (5.3.1)$$

where

- Y_{hji} is the simulated response for the ith SSU in the jth PSU of the hth stratum;
- $\left\{\beta_{h_j}\right\}$, $j = 0, ..., p$, are the coefficients for the hth stratum;
- I_{hji} is an indicator variable indicating to which stratum the ith SSU belongs;
- $X_{hji_1}, ..., X_{hji_4}$ are the variables simulated for each stratum as explained above;
  - $X_{hji_3}^{D_2}$ and $X_{hji_4}^{D_2}$ are, respectively, the dummy variables of the second category of X_3 and X_4;
  - $X_{hji_3}^{D_3}$ and $X_{hji_4}^{D_3}$ are, respectively, the dummy variables of the third category of X_3 and X_4; and
  - $X_{hji_3}^{D_4}$ is the dummy variable of the fourth category of X_3;
- e_h is the random effect of the hth stratum;
- e_{hj} is the random effect of the jth PSU in the hth stratum; and
- e_{hji} is the random effect of the ith SSU in jth PSU in the hth stratum.

A diagram of the simulation process is provided in the following figure. Take note of the following counter notation used in the diagram:

- H_0, the stratum number counter;
- N_0, the PSU number counter in the stratum; and
- M_0, the SSU number counter in the PSU in the stratum.

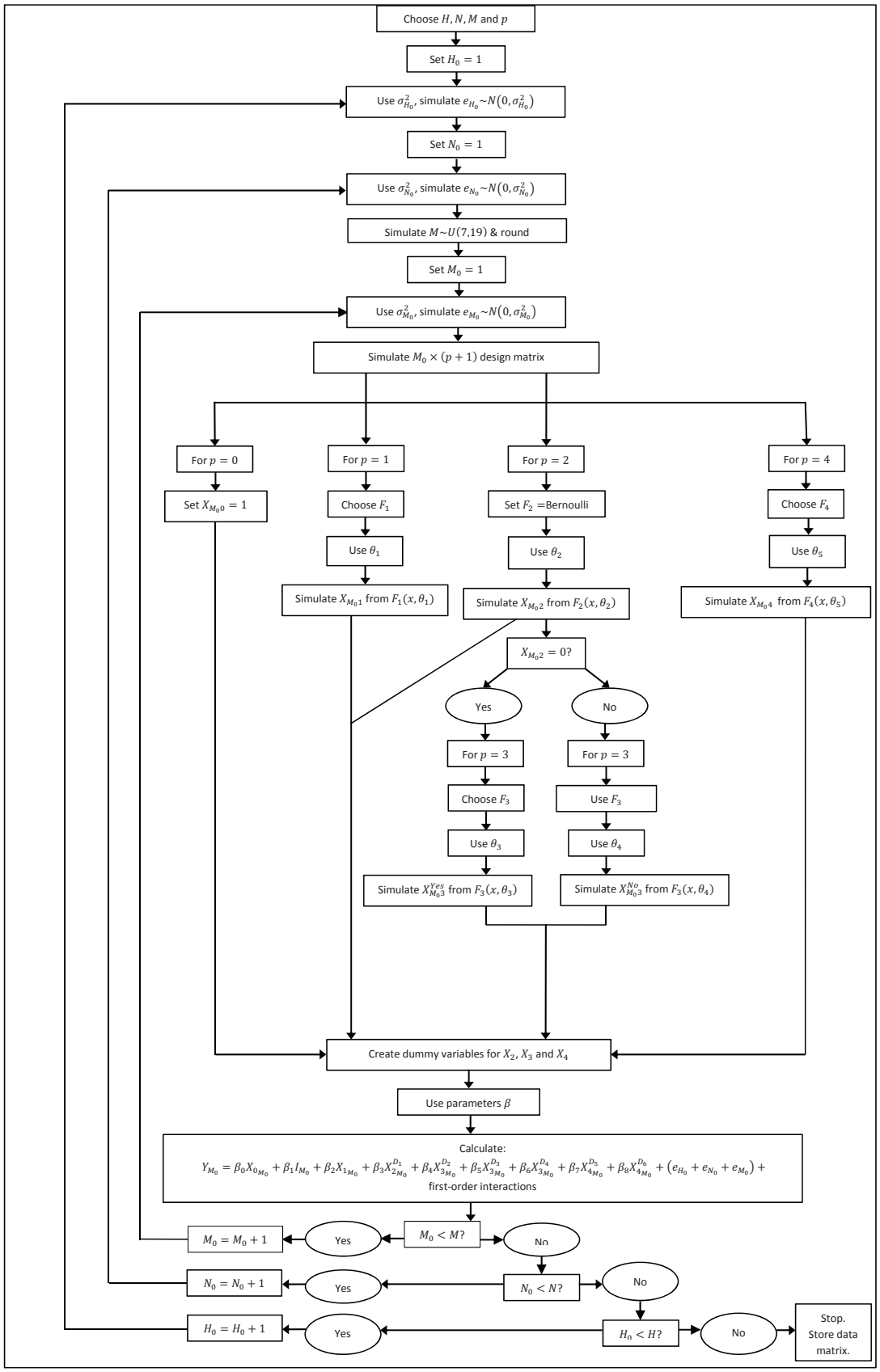


Figure 5.3.5: Diagram of Simulation Process

The table below is simply provided as a summary of the distributions and parameters selected to simulate this first multilevel population.

				STRATA			
				1		2	
Nr. of PSU's				1800		1200	
Nr. of Covariates				4		4	
Distributions	Continuous Variable X_1	Gamma					
		Parameters	α	2		2	
			β	1		1	
	Categorical Variable X_2	Bernoulli					
		Parameter	p	0.45		0.55	
	Categorical Variable X_3	Multinomial					
		Parameters		$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
			p_1	0.21	0.15	0.16	0.22
			p_2	0.5	0.6	0.6	0.5
			p_3	0.1	0.1	0.1	0.1
			p_4	0.19	0.15	0.14	0.18
	Categorical Variable X_4	Multinomial					
		Parameters	p_1	0.1		0.5	
			p_2	0.3		0.3	
			p_3	0.6		0.2	
Random Effects	Stratum Effect	Normal					
		Parameters	μ	0		0	
			σ^2	1.31		0.28	
	PSU Effect	Normal					
		Parameters	μ	0		0	
			σ^2	1.65		3.56	
	SSU Effect	Normal					
		Parameters	μ	0		0	
			σ^2	48.95		12.44	
Regression Parameters	Intercept			-0.31		-0.31	
	Stratum Indicator			0		-0.10	
	X_1			0.01		0.01	
	X_2						
		$X_2 = 1$		0.07		0.07	
	X_3						
		$X_3 = 2$		0.00		0.00	
		$X_3 = 3$		0.09		0.09	
		$X_3 = 4$		0.19		0.19	
	X_4						
		$X_4 = 2$		0.05		0.05	
		$X_4 = 3$		0.07		0.07	
	$X_2 \cdot X_3$						
		$X_3 = 2$		0.02		0.02	
		$X_3 = 3$		0.13		0.13	
		$X_3 = 4$		0.63		0.63	
	$X_2 \cdot X_4$						
		$X_4 = 2$		0.02		0.02	
		$X_4 = 3$		0.06		0.06	
	$X_3 \cdot X_4$						
		$X_3 = 2,\ X_4 = 2$		-0.04		-0.04	
		$X_3 = 2,\ X_4 = 3$		-0.01		-0.01	
		$X_3 = 3,\ X_4 = 2$		-0.09		-0.09	
		$X_3 = 3,\ X_4 = 3$		-0.10		-0.10	
		$X_3 = 4,\ X_4 = 2$		0.25		0.25	
		$X_3 = 4,\ X_4 = 3$		-0.59		-0.59	

Table 5.3.10: Summary of Distributions and Parameters used in WCEC Simulation Process

5.3.2 ECKZN Simulation Process

The same process described in the previous section was repeated, but the parameter values were changed to mimic the characteristics of the Eastern Cape and Kwa-zulu Natal. These two provinces were selected since their characteristics are typically very similar. Thus the second simulated population is simulated as a contrast to the first population.

Now, the following distributions and initial parameter values were selected after inspecting various descriptive measures of the two provinces based on their information in the IES. For stratum $h = 1$:

- e_1, simulated from $N(0, \sigma_1^2)$ with $\sigma_1^2 = 0.28$;
- N_1, the number of PSU's to simulate for stratum 1, $N_1 = 1800$;
- e_{N_1}, simulated from $N\left(0, \sigma_{1_{between}}^2\right)$ with $\sigma_{1_{between}}^2 = 3.56$;
- p, the number of explanatory variables, $p = 4$ (1 continuous, 3 categorical);
- N_{1j}, the number of SSU's in the jth PSU, a number to be generated as $N_{1j} \sim U(7, 19)$ and then rounded to be between 7 and 19;
- $e_{N_{1j}}$, simulated from $N\left(0, \sigma_{1_{within}}^2\right)$ with $\sigma_{1_{within}}^2 = 12.44$;
- B_1, the vector of regression parameters for stratum 1. Notice that the second coefficient is zero since stratum 1 is considered the reference category of the stratification variable,

$$B_1 = \begin{pmatrix} 0.02 & 0 & 0.01 & 0.02 & -0.01 & -0.01 & -0.06 & 0.05 & 0.61 & 0.05 \\ 0.14 & 0.41 & 0.05 & 0.12 & 0.10 & 0.25 & 0.16 & 0.35 & 0.60 & 0.47 \end{pmatrix};$$

- $\theta_1 = \begin{pmatrix} 3 & 1 \end{pmatrix}$, the vector of parameter values for distribution F_1 from which the continuous variable will be simulated,

$$F_1 \sim gamma(\theta_1);$$

- $\theta_2 = 0.46$, the parameter value for distribution F_2 from which the 2-level categorical variable will be simulated,

$$F_2 \sim bernoulli(\theta_2);$$

- $\theta_3 = \begin{pmatrix} 0.70 & 0.10 & 0.05 & 0.15 \end{pmatrix}$, the vector of parameter values for distribution F_3 given that $X_2 = 0$.

- $\theta_4 = \begin{pmatrix} 0.70 & 0.10 & 0.05 & 0.15 \end{pmatrix}$, the vector of parameter values for distribution F_3 given that $X_2 = 1$.
- $\theta_5 = \begin{pmatrix} 0.20 & 0.70 & 0.10 \end{pmatrix}$, the vector of parameter values for distribution F_4.

The following distributions and initial parameter values will be used for stratum $h = 2$:

- e_2, simulated from $N(0, \sigma_2^2)$ with $\sigma_2^2 = 0.27$;
- N_2, the number of PSU's to simulate for stratum 2, $N_2 = 1200$;
- e_{N_2}, simulated from $N\left(0, \sigma^2_{2_{between}}\right)$ with $\sigma^2_{2_{between}} = 1.04$;
- p, the number of explanatory variables, $p = 4$ (1 continuous, 3 categorical);
- N_{2j}, the number of SSU's in the jth PSU, a number to be generated as $N_{2j} \sim U(7, 19)$ and then rounded to be between 7 and 19;
- $e_{N_{2j}}$, simulated from $N\left(0, \sigma^2_{2_{within}}\right)$ with $\sigma^2_{2_{within}} = 10.7$;
- B_2, the vector of regression parameters for stratum 2,

$$B_2 = \begin{pmatrix} 0.02 & 0.02 & 0.01 & 0.02 & -0.01 & -0.01 & -0.06 & 0.05 & 0.61 & 0.05 \\ 0.14 & 0.41 & 0.05 & 0.12 & 0.10 & 0.25 & 0.16 & 0.35 & 0.60 & 0.47 \end{pmatrix};$$

- $\theta_1 = \begin{pmatrix} 3 & 1 \end{pmatrix}$, the vector of parameter values for distribution F_1 from which the continuous variable will be simulated,

$$F_1 \sim gamma(\theta_1),$$

  which is the same as for stratum 1;

- $\theta_2 = 0.43$, the parameter value for distribution F_2 from which the 2-level categorical variable will be simulated,

$$F_2 \sim bernoulli(\theta_2);$$

- $\theta_3 = \begin{pmatrix} 0.75 & 0.05 & 0.15 & 0.10 \end{pmatrix}$, the vector of parameter values for distribution F_3 given that $X_2 = 0$.
- $\theta_4 = \begin{pmatrix} 0.75 & 0.05 & 0.15 & 0.10 \end{pmatrix}$, the vector of parameter values for distribution F_3 given that $X_2 = 1$.

- $\theta_5 = \begin{pmatrix} 0.20 & 0.70 & 0.10 \end{pmatrix}$, the vector of parameter values for distribution F_4.

Exactly the same simulation model, given in equation 5.3.1, and simulation process described in figure 5.3.5 are followed and a summary of the parameters used in this second simulation is presented in the table below.

				STRATA			
				1		2	
Nr. of PSU's				1800		1200	
Nr. of Covariates				4		4	
Distributions	Continuous Variable X_1	Gamma					
		Parameters	α	3		3	
			β	1		1	
	Categorical Variable X_2	Bernoulli					
		Parameter	p	0.46		0.43	
	Categorical Variable X_3	Multinomial					
		Parameters		$X_2 = 0$	$X_2 = 1$	$X_2 = 0$	$X_2 = 1$
			p_1	0.7	0.7	0.75	0.75
			p_2	0.1	0.1	0.05	0.05
			p_3	0.05	0.05	0.15	0.15
			p_4	0.15	0.15	0.1	0.1
	Categorical Variable X_4	Multinomial					
		Parameters	p_1	0.2		0.2	
			p_2	0.7		0.7	
			p_3	0.1		0.1	
Random Effects	Stratum Effect	Normal					
		Parameters	μ	0		0	
			σ^2	0.28		0.27	
	PSU Effect	Normal					
		Parameters	μ	0		0	
			σ^2	3.56		1.04	
	SSU Effect	Normal					
		Parameters	μ	0		0	
			σ^2	12.44		10.70	
Regression Parameters	Intercept			0.02		0.02	
	Stratum Indicator			0		0.02	
	X_1			0.01		0.01	
	X_2						
		$X_2 = 1$		0.02		0.02	
	X_3						
		$X_3 = 2$		-0.01		-0.01	
		$X_3 = 3$		-0.01		-0.01	
		$X_3 = 4$		-0.06		-0.06	
	X_4						
		$X_4 = 2$		0.05		0.05	
		$X_4 = 3$		0.61		0.61	
	$X_2 \cdot X_3$						
		$X_3 = 2$		0.05		0.05	
		$X_3 = 3$		0.14		0.14	
		$X_3 = 4$		0.41		0.41	
	$X_2 \cdot X_4$						
		$X_4 = 2$		0.05		0.05	
		$X_4 = 3$		0.12		0.12	
	$X_3 \cdot X_4$						
		$X_3 = 2,\ X_4 = 2$		0.10		0.10	
		$X_3 = 2,\ X_4 = 3$		0.25		0.25	
		$X_3 = 3,\ X_4 = 2$		0.16		0.16	
		$X_3 = 3,\ X_4 = 3$		0.35		0.35	
		$X_3 = 4,\ X_4 = 2$		0.60		0.60	
		$X_3 = 4,\ X_4 = 3$		0.47		0.47	

Table 5.3.11: Summary of Distributions and Parameters used in ECKZN Simulation Process

5.4 The Simulated Complex Sampling Data

At this point the two populations, WCEC and ECKZN, have been simulated according to the diagram in figure 5.3.5 using the distributions and parameter values summarized in table 5.3.10, for WCEC, and table 5.3.11 for ECKZN. Both populations comprise two strata and in both populations the first stratum contains 1800 PSU's while the second stratum contains 1200. Thus, 3000 PSU's were simulated for both populations. Furthermore, recall that the PSU's contain observations, SSU's, and that the number of observations in each PSU will be a random number simulated to fall between a minimum of 7 and a maximum of 19. Summary statistics of the PSU sizes are presented in table 5.4.1 below and in both populations the average PSU size per stratum is approximately 13. The minima and maxima are also as intended.

WCEC						
STRATUM	Min	Q_1	Q_2	Q_3	Max	Mean
1	7	10	13	16	19	13
2	7	10	13	16	19	13.21
ECKZN						
STRATUM	Min	Q_1	Q_2	Q_3	Max	Mean
1	7	10	13	16	19	13
2	7	10	13	16	19	13.05

Table 5.4.1: Summary Statistics of PSU Sizes

If the average PSU size is 13, then it is expected that the population sizes will be approximately 39000. The actual sizes achieved after simulation and due to the varying PSU sizes, are 39255 for WCEC and 39054 for ECKZN.

Multilevel models made it possible to specify different random effects for the different levels intended in the hierarchical population. Firstly consider the stratum random effect. The WCEC random effect choices intended for the strata in this population to be different from each other while ECKZN parameter choices intended for the strata to be more similar. Figure (5.4.1) presents, on the left, the probability density function of the stratum effects for WCEC while that of ECKZN is presented on the right.

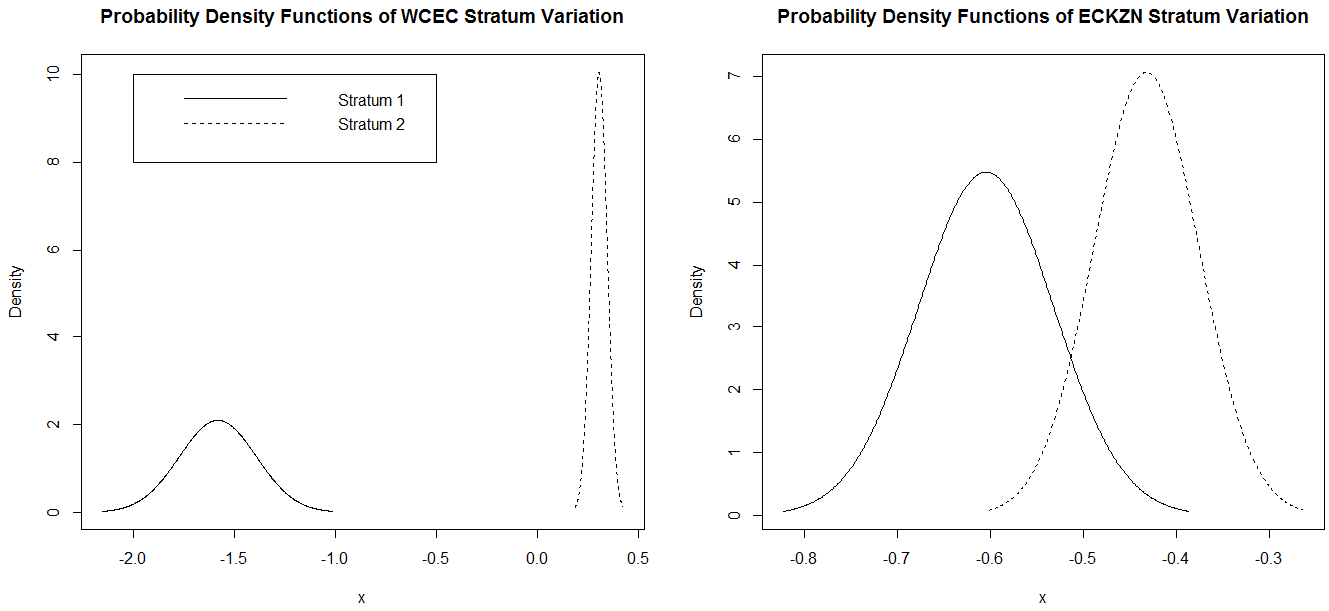


Figure 5.4.1: Achieved Stratum Variation

It is noticed in both density functions that the stratum variations follow a bell-shaped curve, indicative of the normal distribution from which it was simulated. Furthermore, the plot on the left clearly shows that the strata are quite different while the curves on the right overlap. Thus, it appears as if the desired stratum variations for the two populations are achieved.

Next, the variation between PSU's was included in the simulation model and the parameter choices here intended for the effect to be reversed between the two populations. In other words, for WCEC the intention was for the between PSU variation in stratum 1 to be less than in stratum 2 to mimic the relationship observed in the IES, while for ECKZN the variation in stratum 1 was to be larger than in stratum 2. The achieved between PSU variation, by stratum, for the two populations, is presented in figure 5.4.2 below.

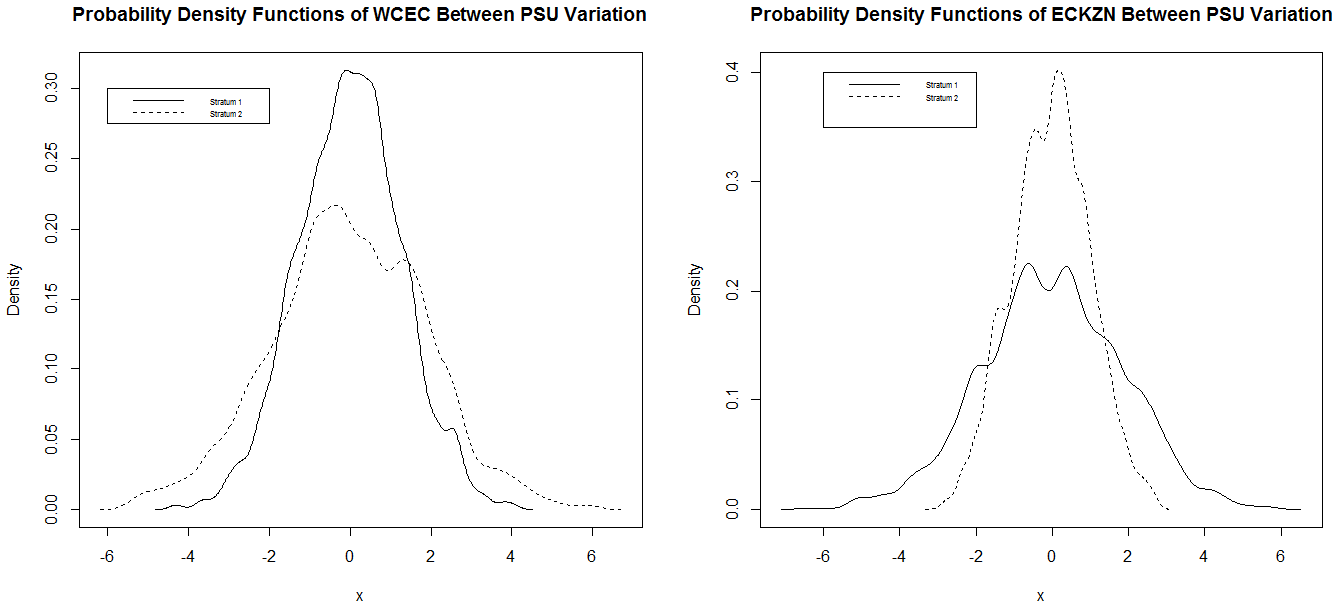


Figure 5.4.2: Achieved Between Cluster Variation

The density function on the left presents the between PSU variation achieved in WCEC. It is clear that stratum 1, denoted by the solid line, has smaller variation than stratum 2. On the right it is seen that the between PSU variation in stratum 2, the dashed line, is smaller than the stratum 1 variation. Both populations portray the intended between PSU variations.

The third random effect specified in the simulation model is with regard to the within-PSU variation. According to the descriptive statistics obtained from the IES the within-PSU variation in the first stratum of WCEC should be larger than that of the second stratum while the within-PSU variation of the strata in ECKZN should be approximately the same. Consider figure 5.4.3 below.

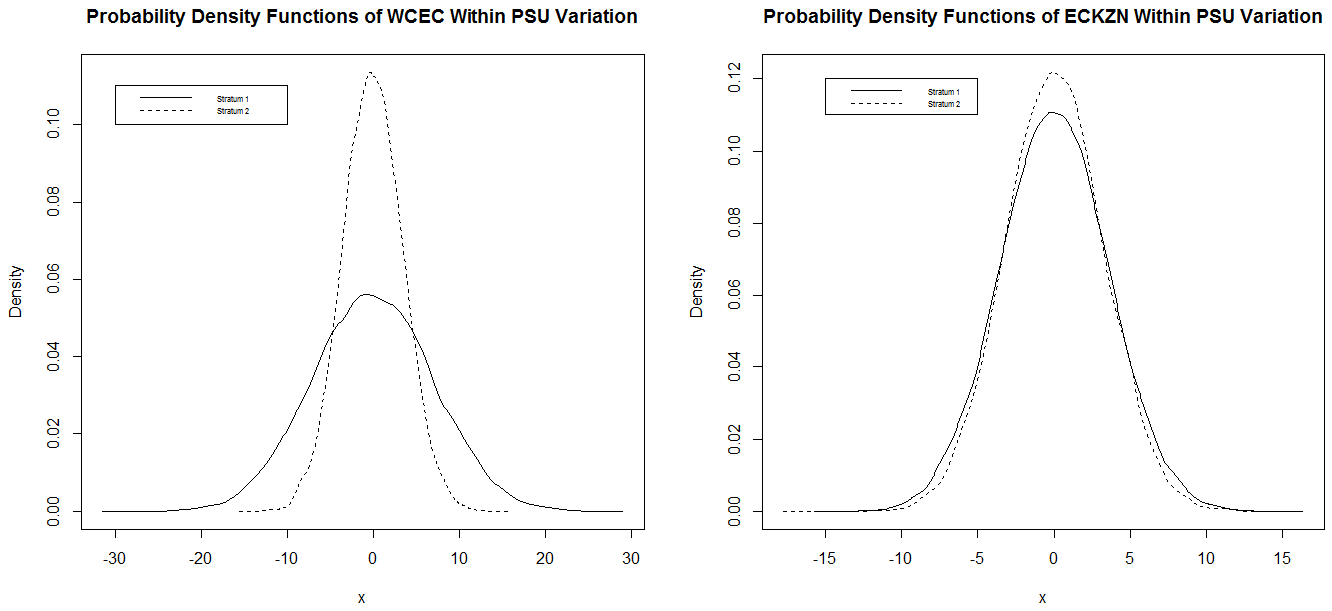


Figure 5.4.3: Achieved Within Cluster Variation

From the above figure it is clear that the achieved within-PSU variation in both strata of both populations is as desired.

The last two figures presented in this section represent the distribution of the simulated response and the simulated continuous variable, respectively.

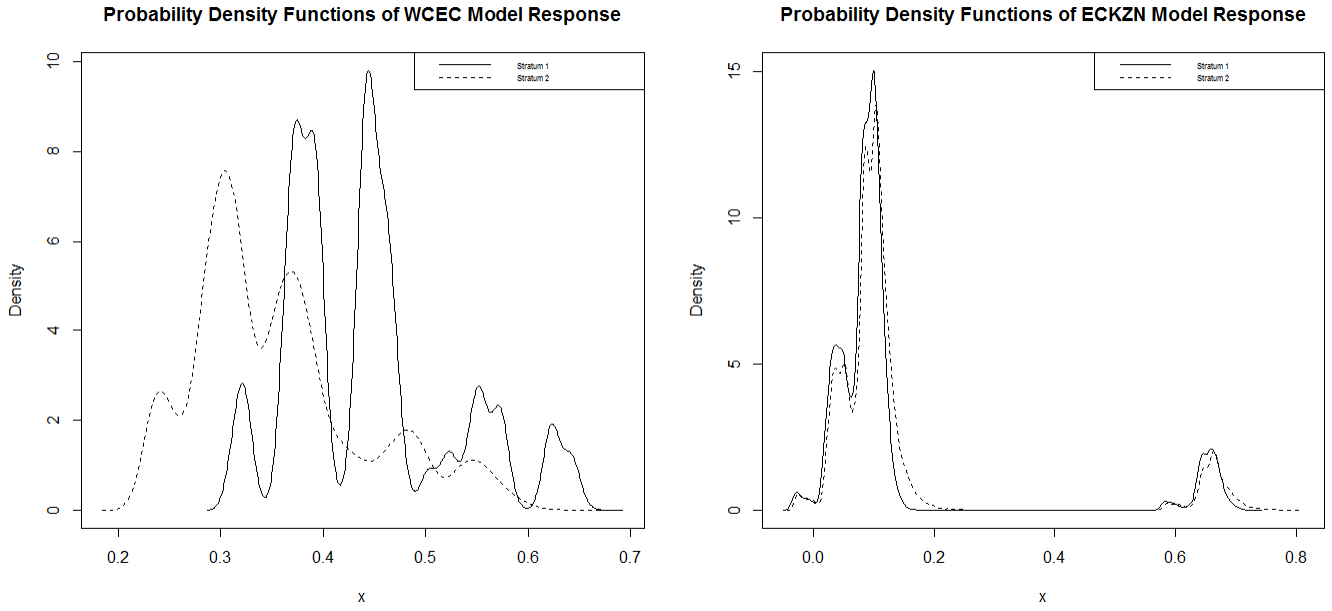


Figure 5.4.4: Model Response Distributions

The density function on the left represents the simulated response of the WCEC population.

Recall that the intention with this population was to simulate two strata that differ from each other. Although the solid and dashed lines overlap it is clear that the dashed line has a wider range than the solid line and its highest peak is flatter than that of the solid line. The density function on the right represents the ECKZN simulated response. Here the intention was to simulate strata that are similar and the solid and dashed lines do seem to overlap quite substantially.

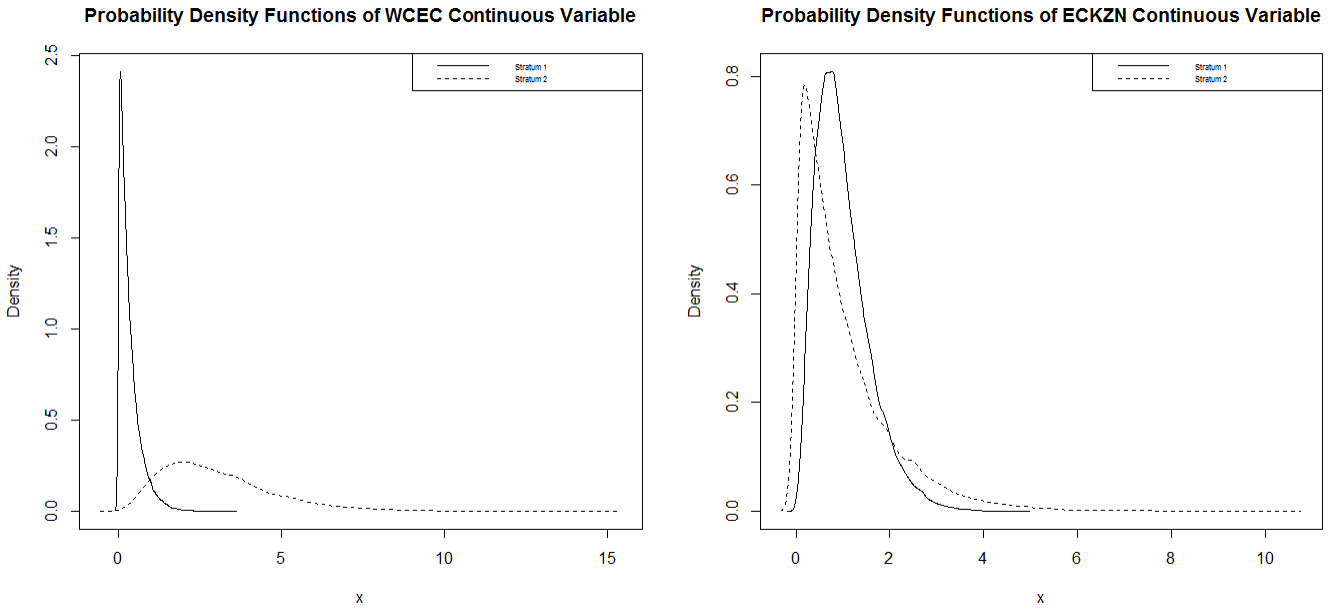


Figure 5.4.5: Continuous Variable Distributions

This figure also portrays the WCEC continuous variable densities on the left and the ECKZN continuous variable densities on the right. The curves on both sides are positively skewed, as intended, and the desired population similarities and differences are once again evident.

Taking into consideration the summary statistics presented in this section it can be concluded that the achieved populations seem to possess the characteristics they were intended to have.

This chapter started with a discussion of the two-level model and then related this model to a CS design. It continued with a discussion of how multilevel models could be used to simulate a hierarchical population from which a CS can be selected. The intention was to simulate a population that bares some resemblance to a real-world survey. The IES was studied and different probability distributions and parameter values were deduced from the descriptive measures of this survey. A decision was made to simulate two populations, each with two strata. The strata of the first population would differ in terms of their characteristics and random effects while the strata of the second population would be similar. Summaries of the distributions and parameters are presented in tables 5.3.10 and 5.3.11 and a diagram of the simulation process was presented in figure 5.3.5. The chapter was concluded with a selection of descriptive statistics of the two

simulated populations and the conclusion was reached that the populations bare the intended characteristics.

The next chapter will discuss the CS design and selection of the sample from the populations. It will also describe the analyses to be conducted using the simulated data.

Chapter 6

Simulation Data and Analysis

The previous chapter discussed the simulation of two hierarchical populations, namely WCEC and ECKZN, through multilevel modeling and broadly based on the characteristics of the IES (see chapter 7). A diagram of the simulation process was provided (see figure 5.3.5) as well as the distributional and parameter value choices in tables 5.3.10 and 5.3.11. These form the populations from which samples were drawn for the simulation. This chapter will explain how the sampling was conducted and also provide the outlines of the analyses conducted using these samples. The chapter will be broken down into two main parts, namely a section considering the analysis pertaining to the evaluation of the fitted model and a section for the model parameter analysis. In each section the results of the analyses will be presented and discussed after which the sections will be concluded with a summary of the main findings.

6.1 Sampling Scheme and Simulation Study Outline

The observations in the two simulated populations, WCEC and ECKZN, hereafter referred to as the simulated populations, are grouped into 3000 PSU's over two strata. The simulation process was set up such that stratum 1 contains 60% of the PSU's. Recall that the PSU's were simulated to be of varying sizes, the range being from 7 to 19 SSU's in a PSU. In the end, WCEC had 39255 observations and ECKZN had 39054. The table below summarizes this information. Note that the SSU totals are presented in italics.

	TOTAL PSU's (*SSU's*)		
	STRATUM 1	STRATUM 2	TOTAL
WCEC	1800 (*23401*)	1200 (*15854*)	3000 (*39255*)
ECKZN	1800 (*23397*)	1200 (*15657*)	3000 (*39054*)

Table 6.1.1: Population Totals

For sampling purposes it was important that each PSU and SSU be uniquely identifiable. The

PSU unique number was constructed as a combination of the stratum number to which the PSU belongs and the number of the PSU within that stratum. The SSU unique number is a combination of the unique number of the PSU to which it belongs as well as its number in that PSU. Consider the number 1800112. The figure below illustrates how from this number it can be deduced that this observation belongs to stratum 1, PSU 1800 and is the twelfth SSU in this PSU.

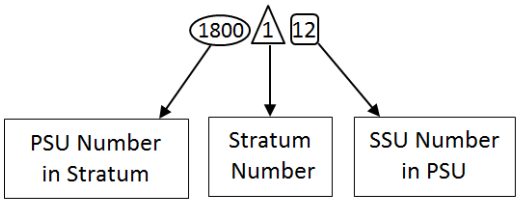


Figure 6.1.1: SSU Unique Number Example

Consider a stratified two-stage sample to be selected from any of the populations. To achieve an approximate 10% sample it was decided to sample a total of 750 PSU's and then 4 SSU's from each selected PSU. This results in a sample of size 3000 which, for both surrogate populations, achieves approximately 8% samples. Since this is not too different from the intended 10%, it was decided to go ahead with the sampling.

The sample of size 3000 was obtained by sampling 375 PSU's, with equal probability (EPSEM), from each stratum followed by sampling 4 SSU's, also EPSEM, from each selected PSU. The application of the weight trimming methods discussed in chapter 3 requires the achieved weight distributions, whether it be the design weights of the final sampling weights, to have large variability. Although the largest variation in the sampling weights is usually observed once the non-response adjusted design weights have been benchmarked, it was important, for illustration and comparison purposes, to try to increase the variability within the design weights as well. This necessity lead to the decision to use EPSEM sampling at both PSU and SSU level such that the weight trimming methods are actually employed in the analyses.

Consider the ith observation in the jth PSU in the hth stratum. The stratum contains N_h PSU's of which n_h have been sampled (EPSEM). If the number of SSU's in the jth selected PSU is denoted by N_{hj} and the number of SSU's sampled, also EPSEM, from this PSU by n_{hj}, then the inclusion probability of the ith observation is given by

$$\pi_{hji} = \left(\frac{n_h}{N_h}\right) \cdot \left(\frac{n_{hj}}{N_{hj}}\right),$$

and its associated design weight, $d_{CS_{hji}}$, is calculated as the inverse of π_{hji}. Recalling that N_{hj} has been simulated to vary between 7 and 19, the table below presents the possible range in design weights for WCEC and ECKZN.

	STRATUM 1		STRATUM 2	
	Min	Max	Min	Max
WCEC	8.4	22.8	5.6	15.2
ECKZN	8.4	22.8	5.6	15.2

Table 6.1.2: Design Weight Range (d_{CS})

Note that differential non-response, i.e. the over-/ under-representation of certain groups, is often found in practical situations. Thus to keep the samples as realistic as possible and to be able to determine this type of non-response error, it was simulated in the design of the samples through the use of auxiliary variables. This was done to evaluate the weighting procedures under non-perfect circumstances. When benchmarking to correct for this phenomenon, it is customary to use the totals of two sets of auxiliary variables to benchmark the design weights. However, for the simulated populations a single set of auxiliary variables were used in the simulation to aid in determining which weighting technique would be best under these circumstances. The auxiliary variables used are listed below:

- the strata;
- the continuous variable, X_1, that has been grouped into four categories based on its quartiles;
- the two-level categorical variable, X_2; and
- the four-level categorical variable, X_3.

A distance measure is defined when benchmarking design weights and in this thesis two distance measures were employed, namely the linear, denoted by the superscript pp_1, and the exponential (raking ratio), denoted by the superscript pp_2. The post-benchmarking sampling weight of the ith observation in the jth PSU in the hth stratum is denoted by $w_{CS_{hji}}$ and is the final sampling weight associated with the observation. The sampling weight ranges for the first sample selected from WCEC and ECKZN by stratum, respectively, are presented in the table below.

	STRATUM 1				STRATUM 2			
	$w_{CS}^{pp_1}$		$w_{CS}^{pp_2}$		$w_{CS}^{pp_1}$		$w_{CS}^{pp_2}$	
	Min	Max	Min	Max	Min	Max	Min	Max
WCEC	7.55	25.69	7.58	25.77	5.07	17.03	5.08	17.13
ECKZN	5.71	31.42	6.28	32.23	3.64	21.88	4.07	22.91

Table 6.1.3: Final Sampling Weight Range: d_{CS}

Notice how the distance between the minimum and maximum weights, especially under the raking ratio, is greater post-benchmarking than pre-benchmarking. Also, the variation within the ECKZN benchmarked weights seems larger than for the WCEC weights.

It has happened that some statisticians do not benchmark the design weights as explained until now, but rather benchmark the raw data. This implies that the design weight is assumed to be

$$d_i = \frac{N}{n},$$

i.e. when making use of SRS. The (hj) is dropped from the subscript to emphasize that the design is not taken into account here. Let this design weight be denoted as d_{SRS} with benchmarked weights $w_{SRS}^{pp_1}$ and $w_{SRS}^{pp_2}$. In this case the design weight, when sampling from WCEC, equals 13.085, and when sampling from ECKZN it equals 13.018. The table below summarizes the ranges of the weights after benchmarking d_{SRS} for a single sample.

	STRATUM 1				STRATUM 2			
	$w_{SRS}^{pp_1}$		$w_{SRS}^{pp_2}$		$w_{SRS}^{pp_1}$		$w_{SRS}^{pp_2}$	
	Min	Max	Min	Max	Min	Max	Min	Max
WCEC	13.68	17.37	13.56	17.73	9.45	12.47	9.67	12.09
ECKZN	10.26	21.28	10.15	23.66	4.93	15.94	6.68	15.58

Table 6.1.4: Final Sampling Weight Range: d_{SRS}

Clearly the variation within the d_{SRS} benchmarked weights is smaller than within the d_{CS} benchmarked weights. Both design weights as well as their respective benchmarked weights will be considered in the analyses.

The simulation of these samples was conducted in SAS since the program used to do the benchmarking of the design weights is part of that software. Let the number of samples selected from a population, be denoted by R. In this part of the application $R = 100$. Due to the size of the populations and the samples, as well as the number of samples selected from the populations, this process was quite time consuming.

Now, consider the rth sample as a bootstrap population and let the parameter of interest be the jth regression coefficient, β_j, to be estimated by OLS, WLS or SWLS using the bootstrap population. Let this estimator be denoted by $\hat{\beta}_{j_r}$. Bootstrap resampling needed to be applied within each bootstrap population. The bootstrap sampling was also carried out in SAS such that the benchmarking of the bootstrap sampling weights could be carried out using the SAS calmar function. Initially B_1, the number of first-level bootstrap samples, was chosen to be 500. Trial runs of the resampling and benchmarking soon made it clear that the magnitude of the resampling and benchmarking far surpassed the computer time required for the sampling and benchmarking of the bootstrap populations and thus a decision was made to reduce the number of bootstrap samples. Various opinions exist regarding the choice of the number of bootstrap samples to use, but Efron and Tibshirani (1998) usually choose 200 and thus B_1 was set equal to this number. Although there was a significant improvement in the amount of computer time used, it was clear that this level in the simulation study was going to require many hours to complete. This presented the first

hurdle regarding the computer work and lead to an introduction to high-performance computing (HPC).

HPC is a method that uses supercomputers and computer clusters to solve advanced computing problems. Supercomputers, although cutting edge, are quite expensive. Cluster computers consist of a group of computers that are closely linked and that work in parallel. Since the cluster can consist of everyday off-the-shelf computers it is a relatively inexpensive setup. The computers are typically connected through local area networks (LAN's) and provide improved computing performance above what is typically available from a single computer. The figure below presents an example of a typical Linux cluster layout.

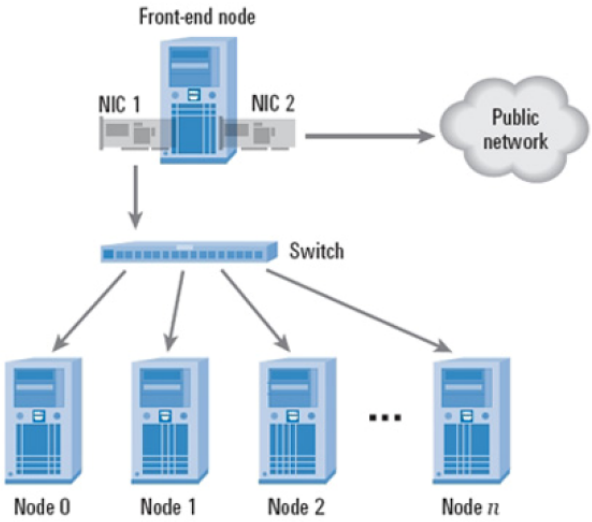


Figure 6.1.2: Linux Cluster Layout

Stellenbosch University has such a Linux cluster which is called "Rhasatsha" or referred to as the HPC1. A "rhasatsha" is versatile SSU or object that completes any task promptly and successfully. A new job is started by connecting to the head node followed by setting up the job which entails the copying of data and source code and also the specification of a submit script for the scheduler. Once the job is submitted the scheduler decides when and where to run the job. The jobs can be classified in different ways, but this simulation study mostly made use of parallel array jobs. This implies that multiple instances of the same executable are run and each instance processes different input files.

Currently the HPC1 has 1328 cores available. Submitted jobs are classified into different queues depending on the walltime specified for a job. To ensure that short jobs are not blocked by longer running jobs, each queue is allocated a maximum number of cores. Hence, it is quite important to set up jobs in such a way that the specified walltime falls in a short queue. Up to a walltime of a week, or 168 hours, there is no limit on the number of jobs that a user can submit, but the number of cores available changes from being unlimited (up to 24 hours) to a maximum of 800 for a week. However, when a job continues for longer than a week the user is limited to 10 jobs and

a maximum of 200 cores per user. Most of the jobs submitted for this simulation study could be set up such that a 168 hour walltime was sufficient. The SWLS jobs, however, had to move into the month queue, 744 hours. This was unfortunate since many researchers had already submitted jobs to this queue.

Since the HPC1 runs on a Linux OS, Windows programs cannot be run on it. However, the cluster can still be accessed from a Windows OS through two Windows clients, namely *ssh* and *scp*. Putty and WinSCP, respectively, are necessary for these clients and are available for researchers to download. Unfortunately SAS does not run on a Linux (OS) and thus the resampling could not be submitted to the HPC1, but R is Linux compatible and thus all programming conducted in R (Lumley, 2014) could be submitted to the HPC1.

Consider the bth bootstrap sample selected from the rth bootstrap population. To recap the bootstrap method applied to CS data, consider the hth stratum with n_h PSU's. The bth bootstrap sample is obtained by taking a with-replacement sample of $n_h - 1$ PSU's independently from each stratum. Define m^*_{hj} as the number of times the jth PSU of the hth stratum is selected for the bootstrap sample. The bootstrap design weights are then calculated, using m^*_{hj}, as given in equation (4.3.8) and then benchmarked to the population totals.

After approximately two weeks the first-level bootstrap sampling from each of the 100 bootstrap populations was complete, including the benchmarking, and ready to be used for the analyses outlined in the sections below.

Finally, the simulated populations, WCEC and ECKZN, and their respective sub-samples are used for the application of the techniques discussed in chapter 4 that were categorized according to parameter estimation, variance estimation, model evaluation and parameter inference. The results of these analyses will be presented in the two subsections set out below, namely "model evaluation analysis" and "model parameter analysis". The output from each subsection will be used for the investigation of the following comparisons:

- OLS versus WLS versus SWLS;
- d_{CS} benchmarked weights' output versus d_{SRS} benchmarked weights' output; and
- untrimmed versus trimmed sampling weights' output.

The relationship between the main objectives and the simulation study is illustrated in the diagram below:

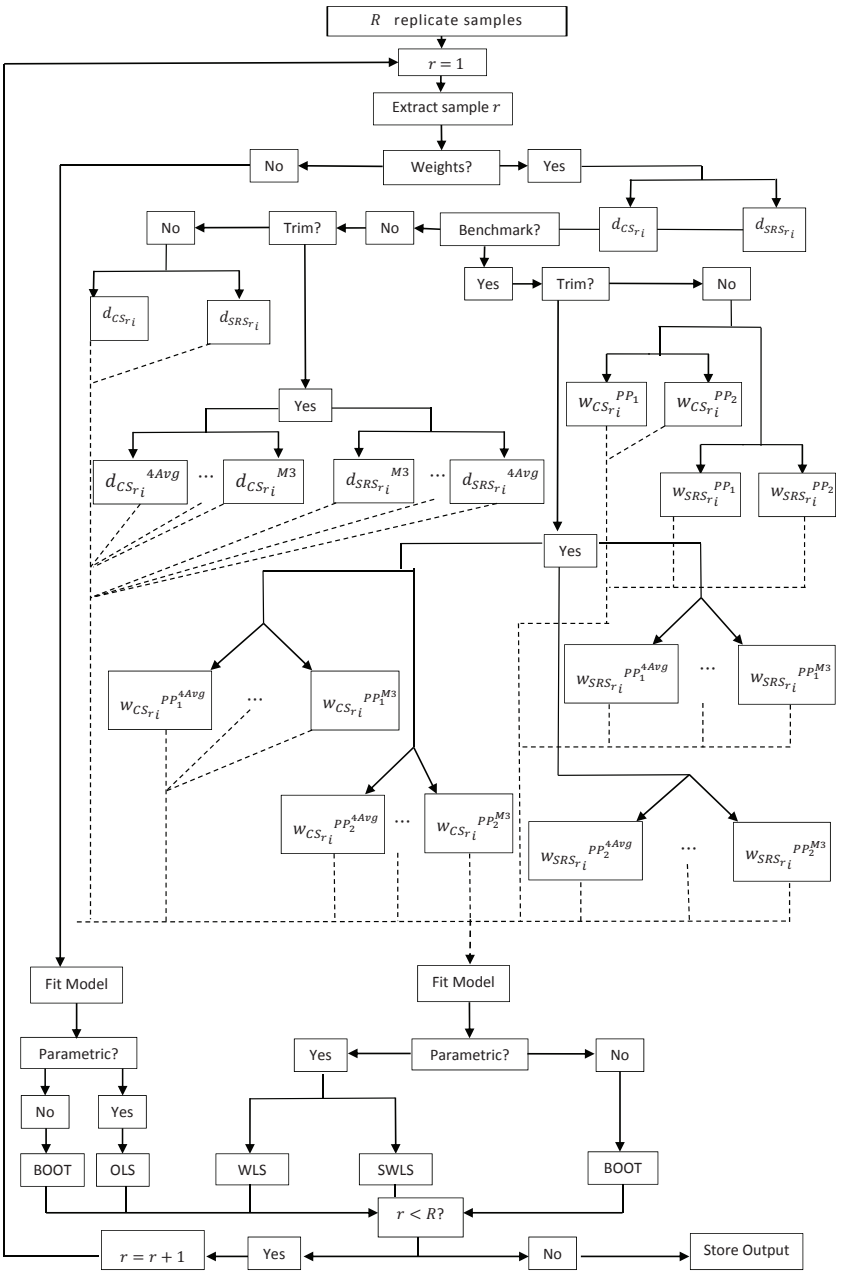


Figure 6.1.3: Diagram of the Simulation Study

This diagram is repeated for the inference concerning the model parameters, i.e. point and interval estimation, as well as the evaluation of the model fit. The process will be repeated for both populations, WCEC and ECKZN, but only the WCEC results will be discussed while the ECKZN results will be available for perusal from the author.

6.2 Model Evaluation Analysis

This section presents the outlines of the model evaluation analyses. Recall from the discussion in section 4.4 that the models are evaluated according to three diagnostics, namely

1. the coefficient of multiple determination;
2. prediction error; and
3. outlier diagnostics.

The analyses are outlined in sections 6.2.1 to 6.2.3 and results will be presented as part of each section. The model evaluation analyses will be concluded by a summary of the main findings in section 6.2.4.

6.2.1 Coefficient of Multiple Determination

The coefficient of multiple determination, discussed in section 4.4.1, is a first indication of how well the explanatory variables explain the variation in the model response. Let the coefficient of multiple determination be denoted by R^2. This measure will be calculated for each linear model fitted to each replicate sample, $\{R_r^2\}$, $1, ..., R$, and the following statistics will be reported:

- average, $R_{avg}^2 = \frac{1}{R}\sum_r R_r^2$; and
- standard deviation, $R_{sd}^2 = \sqrt{\frac{1}{R-1}\sum_r \left(R_r^2 - R_{avg}^2\right)^2}$.

Consider the averages and standard deviations of the R^2's obtained from OLS, WLS and SWLS using the different sampling weights,

1. theoretical design weights, d_{CS};
2. alternative design weights, d_{SRS};
3. SSU/person-level benchmarked theoretical design weights, linear distance function, $w_{CS}^{pp_1}$;
4. SSU/person-level benchmarked alternative design weights, linear distance function, $w_{SRS}^{pp_1}$;
5. SSU/person-level benchmarked theoretical design weights, exponential distance function, $w_{CS}^{pp_2}$; and
6. SSU/person-level benchmarked alternative design weights, exponential distance function, $w_{SRS}^{pp_2}$.

The averages and standard deviations are given in table 6.2.1 and the largest means are highlighted in green with their respective standard deviations highlighted in blue.

		MEAN							STANDARD DEVIATION					
		No	Avg	IQR	Med	Hill	M3		No	Avg	IQR	Med	Hill	M3
OLS		0.8636	0.8636	0.8636	0.8636	0.8636	0.8636		0.0053	0.0053	0.0053	0.0053	0.0053	0.0053
WLS	d_{CS}	0.8772	0.8772	0.8772	0.8772	0.8772	0.8772		0.0054	0.0054	0.0054	0.0054	0.0054	0.0054
	d_{SRS}	0.8636	0.8636	0.8636	0.8636	0.8636	0.8636		0.0053	0.0053	0.0053	0.0053	0.0053	0.0053
	$w_{CS}^{pp_1}$	0.8791	0.8791	0.8791	0.8791	0.8788	0.8788		0.0051	0.0051	0.0051	0.0051	0.0052	0.0050
	$w_{SRS}^{pp_1}$	0.8785	0.8785	0.8785	0.8785	0.8783	0.8784		0.0048	0.0048	0.0047	0.0048	0.0048	0.0047
	$w_{CS}^{pp_2}$	**0.8791**	**0.8791**	**0.8791**	**0.8791**	0.8786	0.8788		**0.0051**	**0.0051**	**0.0051**	**0.0051**	**0.0052**	0.0050
	$w_{SRS}^{pp_2}$	**0.8792**	**0.8792**	**0.8792**	**0.8792**	0.8791	0.8791		**0.0047**	**0.0047**	**0.0047**	**0.0047**	**0.0046**	0.0047
SWLS	d_{CS}	0.8772	0.8772	0.8772	0.8772	0.8772	0.8772		0.0054	0.0054	0.0054	0.0054	0.0054	0.0054
	d_{SRS}	0.8636	0.8636	0.8636	0.8636	0.8636	0.8636		0.0053	0.0053	0.0053	0.0053	0.0053	0.0053
	$w_{CS}^{pp_1}$	0.8791	0.8791	0.8791	0.8791	0.8788	0.8788		0.0051	0.0051	0.0051	0.0051	0.0052	0.0050
	$w_{SRS}^{pp_1}$	0.8785	0.8785	0.8785	0.8785	0.8783	0.8784		0.0048	0.0048	0.0047	0.0048	0.0048	0.0047
	$w_{CS}^{pp_2}$	**0.8791**	**0.8791**	**0.8791**	**0.8791**	0.8786	0.8788		**0.0051**	**0.0051**	**0.0051**	**0.0051**	**0.0052**	0.0050
	$w_{SRS}^{pp_2}$	**0.8792**	**0.8792**	**0.8792**	**0.8792**	0.8791	0.8791		**0.0047**	**0.0047**	**0.0047**	**0.0047**	**0.0046**	0.0047

Table 6.2.1: WCEC R^2 Mean and Standard Deviation over Replicate Samples

The range of the averages is not very wide. However, including the sampling weights in the linear model, i.e. WLS and SWLS, resulted in an improvement in R^2 with further increases in R^2 observed once the respective design weights, d_{CS} and d_{SRS} were benchmarked. It should be noted that the theoretical sampling weights, i.e. d_{CS}, $w_{CS}^{pp_1}$ and $w_{CS}^{pp_2}$, consistently achieved a slightly higher R^2 although their associated standard deviations were consistently larger. However, as shown in tables 6.1.3 and 6.1.4, the "CS" sampling weights have larger variation than the "SRS" sampling weights which could be the cause of the slightly larger variation in the "CS" R^2's. Weight trimming does not seem to have had a significant effect on the R^2's.

6.2.2 Prediction Error Estimation

To determine which of the prediction error (PE) estimation methods perform "best" one needs to be able to compare the obtained estimates of PE to the "true" PE. Since the "true" PE is unknown it also needs to be estimated. For this purpose the simulated population will be considered as the population from which the "truth" can be deduced. Hence, the simulation study for the evaluation of the linear model PE consists of two phases, namely

1. the calculation of the "true" PE; and
2. the comparison of the PE estimation methods to the "true" PE through the evaluation of diagnostic measures.

Consider the following terminology and notation that will be used in the simulation study:

- population, which refers to the simulated population;
- replicate, which refers to a sample taken from the population;
  - each replicate sample is considered a bootstrap population.
- N, the number of observations in the population;
- n_r, the number of observations in the rth replicate sample, $r = 1, .., R$;
- y, the observed response;
- $\hat{y}$, the predicted response;
- w, the sampling weight of an observation;
- $\tilde{PE}_r$, the "true" prediction error estimated from the rth replicate;
  - $\tilde{PE}$, the overall "true" prediction error.
- $\hat{PE}^{Apparent}$, the apparent prediction error;
  - $\hat{PE}^{Apparent}_{SWLS}$, the apparent prediction error under complex sampling;
- $\hat{PE}^{LOOCV}$, leave-one-out cross-validation (LOOCV) estimated prediction error;
  - $\hat{PE}^{LOOCV}_{SWLS}$, leave-one-out cross-validation (LOOCV) estimated prediction error under complex sampling;
- $\hat{PE}^{BS}$, bootstrap estimated prediction error;
  - $\hat{PE}^{BS}_{SWLS}$, bootstrap estimated prediction error under complex sampling; and
- $\hat{PE}^{.632}$, .632 bootstrap estimated prediction error;
  - $\hat{PE}^{.632}_{SWLS}$, .632 bootstrap estimated prediction error under complex sampling.

Consider the first phase where the "true" PE is estimated and let the R replicate samples, as discussed in section 6.1, denote R learning sets. Let the population size be denoted by N. Now, consider the rth replicate with n_r observations and let it denote the learning set on which the linear model is fitted. It should be pointed out that, since the replicate samples have been selected based on a complex sample design, an SWLS model is fitted to the learning set. The test set thus consists of the remaining $N - n_r$ observations to be predicted by the fitted SWLS model,

$$\underline{\hat{y}} = \mathbf{X}\underline{\hat{\beta}}_{SWLS_r},$$

where $\mathbf{X}$ is the matrix of predictor variables for the observations in the test set and $\underline{\hat{\beta}}_{SWLS_r}$ is the vector of estimated regression coefficients obtained from the learning set. The "true" prediction error is then calculated as

$$\tilde{PE}_r = \frac{1}{N - n_r} \sum_{i=1}^{N-n_r} (y_i - \hat{y}_i)^2 .$$

Note that the sampling weights are not used in the calculation of $\tilde{PE}_r$. It is important to use the sampling weights when fitting a linear model to the learning set since the learning set is a complex sample from the population. However, the test set contains the remainder of the population units that are not included in the learning set. Thus, no sampling weights are in question when calculating $\tilde{PE}_r$ from the test set.

This is repeated for all R replicates and results in R estimates of the "true" prediction error, $\left\{\tilde{PE}_r\right\}$, $r = 1, ..., R$. The overall estimate of the "true" PE can thus be calculated as the average of the R estimates,

$$\tilde{PE} = \frac{1}{R} \sum_r \tilde{PE}_r. \tag{6.2.1}$$

Alternatively, as described in Molinaro et. al (2005), the R estimates of the "true" PE can be seen as R individual PE's, one for each replicate. Both approaches to the estimation of the "true" PE will be considered. Let the first approach to the estimation of the "true" PE, $\tilde{PE}$, be referred to as the Luus approach while the second approach, $\left\{\tilde{PE}_r\right\}$, $r = 1, ..., R$, be referred to as the Molinaro approach. Figure 6.2.1 below presents diagrams of the two approaches such that the difference between them is clear.

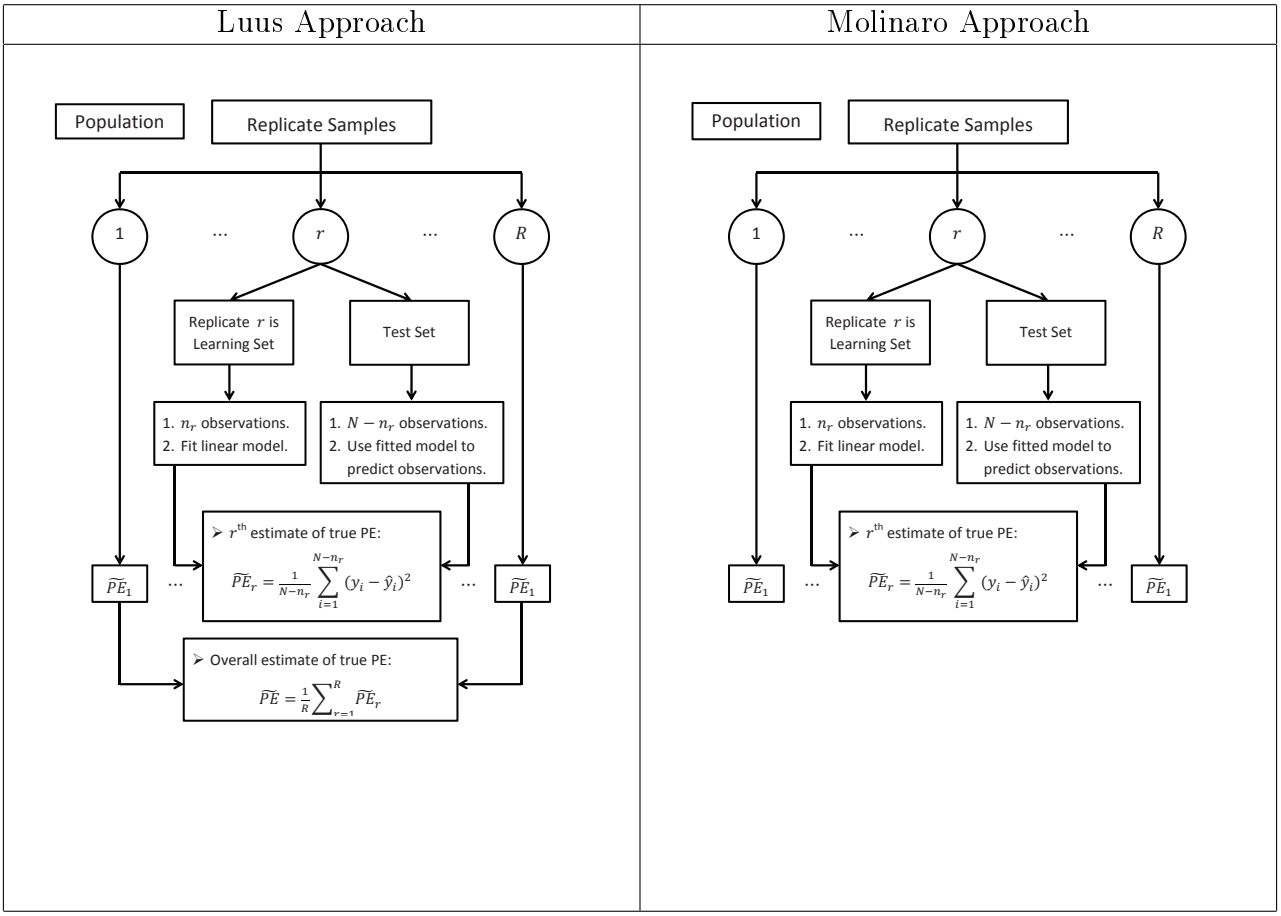


Figure 6.2.1: Luus versus Molinaro True PE Estimation

The replicates have a second purpose in the simulation study, namely as a sample from which the PE can be estimated by the methods discussed in section 4.4.2, namely

1. leave-one-out cross-validation (LOOCV);
2. bootstrap estimation of PE (BS); and
3. .632 bootstrap estimation of PE (.632).

Each of the methods discussed will be evaluated for OLS, WLS and SWLS regression. Furthermore, to determine the effectiveness of weight trimming on estimation, each of the regression methods will be carried out using the untrimmed sampling weights as well as the various trimmed weights. This will be repeated for unbenchmarked as well as benchmarked sampling weights.

Note that all results have been transformed,

$$x^* = \frac{x - x_{min}}{x_{max} - x_{min}},$$

where x represents the analysis output, x_{min} and x_{max} represent, respectively, the minimum and maximum of a given array of output, and x^* denotes the transformed output. This is done to ensure that the results are on the same scale.

Only the results based on WCEC will be presented while ECKZN results are available with the author for perusal.

6.2.2.1 Leave-one-out Cross-Validation Estimator of Prediction Error

The LOOCV method of PE estimation, discussed in section 4.4.2.1, will be applied in each of the R replicate samples selected from the population. Consider the rth replicate to be divided into a learning set and a test set. An OLS, WLS or SWLS linear model is fitted to the test set and dependent on which type is used the application of the LOOCV will differ. The different applications will be explained below for a single repetition of the LOOCV.

- OLS

  1. The n_r observations (SSU's) are split into a learning set, size $n_r - 1$, and a test set, size 1.
  2. An OLS is fitted to the learning set and the model is used to predict the observation in the test set.
  3. Calculate a single LOOCV estimate of the PE as $\widehat{PE}_{OLS_1} = (y_1 - \hat{y}_1)^2$.
  4. Repeat the first 3 steps for all n_r observations, $\left\{\widehat{PE}_{OLS_i}\right\}, i = 1, ..., n_r$.

  The rth OLS estimated PE under LOOCV will be calculated as

$$\widehat{PE}_{OLS_r}^{LOOCV} = \frac{1}{n_r}\sum_i \widehat{PE}_{OLS_i}.$$

- WLS

  1. The LOOCV is applied in each stratum. Suppose the hth stratum of the replicate contains n_{h_r} PSU's. Let the learning set contain $n_{h_r} - 1$ PSU's and the test set a single PSU. Since a PSU has been removed to the test set, the correct approach would be to adjust the sampling weights of the observations in the learning set as one would do when applying the jackknife method. See section 4.3 for an explanation of the jackknife applied to CS data. An assumption is made that statisticians whom apply WLS to CS data might know that the clustering structure needs to be respected and as such might know to leave out a PSU at a time. However, they might not know that the sampling weights of the remaining PSU's need to be adjusted such that the sum of the weights still equal the population total. As such, under WLS the PSU weights will not be adjusted in an attempt to assess the effect of only partially correctly applying LOOCV to CS data. The sampling weights of the observations in the test set will be retained since uninformed statisticians might move the sampling weights with the PSU to the test set.

2. Fit a WLS model to the data in the learning set and use the model to predict the data in the test set.

3. Recall that a PSU in the test set contains n_{h1_r} SSU's and let the observed responses for these SSU's be denoted by y_{h1i_r}, $i = 1, ..., n_{h1}$, $h = 1, ..., H$. The estimated PE after one repetition of the LOOCV will be calculated as

$$\widehat{PE}_{WLS_{(h1)_r}} = \frac{1}{n_{h1_r}} \sum_{i=1}^{n_{h1_r}} (y_{h1i_r} - \hat{y}_{h1i_r})^2 ,$$

where w_{h1i_r} is the sampling weight of the ith observation in the first PSU of stratum h.

4. Repeat the first 3 steps for all n_{h_r} PSU's in all H strata.
The rth WLS estimated PE under LOOCV will be calculated as

$$\widehat{PE}^{LOOCV}_{WLS_r} = \sum_h \frac{1}{n_{h_r}} \sum_j \widehat{PE}_{WLS_{(hj)_r}}.$$

- SWLS

1. Consider the hth stratum of the rth replicate which contains n_{h_r} PSU's. Let the learning set contain $n_{h_r} - 1$ PSU's and the test set a single PSU. Since a PSU has been removed from the stratum the sampling weights of the observations in the remaining PSU's need to be adjusted to compensate for this. It will be done using the weight adjustment proposed under the jackknife,

$$w_{i(hj)_r} = \begin{cases} w_{hji_r},\ i \notin h \\ w_{hji_r} \cdot \frac{n_{h_r}}{(n_{h_r}-1)},\ i \in h,\ i \notin j \\ 0,\ i \in (h,j)_r \end{cases} ,$$

where the (hj) notation indicates that the jth PSU in the hth stratum has been assigned to the learning set. The sampling weights of the SSU's in the PSU allocated to the test set will not be adjusted since the above adjustment accounts for these SSU's.

2. Fit an SWLS model to the data in the learning set and use the model to predict the data in the test set.

3. Recall that the PSU in the test set contains n_{h1_r} SSU's and let the observed responses for these SSU's be denoted by y_{h1i_r}, $i = 1, ..., n_{h1}$, $h = 1, ..., H$. If the predicted responses are denoted by $\{\hat{y}_{h1i_r}\}$, then the estimated PE after one repetition of the LOOCV will be calculated as

$$\widehat{PE}_{SWLS_{(h1)_r}} = \frac{1}{n_{h1_r}} \sum_{i=1}^{n_{h1_r}} (y_{h1i_r} - \hat{y}_{h1i_r})^2 .$$

The rth SWLS estimated PE under LOOCV will be calculated as

$$\widehat{PE}_{SWLS_r}^{LOOCV} = \sum_h \frac{N_h}{N} \sum_j \widehat{PE}_{SWLS_{(hj)_r}},$$

where N_h is the population number of PSU's and N is the total number of PSU's in the population.

Consider the diagram for a summary and comparison of the implementation of this method for OLS and SWLS, per replicate.

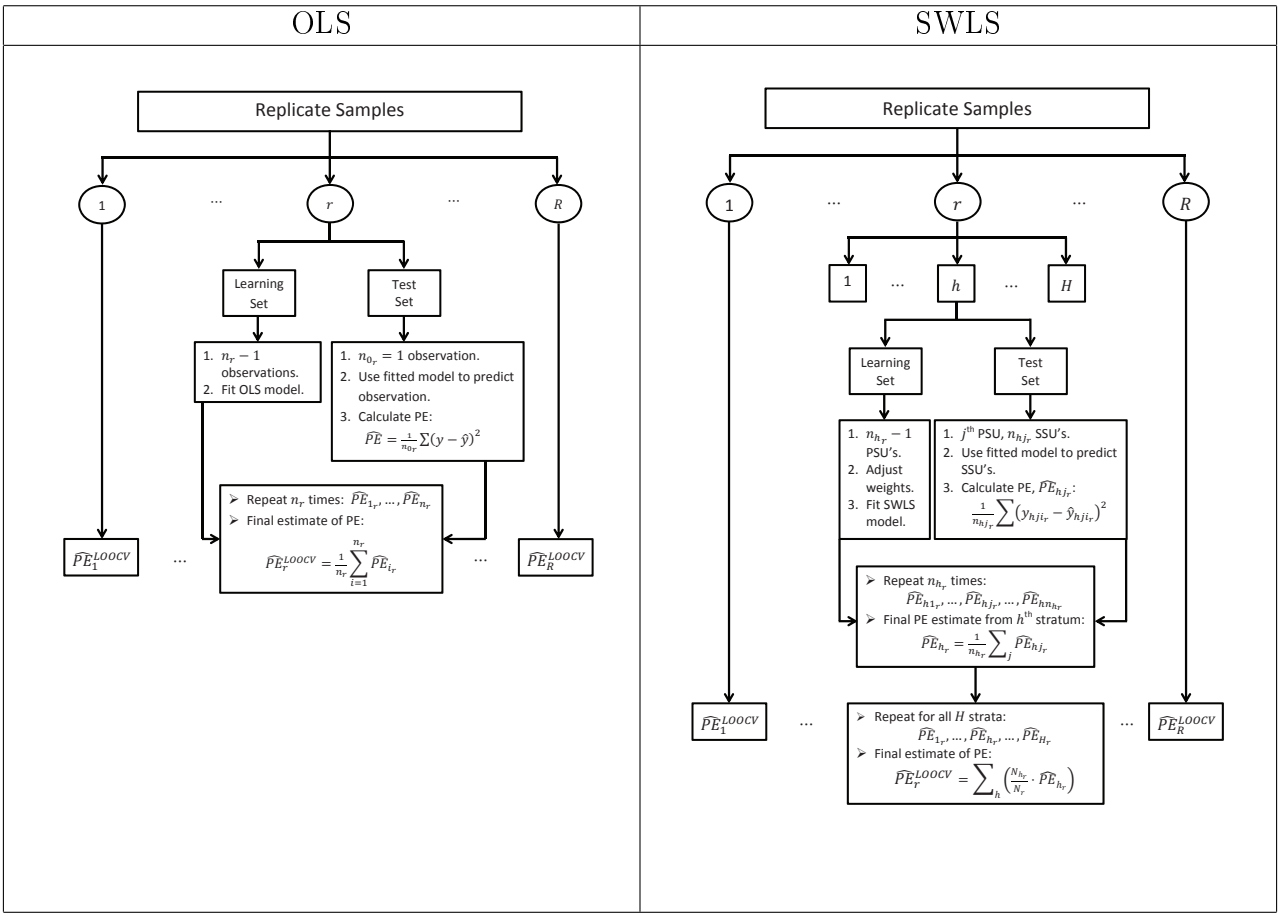


Figure 6.2.2: LOOCV implementation

The R estimates of prediction error, $\hat{PE}_1^{LOOCV}, \ldots, \hat{PE}_R^{LOOCV}$, will be used to calculate the estimated mean squared error, bias and standard deviation of the prediction error estimator of the "true" prediction error. The estimated mean squared error under the Luus approach, $\widehat{MSE}^L$, is given by

$$\widehat{MSE}_{LOOCV}^L\left(\hat{PE}\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{PE}_r^{LOOCV} - \tilde{PE}\right)^2, \tag{6.2.2}$$

where $\tilde{PE}$ is the Luus estimate of the "true" prediction error. Alternatively, the estimated mean squared error under the Molinaro approach, $\widehat{MSE}^{M}$, can be calculated as

$$\widehat{MSE}^{M}_{LOOCV}\left(\hat{PE}\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{PE}_r^{LOOCV} - \tilde{PE}_r\right)^2, \tag{6.2.3}$$

where $\tilde{PE}_r$ is the Molinaro estimated "true" PE of the rth replicate. The estimated bias by the Luus approach, $\widehat{Bias}^{L}$, is calculated as

$$\widehat{Bias}^{L}_{LOOCV}\left(\hat{PE}\right) = \left(\frac{1}{R}\sum_{r=1}^{R}\hat{PE}_r^{LOOCV}\right) - \tilde{PE}, \tag{6.2.4}$$

and by the Molinaro approach, $\widehat{Bias}^{M}$, as

$$\widehat{Bias}^{M}_{LOOCV}\left(\hat{PE}\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{PE}_r^{LOOCV} - \tilde{PE}_r\right). \tag{6.2.5}$$

Note that for bias the Luus and Molinaro approaches give the same result.

Consider the LOOCV PE estimation results below.

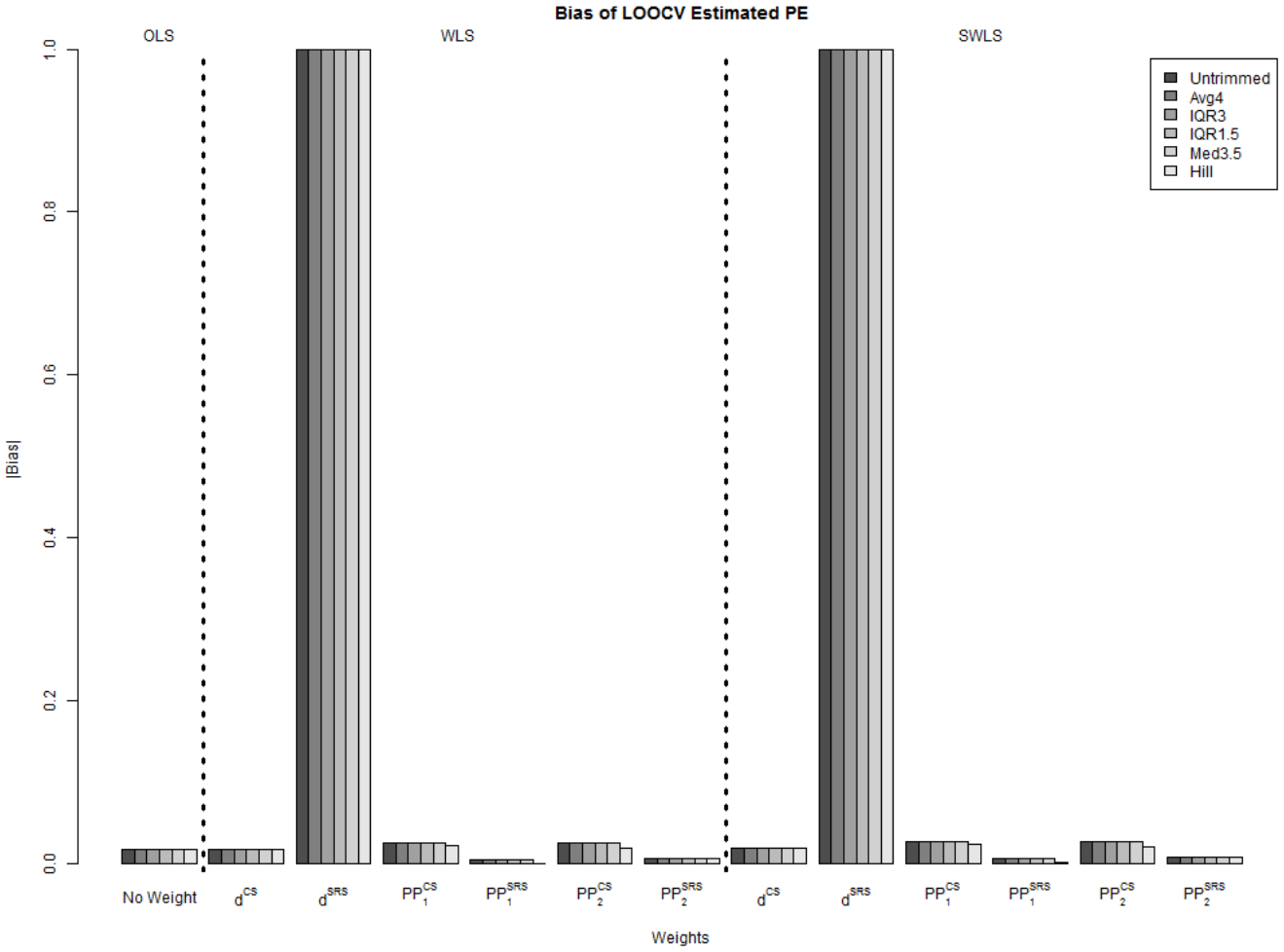


Figure 6.2.3: "True" Bias of LOOCV Estimated PE: Luus approach

Figure 6.2.3 shows that the inclusion of sampling weights (WLS and SWLS) slightly improves the fitted model's prediction error, especially once the weights have been benchmarked ($w_{CS}^{pp_1}$, $w_{SRS}^{pp_1}$, $w_{CS}^{pp_2}$ and $w_{SRS}^{pp_2}$). The alternative design weights, d_{SRS}, clearly increases the estimated prediction error. However, once these have been benchmarked, the associated estimated prediction errors appear slightly smaller than the estimated prediction errors based on the benchmarked theoretical design weights. With regards to the weight trimming methods, the Hill trimmed sampling weights resulted in a smaller estimated prediction error throughout.

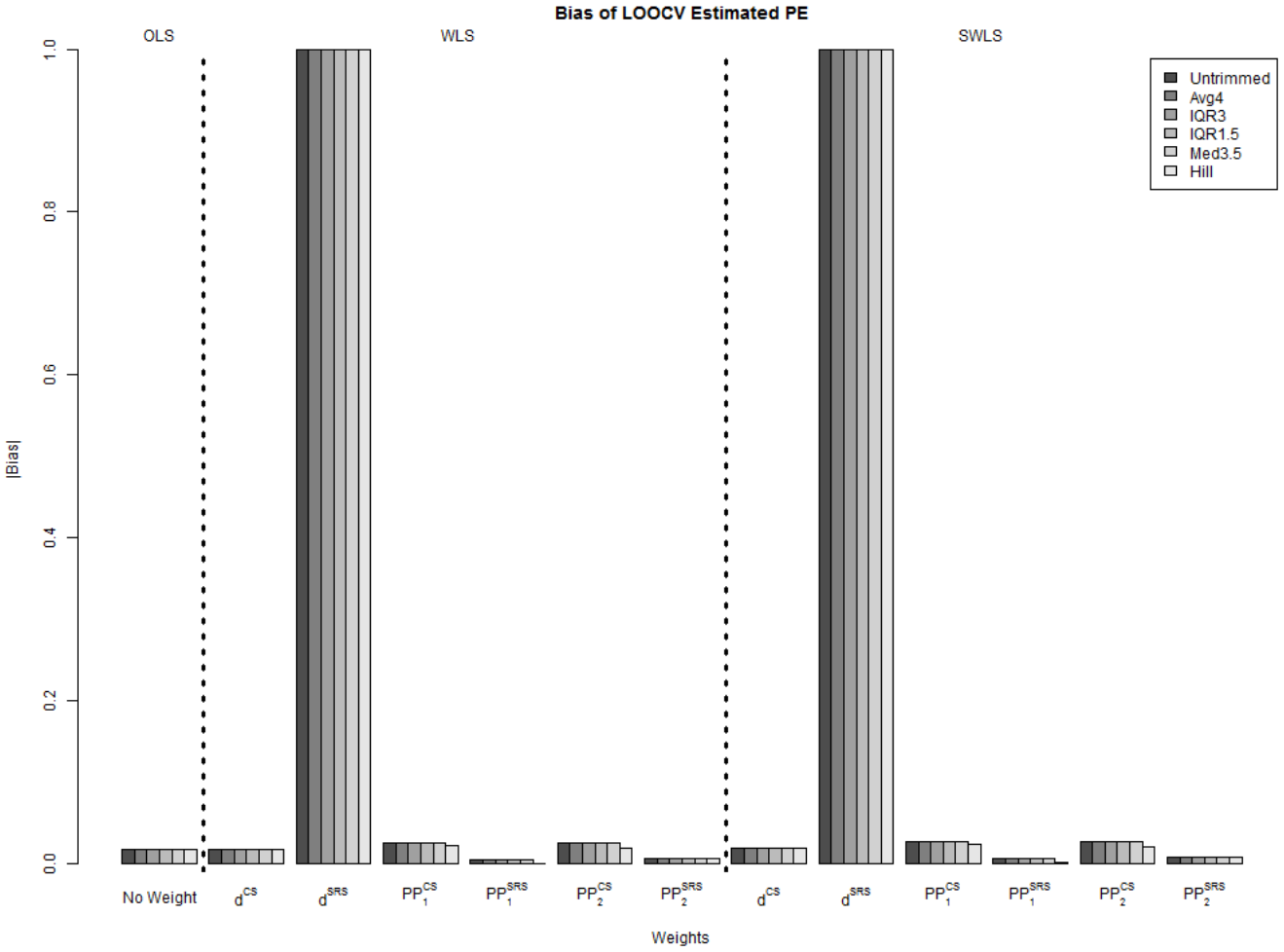


Figure 6.2.4: "True" Bias of LOOCV Estimated PE: Molinaro approach

Figure 6.2.3 was based on the Luus approach to the "true" prediction error. However, for the bias, both the Luus and Molinaro approaches give the same result as is clear from figure 6.2.4.

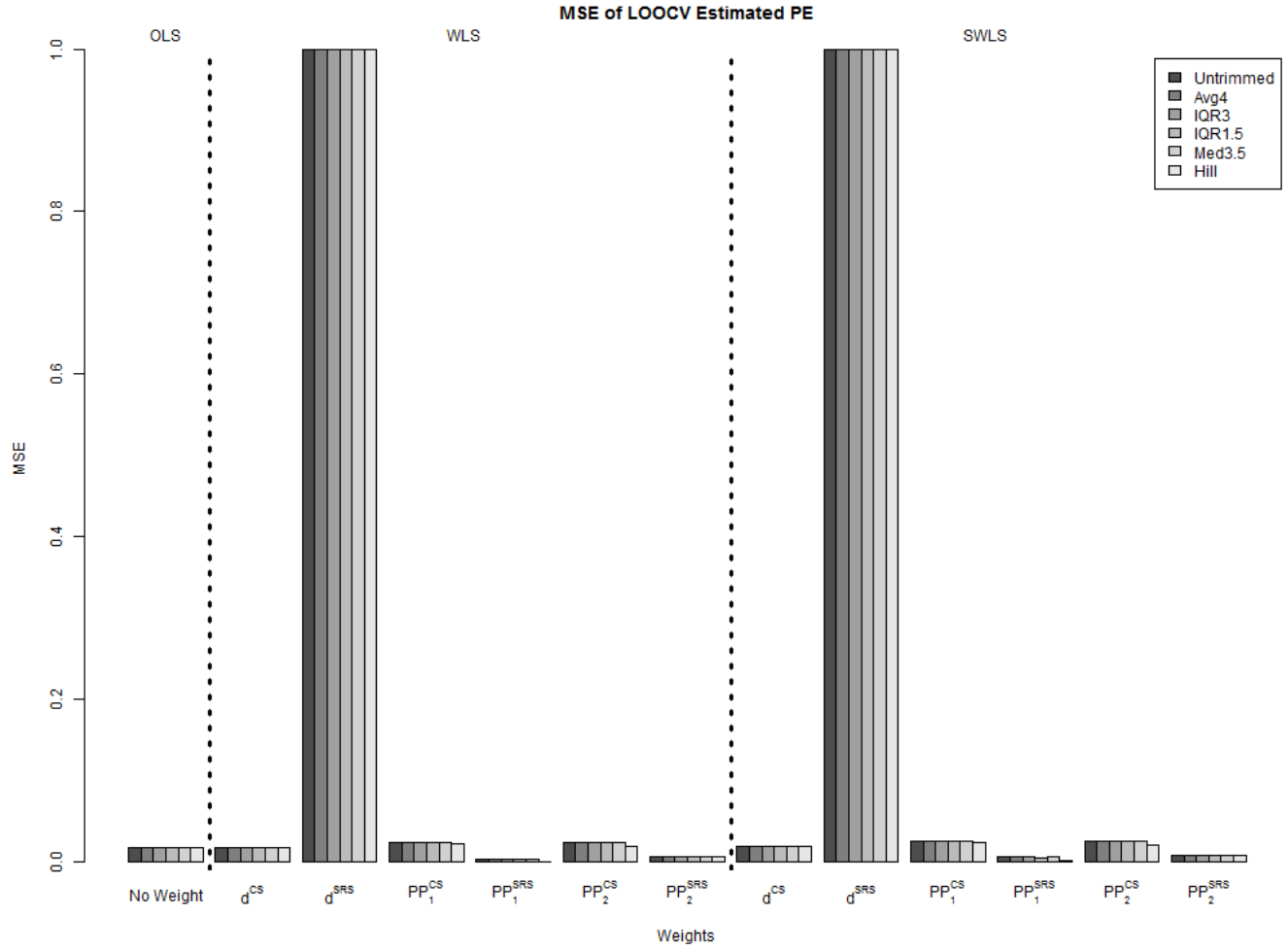


Figure 6.2.5: "True" MSE of LOOCV Estimated PE: Luus approach

In figure 6.2.5 similar conclusions about the "true" MSE, based on the Luus approach, can be made as for "true" bias in figure 6.2.3.

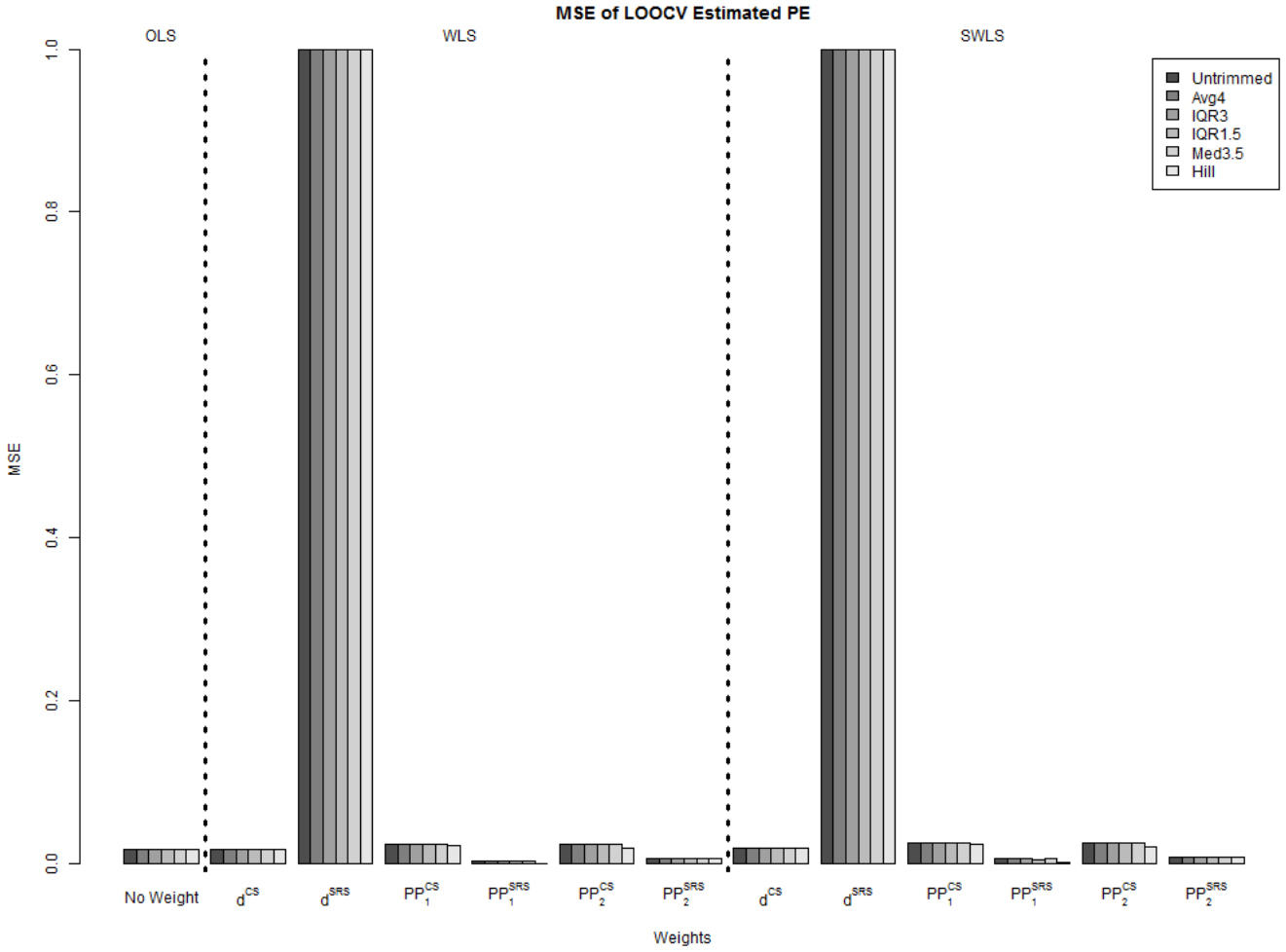


Figure 6.2.6: "True" MSE of LOOCV Estimated PE: Molinaro approach

Although the Luus approach and Molinaro approach give different MSE's, figures 6.2.5 and 6.2.6 are very similar. The use of sampling weights in the linear models improve the estimated prediction errors of the models. Furthermore, the benchmarked alternative weights, i.e. $w_{SRS}^{pp_1}$ and $w_{SRS}^{pp_2}$, result in the smallest estimated prediction errors, especially based on the Hill trimmed weights.

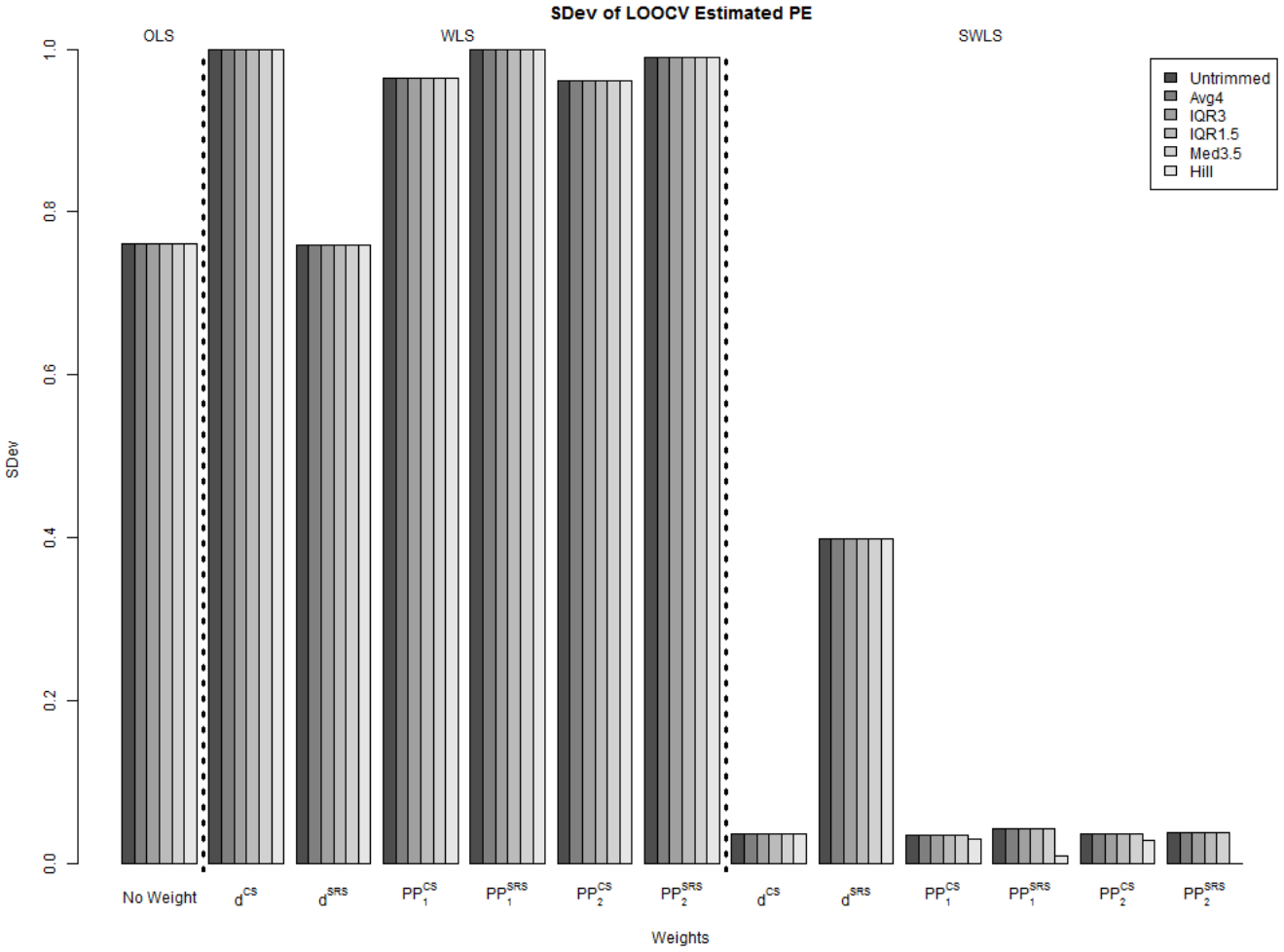


Figure 6.2.7: Estimated Standard Deviation of LOOCV Estimated PE: Luus approach

In figure 6.2.7, it is clear that the SWLS prediction errors vary less than the OLS and WLS prediction errors. This is the case for both the design weights as well as their benchmarked weights, especially under the Hill trimmed benchmarked weights.

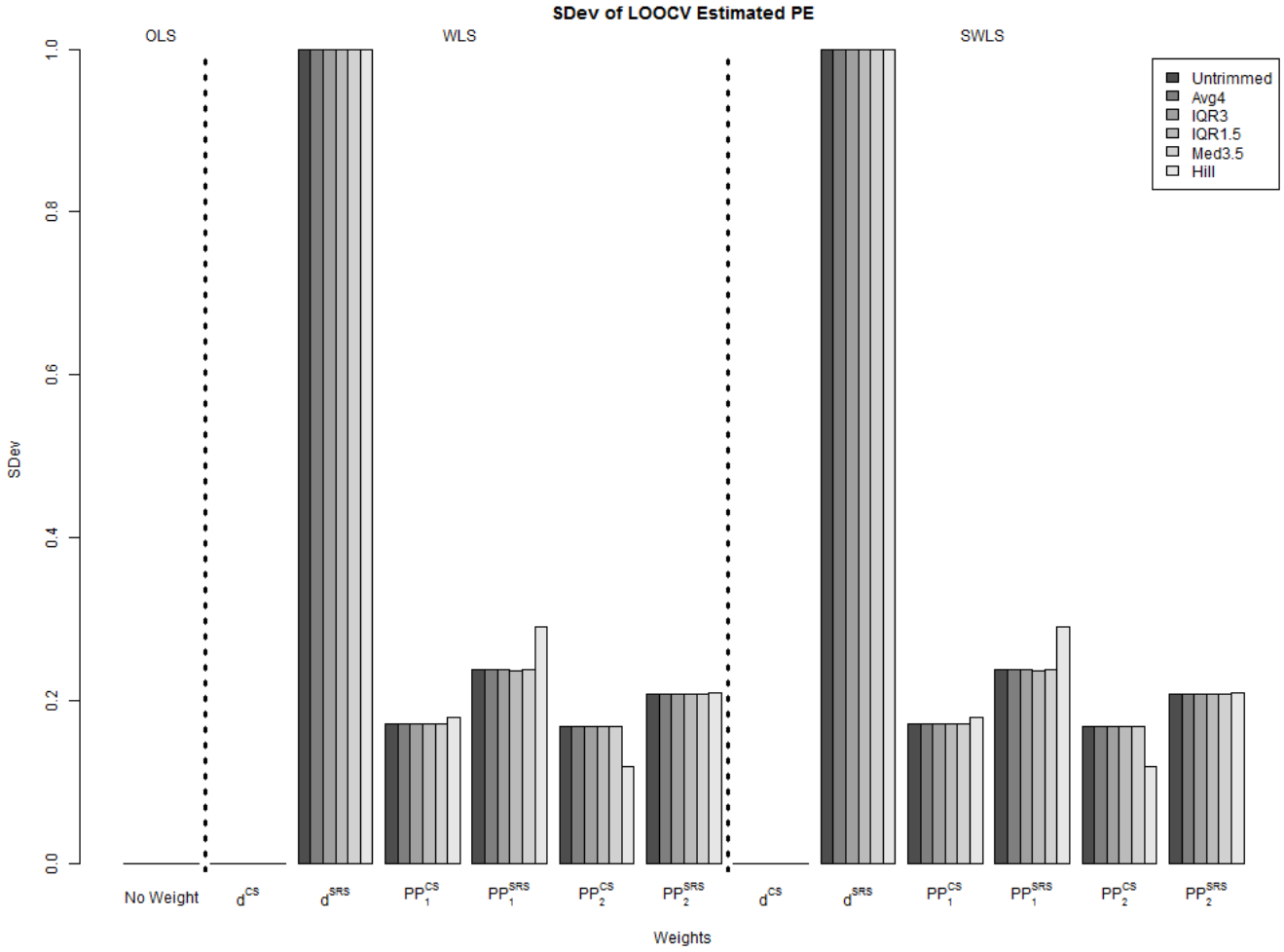


Figure 6.2.8: Estimated Standard Deviation of LOOCV Estimated PE: Molinaro approach

Figure 6.2.8 presents similar conclusions to figure 6.2.7, but here the smallest standard deviation under SWLS is observed under the Hill trimmed $w_{CS}^{pp_2}$.

6.2.2.2 Bootstrap Estimation of Prediction Error

The bootstrap estimation of PE was discussed in section 4.4.2.2. Now, consider the rth replicate sample from which $B = 200$ samples were selected. The implementation of OLS, WLS and SWLS linear models in this PE estimation approach will be described below.

- OLS

  1. The rth replicate contains n_r observations. Fit an OLS and calculate the apparent PE, $\hat{PE}_{OLS_r}^{Apparent}$.
  2. Draw B bootstrap samples from r. Consider the r_bth bootstrap sample with n_b observations.

(a) Fit an OLS to the bootstrap sample and use the model to predict the responses of the replicate to be used to calculate the simple PE estimate, $\hat{PE}_{r_b}^{simple}$.

(b) Use the same model to estimate the responses of the bootstrap sample and use these to calculate the improved PE estimate, $\hat{PE}_{r_b}^{improved}$.

(c) Calculate the difference, $Diff_{r_b} = \hat{PE}_{r_b}^{simple} - \hat{PE}_{r_b}^{improved}$.

3. Repeat (a) to (c) for all B bootstrap samples and use the B differences to calculate the optimism, $Optimism_{OLS_r} = \frac{1}{B}\sum_{r_b} Diff_{r_b}$.

4. The bootstrap estimated PE for the rth replicate is then calculated as $\hat{PE}_{OLS_r}^{BS} = \hat{PE}_{OLS_r}^{Apparent} + Optimism_{OLS_r}$.

See the diagram on the left in figure 6.2.9 as a summary of this application.

- WLS

1. The rth replicate contains n_r observations. Fit a WLS and calculate the apparent PE, $\hat{PE}_{WLS_r}^{Apparent} = \frac{1}{n_r}\sum_i (y_{i_r} - \hat{y}_{i_r})^2$.

2. The BS is applied in each stratum. Suppose the hth stratum contains n_{h_r} PSU's. The same assumption is made here as under LOOCV, namely that statisticians whom apply WLS to CS data might know that the clustering structure needs to be respected. Hence, they might know to sample the PSU's with-replacement for the bootstrap sample, but they might not know that the sampling weights of the remaining PSU's need to be adjusted such that the sum of the weights still equal the population total. See section 4.3 for an explanation of the bootstrap applied to CS data. Hence, here too the PSU weights will not be adjusted in an attempt to assess the effect of only partially correctly applying BS to CS data.

   Use the same B bootstrap samples as before and consider the r_bth bootstrap sample.

   (a) Fit a WLS to the bootstrap sample and use the model to predict the responses of the replicate to be used to calculate the simple PE estimate, $\hat{PE}_{r_b}^{simple}$.

   (b) Use the same model to estimate the responses of the bootstrap sample and use these to calculate the improved PE estimate, $\hat{PE}_{r_b}^{improved}$.

   (c) Calculate the difference, $Diff_{r_b} = \hat{PE}_{r_b}^{simple} - \hat{PE}_{r_b}^{improved}$.

3. Repeat (a) to (c) for all B bootstrap samples and use the B differences to calculate the optimism, $Optimism_{WLS_r} = \frac{1}{B}\sum_{r_b} Diff_{r_b}$.

4. The bootstrap estimated PE for the rth replicate under WLS is then calculated as $\hat{PE}_{WLS_r}^{BS} = \hat{PE}_{WLS_r}^{Apparent} + Optimism_{WLS_r}$.

- SWLS

1. The rth replicate contains n_r observations. Fit an SWLS and calculate the apparent PE, $\hat{PE}^{Apparent}_{SWLS_r}$.

2. See section 4.3 for an explanation of the bootstrap applied to CS data. Here, as opposed to the WLS implementation, the PSU weights will be adjusted as they should be. Define $m^*_{r_{hj}}$ to be the number of times the jth PSU in stratum h is included in the bootstrap sample. The bootstrap sampling weight of the ith SSU in the jth PSU in the hth stratum is then calculated as

$$w^*_{r_{hji}} = w_{r_{hji}} \cdot \left[\left(\frac{n_{r_h}}{n_{r_h} - 1} \right) \cdot m^*_{r_{hj}} \right],$$

   where $w_{r_{hji}}$ is the original sampling weight of the ith observation in the jth PSU in the hth stratum, $h = 1, ..., H$, $j = 1, ..., n_{r_h}$ and $i = 1, ..., n_{r_{hj}}$.

   Using the same B bootstrap samples as before, consider the r_bth bootstrap sample.

   (a) Fit an SWLS to the bootstrap sample and use the model to predict the responses of the replicate to be used to calculate the simple PE estimate, $\hat{PE}^{simple}_{r_b}$.

   (b) Use the same model to estimate the responses of the bootstrap sample and use these to calculate the improved PE estimate, $\hat{PE}^{improved}_{r_b}$.

   (c) Calculate the difference, $Diff_{r_b} = \hat{PE}^{simple}_{r_b} - \hat{PE}^{improved}_{r_b}$.

3. Repeat (a) to (c) for all B bootstrap samples and use the B differences to calculate the optimism, $Optimism_{SWLS_r} = \frac{1}{B} \sum_{r_b} Diff_{r_b}$.

4. The bootstrap estimated PE for the rth replicate under SWLS is then calculated as $\hat{PE}^{BS}_{SWLS_r} = \hat{PE}^{Apparent}_{SWLS_r} + Optimism_{SWLS_r}$.

See the diagram on the right in figure 6.2.9 as a summary of this application.

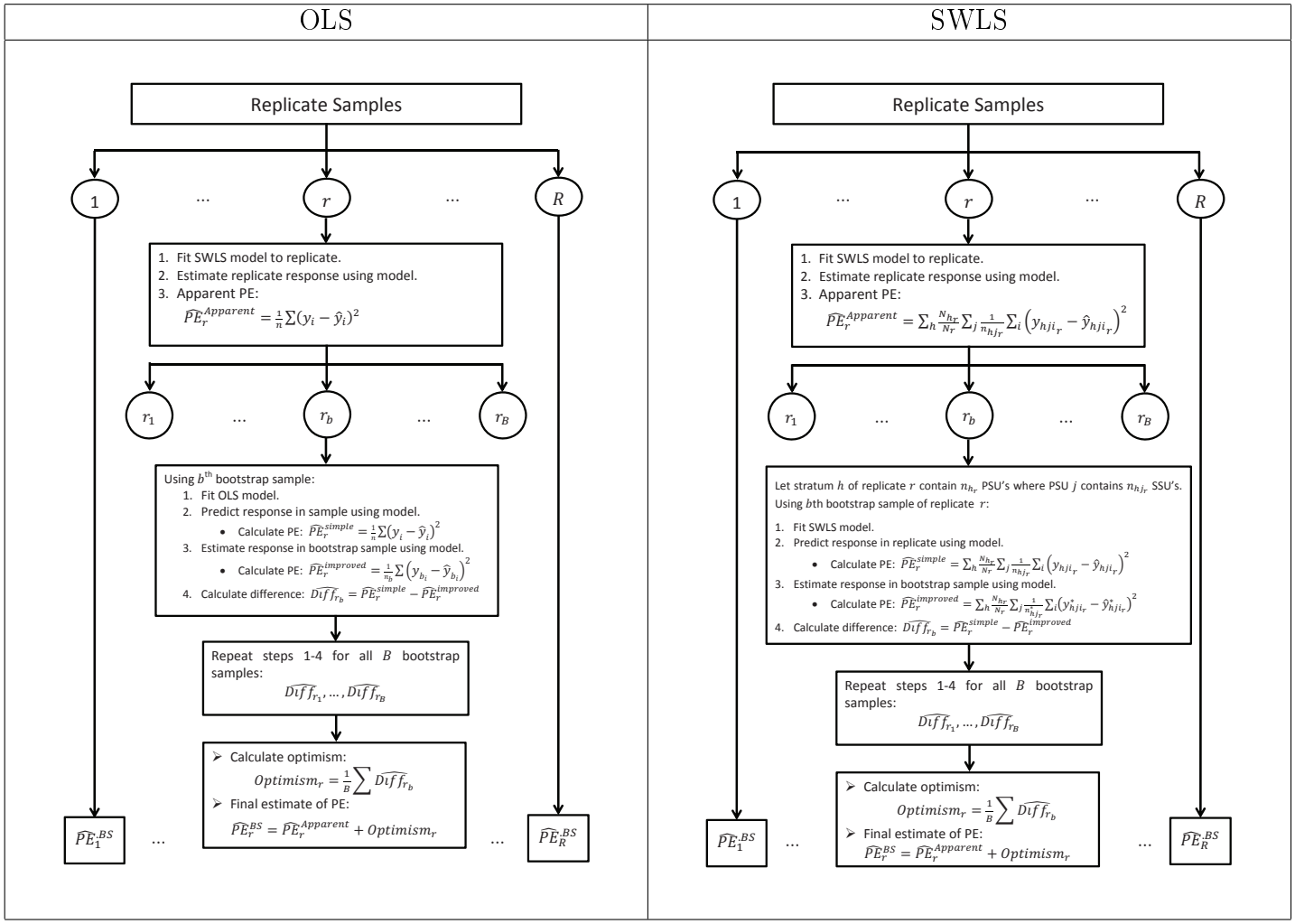


Figure 6.2.9: BS implementation

The R estimates of prediction error, $\hat{PE}_1^{BS}, ..., \hat{PE}_R^{BS}$, will be used to calculate the estimated mean squared error, bias and standard deviation of the prediction error estimator of the "true" prediction error in the same way as outlined above for LOOCV. The resultant diagnostic measures will be $\widehat{MSE}_{BS}^{L}\left(\hat{PE}\right)$, $\widehat{MSE}_{BS}^{M}\left(\hat{PE}\right)$, $\widehat{Bias}_{BS}^{L}\left(\hat{PE}\right)$, and $\widehat{Bias}_{BS}^{M}\left(\hat{PE}\right)$. These results are given in the figures below.

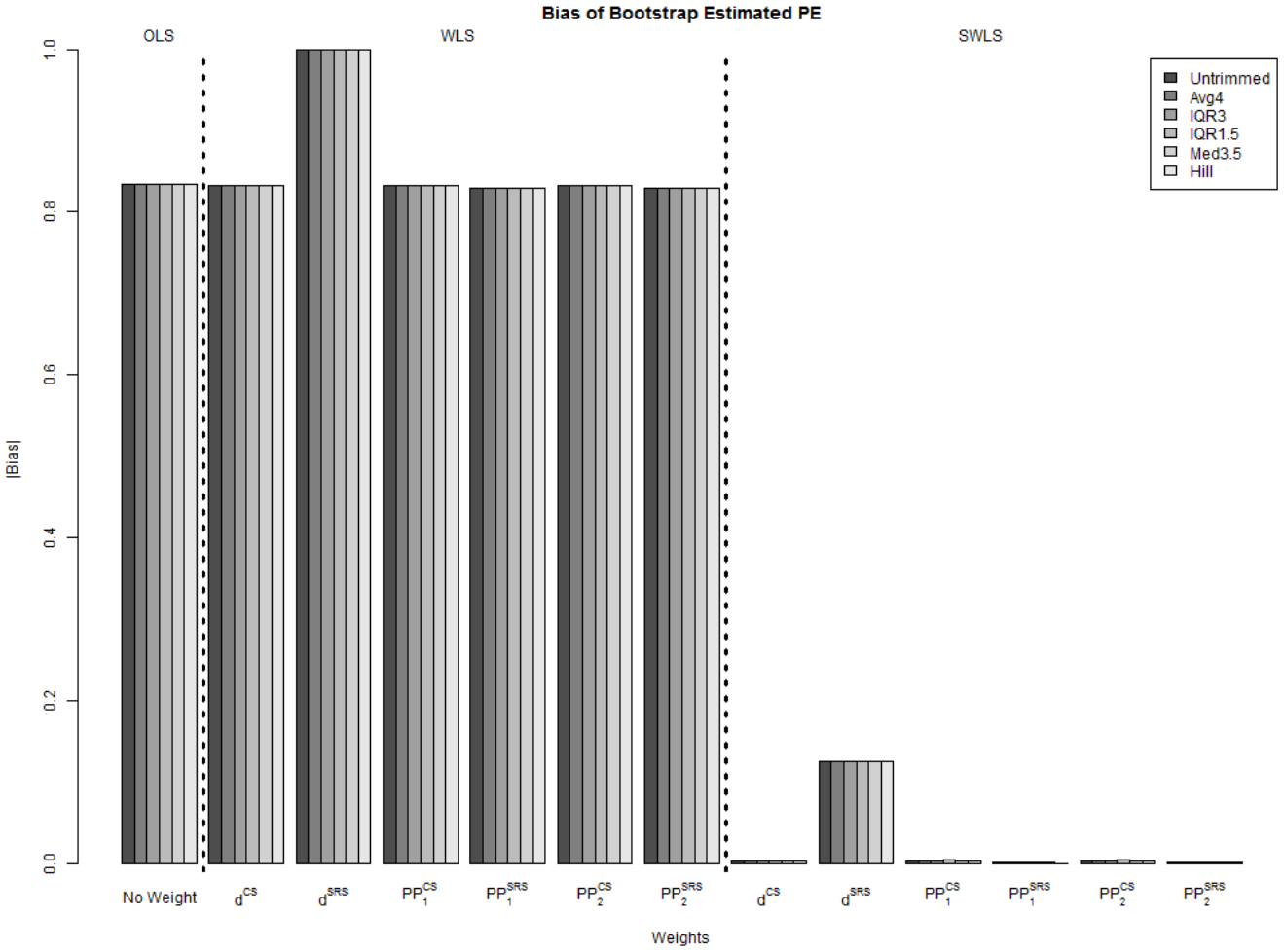


Figure 6.2.10: "True" Bias of Bootstrap Estimated PE: Luus approach

From figure 6.2.10 it is clear that SWLS bootstrap estimated prediction error outperformed the corresponding OLS and WLS estimated prediction errors. It is furthermore clear, for SWLS, that the alternative design weights, d_{SRS}, do not perform well in comparison to the theoretical design weigths, d_{CS}, or the associated benchmarked weights. Here the trimming methods do not appear to have a significant effect on the estimated prediction errors in comparison to the estimated prediction errors based on the untrimmed weights. The "true" bias based on the Molinaro approach is not given here since the result is the same as the result given here.

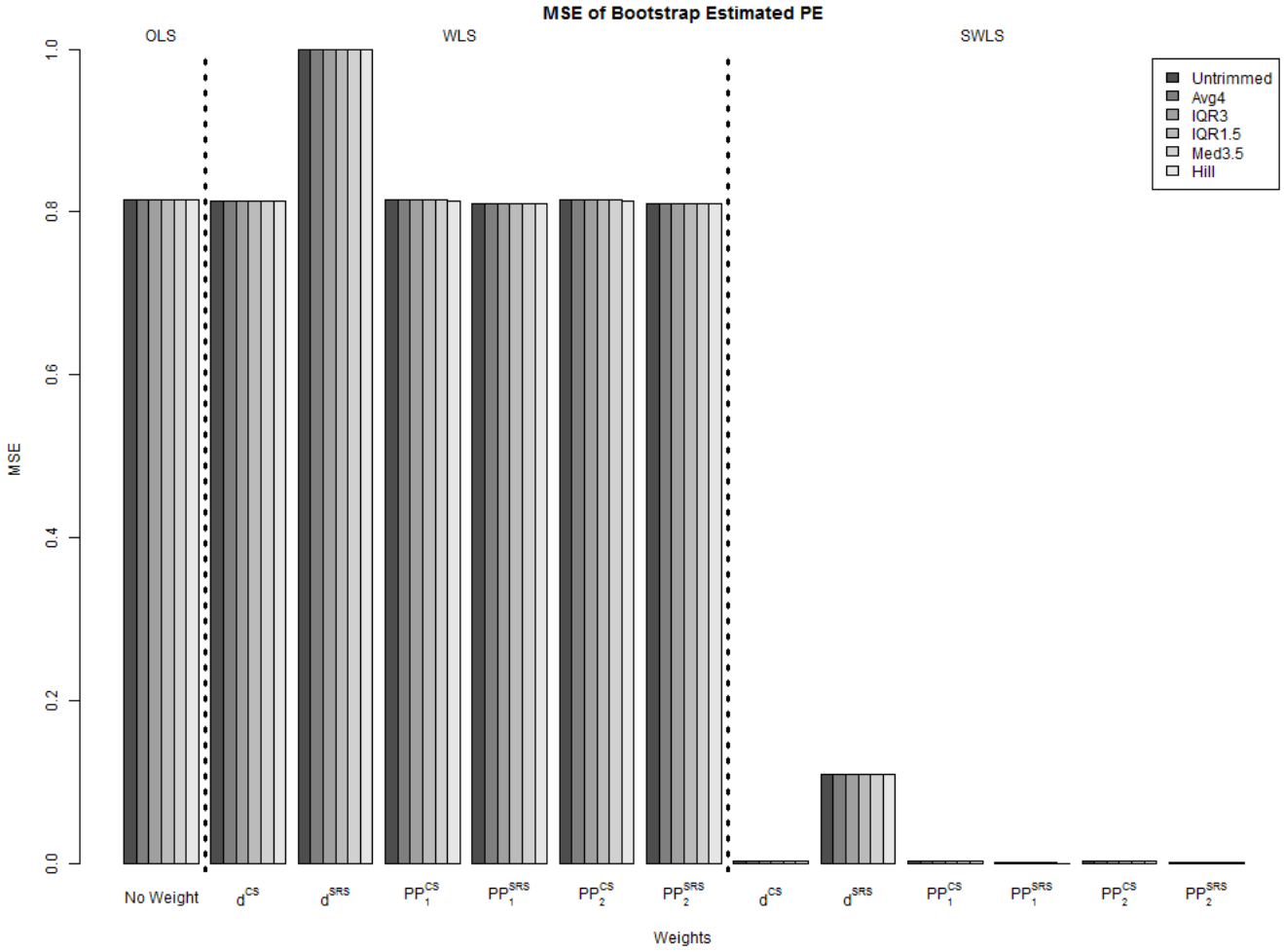


Figure 6.2.11: "True" MSE of Bootstrap Estimated PE: Luus approach

Very small differences between the "true" bias, given figure 6.2.10, and "true" MSE, given in figure 6.2.11, are observed. Note that mostly the "true" MSE based on the Luus approach concurs with the conclusions made from figure 6.2.10. The "true" MSE based on the Molinaro approach is very similar to the "true" MSE presented in figure 6.2.11 and thus is not shown here.

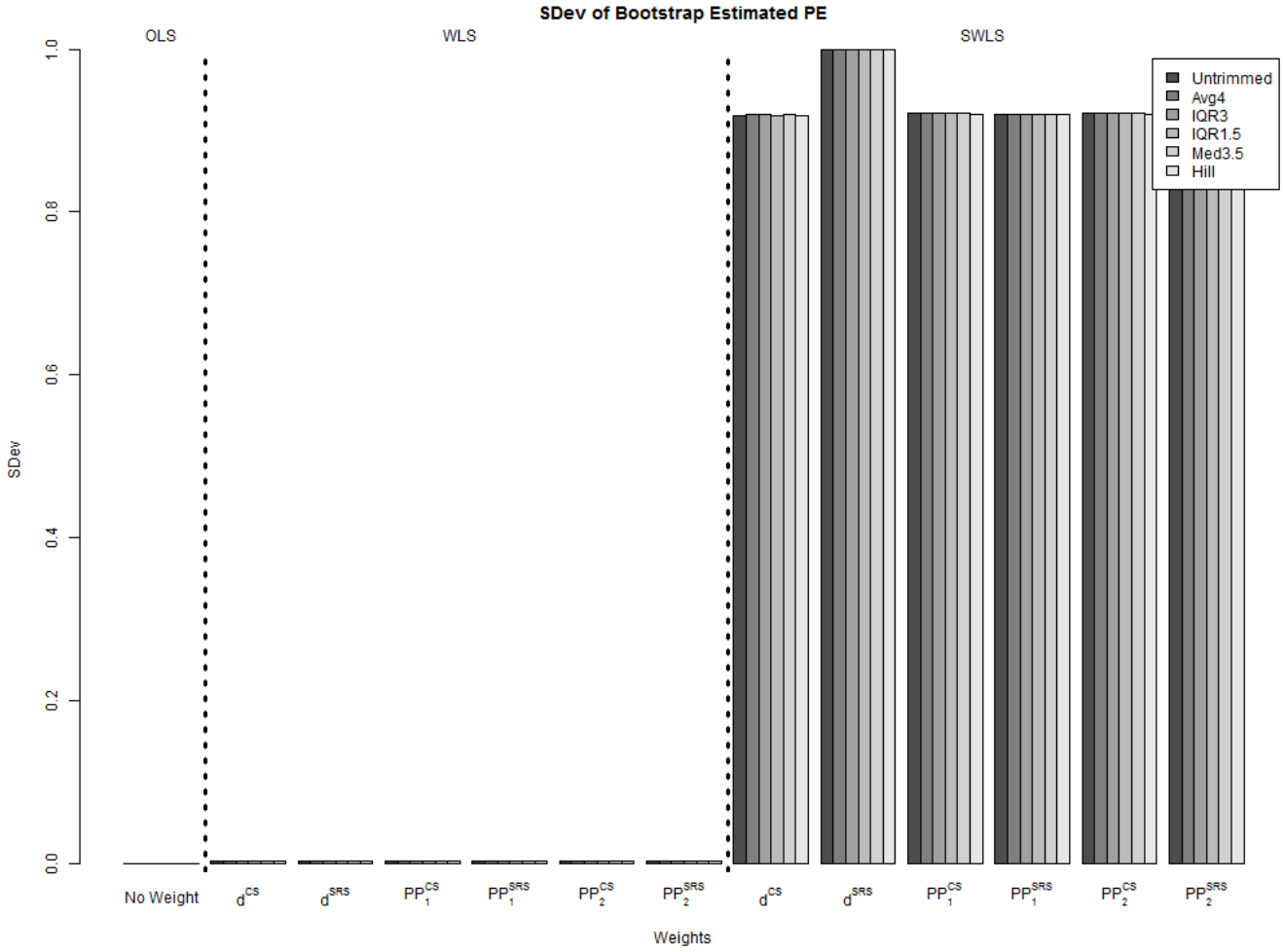


Figure 6.2.12: Estimated Standard Deviation of Bootstrap Estimated PE: Luus approach

Clearly, as seen from figure 6.2.12, the variability in the SWLS estimated prediction errors is much larger than for OLS and WLS. A possible reason that could be given for this is that the SWLS estimated prediction error is much more sensitive to the number of bootstrap samples used in the simulation study than the OLS and WLS results. The estimated standard error based on the Molinaro approach looks similar to the figure presented here.

6.2.2.3 .632 Bootstrap Estimation of Prediction Error

Consider the rth replicate sample from which $B = 200$ samples were selected. The implementation of OLS, WLS and SWLS models in this PE estimation approach will be described below.

- OLS

  1. The rth replicate contains n_r observations. Fit an OLS and calculate the apparent PE, $\hat{PE}_{OLS_r}^{Apparent}$.

2. Consider the ith observation of replicate sample r.

   (a) Determine the bootstrap samples that do not contain this observation and let this number be denoted by r_{B_i}.
   Consider sample r_{b_i} of these r_{B_i} bootstrap samples.

   i. Fit an OLS model to bootstrap sample r_{b_i}.

   ii. Use this model to predict the ith observation.

   iii. Calculate $\hat{PE}_{r_{b_i}} = (y_{r_i} - \hat{y}_{r_i})^2$.

   (b) Repeat steps (i) - (ii) for $r_{b_i} = 1, ..., r_{B_i}$.

   (c) Calculate the prediction error of the ith observation as

$$\hat{PE}_{OLS_{r_i}} = \frac{1}{r_{B_i}} \sum_{r_{b_i}} \hat{PE}_{r_{b_i}}.$$

   .

3. Repeat step 2 for all n_r observations and obtain $\hat{PE}_{OLS_{r_1}}, ..., \hat{PE}_{OLS_{r_n}}$.

4. The average estimated error rate is then calculated as

$$\hat{\varepsilon}_{OLS_{r_0}} = \frac{1}{n_r} \sum_i \hat{PE}_{OLS_{r_i}},$$

   and the .632 estimate of optimism is given by

$$Optimism^{.632}_{OLS_r} = 0.632 \left[\hat{\varepsilon}_{OLS_{r_0}} - \hat{PE}^{Apparent}_{OLS_r} \right].$$

5. Finally, the .632 bootstrap estimated prediction error is calculated as

$$\hat{PE}^{.632}_{OLS_r} = \hat{PE}^{Apparent}_{OLS_r} + Optimism^{.632}_{OLS_r}.$$

See the diagram on the left in figure 6.2.13 as a summary of this application.

- WLS

1. The rth replicate contains n_r observations. Fit a WLS and calculate the apparent PE, $\hat{PE}^{Apparent}_{WLS_r}$.

   The same assumption regarding the handling of the PSU's and sampling weight adjustment, as described under bootstrap PE estimation, is made here. Recall that the hth stratum of the rth replicate contains n_{r_h} PSU's.

2. Consider the jth PSU in the hth stratum of replicate sample r.

(a) Determine the bootstrap samples that do not contain this observation and let this number be denoted by $r_{B_{hj}}$.

Consider sample $r_{b_{hj}}$ of these $r_{B_{hj}}$ bootstrap samples.

i. Fit a WLS model to bootstrap sample $r_{b_{hj}}$.

ii. Suppose that the (hj)th PSU contains $n_{r_{hj}}$ observations. Use this model to predict the observations in the PSU.

iii. Calculate $\hat{PE}_{r_{b_{hj}}} = \frac{1}{n_{r_{hj}}} \sum_i \left(y_{r_{hji}} - \hat{y}_{r_{hji}}\right)^2$.

(b) Repeat steps (i) - (ii) for $b_{r_{b_{hj}}} = 1, ..., r_{B_{hj}}$.

(c) Calculate the prediction error of the hjth PSU as

$$\hat{PE}_{WLS_{r_{hj}}} = \frac{1}{r_{B_{hj}}} \sum_{r_{b_{hj}}} \hat{PE}_{r_{b_{hj}}}.$$

.

3. Repeat step 2 for all PSU's in all strata.

4. Calculate the average estimated error rate for stratum h,

$$\hat{\varepsilon}_{r_{h_0}} = \frac{1}{n_{r_{hj}}} \sum_j \hat{PE}_{WLS_{r_{hj}}}.$$

5. The overall average estimated error rate is then calculated as

$$\hat{\varepsilon}_{WLS_{r_0}} = \sum_h \frac{N_{r_h}}{N_r} \cdot \hat{\varepsilon}_{r_{h_0}},$$

and the .632 estimate of optimism is given by

$$Optimism^{.632}_{WLS_r} = 0.632 \left[\hat{\varepsilon}_{WLS_{r_0}} - \hat{PE}^{Apparent}_{WLS_r}\right].$$

6. Finally, the .632 bootstrap estimated prediction error under WLS is calculated as

$$\hat{PE}^{.632}_{WLS_r} = \hat{PE}^{Apparent}_{WLS_r} + Optimism^{.632}_{WLS_r}.$$

- SWLS

1. Fit an SWLS to replicate r and calculate the apparent PE, $\hat{PE}^{Apparent}_{SWLS_r}$.

See section 4.3 for an explanation of the bootstrap applied to CS data. Here, as opposed to the WLS implementation, the PSU weights will be adjusted as they should be. Define $m^*_{r_{hj}}$ to be the number of times the jth PSU in stratum h is included in the bootstrap

sample. The bootstrap sampling weight of the ith SSU in the jth PSU in the hth stratum is then calculated as

$$w^*_{r_{hji}} = w_{r_{hji}} \cdot \left[\left(\frac{n_{r_h}}{n_{r_h} - 1} \right) \cdot m^*_{r_{hj}} \right],$$

where $w_{r_{hji}}$ is the original sampling weight of the ith observation in the jth PSU in the hth stratum, $h = 1, ..., H$, $j = 1, ..., n_{r_h}$ and $i = 1, ..., n_{r_{hj}}$.

2. Consider the jth PSU in the hth stratum of replicate sample r.

   (a) Determine the bootstrap samples that do not contain this observation and let this number be denoted by $r_{B_{hj}}$.
   Consider sample $r_{b_{hj}}$ of these $r_{B_{hj}}$ bootstrap samples.

      i. Fit an SWLS model to bootstrap sample $r_{b_{hj}}$.

      ii. Suppose that the (hj)th PSU contains $n_{r_{hj}}$ observations. Use this model to predict the observations in the PSU.

      iii. Calculate $\hat{PE}_{r_{b_{hj}}} = \frac{1}{n_{r_{hj}}} \sum_i \left(y_{r_{hji}} - \hat{y}_{r_{hji}} \right)^2$.

   (b) Repeat steps (i) - (ii) for $r_{b_{hj}} = 1, ..., r_{B_{hj}}$.

   (c) Calculate the prediction error of the hjth PSU as

$$\hat{PE}_{SWLS_{r_{hj}}} = \frac{1}{r_{B_{hj}}} \sum_{r_{b_{hj}}} \hat{PE}_{r_{b_{hj}}}.$$

   .

3. Repeat step 2 for all PSU's in all strata.

4. Calculate the average estimated error rate for stratum h,

$$\hat{\varepsilon}_{r_{h_0}} = \frac{1}{n_{r_{hj}}} \sum_j \hat{PE}_{SWLS_{r_{hj}}}.$$

5. The overall average estimated error rate is then calculated as

$$\hat{\varepsilon}_{SWLS_{r_0}} = \sum_h \frac{N_{r_h}}{N_r} \cdot \hat{\varepsilon}_{r_{h_0}},$$

   and the .632 estimate of optimism is given by

$$Optimism^{.632}_{SWLS_r} = 0.632 \left[\hat{\varepsilon}_{SWLS_{r_0}} - \hat{PE}^{Apparent}_{SWLS_r} \right].$$

6. Finally, the .632 bootstrap estimated prediction error under SWLS is calculated as

$$\hat{PE}^{.632}_{SWLS_r} = \hat{PE}^{Apparent}_{SWLS_r} + Optimism^{.632}_{SWLS_r}.$$

See the diagram on the right in figure 6.2.13 as a summary of this application.

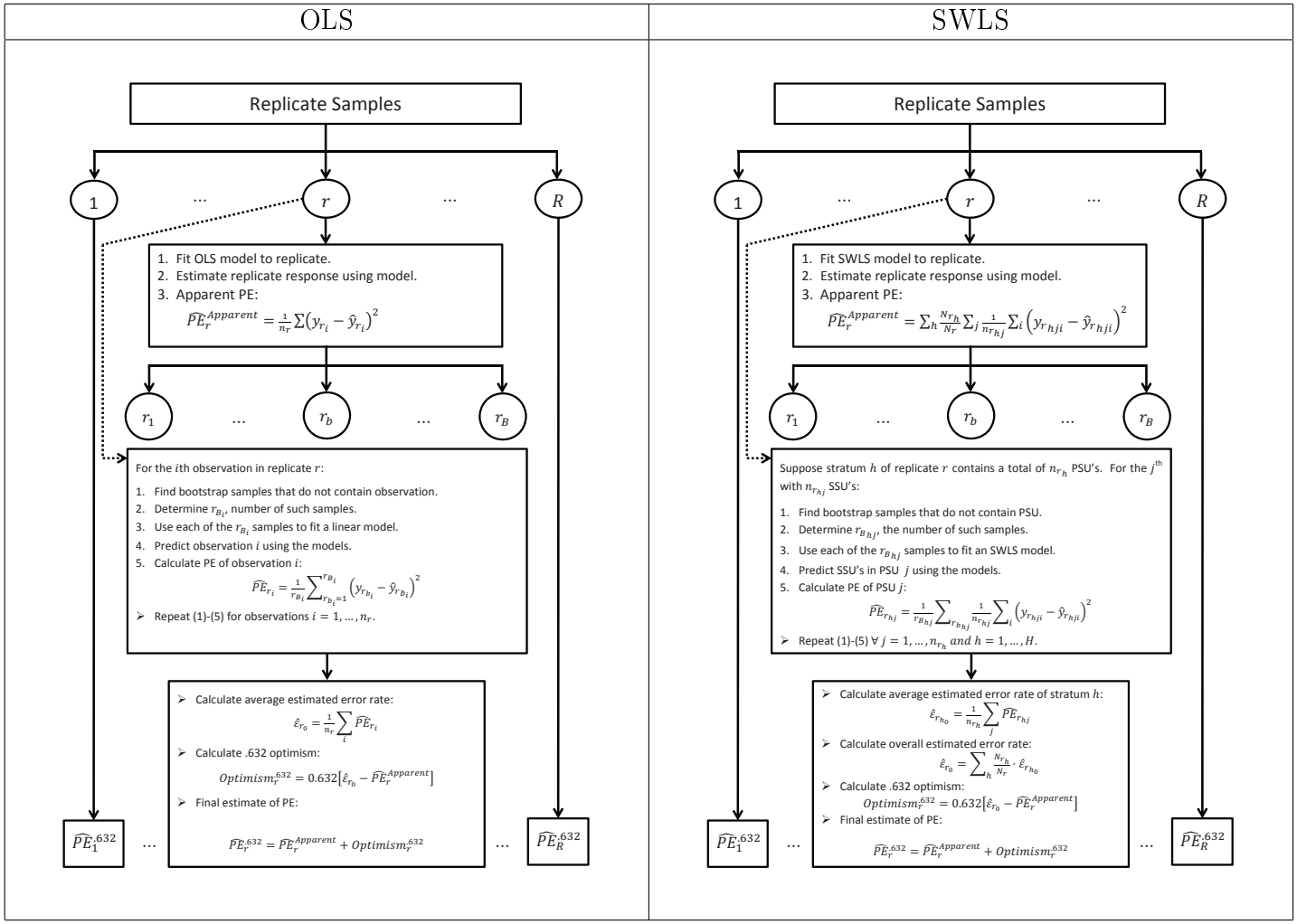


Figure 6.2.13: .632 implementation

As before the R estimates of prediction error, $\hat{PE}_1^{.632}, ..., \hat{PE}_R^{.632}$, obtained for OLS, WLS and SWLS, will be used to calculate the estimated mean squared error, bias and standard deviation of the prediction error estimator of the "true" prediction error. The resultant diagnostic measures will be $\widehat{MSE}^{L}_{.632}\left(\hat{PE}\right)$, $\widehat{MSE}^{M}_{.632}\left(\hat{PE}\right)$, $\widehat{Bias}^{L}_{.632}\left(\hat{PE}\right)$, and $\widehat{Bias}^{M}_{.632}\left(\hat{PE}\right)$. The results are given in the below figures.

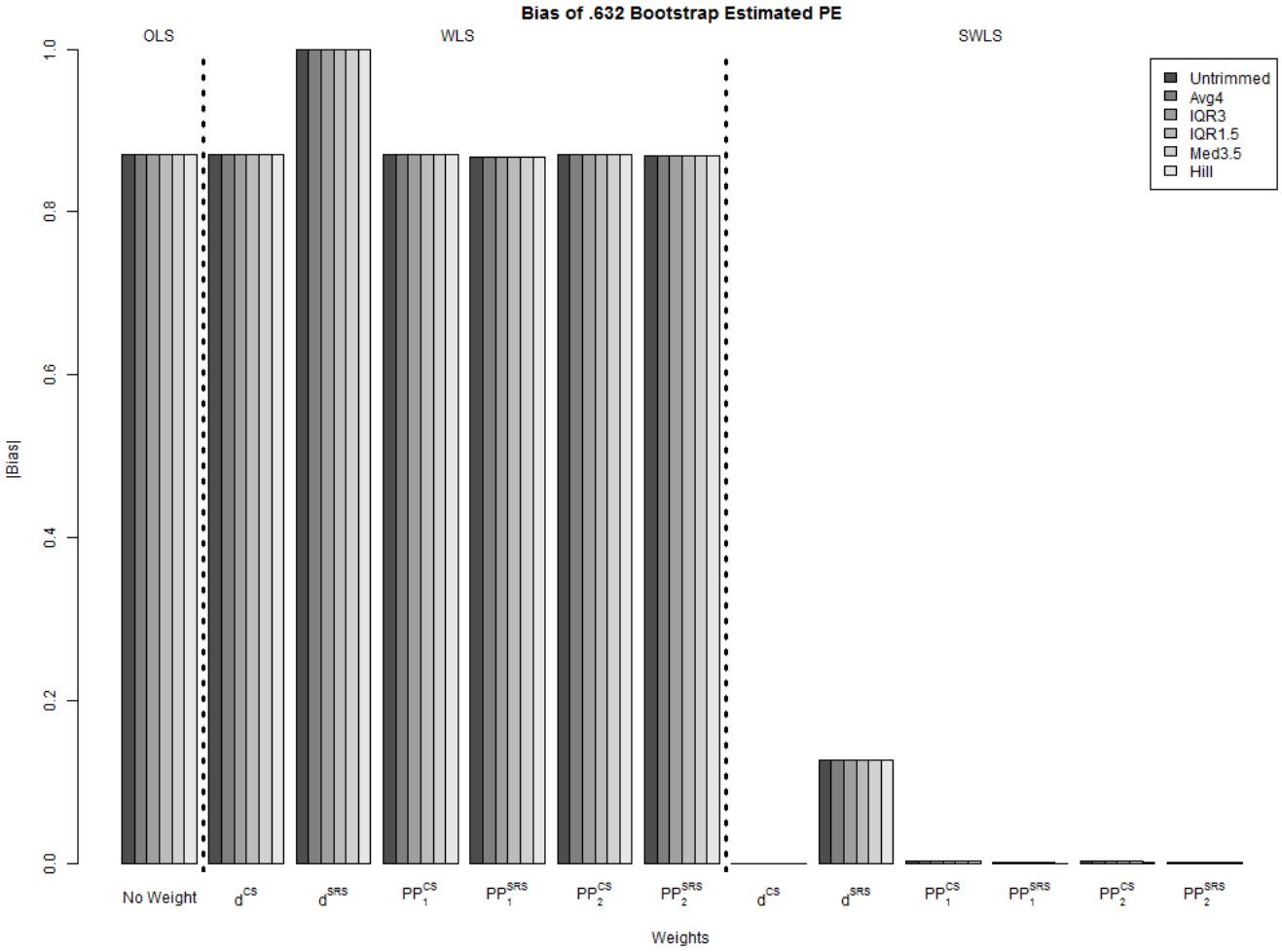


Figure 6.2.14: "True" Bias of .632 Bootstrap Estimated PE: Luus approach

Figure 6.2.14 shows that based on the .632 bootstrap estimated prediction error, the SWLS model performs very well. It is once again apparent that the alternative design weights do not perform well in comparison to the SWLS estimated prediction errors based on the other sampling weights. Also, weight trimming did not appear to further improve the prediction error of the SWLS model. The "true" bias based on the Molinaro approach is the same as for the Luus approach and hence is not presented here.

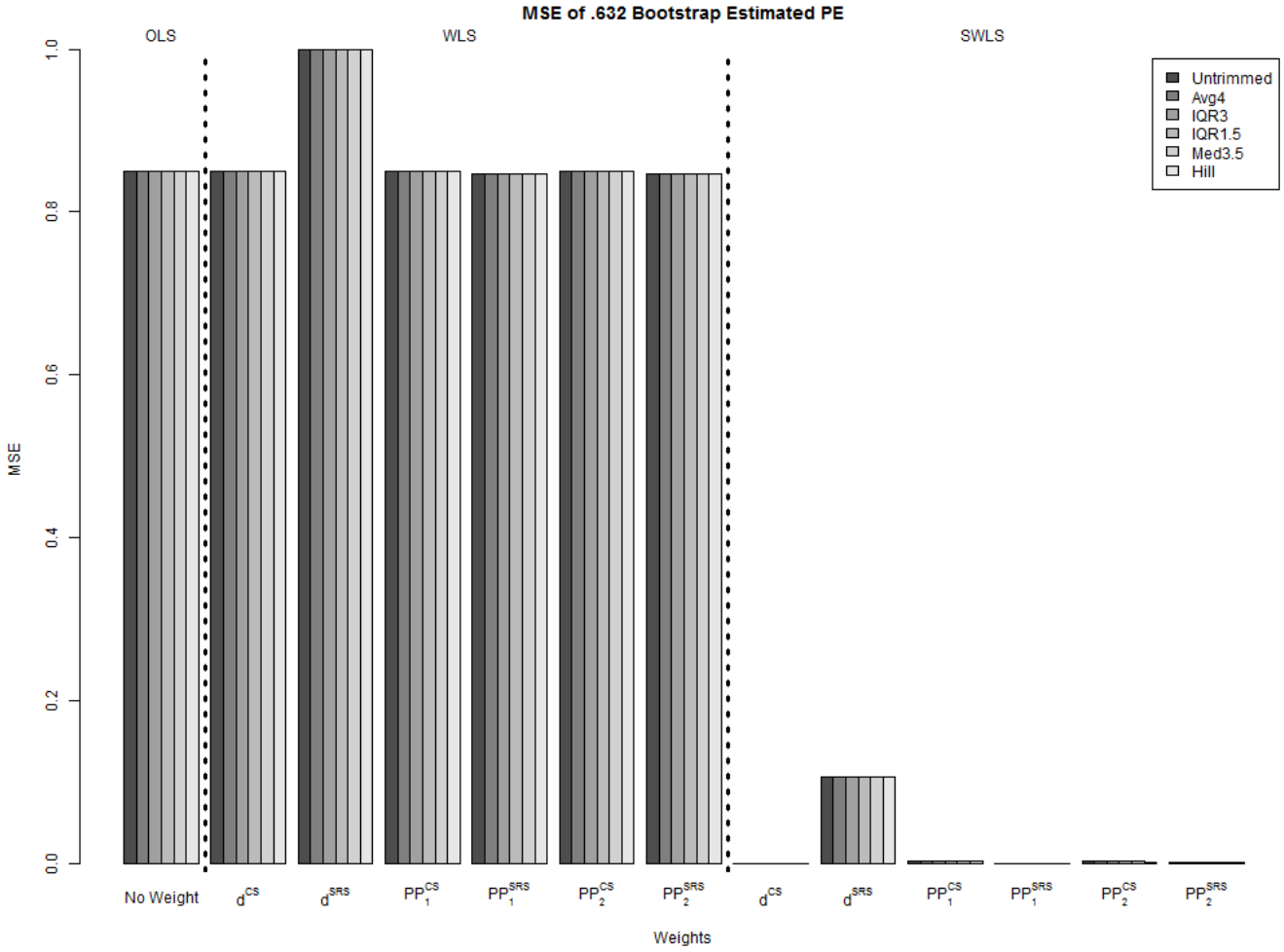


Figure 6.2.15: "True" MSE of .632 Bootstrap Estimated PE: Luus approach

Consideration of figure 6.2.15 leads to the same conclusion as the "true" bias presented in figure 6.2.14. These two diagnostics of the .632 bootstrap estimated prediction error agree that the SWLS model performs "best" in terms of its predictive ability. The Molinaro-based "true" MSE arrives at the same result.

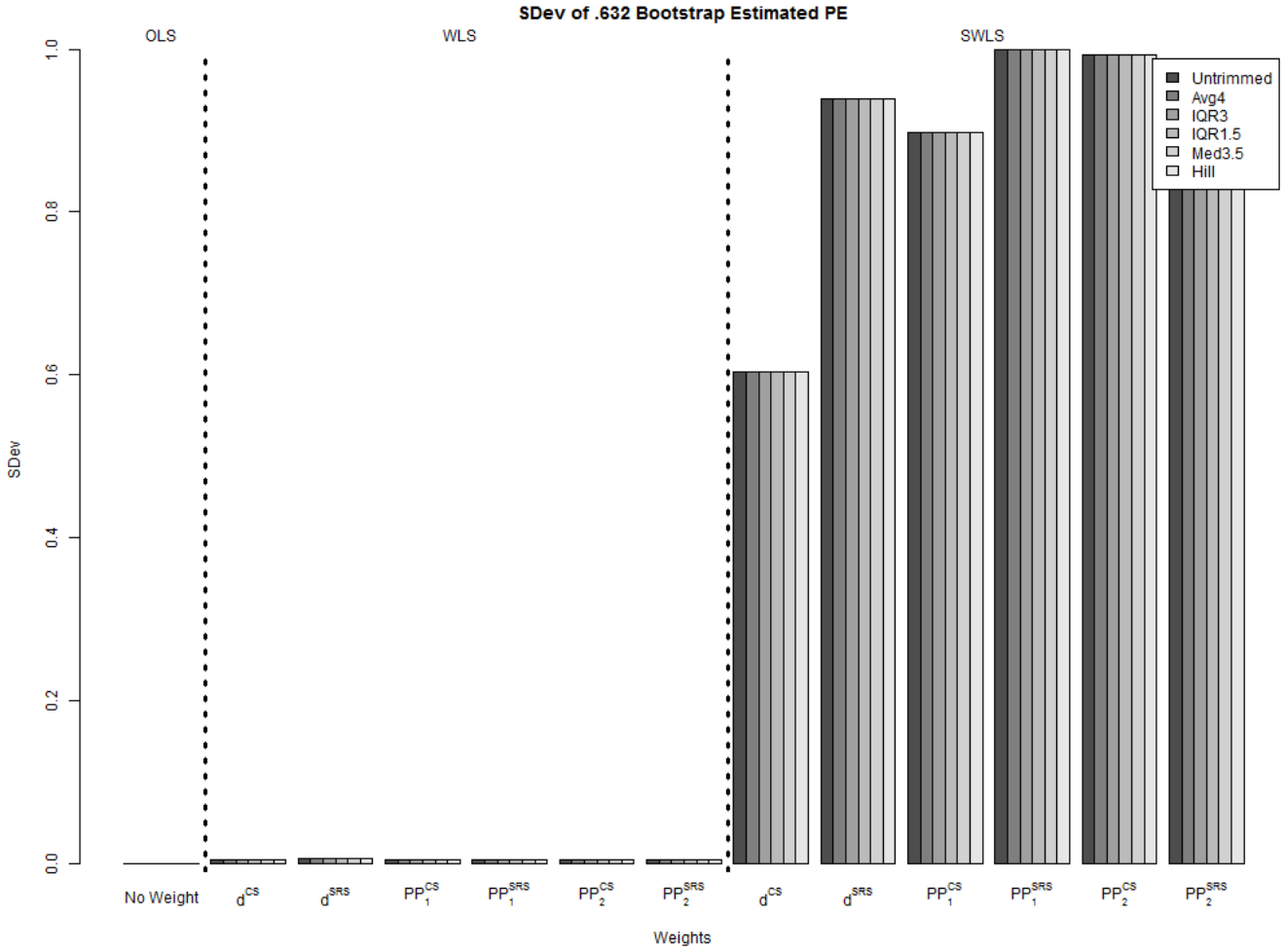


Figure 6.2.16: Estimated Standard Deviation of .632 Bootstrap Estimated PE: Luus approach

As with the bootstrap estimated prediction error, it is seen from figure 6.2.16 that the variability in the SWLS estimated prediction errors is much larger than for OLS and WLS and the same reason as given before, holds here as well. However, it is noted that, irrespective of the large estimated standard deviations, the estimated standard deviations increase further under the alternative design weights, d_{SRS}, and their associated benchmarked weights, $w_{SRS}^{pp_1}$ and $w_{SRS}^{pp_2}$.

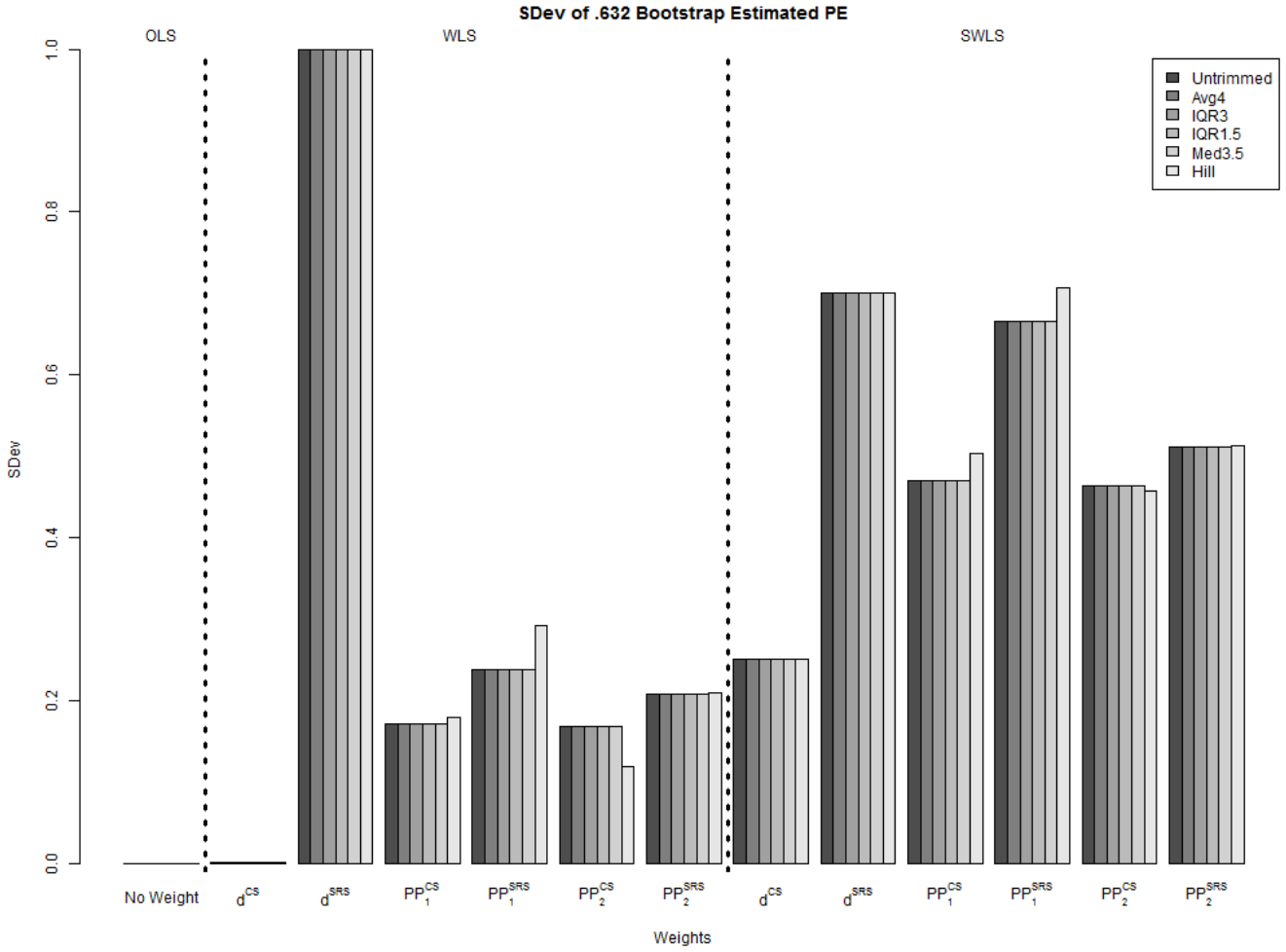


Figure 6.2.17: Estimated Standard Deviation of .632 Bootstrap Estimated PE: Molinaro approach

Compared to figure 6.2.16, figure 6.2.17 shows that the differences between the OLS, WLS and SWLS estimated standard deviations are is not as severe as under the Luus approach. Considering the description of the Luus approach versus the Molinaro approach, one could reason that averaging over a small number of individual "true" prediction errors (Luus) and then using this average in the calculation of the estimated standard deviation, could possibly inflate the standard deviations in comparison to the alternative.

This subsection presented the results of the newly developed CS prediction error estimation methods. The next subsection presents the outlier diagnostics as a final evaluation of the model after which overall conclusions regarding the model evaluation methods will be presented in section 6.2.4.

6.2.3 Outlier Detection Diagnostics

A selection of outlier diagnostics were discussed in section 4.4.3 and this section introduces how these measures will be presented for the comparison of OLS, WLS, and SWLS. It should be mentioned that the model used to simulate WCEC and ECKZN did not accommodate the simulation of outliers. Thus, only a selection of the outlier diagnostics will be included in this section. The IES results presented in chapter 8, however, do include all of the outlier diagnostics due to outliers automatically occurring in the real-world survey data. The simulation of outliers in CS data will form part of further research.

Consider a table containing summary quantiles of the response (Y), the continuous variable (X_1), as well as the sampling weights.

	QUANTILES				
	0%	25%	50%	75%	100%
Y	0.2151	0.3215	0.3795	0.4470	0.6523
X_1	0.0002	0.2304	0.9278	2.6855	12.3895
d_{CS}	5.6000	9.6000	12.3673	15.6145	22.8000
d_{SRS}	13.0850	13.0850	13.0850	13.0850	13.0850
$w_{CS}^{pp_1}$	5.0002	9.7303	12.4856	15.7712	25.9174
$w_{SRS}^{pp_1}$	8.9023	10.4296	13.1259	15.6185	17.5651
$w_{CS}^{pp_2}$	5.0211	9.7287	12.4877	15.7783	26.0566
$w_{SRS}^{pp_2}$	9.3059	10.4481	12.9122	15.5941	17.9792

Table 6.2.2: Quantiles of Variables in WCEC Regression

From this quantiles presented in the table a slight positive skewness is observed in the response and a much larger positive skewness in X_1. Concerning the sampling weights, d_{CS} is positively skewed and the skewness is increased by the benchmarking, while d_{SRS} has no variation and the benchmarked "SRS" weights only possess slight positive skewness. Thus, when an observation is flagged as an outlier, one should consider whether it is flagged due to the size of its associated sampling weight or whether the observed value is simply an outlier. To aid in this the outlier diagnostics will be presented in bubble plots where the size of the bubble is proportional to the sampling weight of the observations. Furthermore, the bubble plots will show the OLS diagnostic on the x-axis and the SWLS diagnostic on the y-axis. Hence, the OLS outlier cut-off will be given by a vertical line and the SWLS outlier cut-off by a horizontal line.

6.2.3.1 Leverages

Consider the bubble plots of the OLS versus SWLS leverages where the theoretical design weights, d_{CS}, were included in the SWLS linear model. The first plot portrays the results for the untrimmed weights while the next plots portray the results for 1.5IQR, Hill and M3 trimmed d_{CS}. Note that the horizontal and vertical cut-offs divide each plot into four blocks. The upper left block is where

all outliers identified only by SWLS leverages, will be found. The lower right block is where all outliers identified only by OLS leverages, will be found.

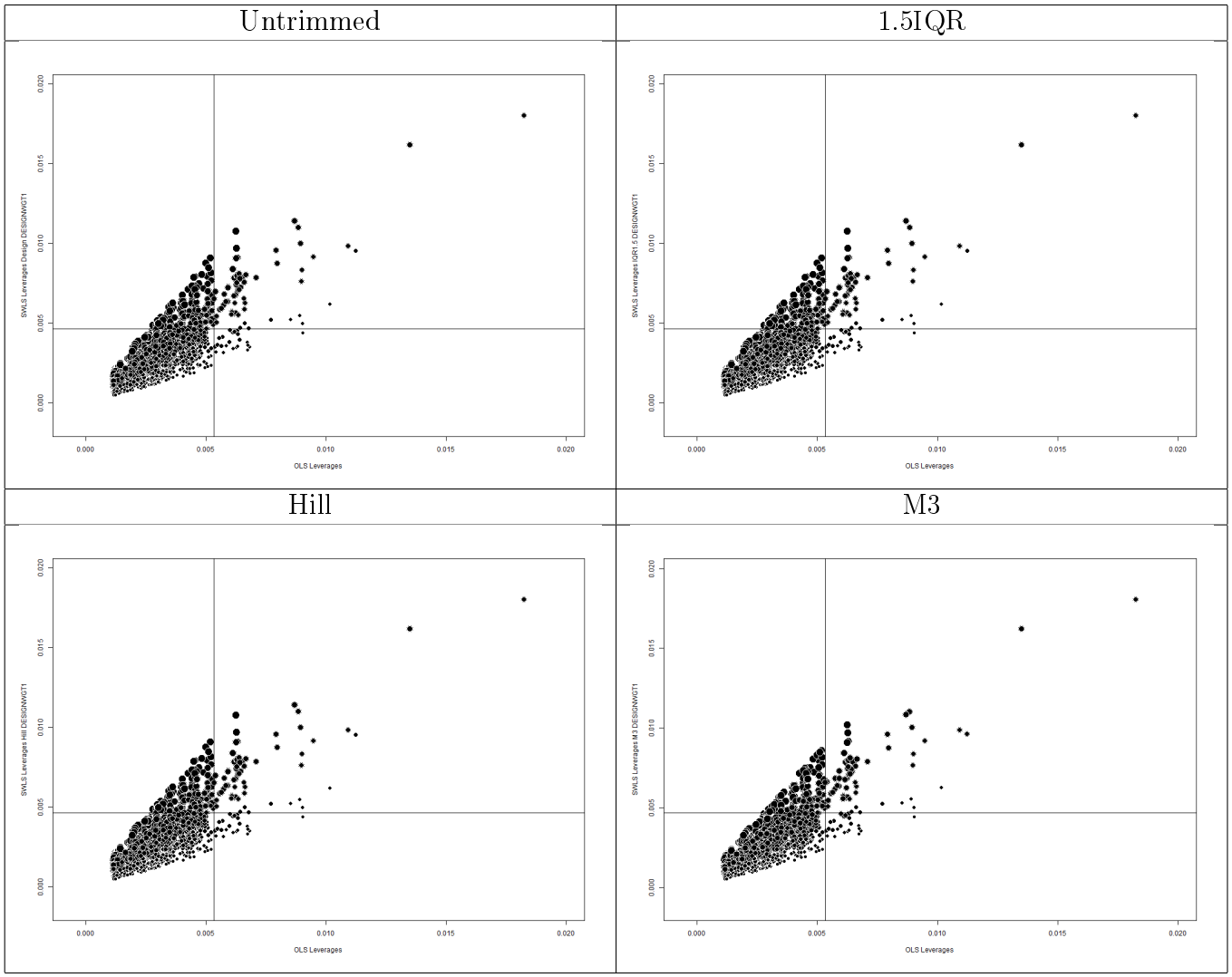


Figure 6.2.18: WCEC Bubble plots of OLS versus SWLS Leverages: d_{CS}

First consider the OLS versus SWLS leverages using the untrimmed d_{CS}. According to the cut-offs it is clear that there is a group of observations flagged only by OLS leverages (lower right block) and a group flagged only by SWLS leverages (upper left block). If one considers the flagged observations in each of these blocks it can be seen from their respective bubble sizes that the OLS leverages typically only flag observations with smaller weights while the SWLS leverages tend to identify observations with varying weight sizes. In figure 6.2.18 the trimmed weight leverages do not appear to have a great influence on which observations are flagged. However, the bubble sizes of the SWLS outliers seem less varied and thus one could see this as a sign that those observations are flagged due to their captured value and not their weight size.

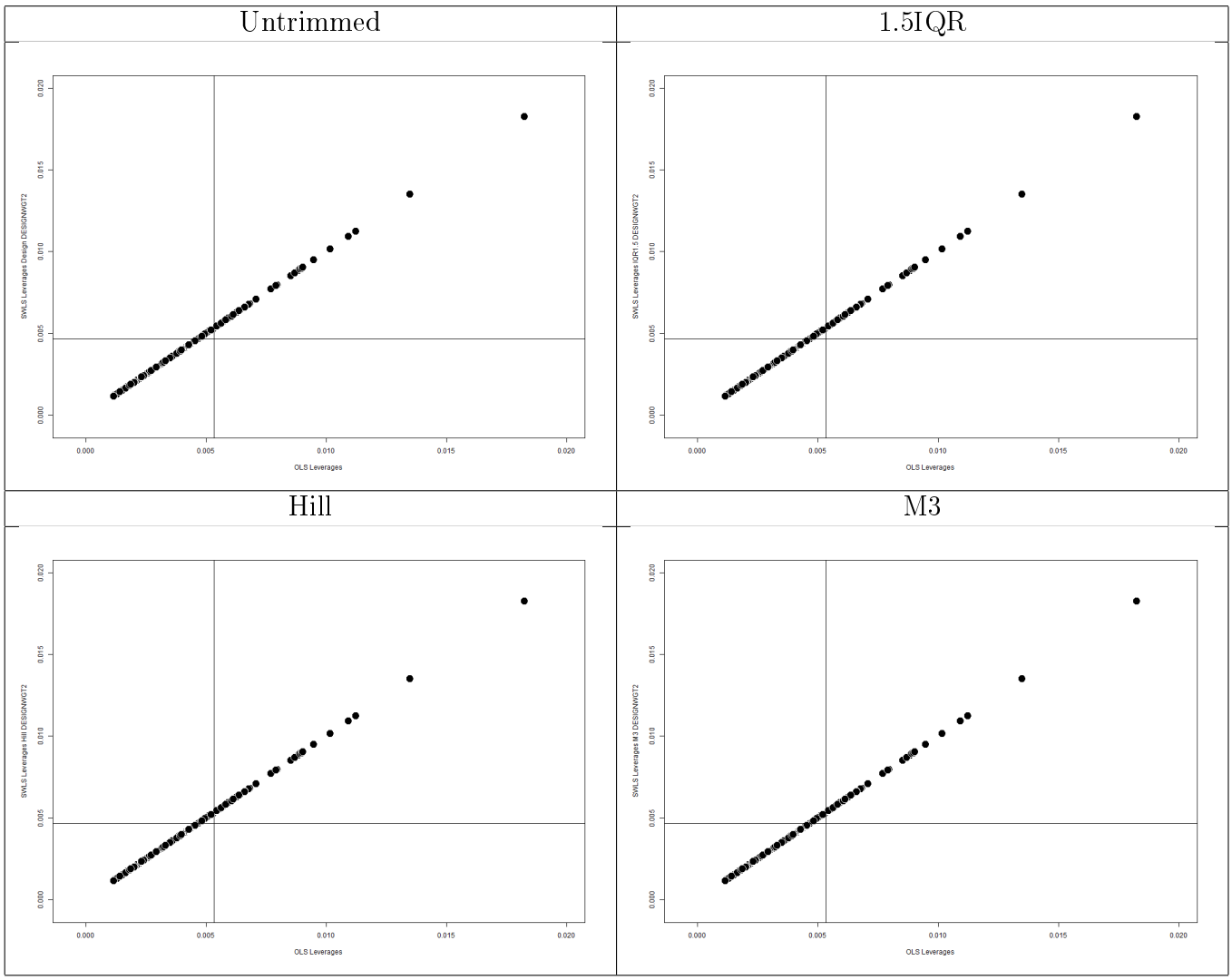


Figure 6.2.19: WCEC Bubble plots of OLS versus SWLS Leverages: d_{SRS}

What is clear from the d_{SRS} leverage plots in figure 6.2.19 is that these weights under SWLS identify only those observations already flagged by the OLS leverages as well.

It is quite well-known among survey statisticians that design weights, based on the inverse of the inclusion probability, is simply the first phase of the calculation of final sampling weights and that design weights should be benchmarked to correct for certain discrepancies between the achieved sample and the target population. Thus, the remainder of the outlier diagnostic plots will only be presented for $w_{CS}^{pp_2}$, the benchmarked d_{CS} weights using the exponential distance measure. Furthermore, the "SRS" sampling weights will not be included since it is clear from figure 6.2.19 that these weights do not work well.

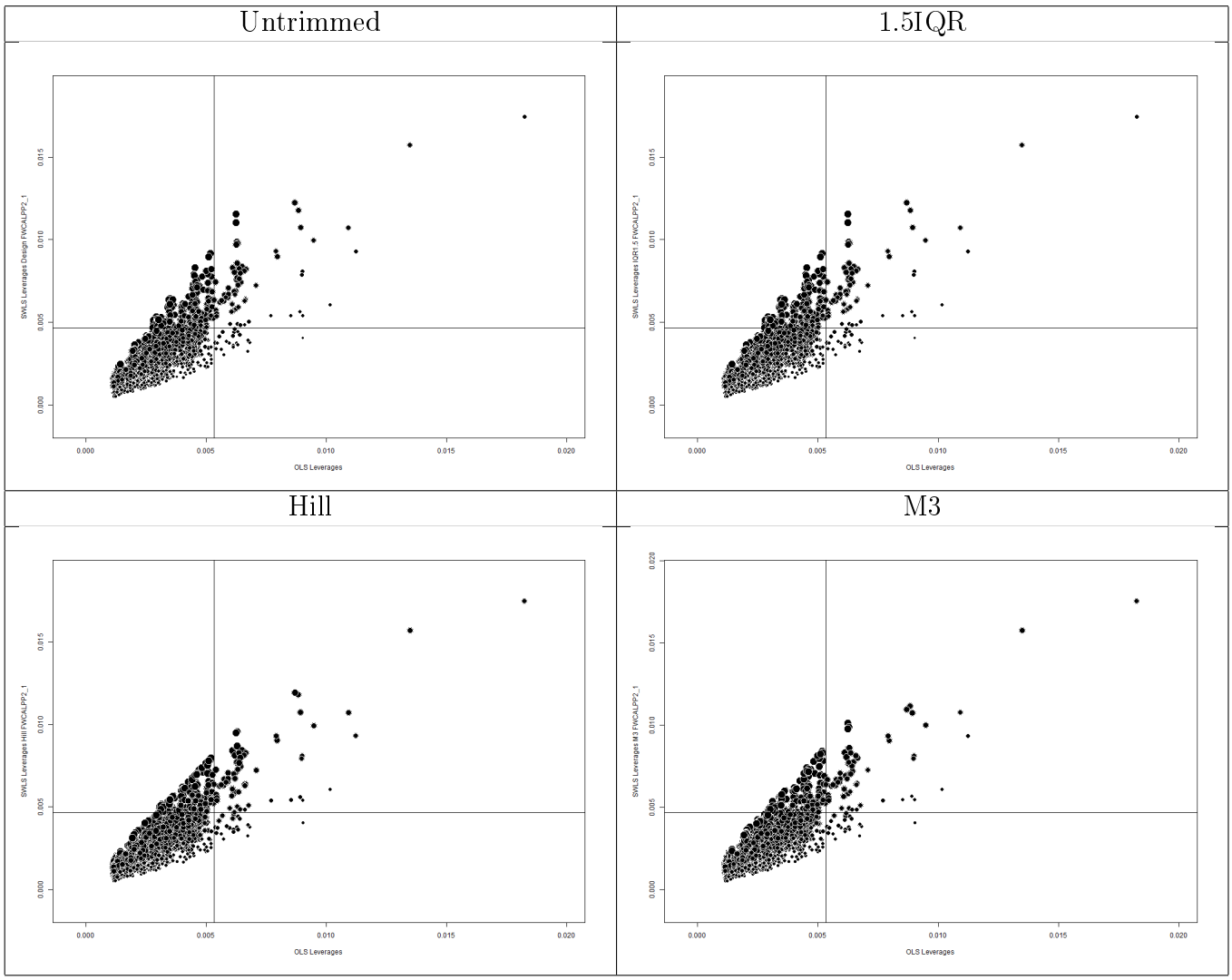


Figure 6.2.20: WCEC Bubble plots of OLS versus SWLS Leverages: $w_{CS}^{pp_2}$

Considering the leverage plot based on the untrimmed $w_{CS}^{pp_2}$ weights it is seen, as in figure 6.2.18, that there are observations flagged by SWLS leverages (upper left block) that are not flagged by the OLS leverages (lower right block). It is also clear that those observations in the OLS block have much smaller weights than those in the SWLS block. The difference between the untrimmed leverages and the trimmed leverages is again the decreased variability in the bubble sizes, due to the weight trimming. It also seems as if fewer observations are flagged by SWLS under trimmed weights than for the untrimmed weights.

6.2.3.2 DFBetas of Predictor X_1

Consider below the bubble plots of the OLS versus SWLS DFBetas of the continuous predictor, X_1. Here too the bubble sizes are proportional to the sizes of the sampling weights with the OLS diagnostic being denoted on the x-axis and the SWLS diagnostic on the y-axis. The OLS cut-off is consequently represented by the vertical lines and the SWLS cut-off by the horizontal cut-off

lines.

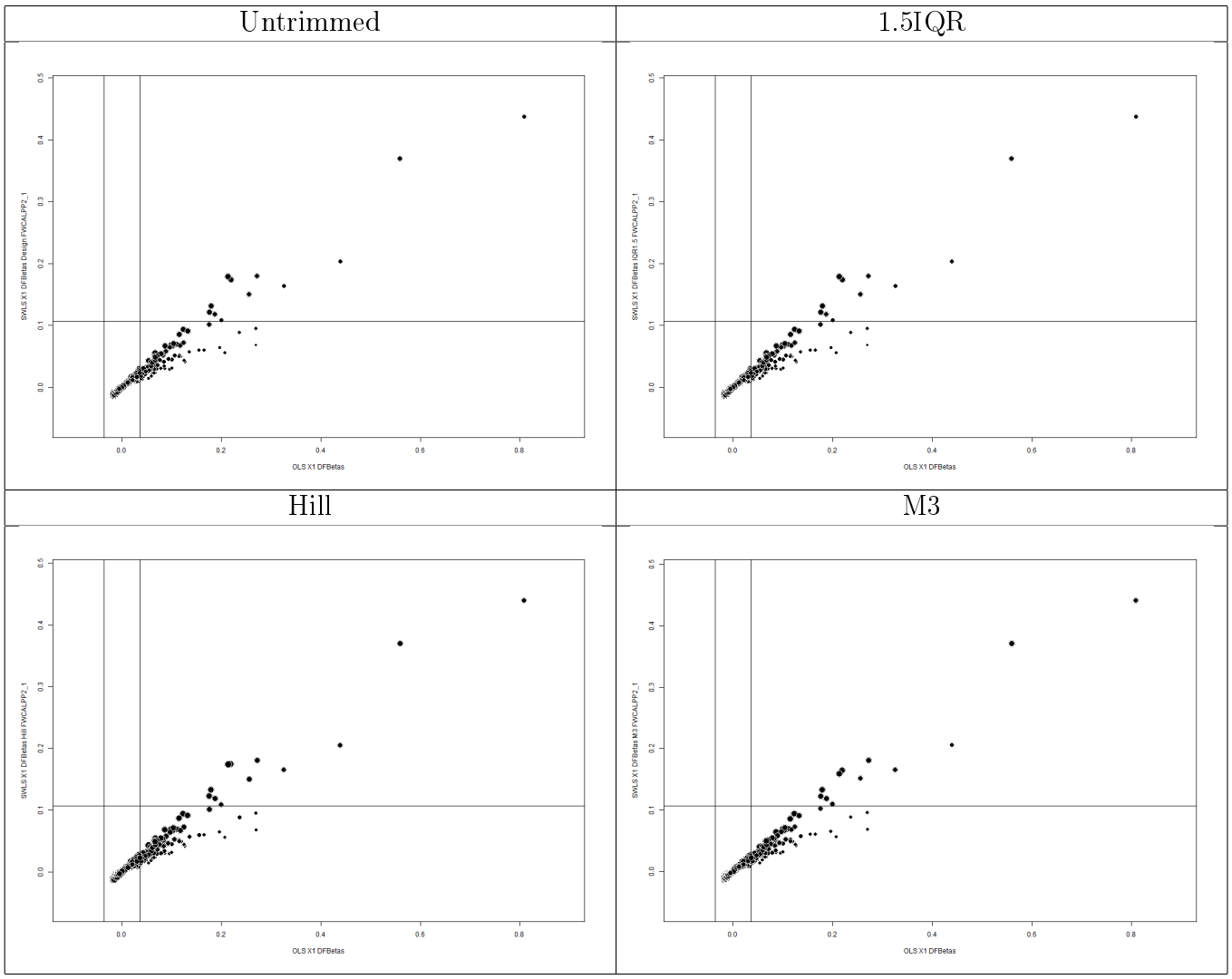


Figure 6.2.21: WCEC Bubble plots of OLS versus SWLS X1 DFBetas: $w_{CS}^{ph_2}$

Recall that it was mentioned that the simulation model did not make provision for the simulation of outliers and from the bubble plots of the DFBetas this is seen. According to the untrimmed weight bubble plot there are no observations that were flagged only by the SWLS diagnostic (upper and lower middle blocks), but some were flagged by the OLS diagnostic (lower left and right blocks). Note that the second horizontal line does not show on the plots since it's value lies below the y-range. The OLS flagged observations again have small weights as seen from their bubble sizes. Of those observations flagged by both OLS and SWLS (upper right block) it seems as if the bubble sizes do not vary much. The weight trimming methods did not change the outcome much, but the bubble sizes did become more uniform.

6.2.3.3 DFFits

Here the bubble plots of the DFFits diagnostic are shown. The OLS diagnostic is again on the x-axis and the SWLS diagnostic on the y-axis. The bubble sizes are proportional to the sampling weights and the cut-offs are presented by the vertical, for OLS, and the horizontal, for SWLS, lines.

The four cut-off lines divide each plot into nine blocks. To find the observations flagged by SWLS, consider the upper middle block and the lower middle block. To find the observations flagged by OLS, consider the middle left block and the middle right block.

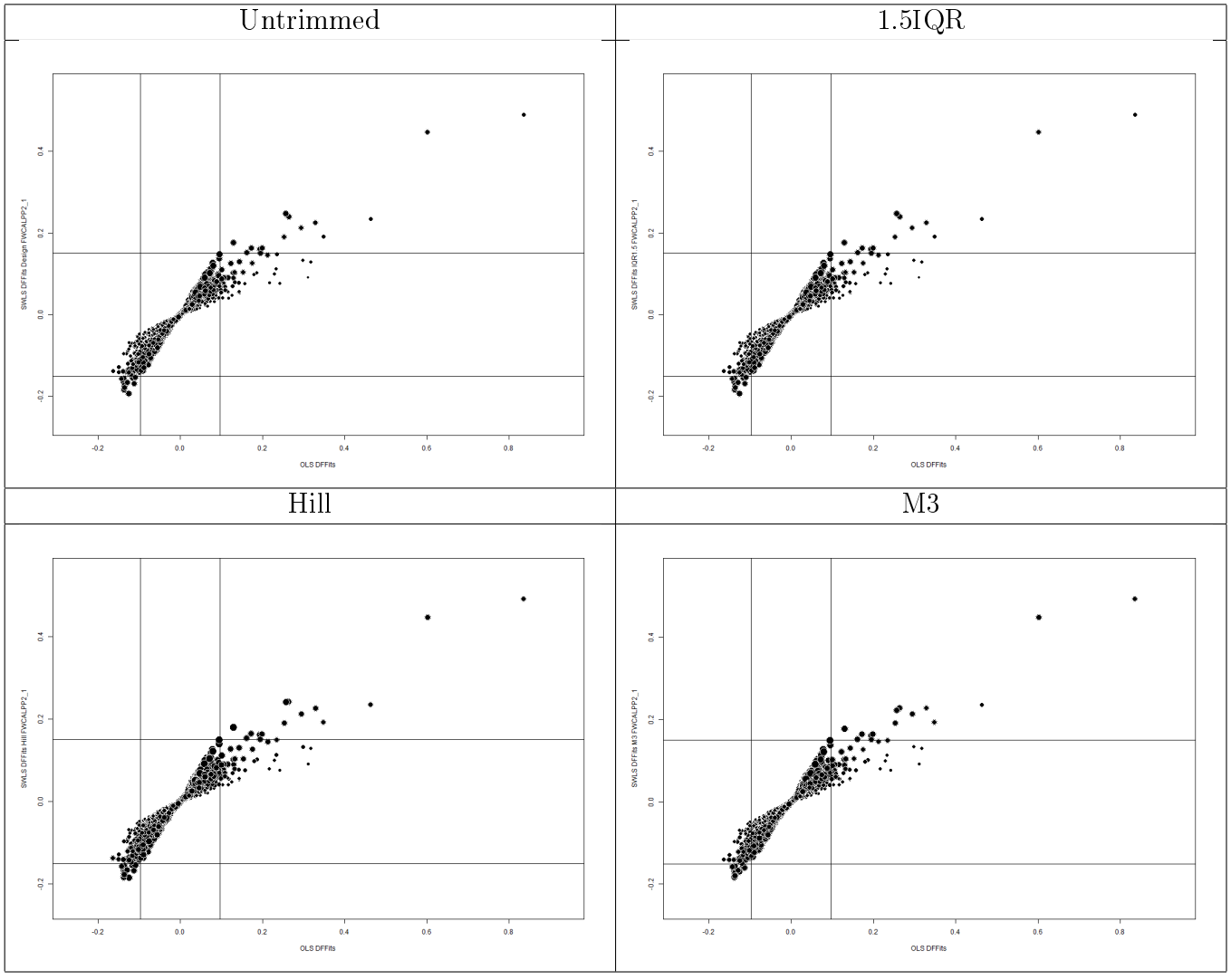


Figure 6.2.22: WCEC Bubble plots of OLS versus SWLS DFFits: w_{CS}^{ph2}

With the exception of one or two observations, SWLS does not appear to flag any observations as outliers as opposed to OLS that does. The possible SWLS outliers do have large sampling weights while those flagged by OLS have smaller sampling weights. The weight trimming methods contributed marginal changes in the bubble plots, mostly changing the size of the bubble.

6.2.4 Summary and Conclusions

The previous section displayed the outlier diagnostics using bubble plots where the bubble sizes were proportional to the benchmarked theoretical weights, w_{CS}^{pp2}. Now the output is summarized into the table below and compared to the summarized output of the benchmarked alternative weights, w_{SRS}^{pp2}. The table has been grouped by trimming method using the colours green (no trimming), blue (1.5IQR), red (Hill), and black (M3). The "count" column shows how many observations were flagged by each linear model under each diagnostic. The weight range columns give the minimum and maximum weights associated with each outlier flagged by each diagnostic under each linear model.

			Linear Model								
			OLS			WLS			SWLS		
Weight	Trimming	Diagnostic	Count	Weight Range		Count	Weight Range		Count	Weight Range	
w_{CS}^{pp2}	No	Leverages	92	5.1072	25.7704	203	7.0535	25.7704	324	7.0535	25.7704
		DFBetas	113	5.1072	16.7061	116	5.1072	18.7964	12	9.4046	16.7061
		DFFits	158	5.1072	18.7964	154	5.1072	22.4925	14	9.4046	18.7964
	1.5IQR	Leverages	92	5.1072	25.7704	203	7.0535	25.7704	324	7.0535	25.7704
		DFBetas	113	5.1072	16.7061	116	5.1072	18.7964	12	9.4046	16.7061
		DFFits	158	5.1072	18.7964	154	5.1072	22.4925	14	9.4046	18.7964
	Hill	Leverages	92	5.1072	25.7704	203	7.0535	25.7704	301	7.0535	25.7704
		DFBetas	113	5.1072	16.7061	116	5.1072	18.7964	12	9.4046	16.7061
		DFFits	158	5.1072	18.7964	154	5.1072	22.4925	15	9.4046	18.7964
	M3	Leverages	92	5.1072	25.7704	203	7.0535	25.7704	317	7.0535	25.7704
		DFBetas	113	5.1072	16.7061	116	5.1072	18.7964	12	9.4046	16.7061
		DFFits	158	5.1072	18.7964	154	5.1072	22.4925	15	9.4046	18.7964
w_{SRS}^{pp2}	No	Leverages	92	9.8567	17.7296	175	9.8567	17.7296	267	9.8567	17.7296
		DFBetas	113	9.8567	11.8796	115	9.8567	11.8796	14	9.8567	11.8796
		DFFits	158	9.8567	16.6505	125	9.8567	16.6505	15	9.8567	16.6505
	1.5IQR	Leverages	92	9.8567	17.7296	175	9.8567	17.7296	267	9.8567	17.7296
		DFBetas	113	9.8567	11.8796	115	9.8567	11.8796	14	9.8567	11.8796
		DFFits	158	9.8567	16.6505	125	9.8567	16.6505	15	9.8567	16.6505
	Hill	Leverages	92	9.8567	17.7296	175	9.8567	17.7296	265	9.8567	17.7296
		DFBetas	113	9.8567	11.8796	115	9.8567	11.8796	15	9.8567	11.8796
		DFFits	158	9.8567	16.6505	125	9.8567	16.6505	15	9.8567	16.6505
	M3	Leverages	92	9.8567	17.7296	175	9.8567	17.7296	265	9.8567	17.7296
		DFBetas	113	9.8567	11.8796	115	9.8567	11.8796	15	9.8567	11.8796
		DFFits	158	9.8567	16.6505	125	9.8567	16.6505	15	9.8567	16.6505

Table 6.2.3: WCEC Number of Outliers Identified and Associated Weight Ranges (w_{CS}^{pp2} versus w_{SRS}^{pp2})

Consider the w_{CS}^{pp2} results starting with the untrimmed weight results highlighted in green. Across the diagnostics it is seen that the minimum weights associated with the OLS outliers are

smaller than those of the SWLS outliers. SWLS leverages flag many more outliers than OLS leverages while the opposite is true of DFBetas and DFFits. Further changes in number of outliers flagged by SWLS are observed once the weights are trimmed. However, the weight ranges of the flagged observations did not change after being trimmed. It is possible to conclude from this that those observations that were perhaps flagged due to extreme sampling weights were unflagged after their weights were trimmed, and only the outliers flagged due to their captured values remained. If the $w_{SRS}^{pp_2}$ results are compared in the same way, it is seen that the number of outliers flagged across the untrimmed and trimmed results, is not changed much.

Three prediction error estimation methods were introduced and their results presented here, viz. the leave-one-out cross-validation, bootstrap, and .632 bootstrap methods. All three methods presented encouraging results and mostly concurred that the SWLS, at least according to the bias and MSE diagnostics, resulted in models that will make "good" predictions. Mostly the results were very similar whether based on the theoretical design weights and their associated benchmarked weights, or based on the alternative design weights and their associated benchmarked weights. The application of the weight trimming methods did not appear to further improve the model prediction errors, however the Hill trimming methods did show some promise. The small number of replicate samples and bootstrap samples within each replicate sample, can be considered valid reasons for the not very conclusive results. This is a point that will form part of further research.

From the results presented here it can simply be said that there is a difference between which observations and how many are flagged by OLS and SWLS using the "CS" or "SRS" sampling weights in their untrimmed or trimmed form. To be able to say whether the $w_{CS}^{pp_2}$ or $w_{SRS}^{pp_2}$ results, untrimmed or trimmed, are better, would only be possible after the flagged observations under each scenario were removed and the model fit results compared again after re-fitting the model to the reduced data. This is another area that forms part of further research.

6.3 Outline of Model Parameter Analysis

The parameters of interest in linear modeling are the regression coefficients, β_j, $j = 1, ..., p$. This research considers, among others, the estimation of these parameters by the methods of OLS, $\underline{\hat{\beta}}_{OLS}$, WLS, $\underline{\hat{\beta}}_{WLS}$, and SWLS, $\underline{\hat{\beta}}_{SWLS}$. The purpose of this section is to investigate the properties of the estimators through the calculation of various diagnostic measures. The point estimators of the parameters will be assessed by considering their standard error, bias, mean squared error (MSE) and median absolute deviation (MAD). These measures are discussed in section 6.3.1.

The point estimation is followed by the interval estimation of the parameter. Section 6.3.2 includes: the standard (asymptotic) interval using model estimated and the bootstrap estimated variance, respectively; percentile interval; bootstrap-t interval based on, respectively, a second level bootstrap and second level jackknife estimated variance; the BCa interval using, respectively, a jackknife estimated and a bootstrap estimated acceleration constant. These interval estima-

tors will be assessed according to their respective non-coverage probabilities (NCP), lengths and standardized lengths.

The results are presented in sections 6.3.1 and 6.3.2 and the main findings will be summarized and discussed in section 6.3.3.

6.3.1 Model Parameter Estimation Diagnostics

Since sampling is from a (known) simulated population, the β_j's are known. Using R samples from the simulation distributions, the parameter estimation diagnostics can be approximated as follows. For each of the estimators the following properties are investigated (Neethling, 2004; Kovar et al., 1988):

- The bias of the estimator with respect to the population parameter, β_j, is calculated in two ways, namely

$$bias^1\left(\hat{\beta}_j\right) = \left[\left(\frac{1}{R}\sum_{r=1}^{R}\hat{\beta}_{r_j}\right) - \beta_j\right], \tag{6.3.1}$$

  where R is the number of replicate samples and $\hat{\beta}_{r_j}$ is the estimator calculated on the rth replicate sample using OLS, WLS and SWLS. This will result in $bias^1\left(\hat{\beta}_{OLS_j}\right)$, $bias^1\left(\hat{\beta}_{WLS_j}\right)$, and $bias^1\left(\hat{\beta}_{SWLS_j}\right)$. Also,

$$bias^2\left(\hat{\beta}_j\right) = \underset{1\leq r\leq R}{median}\left\{\hat{\beta}_{r_j}\right\} - \beta_j. \tag{6.3.2}$$

  Similar notation will be used for the other measures.

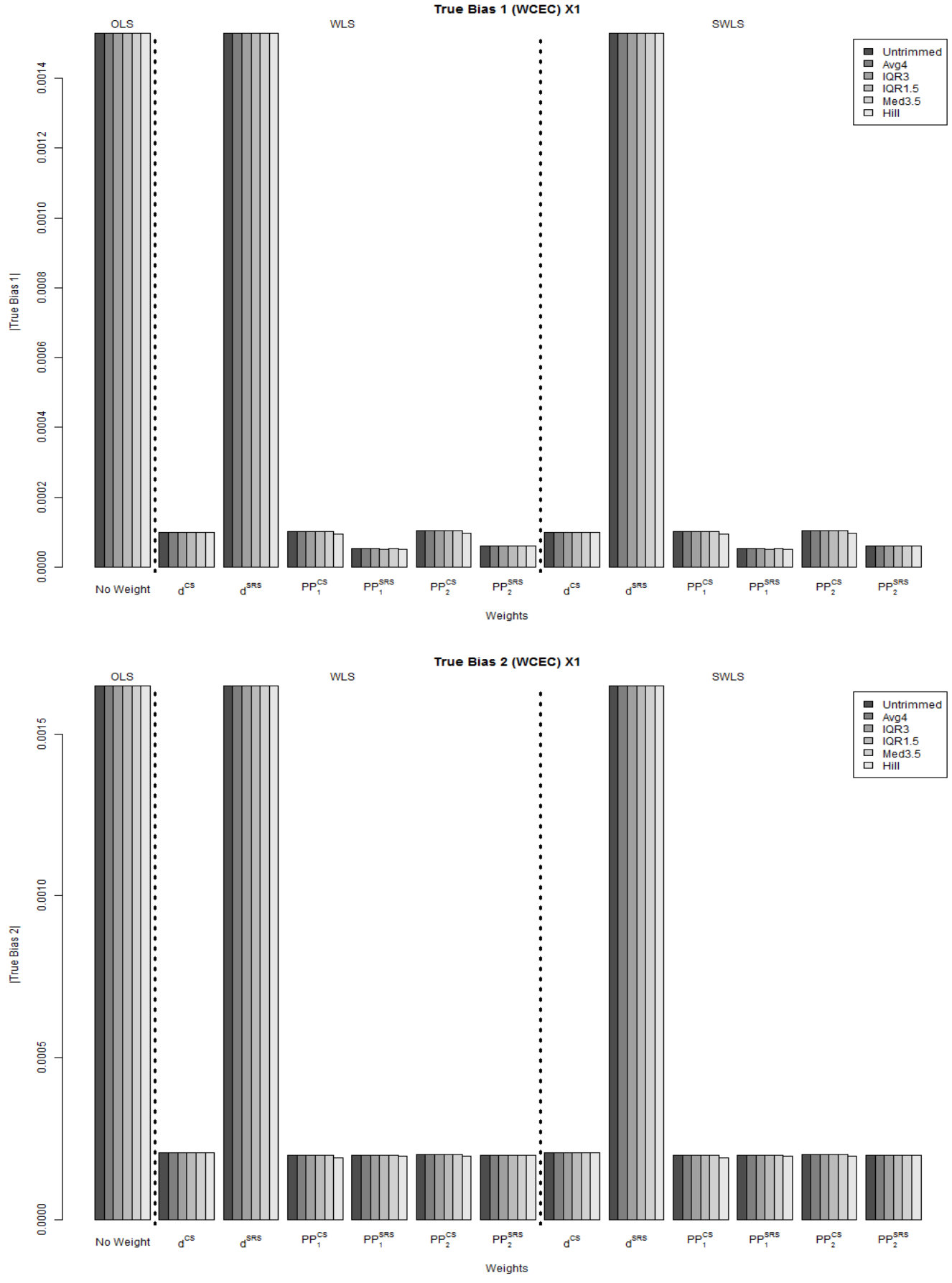


Figure 6.3.1: WCEC Absolute Value of "True" Bias 1 and 2 of predictor X_1

Figure 6.3.1 displays a small difference between the "true" bias based on equation (6.3.1) and equation (6.3.2). This could imply that the median is a more reliable estimator of the midpoint than the average as would be expected when using only a limited number of samples. Hence, consider the figure based on equation (6.3.2). It is clear that the use of the theoretical design weights and their associated benchmarked weights, i.e. d_{CS}, $w_{CS}^{pp_1}$ and $w_{CS}^{pp_2}$, brings the estimator closer to the "true" parameter than when using no weights. Furthermore, Hill trimmed weights improved the distance slightly further by reaching the smallest bias under $w_{CS}^{pp_1}$ and $w_{CS}^{pp_2}$.

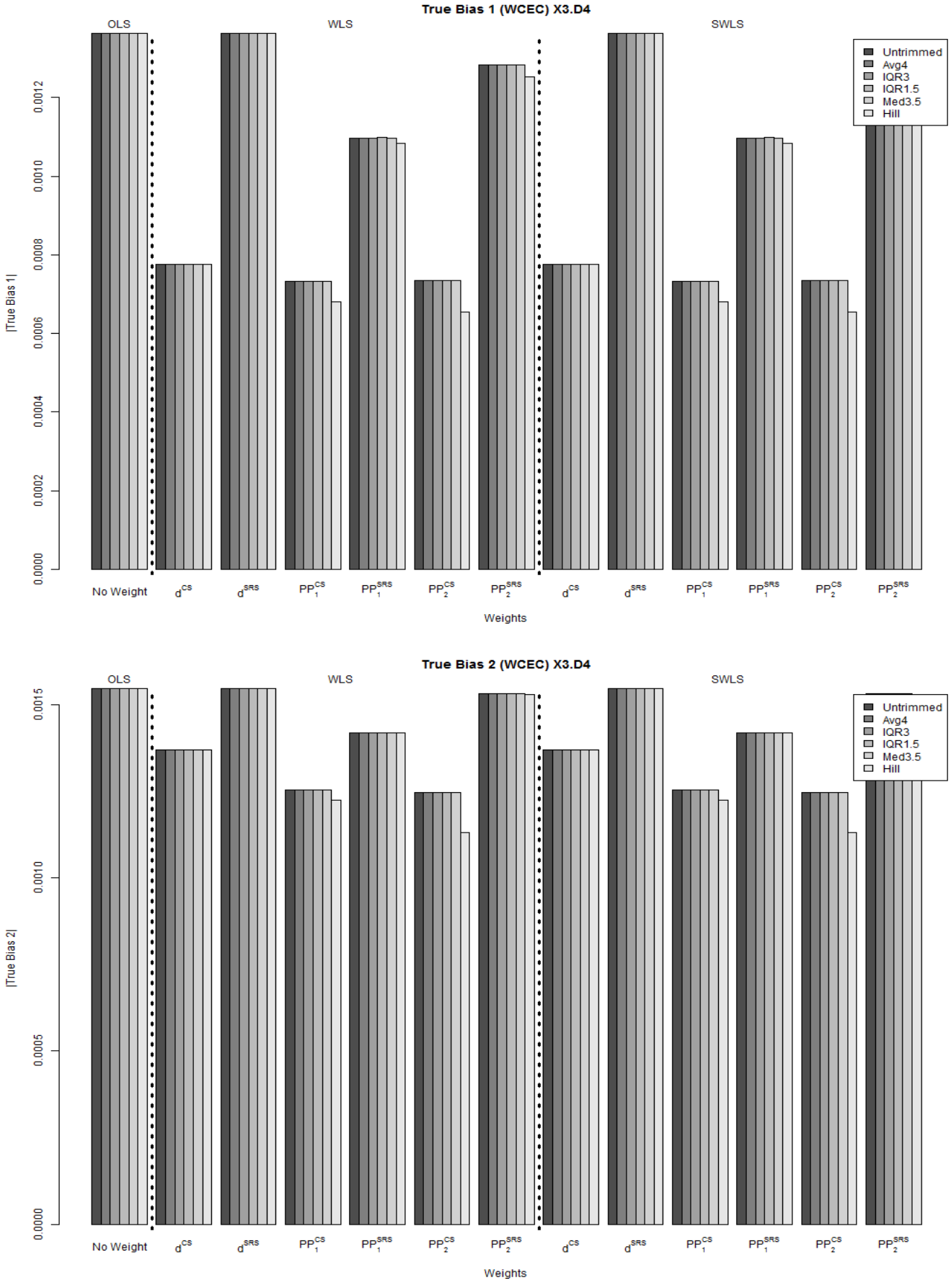


Figure 6.3.2: WCEC Absolute Value of "True" Bias 1 and 2 of predictor category $X_3 = 4$

In figure 6.3.2 both "true" biases portray very similar patterns. From both it is clear, however, that the Hill-trimmed benchmarked theoretical weights, $w_{CS}^{pp_2}$, achieved the smallest bias.

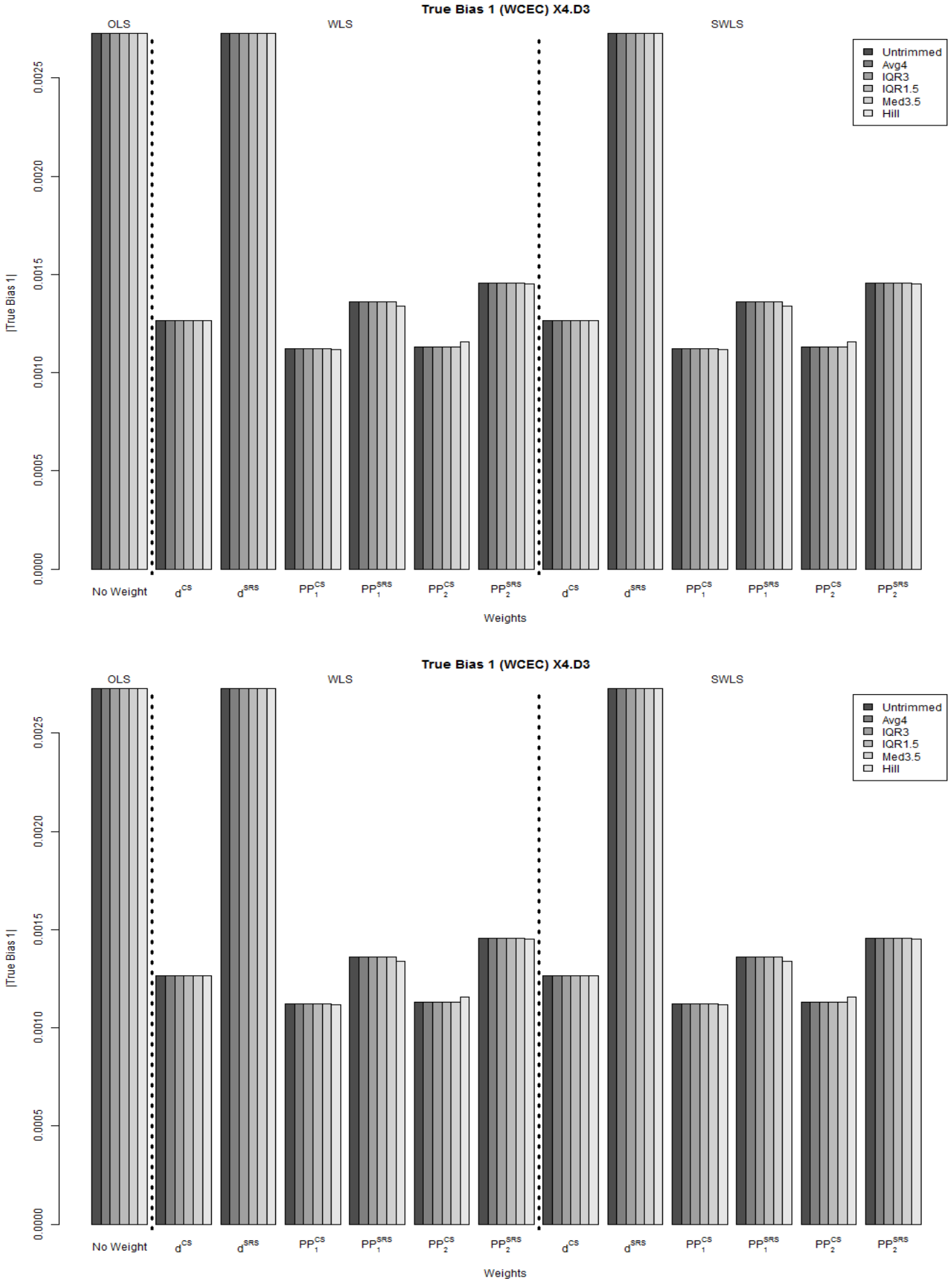


Figure 6.3.3: WCEC Absolute Value of "True" Bias 1 and 2 of predictor category $X_4 = 3$

Figure 6.3.3 is quite similar to figures 6.3.1 and 6.3.2 and thus leads to the same conclusions.

- The mean squared error (MSE) of the estimator with respect to the population parameter will also be calculated in two ways, namely as

$$MSE^1\left(\hat{\beta}_j\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\beta}_{r_j} - \beta_j\right)^2, \tag{6.3.3}$$

and as

$$MSE^2\left(\hat{\beta}_j\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\beta}_{r_j} - \bar{\hat{\beta}}_{r_j}\right)^2, \tag{6.3.4}$$

where $\bar{\hat{\beta}}_{r_j}$ is the average of the R estimates of $\hat{\beta}_j$.

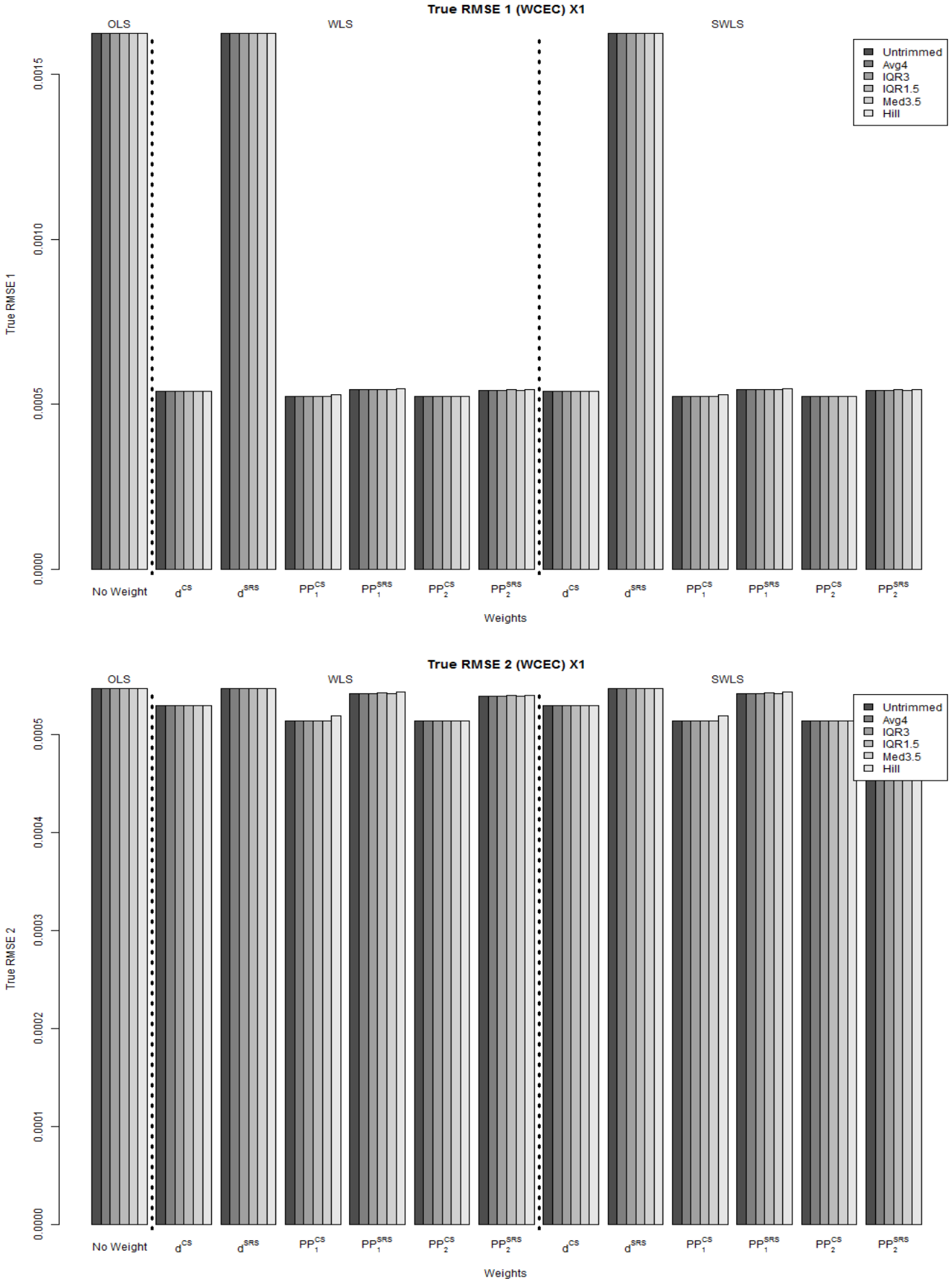


Figure 6.3.4: WCEC "True" RMSE 1 and 2 of predictor category X_1

A first glance of figure 6.3.4 seems to indicate that the figure based on equation (6.3.3) is more successful than the figure based on equation (6.3.4). However, if one considers the range of the y-axis it is clear that equation (6.3.4) resulted in the smaller RMSE's. It can thus be said that, as with the "true" biases, the theoretical design weights and their associated benchmarked weights, i.e. d_{CS}, $w_{CS}^{pp_1}$ and $w_{CS}^{pp_2}$, bring the estimator closer to the "true" parameter than when using no weights or the alternative design weights and their associated benchmarked weights. However, according to the RMSE, the weight trimming methods do not really improve the diagnostic further.

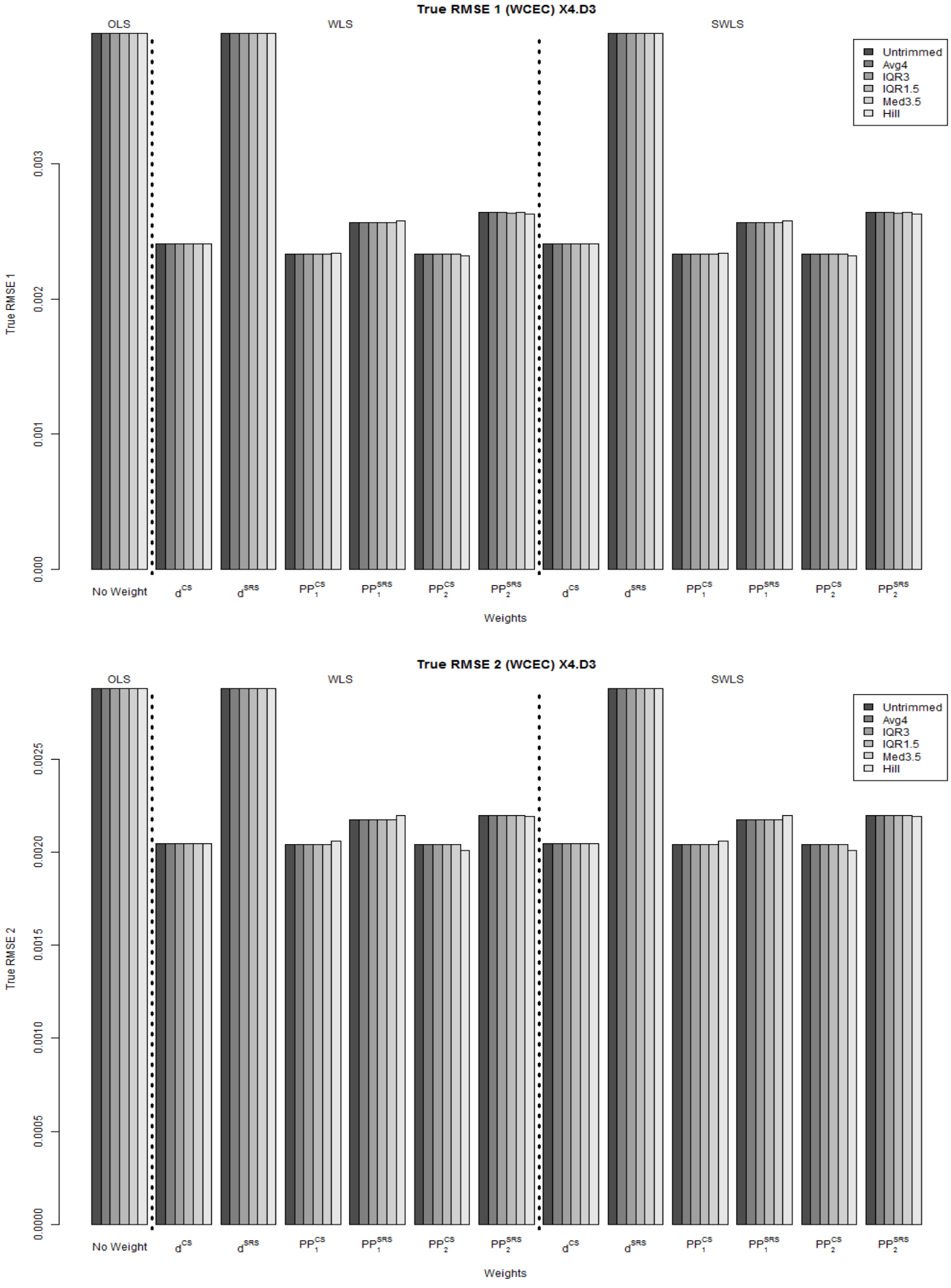


Figure 6.3.5: WCEC "True" RMSE 1 and 2 of predictor category X_4= 3

Similar conclusions can be made from figure 6.3.5, but with the difference of the Hill trimmed $w_{CS}^{pp_2}$ achieving the smallest RMSE.

- Another informative diagnostic to consider is the median absolute deviation, or MAD. This can also be defined in two ways, namely as

$$MAD^1\left(\hat{\beta}_j\right) = median\left|\hat{\beta}_{r_j} - \beta_j\right|, \tag{6.3.5}$$

and

$$MAD^2\left(\hat{\beta}_j\right) = median\left|\hat{\beta}_{r_j} - median\left\{\hat{\beta}_{r_j}\right\}\right|. \tag{6.3.6}$$

The results will not be presented here since conclusions drawn from the median absolute deviation are in line with those based on the RMSE. They are available on the accompanying CD. Conclusions that can be made from them are that the correct use of the sampling weights (SWLS) improves the precision of the estimator quite significantly. The alternative design weights, d_{SRS}, performed as badly as OLS, but some improvement was observed once the "SRS" design weights were benchmarked. In some cases the precision was improved even further by the application of the trimming methods.

- For each of the R bootstrap populations the bootstrap estimated bias will be considered using both definitions given above,

$$\widehat{bias}_B^1\left(\hat{\beta}_{r_j}\right) = \left(\frac{1}{B}\sum_{r_b=1}^{B}\hat{\beta}_{r_{b_j}}^*\right) - \hat{\beta}_{r_j}, \tag{6.3.7}$$

where $\hat{\beta}_{r_{b_j}}^*$ is the bth bootstrap estimate of the estimator of the rth replicate of the jth regression parameter, and

$$\widehat{bias}_B^2\left(\hat{\beta}_{r_j}\right) = median\left\{\hat{\beta}_{r_{b_j}}^*\right\} - \hat{\beta}_{r_j}. \tag{6.3.8}$$

This results in R bootstrap estimated biases for both defined biases. Thus, the overall bootstrap estimate of bias of $\hat{\beta}_j$, for the first definition, is given by

$$\widehat{bias}_B^1\left(\hat{\beta}_j\right) = \left[\frac{1}{R}\sum_{r=1}^{R}\widehat{bias}_B^1\left(\hat{\beta}_{r_j}\right)\right], \tag{6.3.9}$$

and for the second defined bias,

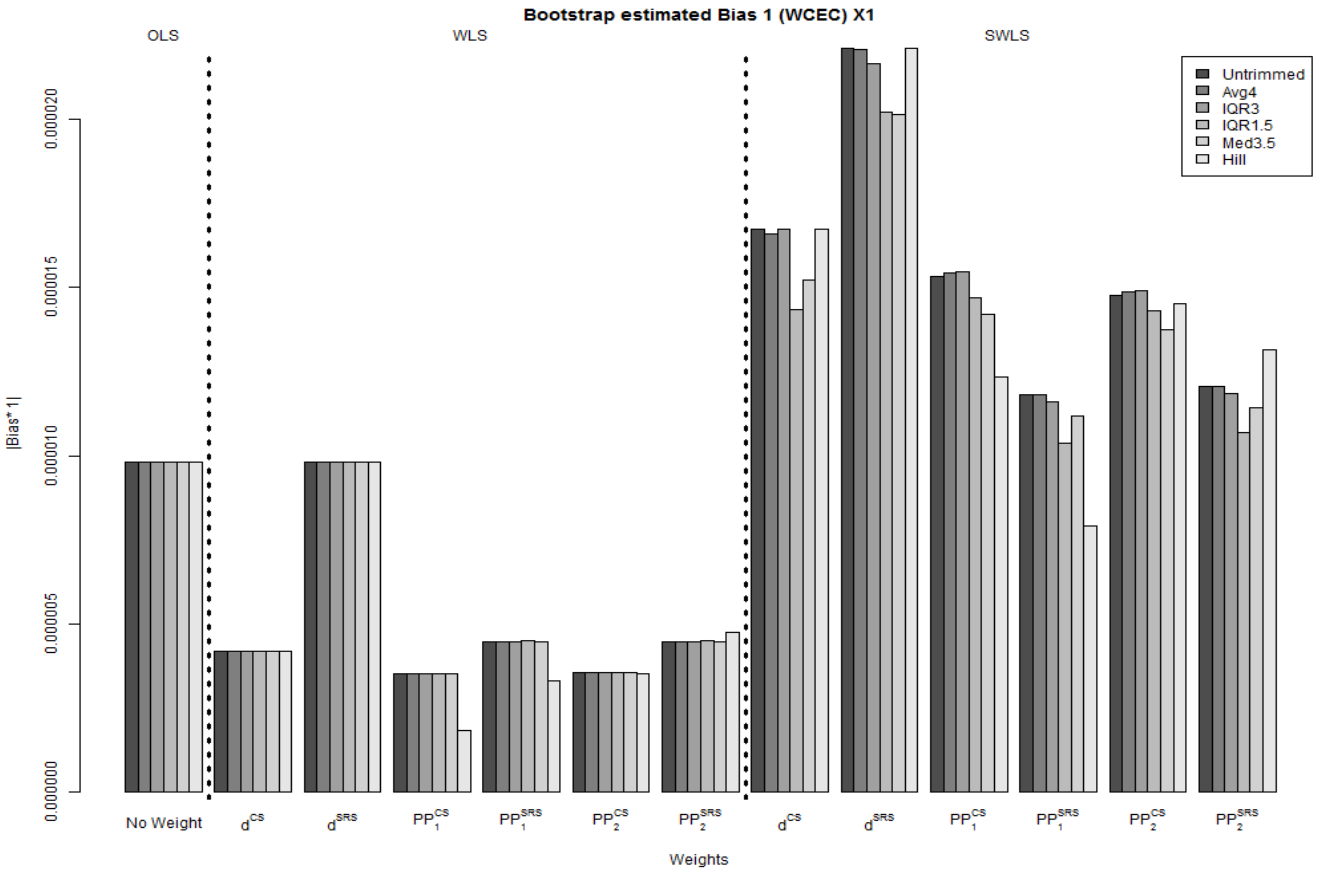
$$\widehat{bias}_B^{\,2}\left(\hat{\beta}_j\right) = \left[\frac{1}{R}\sum_{r=1}^{R}\widehat{bias}_B^{\,2}\left(\hat{\beta}_{r_j}\right)\right]. \tag{6.3.10}$$

The difference between the bootstrap estimates of the two biases and the associated "true" bias, defined, respectively, in 6.3.1 and 6.3.2, can then calculated as

$$Dev^1_{bias}\left(\hat{\beta}_j\right) = \widehat{bias}_B^{\,1}\left(\hat{\beta}_j\right) - bias^1\left(\hat{\beta}_j\right), \tag{6.3.11}$$

and

$$Dev^2_{bias}\left(\hat{\beta}_j\right) = \widehat{bias}_B^{\,2}\left(\hat{\beta}_j\right) - bias^2\left(\hat{\beta}_j\right).$$



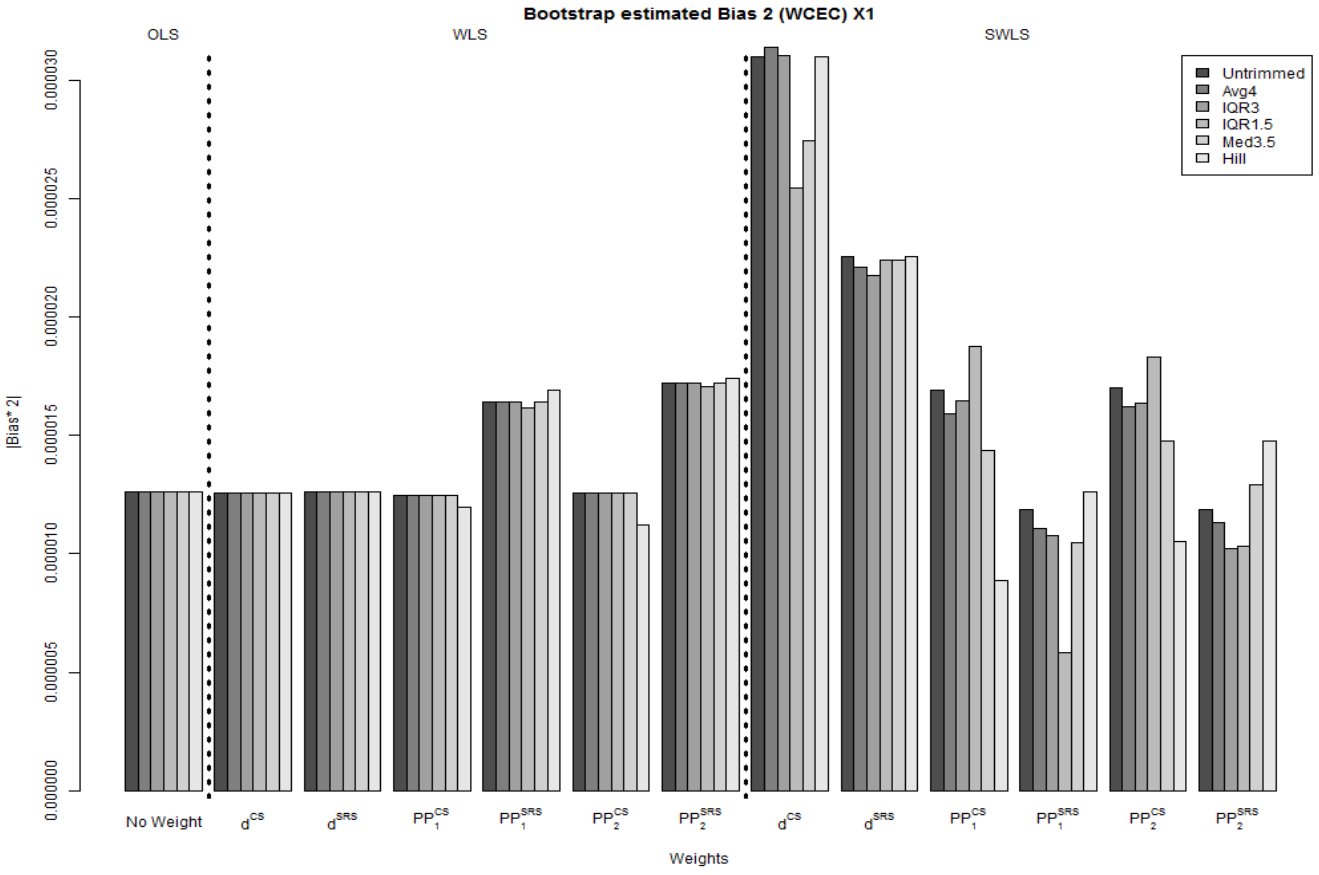


Figure 6.3.6: WCEC Bootstrap estimated Bias 1 and 2 for coefficient of predictor X_1

At first glance it seems as if the OLS bootstrap regression as well as the incorrect WLS bootstrap regression performed very well when compared to the SWLS bootstrap bias. However, the results obtained using the benchmarked weights under SWLS performed even better, especially under 1.5IQR and Hill trimming. Furthermore, comparing the range of the y-axis in figure 6.3.6 to that of figure 6.3.1 it is clear that the bootstrap estimator of the regression parameter is closer to the "true" parameter.

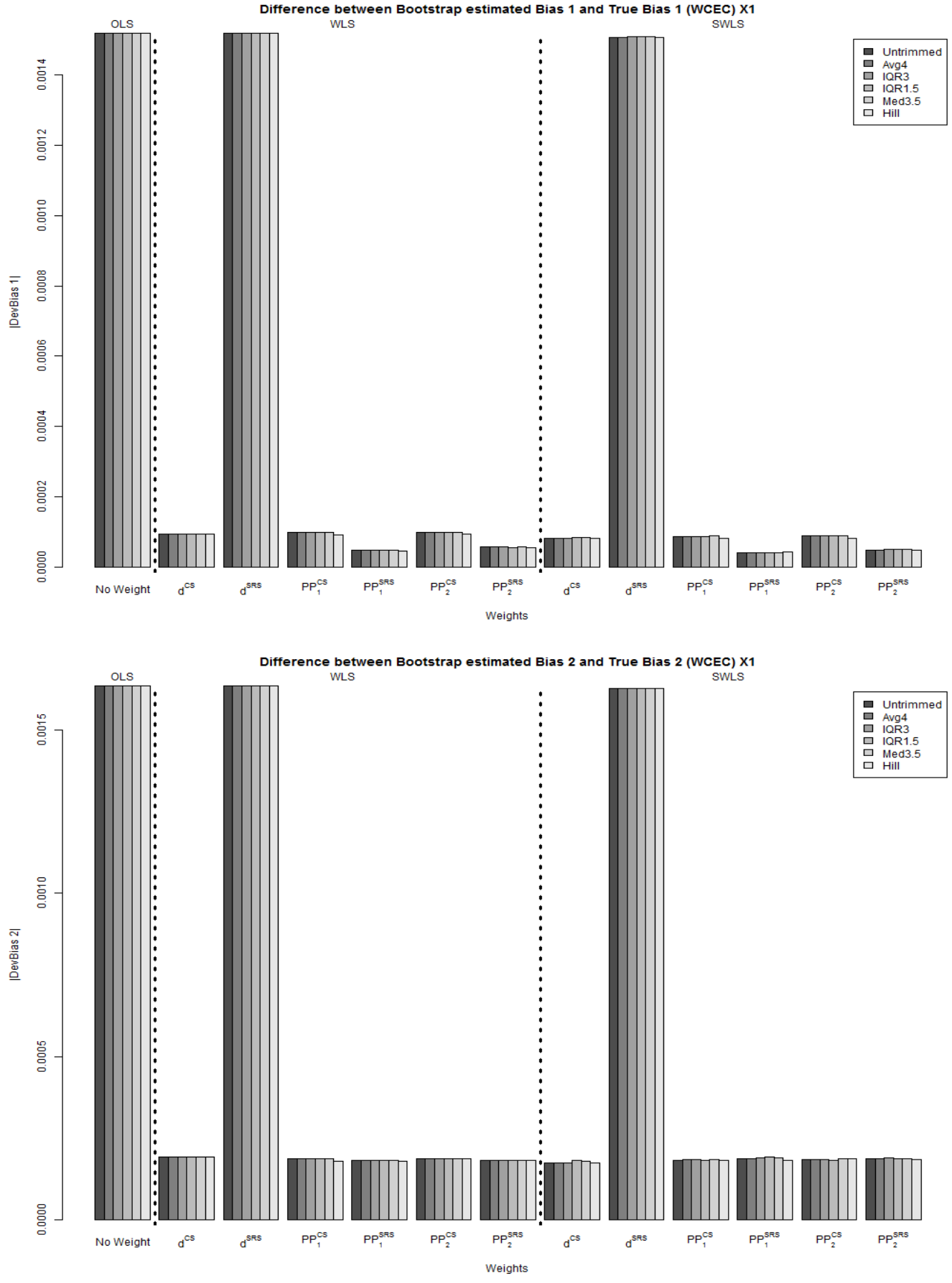


Figure 6.3.7: WCEC Absolute Value of Difference between Bootstrap estimated Bias 1 and 2 and "True" Bias 1 and 2 of X_1

Now consider the difference between the bootstrap estimated bias and the "true" bias of the estimator of a regression parameter. It is quite clear that the OLS application which, from figure 6.3.6, appeared to be doing very well, does not fair well at all in comparison to the "true" bias of the estimator. A significant improvement in precision is observed once the sampling weights are included in the model. In this case the unbenchmarked and benchmarked weights performed very similarly. Furthermore, the WLS results in figure 6.3.7 let it seem as if the WLS approach could be satisfactory. However, as illustrated in the theory discussed in chapter 4, and considering how the bootstrap regression under WLS was carried out (see explanation under section 6.2.2.2), the results can be seen as an example of how wrong the conclusions could be when not modeling CS data correctly.

From this point on only the deviation between the bootstrap estimated and "true" biases will be shown since the trend observed in the above figures remains quite similar throughout.

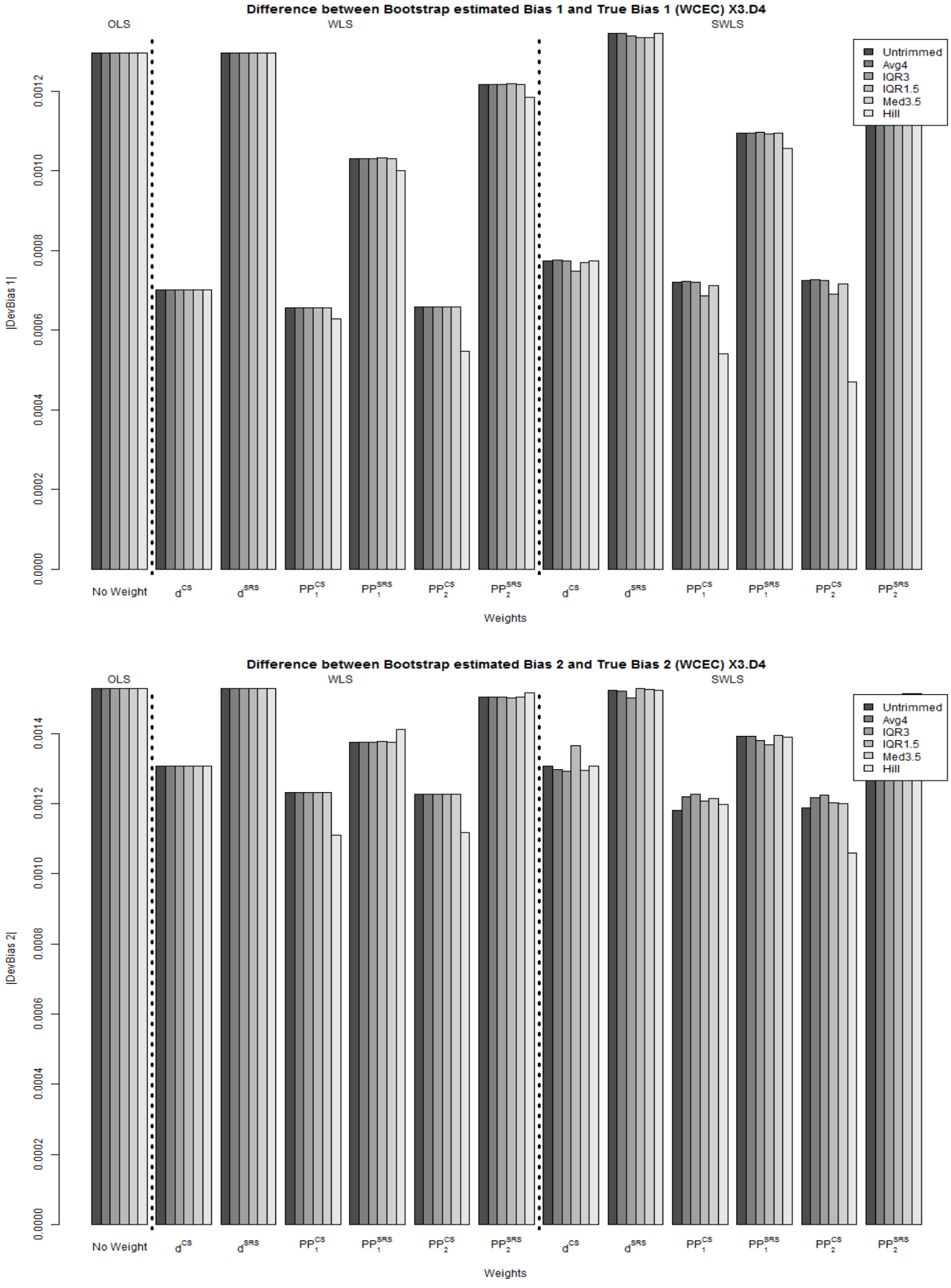


Figure 6.3.8: WCEC Absolute Value of Difference between Bootstrap estimated Bias 1 and 2 and "True" Bias 1 and 2 of X_3= 4

In figure 6.3.8 it is seen how the SWLS estimator of the regression parameter decreases the difference between the bootstrap estimated and "true" biases of the estimator. The "best" precision is obtained when using the $w_{CS}^{pp_2}$ weights that have been trimmed using the Hill trimming method.

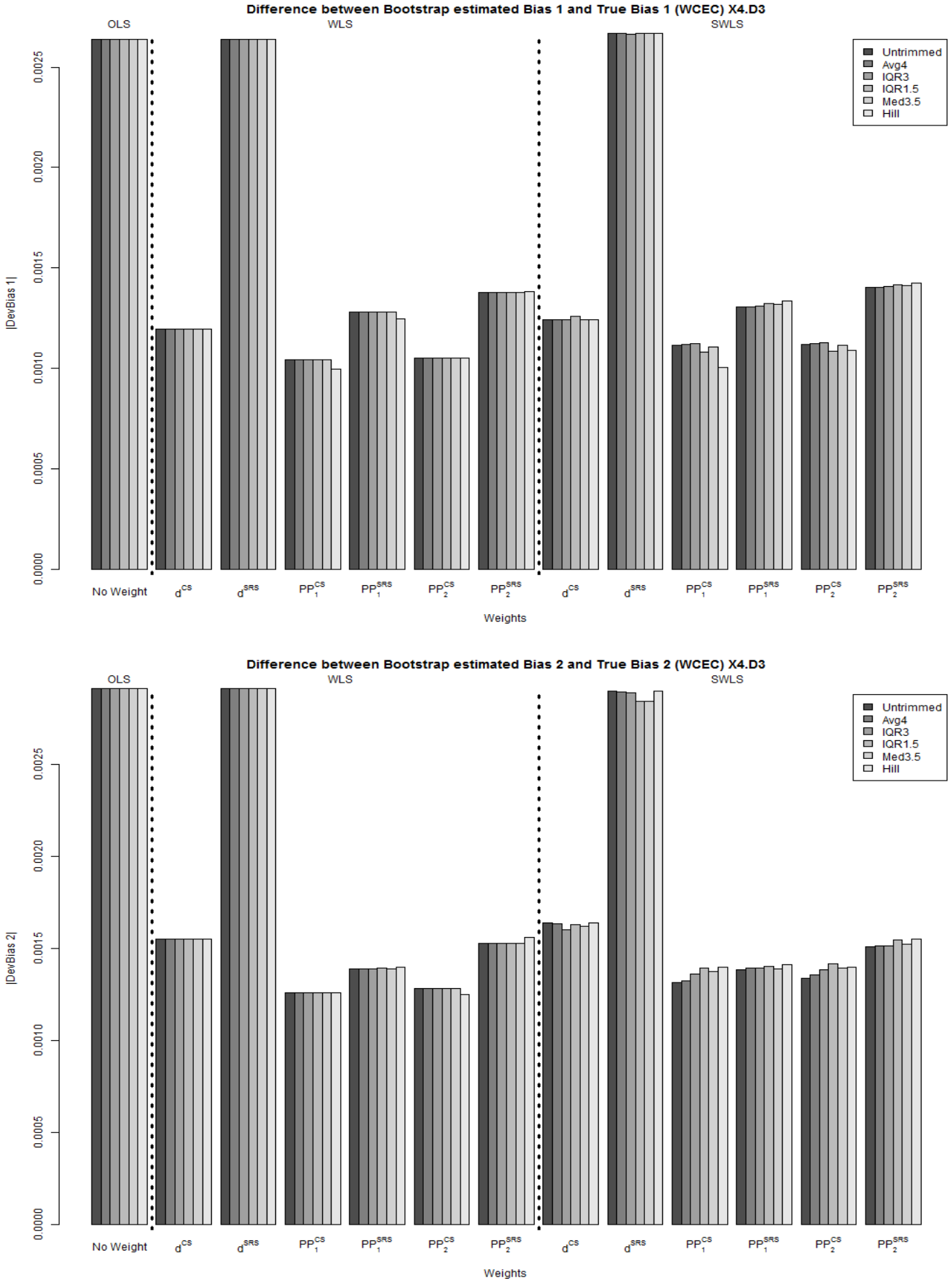


Figure 6.3.9: WCEC Absolute Value of Difference between Bootstrap estimated Bias 1 and 2 and "True" Bias 1 and 2 of X_4= 3

The conclusions made from figure 6.3.8 are further confirmed by the results presented in figure 6.3.9, but the "best" precision here is achieved using the Hill trimmed $w_{CS}^{pp_1}$ weights (see top plot in figure).

- Similar to the two bootstrap estimates of bias, two bootstrap estimates of MSE are calculated for each of the R bootstrap populations,

$$\widehat{MSE}_B^1\left(\hat{\beta}_{r_j}\right) = \frac{1}{B}\sum_{r_b=1}^{B}\left(\hat{\beta}^*_{r_{b_j}} - \hat{\beta}_{r_j}\right)^2, \tag{6.3.12}$$

and

$$\widehat{MSE}_B^2\left(\hat{\beta}_{r_j}\right) = \frac{1}{B}\sum_{r_b=1}^{B}\left(\hat{\beta}^*_{r_{b_j}} - \overline{\hat{\beta}^*_{r_{b_j}}}\right)^2, \tag{6.3.13}$$

where $\overline{\hat{\beta}^*_{r_{b_j}}}$ is the average of the bootstrap estimates of the rth estimator of the jth regression coefficient.

Then the overall bootstrap estimates of MSE of $\hat{\beta}_j$ are, respectively, given by

$$\widehat{MSE}_B^1\left(\hat{\beta}_j\right) = \frac{1}{R}\sum_{r=1}^{R}\widehat{MSE}_B^1\left(\hat{\beta}_{r_j}\right), \tag{6.3.14}$$

and

$$\widehat{MSE}_B^2\left(\hat{\beta}_j\right) = \frac{1}{R}\sum_{r=1}^{R}\widehat{MSE}_B^2\left(\hat{\beta}_{r_j}\right). \tag{6.3.15}$$

The differences between the bootstrap estimates of MSE and the true MSE's defined, respectively, in 6.3.3 and 6.3.4 are then calculated as

$$Dev_{MSE}^1\left(\hat{\beta}_j\right) = \widehat{MSE}_B^1\left(\hat{\beta}_j\right) - MSE^1\left(\hat{\beta}_j\right), \tag{6.3.16}$$

and

$$Dev_{MSE}^2\left(\hat{\beta}_j\right) = \widehat{MSE}_B^2\left(\hat{\beta}_j\right) - MSE^2\left(\hat{\beta}_j\right). \tag{6.3.17}$$

Note that the results presented make use of the square root of the estimated mean squared errors and differences.

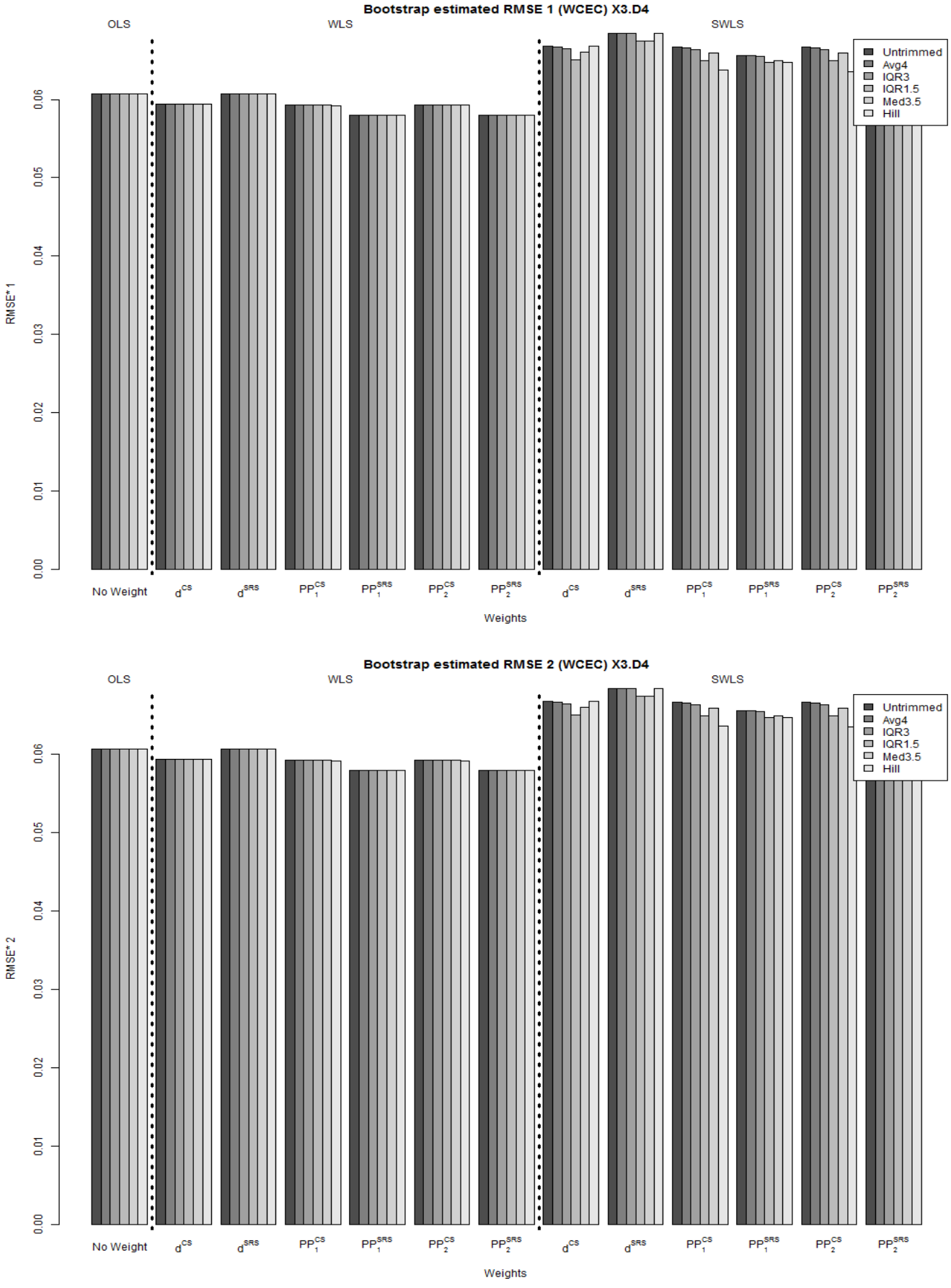


Figure 6.3.10: WCEC Bootstrap estimated RMSE 1 and 2 for coefficient of predictor $X_3 = 4$

Figure 6.3.10 shows that the bootstrap estimated RMSE is slightly larger under SWLS than under OLS and WLS. Recall from chapter 4 that the estimated variance of the estimator of a regression coefficient under SWLS differs from that under WLS which could be the reason for the slightly larger RMSE under SWLS. The Hill trimming achieved a reduction in the RMSE of the estimator, but it is still larger than under the other two models. This is another example of the distortion of results when applying statistical analyses incorrectly to CS data.

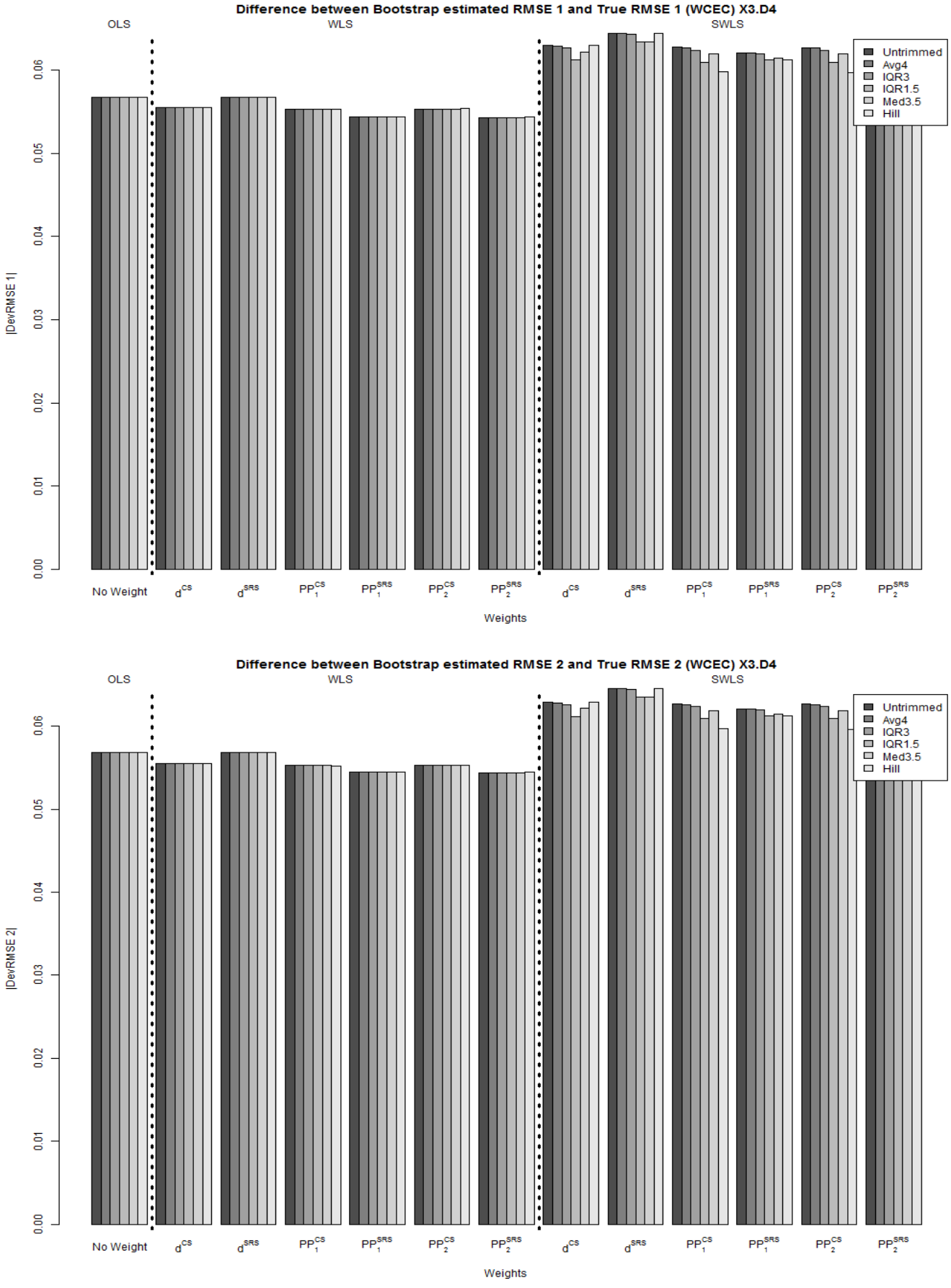


Figure 6.3.11: WCEC Absolute Value of Difference between Bootstrap estimated Bias 1 and 2 and "True" Bias 1 and 2 of X_3= 4

Figure 6.3.11 agrees with the conclusions made from figure 6.3.10.

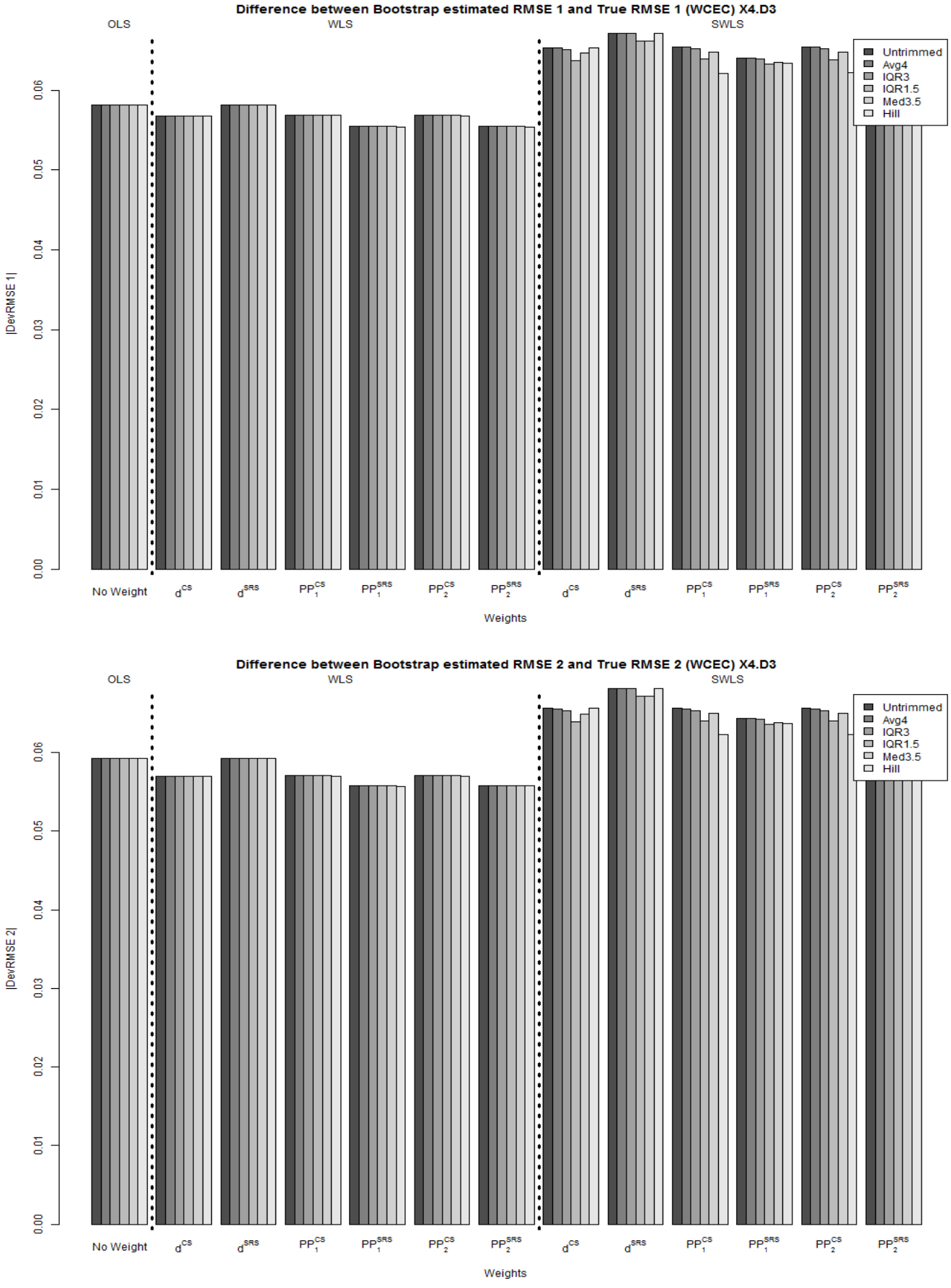


Figure 6.3.12: WCEC Absolute Value of Difference between Bootstrap estimated Bias 1 and 2 and "True" Bias 1 and 2 of X_3= 4

The same conclusions can also be made from figure 6.3.12. Clearly the variance of the estimator has been affected by the correct use of the sampling weights under SWLS, but at least some improvement in precision is obtained when applying the Hill trimming method.

- Two bootstrap estimators of MAD are calculated for each of the R bootstrap populations,

$$\widehat{MAD}_B^1\left(\hat{\beta}_{r_j}\right) = median\left|\hat{\beta}^*_{r_{b_j}} - \hat{\beta}_{r_j}\right|, \tag{6.3.18}$$

and

$$\widehat{MAD}_B^2\left(\hat{\beta}_{r_j}\right) = median\left|\hat{\beta}^*_{r_{b_j}} - median\left\{\hat{\beta}^*_{r_{b_j}}\right\}\right|, \tag{6.3.19}$$

where $median\left\{\hat{\beta}^*_{r_{b_j}}\right\}$ is the median of the bootstrap estimates of the rth estimator of the jth regression coefficient.

Then the overall bootstrap estimated MAD's of $\hat{\beta}_j$ are, respectively, given by

$$\widehat{MAD}_B^1\left(\hat{\beta}_j\right) = \frac{1}{R}\sum_{r=1}^{R}\widehat{MAD}_B^1\left(\hat{\beta}_{r_j}\right), \tag{6.3.20}$$

and

$$\widehat{MAD}_B^2\left(\hat{\beta}_j\right) = \frac{1}{R}\sum_{r=1}^{R}\widehat{MAD}_B^2\left(\hat{\beta}_{r_j}\right). \tag{6.3.21}$$

The differences between the bootstrap estimators of MAD and the true MAD's defined, respectively, in 6.3.5 and 6.3.6 are then calculated as

$$Dev^1_{MAD}\left(\hat{\beta}_j\right) = \widehat{MAD}_B^1\left(\hat{\beta}_j\right) - MAD^1\left(\hat{\beta}_j\right), \tag{6.3.22}$$

and

$$Dev^2_{MAD}\left(\hat{\beta}_j\right) = \widehat{MAD}_B^2\left(\hat{\beta}_j\right) - MAD^2\left(\hat{\beta}_j\right). \tag{6.3.23}$$

The results will also not be presented here since conclusions drawn from the differences between the bootstrap estimated median absolute deviations and the "true" median absolute deviations are in line with those based on the RMSE. They are available available on the accompanying CD.

- The relative bias of the estimated variances of $\hat{\beta}_j$ with respect to, firstly, $MSE^1\left(\hat{\beta}_j\right)$,

$$RelBias^1 = \left[\frac{\sum_r \hat{V}(\hat{\beta}_{r_j})/_R}{MSE^1\left(\hat{\beta}_j\right)}\right] - 1, \tag{6.3.24}$$

and secondly, with respect to $MSE^2\left(\hat{\beta}_j\right)$,

$$RelBias^2 = \left[\frac{\sum_r \hat{V}(\hat{\beta}_{r_j})/_R}{MSE^2\left(\hat{\beta}_j\right)}\right] - 1, \tag{6.3.25}$$

where $\hat{V}\left(\hat{\beta}_{r_j}\right)$ is the estimated variance of the rth estimator of jth regression coefficient (Kovar et al., 1988). These diagnostics will be considered for the estimated variances obtained from the modeling software, denoted by $\hat{V}_M\left(\hat{\beta}_{r_j}\right)$, as well as the bootstrap estimated variances, $\hat{V}_B\left(\hat{\beta}_{r_j}\right)$.

The linear modeling functions that are part of the R statistical software produce estimated variances as part of the linear model output regarding the estimated regression parameters. Let this be known as the model estimated variance. In addition to this variance estimator, which in some statistical software is estimated using the TSL (Taylor series linearization) method, the bootstrap method of variance estimation was also employed.

The figures of the relative bias results will not be given here to curb the length of the thesis document. However, they are available available on the accompanying CD. The figures were constructed for the relative bias of the model estimated variance as well as the bootstrap estimated variance. It was concluded that for both the model and the bootstrap estimated variance the relative bias was considerably reduced by making use of SWLS and specifically with the benchmarked theoretical design weights, namely $w_{CS}^{pp_1}$ and $w_{CS}^{pp_2}$. Furthermore, the alternative design weights and their associated benchmarked weights could not outperform the precision achieved with the "CS" sampling weights.

6.3.2 Model Parameter Confidence Interval Diagnostics

The next part of the regression parameter inference considers the construction of confidence intervals for the jth regression parameter based on the rth replicate sample. The confidence intervals considered in this research were discussed in section 4.5, but are given here for the reader's convenience:

1. The standard confidence interval, calculated as

$$\left[\hat{\beta}_{r_j} - z_{\frac{\alpha}{2}} \cdot \sqrt{\hat{V}\left(\hat{\beta}_{r_j}\right)}; \hat{\beta}_{r_j} + z_{\frac{\alpha}{2}} \cdot \sqrt{\hat{V}\left(\hat{\beta}_{r_j}\right)}\right], \tag{6.3.26}$$

where $\hat{V}\left(\hat{\beta}_{r_j}\right)$ is the estimated variance of $\hat{\beta}_{r_j}$ obtained from the linear modeling function of the statistical software R.

2. The standard confidence interval, calculated as

$$\left[\hat{\beta}_{r_j} - z_{\frac{\alpha}{2}} \cdot \sqrt{\hat{V}_B\left(\hat{\beta}_{r_j}\right)}; \hat{\beta}_{r_j} + z_{\frac{\alpha}{2}} \cdot \sqrt{\hat{V}_B\left(\hat{\beta}_{r_j}\right)}\right], \tag{6.3.27}$$

where $\hat{V}_B\left(\hat{\beta}_{r_j}\right)$ is the bootstrap estimated variance of $\hat{\beta}_{r_j}$.

3. The percentile confidence interval, calculated as

$$\left[\hat{\beta}_{r_{j_{lo}}}, \hat{\beta}_{r_{j_{up}}}\right] = \left[\hat{\beta}^*_{r_{([B\cdot\frac{\alpha}{2}])_j}}, \hat{\beta}^*_{r_{([B\cdot(1-\frac{\alpha}{2})])_j}}\right], \tag{6.3.28}$$

where $\hat{\beta}_{r_{j_{lo}}}$ and $\hat{\beta}_{r_{j_{up}}}$ are the $\left[B \cdot \frac{\alpha}{2}\right]$ largest and $\left[B \cdot \left(1 - \frac{\alpha}{2}\right)\right]$ largest values of the sorted bootstrap replicates, $\left\{\hat{\beta}^*_{r_{b_j}}\right\}$.

4. The bootstrap-t confidence interval, calculated as

$$\left[\hat{\beta}_{r_j} - t^*_{j_U} \cdot \sqrt{\hat{V}_B\left(\hat{\beta}_{r_j}\right)}, \hat{\theta}_r - t^*_{j_L} \cdot \sqrt{\hat{V}_B\left(\hat{\beta}_{r_j}\right)}\right], \tag{6.3.29}$$

where $t^*_{j_U}$ and $t^*_{j_L}$ are respectively the lower and upper $\frac{\alpha}{2}$-points obtained from $t^*_{j_{(1)}}, ..., t^*_{j_{(B)}}$, the ordered values of

$$t^*_{r_{b_j}} = \frac{\hat{\beta}^*_{r_{b_j}} - \hat{\beta}_{r_j}}{\sqrt{\hat{V}_{JK}\left(\hat{\beta}^*_{r_{b_j}}\right)}}, \tag{6.3.30}$$

and $V_{JK}\left(\hat{\beta}^*_{r_{b_j}}\right)$ is the jackknife estimated variance of $\hat{\beta}^*_{r_{b_j}}$.

5. The bootstrap-t confidence interval, calculated as

$$\left[\hat{\beta}_{r_j} - t^*_{j_U} \cdot \sqrt{\hat{V}_B\left(\hat{\beta}_{r_j}\right)}, \hat{\beta}_{r_j} - t^*_{j_L} \cdot \sqrt{\hat{V}_B\left(\hat{\beta}_{r_j}\right)}\right], \tag{6.3.31}$$

where $t^*_{j_U}$ and $t^*_{j_L}$ are respectively the lower and upper $\frac{\alpha}{2}$-points obtained from $t^*_{j_{(1)}}, ..., t^*_{j_{(B)}}$, the ordered values of

$$t^*_{r_{b_j}} = \frac{\hat{\beta}^*_{r_{b_j}} - \hat{\beta}_{r_j}}{\sqrt{\hat{V}_B\left(\hat{\beta}^*_{r_{b_j}}\right)}}, \tag{6.3.32}$$

and $V_B\left(\hat{\beta}^*_{r_{b_j}}\right)$ is the bootstrap estimated variance of $\hat{\beta}^*_{r_{b_j}}$ obtained by performing a second-level bootstrap.

6. The BCa confidence interval, calculated as

$$\left[\hat{\beta}_{r_{j_{lo}}}, \hat{\beta}_{r_{j_{up}}}\right] = \left[\hat{\beta}^*_{r_{\left(B\cdot\frac{\alpha_1}{2}\right)_j}}, \hat{\beta}^*_{r_{\left(B\cdot\frac{\alpha_2}{2}\right)_j}}\right], \tag{6.3.33}$$

where $\hat{\beta}_{r_{j_{lo}}}$ and $\hat{\beta}_{r_{j_{up}}}$ are the $\left[B\cdot\frac{\alpha_1}{2}\right]$ largest and $\left[B\cdot\frac{\alpha_2}{2}\right]$ largest values of the sorted bootstrap replicates, $\left\{\hat{\beta}^*_{r_{b_j}}\right\}$ and α_1 and α_2 are the probabilities defined in 4.5.17 and 4.5.18 and obtained by adjusting the percentiles using the bias-correction and acceleration constants defined in 4.5.15 and 4.5.16, respectively.

The theory surrounding these confidence intervals was discussed in section 4.5.2.1 and it should be noted that the confidence intervals will be constructed based on the OLS, WLS and SWLS estimators of the jth unknown regression parameter. The following summary measures were calculated for the different confidence intervals:

- For each of the confidence intervals their non-coverage probability (NCP), measuring the proportion of times that the interval does not contain the true value of the parameter of interest, is calculated. From each of the R bootstrap populations, one confidence interval is calculated for each of the two standard intervals, the percentile intervals, the two bootstrap-t intervals and the BCa interval. This results in R standard confidence intervals, R percentile confidence intervals, R bootstrap-t confidence intervals and R BCa confidence intervals. Let $\hat{\beta}_{r_{j_{lo}}}$ be the lower limit of the rth confidence interval and let $\hat{\beta}_{r_{j_{up}}}$ be the upper limit of the rth confidence interval. Hence, there are R lower limits

$$\hat{\beta}_{1_{j_{lo}}}, \hat{\beta}_{2_{j_{lo}}}, ..., \hat{\beta}_{R_{j_{lo}}},$$

and R upper limits

$$\hat{\beta}_{1_{j_{up}}}, \hat{\beta}_{2_{j_{up}}}, ..., \hat{\beta}_{R_{j_{up}}},$$

for each of the different confidence intervals methods and for each linear model, namely OLS, WLS and SWLS. Then,

$$NCP_{lo} = \frac{\hat{\beta}_{r_{j_{lo}}} > \hat{\beta}_j}{R},$$

measures the lower non-coverage probability (NCP) and

$$NCP_{up} = \frac{\hat{\beta}_{r_{j_{up}}} < \hat{\beta}_j}{R},$$

measures the upper non-coverage probability of each of the different confidence intervals methods. The total non-coverage probability is then obtained as the sum of the lower and upper non-coverage probabilities,

$$NCP = NCP_{lo} + NCP_{up}. \tag{6.3.34}$$

- The length of the confidence interval is calculated as the difference between the R upper limits and lower limits of each different confidence interval method

$$l_r = \hat{\beta}_{r_{j_{up}}} - \hat{\beta}_{r_{j_{lo}}}, \tag{6.3.35}$$

resulting in R confidence interval lengths

$$l_1, l_2, ..., l_R,$$

for each confidence interval method and each of the linear modeling methods. The average length (AvgLen) of each different confidence interval method,

$$AvgLen = \frac{1}{R}\sum_{r=1}^{R} l_r, \tag{6.3.36}$$

is then plotted for each different weighting method described above to compare the different confidence interval methods. From the length of the confidence intervals their standardized lengths (Std Lenght) are calculated (Kovar et al., 1988),

$$\frac{AvgLen}{2 \cdot z_{\frac{\alpha}{2}}\sqrt{MSE\left(\hat{\beta}_j\right)}}, \tag{6.3.37}$$

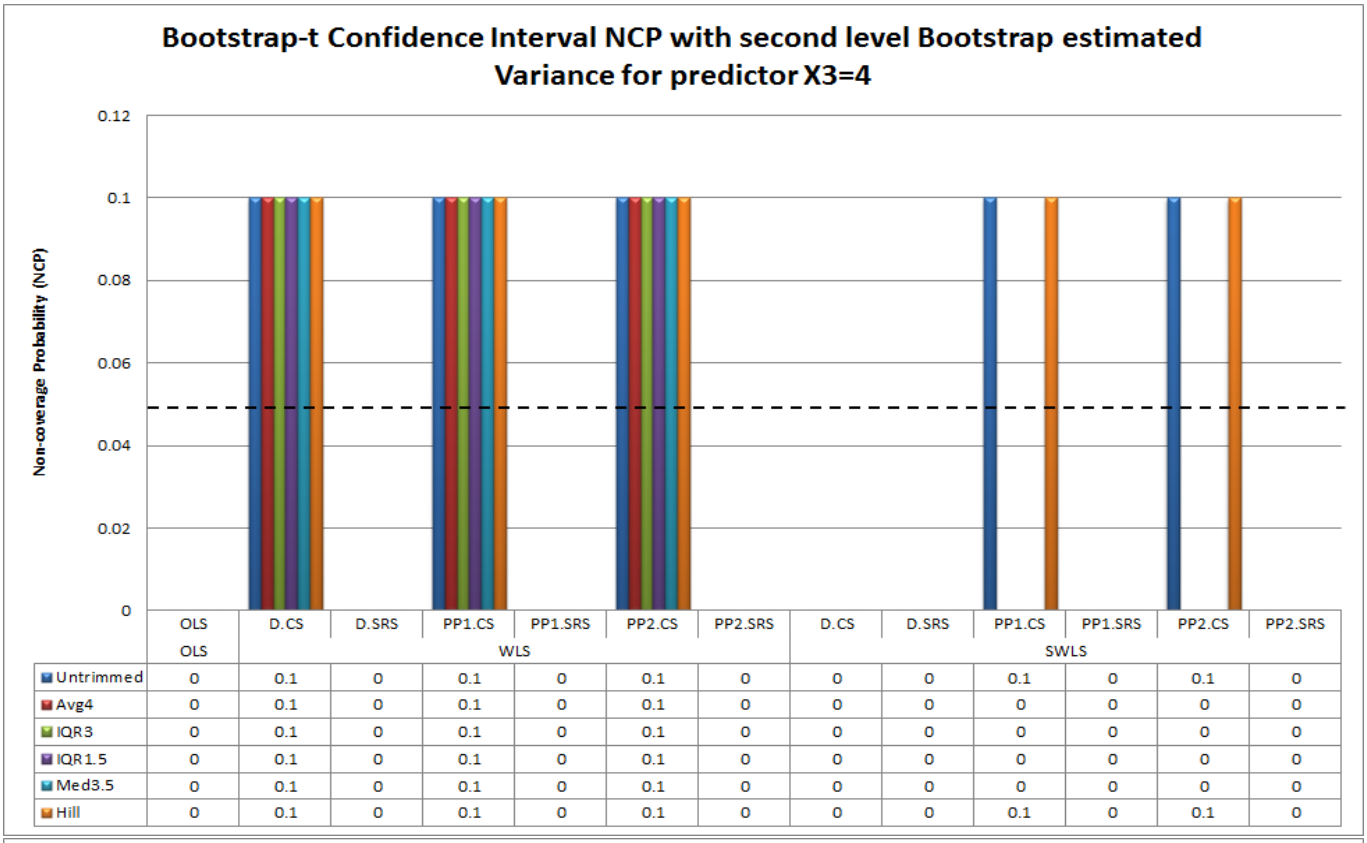
where $MSE\left(\hat{\beta}_j\right)$ is the true MSE as defined in 6.3.3 or 6.3.4.

Note that all of the above diagnostic measures will be calculated and compared for the following:

- OLS, WLS, and SWLS;
- design weights compared to benchmarked weights;
  - design weights based on inclusion probabilities; and
  - SRS design weights.
- untrimmed weights compared to trimmed weights.

The figures presented below show the non-coverage probability (NCP) and the standardized length (RMSE) of the bootstrap-t confidence intervals of a parameter. The bars are grouped at two stages. Firstly they are grouped by OLS, WLS, and SWLS. Then they are grouped by sampling weight and the coloured bars, left to right, in each grouping represent a selection of the weight trimming methods. Recall that the subscript "CS" refers to the theoretical design weights and their associated benchmarked weights while the "SRS" subscript refers to the alternative design weights and their associated benchmarked weights. Furthermore, the NCP figures contain a dotted line which represents the intended significance level, namely 5%. Each figure also contains a table that shows the observed values that correspond to each group of bars.

Note that the results of the other confidence intervals are not presented here, but are available from the author. The standard (asymptotic) intervals either did not include the parameter (OLS and WLS) or over-covered the parameter (SWLS) irrespective of the standardized lengths not being very large. The percentile intervals performed poorly under OLS and under the "SRS" sampling weights. However, some over-coverage occurred under SWLS with the exception of the $w_{CS}^{pp_1}$ Hill trimmed weights. The BCa intervals performed similarly to the bootstrap-t intervals presented below, but over-coverage was still observed.



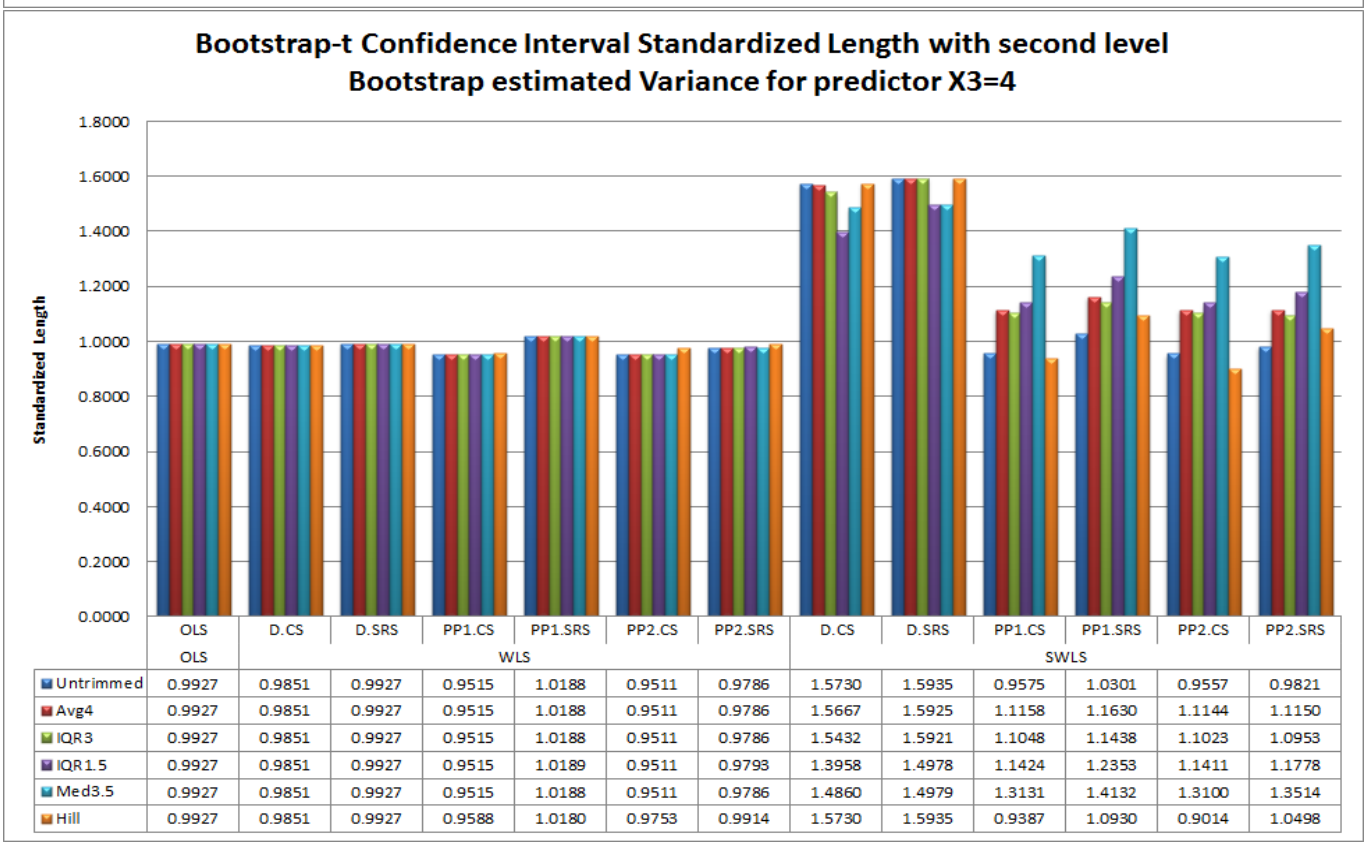
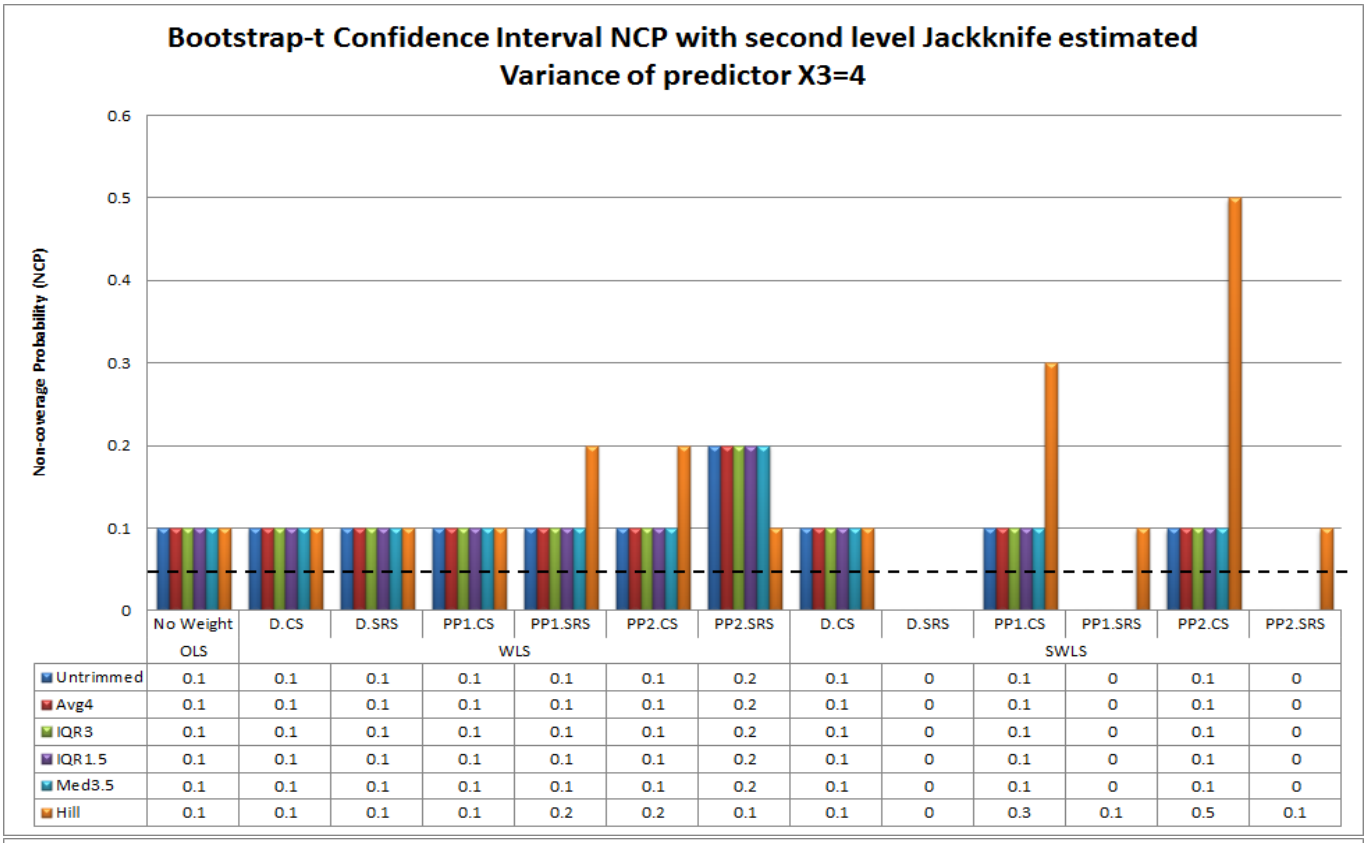
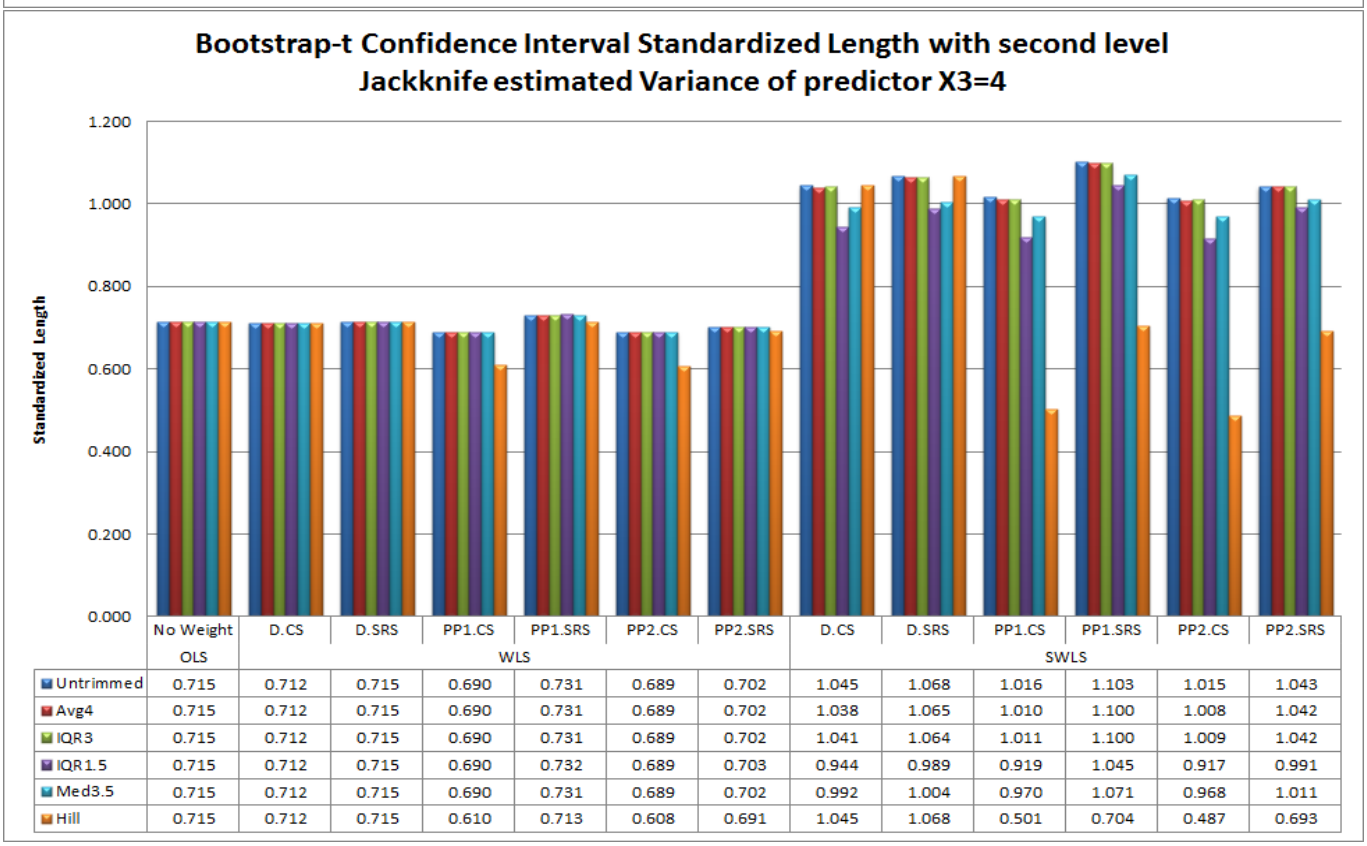


Figure 6.3.13: Bootstrap-t Confidence Interval NCP and Standardized Length with second level Bootstrap estimated Variance for predictor X_3= 4

Figure 6.3.13 presents the bootstrap-t interval where the variances of the bootstrap replicates were estimated by the application of a second level bootstrap. It is clear that the over-coverage also occurred under the bootstrap-t interval. However, consider the NCP and associated standardized length under SWLS with $w_{CS}^{pp_1}$ and $w_{CS}^{pp_2}$. Here an improvement in the NCP was observed irrespective of the standardized lengths being reduced to a minimum (Hill trimming).



	No Weight OLS	D.CS	D.SRS	PP1.CS	PP1.SRS	PP2.CS	PP2.SRS	D.CS	D.SRS	PP1.CS	PP1.SRS	PP2.CS	PP2.SRS
		WLS						SWLS					
Untrimmed	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0	0.1	0	0.1	0
Avg4	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0	0.1	0	0.1	0
IQR3	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0	0.1	0	0.1	0
IQR1.5	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0	0.1	0	0.1	0
Med3.5	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0	0.1	0	0.1	0
Hill	0.1	0.1	0.1	0.1	0.2	0.2	0.1	0.1	0	0.3	0.1	0.5	0.1



	No Weight OLS	D.CS	D.SRS	PP1.CS	PP1.SRS	PP2.CS	PP2.SRS	D.CS	D.SRS	PP1.CS	PP1.SRS	PP2.CS	PP2.SRS
		WLS						SWLS					
Untrimmed	0.715	0.712	0.715	0.690	0.731	0.689	0.702	1.045	1.068	1.016	1.103	1.015	1.043
Avg4	0.715	0.712	0.715	0.690	0.731	0.689	0.702	1.038	1.065	1.010	1.100	1.008	1.042
IQR3	0.715	0.712	0.715	0.690	0.731	0.689	0.702	1.041	1.064	1.011	1.100	1.009	1.042
IQR1.5	0.715	0.712	0.715	0.690	0.732	0.689	0.703	0.944	0.989	0.919	1.045	0.917	0.991
Med3.5	0.715	0.712	0.715	0.690	0.731	0.689	0.702	0.992	1.004	0.970	1.071	0.968	1.011
Hill	0.715	0.712	0.715	0.610	0.713	0.608	0.691	1.045	1.068	0.501	0.704	0.487	0.693

Figure 6.3.14: Bootstrap-t Confidence Interval NCP and Standardized Length with second level Jackknife estimated Variance for predictor X_3= 4

In contrast to figure 6.3.13 the above figure presents the bootstrap-t interval where the jackknife method was used to estimated the variance of a bootstrap replicate. The "SRS" weights under SWLS resulted in over-coverage and increased standardized lengths. Considering the combination of the NCP and the standardized length it is concluded that the interval under SWLS with $w_{CS}^{pp_1}$ and $w_{CS}^{pp_2}$ untrimmed and trimmed weights achieved close to the desired NCP while decreasing the standardized lengths.

The results of the parameter estimators and confidence intervals were presented in this section. Overall conclusions regarding these results will be given in the final section below.

6.3.3 Summary and Conclusions

This part of the chapter contained the results of the parameter estimation and confidence intervals of the parameters. In both parts the results were presented in a manner that would make the comparison of the type of linear model (OLS, WLS, SWLS) possible in terms of the type of sampling weight, namely unbenchmarked and benchmarked "CS" and "SRS" sampling weights, and whether the sampling weights were untrimmed or trimmed. Although a number of trimming methods were considered, only a selection of these could be included in the presentations of the results. Also, the linear model contained many regression parameters and thus only some of the estimators' results could be presented in this chapter.

The estimators of the parameters were assessed on how they compare to the "truth" as well as how their bootstrap estimated versions compare to the "truth". It was clear that a difference exists between whether the CS data are modeled in such way that the design is accounted for (SWLS) versus when the design is only partially accounted for (WLS) or not at all (OLS). Although the nature is to consider the smallest bias, RMSE, etc. as an indication of the best attained precision of an estimator, this is not the golden rule here. One must remember that the use of SWLS as the correct linear model for CS data is considered the golden standard and thus all results need to be compared to the SWLS results. It was seen in many of the figures that SWLS, especially after trimming the sampling weights, did indeed achieve the minimum bias, RMSE, etc. It was also mostly found that the "CS" sampling weights outperformed the "SRS" sampling weights.

The results for the various interval estimators were a fairly mixed bag. This could be attributed to the minimum number of bootstrap samples used, namely 200, as well as the fact that results from only ten samples from the surrogate population were aggregated over. Although the samples from the surrogate population were approximately 10% of the population size, these numbers do perhaps not sufficiently account for the variability due to random sampling. The intervals mostly over-covered the parameters, but some cases were observed where acceptable NCP levels were attained. These occurred mostly under Hill trimmed $w_{CS}^{pp_1}$ and $w_{CS}^{pp_2}$ weights under SWLS. Further research will include increasing the number of samples and bootstrap samples used for the interval estimation.

Simulated data have been used here such that the theory and techniques could be evaluated and compared under "controllable" circumstances. However, this compares to the "learning set" approach used for the evaluation and honing phases of an experiment, but now the methods need to be exposed to a "test set" in the form of real-world data. The next chapter presents the Income and Expenditure Survey (IES) data obtained from a survey conducted by Statistics South Africa in 2005. The IES chapter will be followed by the outlines of and results from the analyses based on the IES data.

Chapter 7

Income and Expenditure Survey 2005/2006

The previous two chapters, firstly, discussed how CS data could be simulated such that the theory and techniques considered in this thesis could be evaluated in a "controlled" environment. This was followed up by a description of the simulation study and analyses conducted using the simulated data. However, a necessity exists to take the methodologies investigated under the simulated data and the conclusions made there and to further evaluate their performance when applied to real-world data. The evaluation will be done by repeating the same analyses outlined in chapter 6 and considering summary diagnostics such as the bias and MSE of estimators, the non-coverage probability and standardized lengths of confidence intervals, model prediction error, etc. Using the diagnostics given in chapter 6, conclusions will be made through the comparison of OLS (no weighting) and WLS to SWLS using untrimmed and trimmed theoretical and alternative design and benchmarked weights.

The survey data identified for this purpose is the Income and Expenditure Survey of 2005/2006, a survey conducted by Statistics South Africa (Stats SA) every five years. This chapter contains a description of the data set as well as how Stats SA conducted the survey. Aspects of the survey that will receive attention include the design and the weighting used.

7.1 Income and Expenditure Survey 2005/2006

The data set that will be used in the analysis and that will act as surrogate population, is the Income and Expenditure survey conducted over the period September 2005 until August 2006, hereafter referred to as IES. The intention of this survey is to examine income and expenditure as well as poverty and inequality in South Africa. Households that were sampled took part in the survey for one month after which new sub-samples of households started taking part in the survey at the beginning of each month (Lehohla, 2008).

As of this IES, Statistics SA changed the methodology used in previous surveys of this kind. Previously the recall method was used, but now a combination of the recall method and the

diary method was used. In a nutshell, a main questionnaire consisting of five interview modules is administered by a fieldworker to a selected household. Each interview was conducted on five different visits. The main questionnaire required households to account for their acquisitions of the following goods and services (Lehohla, 2008):

- Durable

  Items or services that last a long time. For example cars, furniture, etc.

- Semi-durable

  Items that require replacement more often than durable items. For example clothing, shoes, etc.

This information, as well as income acquired by different members of a household, was collected over the eleven months prior to the survey.

The new part of the survey methodology required households to keep a diary of their daily acquisitions over the four weeks of the survey. These diaries were collected on a weekly basis and the purpose was to ensure that the information collected was as close as possible to the period of transaction. Information collection was based on acquisition that takes into account the total value of all goods and services acquired during a given period (Lehohla, 2008).

7.1.1 Data Collection Methods

Three methods were used to collect the survey information (Lehohla, 2008):

1. Main Questionnaire

   It consisted of a booklet of questions administered to respondents during the course of the survey month. As mentioned before, the main questionnaire consisted of five parts. The first part covered household characteristics, the next three parts covered different categories of consumption expenditure and the final part covered household income.

2. Weekly Diary

   Each household had to write down their daily acquisitions according to specific "categories" namely the nature, type, source and purpose of the item acquired.

3. Summary Questionnaire

   The fieldworker had to "summarize" the total value of each item acquired during the week and then had to transfer it to the appropriate section of the questionnaire. This assisted the fieldworker in summarizing the consumption expenditure of each household during the survey month.

7.1.2 Response and Imputation of non-response

As mentioned before, there are two types of non-response, namely unit non-response and item non-response. Unit non-response is taken care of during weighting while item non-response requires imputation at different levels. Here, imputation was done for missing diaries as well as item non-response (Lehohla, 2008).

An imputation method called cell mean imputation is used by Statistics SA. This method divides the data into groups according to variables with no missing values. The mean value is then imputed into the missing values (Lohr, 2010). For the missing diaries, households were divided into groups according to the number of diaries completed within the four weeks of the survey. Those households with less than two diaries or a diary but no main questionnaire were considered non-respondent. The mean expenditure of respondent households, those with two or more completed diaries, were imputed (Lehohla, 2008).

For the item non-response, respondent households with similar characteristics to the non-respondents were grouped together and the average value for these households were imputed (Lehohla, 2008).

7.2 Survey Design

The sampling frame used for IES was a newly designed master sample based on the enumeration areas of the 2001 population census (Lehohla, 2008). The selection of PSU's require the availability of a frame or list of all PSU's. When such a frame is used for multiple surveys or multiple rounds of the same survey, it is known as a master sample frame. A master sample is a sample from which sub-samples can be selected to serve the needs of more than one survey or survey round (Pettersson, 2005). An enumeration area (EA) is the smallest geographical unit (piece of land) into which the country is divided for survey purposes and EA's were used as PSU's (Lehohla, 2008).

The 3000 EA's in the master sample were stratified into four groups of 750 EA's each. A random sample of 250 PSU's were selected each month. From each selected PSU, a systematic sample of 8 dwelling units was chosen. A dwelling unit (DU) is defined as any structure or part of a structure or group of structures occupied or meant to be occupied by one or more than one household. Thus, a stratified two-stage cluster sample was used with the four groups as explicit stratification variable, enumeration areas as PSU's and dwelling units as SSU's. So, 24000 DU's were interviewed over the twelve month period. This design ensured that the sample was evenly spread over the twelve months while being nationally representative in each of the four groups (Lehohla, 2008).

7.3 Weighting

Consider the nine provinces of South Africa as the strata. Note that, in contrast to the method followed by Statistics South-Africa in the IES, the PSU's selected for the samples used in the simulation study based on the IES, were selected with equal probability. Hence, let the inclusion probability of the jth PSU in the hth stratum, $h = 1, ..., 9$, be given by

$$\pi_{hj} = \frac{n_h}{N_h},$$

and let the inclusion probability of a household in the jth PSU be given by $^{n_{hj}}/_{N_{hj}}$ where

- n_h is the number of PSU's selected from stratum h;
- N_h is the population number of PSU's in the hth stratum;
- N_{hj} is the population number of households in the jth PSU in the hth stratum; and
- n_{hj} is the number of households selected from the jth PSU of the hth stratum.

For the purpose of this research, all non-responsive units were deleted from the IES and thus a 100% response rate was assumed. See chapter 8 for an explanation of this decision. Finally, the design weight is given by

$$w_{hji} = \frac{N_h}{n_h} \cdot \frac{N_{hj}}{n_{hj}},$$

where $h = 1, ..., 9$ and $j = 1, ..., n_h$.

7.4 Simulated Data sets

The IES survey described in sections 7.1 and 7.2 formed the basis of the definition of a surrogate population.

A number of adjustments to the original IES had to be made in order to obtain a "clean" surrogate population from which repeated samples could be selected. Firstly, all observations with missing data values, were removed. If the missing values were imputed, this would introduce another level of uncertainty and variability into the data that could affect the precision of the inference. Although imputation as a research area has received some attention, it did not form part of this research and as such the missing values were removed rather than imputed using some simple, but not really recommended, method. Next, only observations for which an age of at least 21 and no older than 65 was captured, were retained. This could be considered a typical working-age interval since individuals that pursue a tertiary education, could start working after a minimum three year bachelor's degree, and 65 is considered as a general retirement age.

Furthermore, starting the interval at an age of 21 still includes those individuals that had only limited or even no education.

At this point the size of the original IES was still in excess of 46000 observations, a number that does not sound very large in an era of big data. However, the functions used in the application of SWLS take many seconds to complete each time an SWLS is fitted. When doing this repeatedly the computer time starts adding up very quickly. Thus, a final adjustment was made based on the decision to truncate values of the model response to be positive, i.e. $y > 0$, where y represents the personal income of an individual. After this final adjustment the surrogate population consisted of 17541 households grouped into 283 EA's which amounted to 25893 individuals. The original IES EA's were grouped to reduce the number to 283.

Monte Carlo simulation was applied to the surrogate population which consisted of drawing 110 samples from the population where each sample has the same design as the IES 2005/2006 survey: a stratified two-stage cluster design with EA's as sampling frame of PSU's and the nine provinces as strata. For the purpose of this research the PSU's of the original IES were re-grouped to form new larger PSU's. A total of 169 of the larger PSU's were selected and from each of these PSU's, 12 dwelling units (SSU's) were selected. This amounted to each sample consisting of 2028 observations. The samples were used in the analyses to, among others,

- compare inference results obtained from applying OLS (no weights), WLS and SWLS to CS data;
- investigate whether improved precision is achieved when addressing the large variability in the sampling weight distribution through the application of weight trimming methods; and
- compare the results obtained when using unbenchmarked and benchmarked theoretical design weights, i.e. d_{CS} and its associated benchmarked weights, $w_{CS}^{pp_1}, ..., w_{CS}^{ph_2}$, as sampling weights versus using sampling weights obtained from benchmarking the raw data, i.e. d_{SRS} and its associated benchmarked weights, $w_{SRS}^{pp_1}, ..., w_{SRS}^{ph_2}$.

Note that differential non-response, for example the under-representation of white people living in urban areas and small households, is found in practical situations in South Africa. Thus to be able to determine this type of non-response error, it was simulated in the design of the samples through the use of auxiliary variables. This was done to evaluate the weighting procedures under non-perfect circumstances. Two sets of auxiliary variables were used in the simulation to aid in determining which weighting technique would be best under these circumstances:

- The first set contains only person level auxiliary variables, indicated by "pp". These are
  1. province, with 9 categories;
  2. gender, with 2 categories;

3. race, with 4 categories; and
4. age, with 4 categories.

- The second set contains person and household level auxiliary variables, indicated by "ph". These are

1. all person level auxiliary variables;
2. area, with 2 categories;
3. dwelling type, with 2 categories; and
4. household size, with 3 categories.

After the selection of the replicate samples the bootstrap and jackknife methods were applied to the samples for the purpose of further examining the questions outlined before, especially in terms of variance estimation, confidence intervals and other measures of accuracy. The application of these simulated data sets as well as any summary measures used to address the outlined research questions, will be discussed in the next chapter.

The Income and Expenditure survey conducted over the period from September 2005 to August 2006 by Statistics South Africa was adjusted to obtain a surrogate population from which smaller data sets could generated by means of Monte Carlo simulation. This chapter introduced the surrogate population and presented a very short discussion of the samples collected from it. The next chapter will return to the samples as well as the repeated samples that will be utilized in the analyses and will also discuss the various inferences conducted using these different levels of samples.

Chapter 8

Income and Expenditure Survey 2005/2006 Analyses

8.1 Sampling Scheme

Chapter 6 considered the application of the theory and techniques discussed in chapters 3 and 4 to samples selected from two populations simulated using multilevel modeling. Making use of samples from simulated populations made it possible to evaluate the models under "controllable" conditions in terms of making sure that the model assumptions are met. However, the real-world data used by survey statisticians do not necessarily adhere to such assumptions and thus it is necessary to conduct the same analyses as in chapter 6, but now making use of real-world data.

The data set identified for this purpose is the Income and Expenditure survey (IES) of 2005, introduced and described in the previous chapter, and the objective is to model personal income, Y, based on a selection of covariates from the IES. Some adjustments were made to the original IES such that a "clean" data set could be obtained, refer to chapter 7 for a description of these adjustments, and this became the surrogate population for the IES analyses. The following covariates were included in the model:

- age, X_1;
- gender (1 = male, 2 = female), X_2;

  A dummy variable was constructed for gender and "female" was chosen as the reference category.

- race (1 = black, 2 = coloured, 3 = indian/asian, 4 = white);

  Since "black" had the largest proportion of observations in the IES, it was used as the reference category. Three dummy variables, RD_2, RD_3, RD_4, were formed for the remaining three race categories. The subscripts are used for identification of the race category with the subscript for the reference category being set to 1.

- level of education (coded from 0 to 26).

  The education levels were grouped into 7 categories according to the code definitions set out in the IES meta data file:

  1. No school: 0, 1 and 26;
  2. Non-completed primary school (grade 1 - 6): 2 - 7;
  3. Completed primary school (grade 7): 8;
  4. Early high school (grade 8 - 9): 9 and 10;
  5. Non-completed high school (grade 10 - 11): 11 and 12;
  6. Completed high school (grade 12): 13; and
  7. Post grade 12: 14 - 25.

  Note that the code "0" indicated that the respondent received no education, "1" indicated pre-school, and "26" was used when a respondent selected the "don't know" option. For the purpose of this research the "26" was regarded as the respondent having received no education. Furthermore, the decision to distinguish between "early high school" and "non-completed high school" was based on the South African school system considering the introduction of a grade 9 school-leaving certificate to offer learners a path other than matric or tertiary education. The "no education" category was selected as the reference category. Six dummy variables, $ED_2, ..., ED_7$, were then constructed for the remaining categories of the new education level predictor.

These predictors comprise the main effects of the IES linear model. After conducting a preliminary investigation the following first-order interactions were included as well:

1. gender by race;
2. gender by education level; and
3. race by education level.

Hence, the IES linear model is given by

$$\begin{aligned}\mathbf{y} &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 RD_2 + \beta_4 RD_3 + \beta_5 RD_4 + \beta_6 ED_2 + \beta_7 ED_3 + \beta_8 ED_4 + \\ &\quad \beta_9 ED_5 + \beta_{10} ED_7 + \beta_{11} ED_7 + (first - order\, interactions) + \underline{\varepsilon},\end{aligned}$$

and thus 39 regression parameters, including the intercept and the interactions, have to be estimated using OLS, WLS and SWLS.

Generally population information is not available since it is too difficult, time consuming or expensive to observe an entire population, but by letting the IES fulfill the role of a surrogate population, it is possible to approximate "true" regression parameters, $\{\beta_j\}$, $j = 0, ..., p$. This enables the comparison of the OLS, WLS and SWLS estimators to the "truth" in an attempt to gauge which approach performed the "best". The "true" values of the main effects model parameters, obtained from the surrogate population, are given in the table below.

Intercept	Age	Gender	Coloured	Indian/Asian	White	Gr. 1 - 6	Gr. 7	Gr. 8 - 9	Gr. 10 - 11	Gr. 12	Post Gr. 12
-13501.44	638.83	-5341.69	2954.25	36002.97	159991.99	9505.35	17070.00	18424.51	19994.45	42385.50	108586.59

Table 8.1.1: IES "true" main effects model parameters

Monte Carlo simulation was applied to the surrogate population to form artificial replicates of the population and each replicate followed the same design as that of the surrogate population. Each sample consists of 9 strata with a total of 169 PSU's (EA's) selected across the strata. Twelve SSU's (households) were selected from each of the sampled PSU's and thus each sample consists of 2028 SSU's. It should be mentioned that since one of the objectives of the analyses is to evaluate the effect of trimming sampling weights, the samples were selected with equal probability at both PSU and SSU in the hope that large weight variability would be achieved.

In this thesis the number of samples was limited to $R = 10$ due to the complex nature of the SWLS linear modeling and the sizes of the samples which require so many hours to complete per sample that ten was the maximum that could be completed within the time frame of this thesis. Let the estimators of the regression coefficients obtained from the rth replicate sample be denoted by $\left\{\hat{\beta}_{r_j}\right\}$, $j = 0, ..., p$ and $r = 1, ..., R$.

The bootstrap resampling technique is also employed here for the purpose of variance estimation as well as for the construction of confidence intervals for the regression parameters. The number of bootstrap samples, B, also had to be limited for the same reasons mentioned before. Thus, B was set equal to 200, a number considered to be the minimum number required, for example, for the bootstrap-t interval (Efron et al., 1998). The same number applied for the second level bootstrap sampling required for the bootstrap-t interval. Let the bootstrap estimator of the jth regression parameters be denoted by $\hat{\beta}_{r_{b_j}}$, $j = 0, ..., p$, $b = 1, ..., B$ and $r = 1, ..., R$.

This chapter commences with the evaluation of the fitted linear model by considering the coefficient of multiple determination, the model prediction error (PE) and the model outlier diagnostics. Recall that the PE estimation methods are the LOOCV, bootstrap PE and .632 bootstrap PE estimation methods. The outlier diagnostics include, for example, the leverages and DFFits. The model evaluation section is followed by the parameter estimation section which consists of a point estimation part and a interval estimation part. The point estimation part includes, viz., "true" bias, bootstrap estimated square root of mean squared error (RMSE), difference between bootstrap estimated and "true" median absolute deviation. The interval part contains, viz., the

standard (asymptotic) interval as well as a selection of bootstrap confidence intervals. The intervals will be evaluated based on their non-coverage probabilities, lengths and standardized lengths. Each part will be concluded with a selection of summaries of the diagnostics presented in that part.

8.2 Model Evaluation Analysis

The same analysis outline presented in section 6.2, will be followed here. This section commences with descriptive measures of the coefficient of multiple determination. Next, the models are evaluated based on their prediction errors (PE's) that are estimated using LOOCV and two bootstrap methods. Finally a selection of outlier diagnostics is presented. All of the results are presented in tables and figures in the below subsections and will be summarized in the conclusion of this section.

8.2.1 Coefficient of Multiple Determination

Consider the $R = 10$ replicate samples and let the corresponding coefficients of multiple determination be denoted by $R_1^2, ..., R_{10}^2$.

An improvement in R^2 is observed from OLS (0.1963 to 0.3725) to SWLS where the largest R^2 is obtained when the M3 trimming method is applied to the person benchmarked sampling weights, $w_{CS}^{pp_1}$. In the weight notation the subscript "CS" is used to denote the benchmarked theoretical design weights and the superscript "pp_1" that the benchmarking was conducted using only person-level auxiliary variables with the linear distance function. These R^2 values might seem disappointing, but it was mentioned in Heeringa et. al (2010) that R^2 between 0.25 and 0.4 when fitting linear models to real-world data is considered an achievement.

Consider a table of the average R^2 over the ten samples by linear model and trimming method accompanied by the standard deviation. The averages highlighted in green are the maxima while the highlighted standard deviations are their associated standard deviations.

		MEAN							STANDARD DEVIATION					
		No	Avg	IQR	Med	Hill	M3		No	Avg	IQR	Med	Hill	M3
OLS		0.2828	0.2828	0.2828	0.2828	0.2828	0.2828		0.0574	0.0574	0.0574	0.0574	0.0574	0.0574
WLS	d_{CS}	0.3035	***0.3035***	0.3003	0.3022	0.2986	0.3004		0.0538	***0.0535***	0.0532	0.0531	0.0568	0.0525
	d_{SRS}	0.2865	0.2864	0.2849	0.2864	0.2861	0.2850		0.0510	0.0509	0.0520	0.0508	0.0508	0.0522
	$w_{CS}^{pp_1}$	0.2975	0.2975	0.2948	0.2955	0.2888	0.2950		0.0542	0.0539	0.0536	0.0531	0.0512	0.0551
	$w_{SRS}^{pp_1}$	0.2813	0.2813	0.2814	0.2814	0.2859	0.2801		0.0528	0.0528	0.0526	0.0525	0.0509	0.0538
	$w_{CS}^{pp_2}$	0.2974	0.2974	0.2946	0.2954	0.2884	0.2944		0.0539	0.0536	0.0533	0.0528	***0.0505***	0.0548
	$w_{SRS}^{pp_2}$	0.2809	0.2809	0.2811	0.2810	0.2826	0.2794		0.0522	0.0521	0.0520	0.0519	0.0527	0.0559
	$w_{CS}^{ph_1}$	0.2975	0.2973	0.2951	0.2952	0.2925	0.2950		0.0547	0.0543	0.0546	0.0534	0.0542	0.0558
	$w_{SRS}^{ph_1}$	0.2824	0.2823	0.2823	0.2822	0.2832	0.2824		0.0519	0.0518	0.0517	0.0516	0.0525	0.0522
	$w_{CS}^{ph_2}$	0.2974	0.2972	0.2950	0.2951	0.2926	0.2949		0.0544	0.0540	0.0544	0.0531	0.0558	0.0557
	$w_{SRS}^{ph_2}$	0.2824	0.2823	0.2822	0.2823	0.2829	0.2822		0.0512	0.0511	0.0510	0.0509	0.0515	0.0515
SWLS	d_{CS}	0.3035	***0.3035***	0.3003	0.3022	0.3015	0.2987		0.0538	*0.0535*	0.0532	0.0531	0.0525	0.0528
	d_{SRS}	0.2865	0.2864	0.2849	0.2864	0.2861	0.2853		0.0510	0.0509	0.0520	0.0508	0.0508	0.0525
	$w_{CS}^{pp_1}$	0.2975	0.2975	0.2948	0.2955	0.2888	0.2950		0.0542	0.0539	0.0536	0.0531	0.0512	0.0551
	$w_{SRS}^{pp_1}$	0.2813	0.2813	0.2814	0.2814	0.2859	0.2801		0.0528	0.0528	0.0526	0.0525	0.0509	0.0538
	$w_{CS}^{pp_2}$	0.2974	0.2974	0.2946	0.2954	*0.2884*	0.2944		0.0539	0.0536	0.0533	0.0528	***0.0505***	0.0548
	$w_{SRS}^{pp_2}$	0.2809	0.2809	0.2811	0.2810	0.2826	0.2794		0.0522	0.0521	0.0520	0.0519	0.0527	0.0559
	$w_{CS}^{ph_1}$	0.2975	0.2973	0.2951	0.2952	0.2925	0.2950		0.0547	0.0543	0.0546	0.0534	0.0542	0.0558
	$w_{SRS}^{ph_1}$	0.2824	0.2823	0.2823	0.2822	0.2832	0.2824		0.0519	0.0518	0.0517	0.0516	0.0525	0.0522
	$w_{CS}^{ph_2}$	0.2974	0.2972	0.2950	0.2951	0.2926	0.2949		0.0544	0.0540	0.0544	0.0531	0.0558	0.0557
	$w_{SRS}^{ph_2}$	0.2824	0.2823	0.2822	0.2823	0.2829	0.2822		0.0512	0.0511	0.0510	0.0509	0.0515	0.0515

Table 8.2.1: R^2 Mean and Standard Deviation over Replicate Samples

The overall highest average R^2 is 0.3035 (0.0535 standard deviation) is obtained using the theoretical design weights, d_{CS}, and the 4Avg trimming method. This being said, the values obtained from the trimmed weights are all quite similar. The standard deviations also do not fluctuate substantially. Furthermore it is also observed that the theoretical weights' R^2 are, on average, better than the those of the alternative weights.

8.2.2 Prediction Error Estimation

Recall the outline of the prediction error estimation in section 6.2.2. The R replicate samples are, firstly, considered to be R learning sets and an SWLS is fitted to each learning set. The SWLS model is used since each replicate sample has been selected using a complex sample design. The test sets that correspond to each learning set are defined to contain all units from the surrogate population that are not included in each learning set. The models fitted to each learning set are used to predict the corresponding test sets and these predictions are used to obtain "true" prediction errors. The calculation of the "true" prediction errors are approached in two ways: using the Luus approach where the "true" prediction error is defined as the average over the R "true" prediction errors; and using the Molinaro approach where the R "true" prediction errors are retained in their individual form. Refer to figure 6.2.1 in section 6.2.2 for a more elaborate explanation of the calculation of the "true" prediction errors. The "true" prediction errors are used

to gauge how close the estimated prediction errors, obtained using leave-one-out cross-validation (section 6.2.2.1 and figure 6.2.2), bootstrap estimation (section 6.2.2.2 and figure 6.2.9), and .632 bootstrap estimation (section 6.2.2.3 and figure 6.2.13), come to the "truth".

The results obtained from the simulated data and presented in section 6.2.2 showed promise. The LOOCV and Bootstrap PE estimation methods applied to the IES data performed similarly to the simulated data and consequently their results are not included in this section. The .632 Bootstrap PE estimation method is thus included here since, out of the three, it performed the best under the IES data. It should be noted that the results have been scaled, using the same transformation as in chapter 6, to ensure that all results are on the same scale. Also, the IES model predicts personal income and by scaling the diagnostic results, the numbers are more legible.

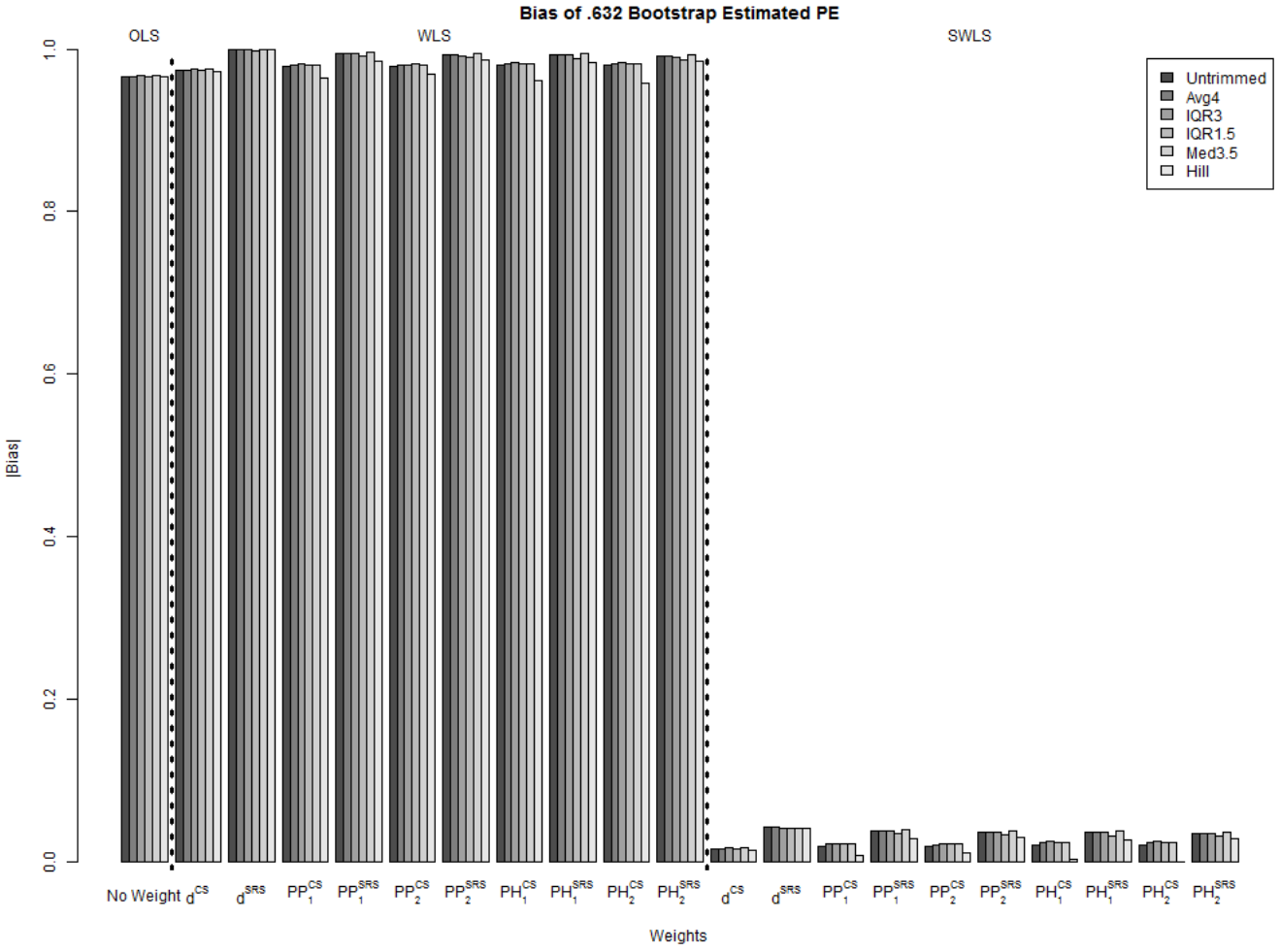


Figure 8.2.1: "True" Bias of .632 Bootstrap Estimated PE: Luus approach

Recall from section 6.2.2 that the Luus and Molinaro "true" biases simplify to be the same results and thus only figure 8.2.1 is included here. It is clear that the estimated prediction error of the SWLS model using the Hill trimmed benchmarked theoretical weights, especially $w_{CS}^{ph_2}$, outperforms using no weights or incorrectly using weights based on distance from the "true" prediction error.

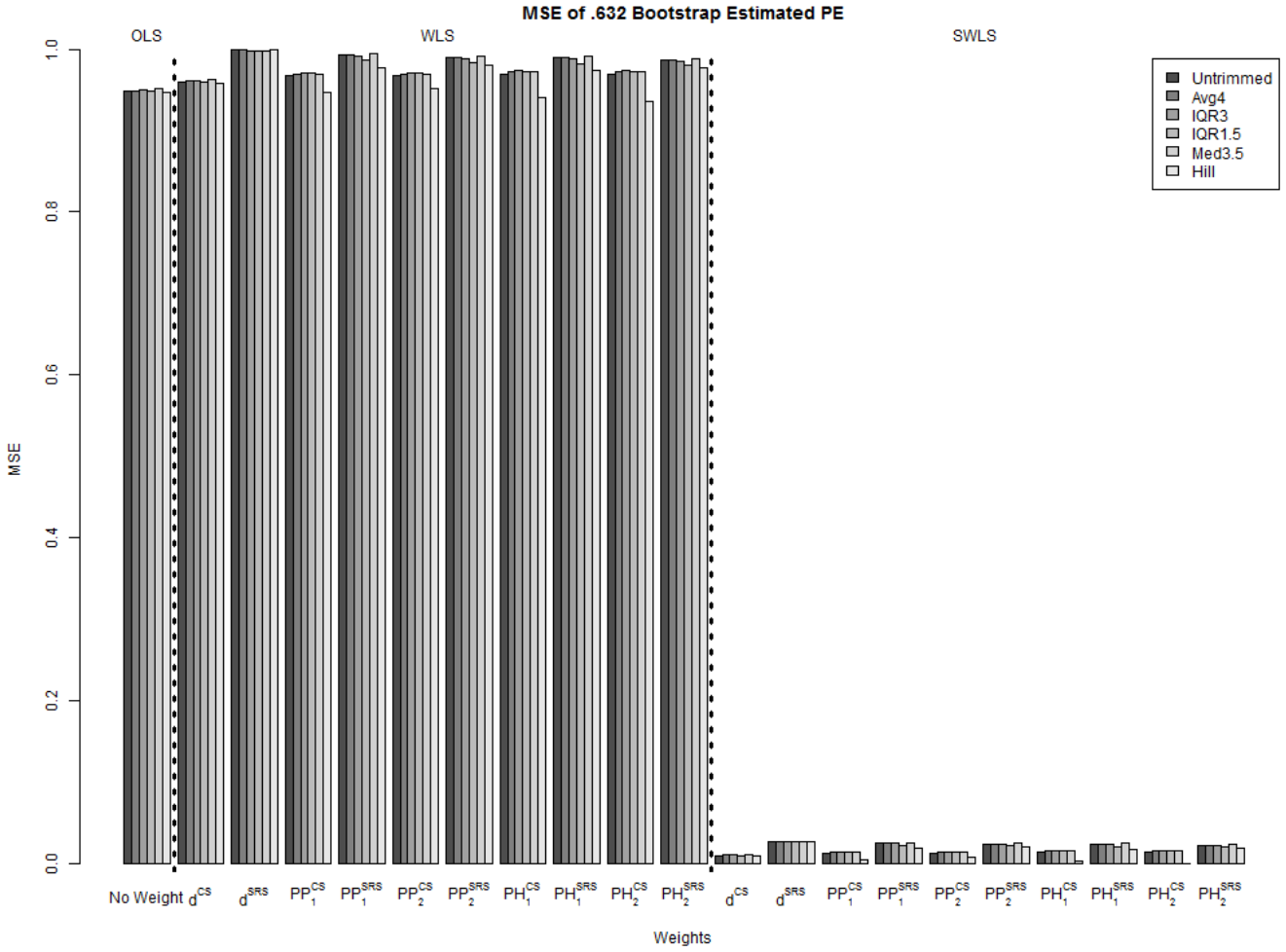


Figure 8.2.2: "True" MSE of .632 Bootstrap Estimated PE: Luus approach

The "true" MSE presented in figure 8.2.2 arrives at the same conclusion as figure 8.2.1. The "true" MSE based on the Molinaro approach achieved results very similar to the results presented in figure 8.2.2 and thus will not be included here.

The estimated standard deviation of the .632 bootstrap estimated prediction errors under SWLS are larger than under WLS or OLS. This could perhaps also be due to the small number of bootstrap samples and small number of replicate samples used in this simulation study. The same outcome was observed using the simulated data and thus will have to be researched further to determine whether the number of samples does in fact have an effect on the results.

This section showed a small selection of the prediction error estimation results obtained from the IES data. The final part of the model evaluation, namely outlier detection diagnostics, follows in the next section.

8.2.3 Outlier Detection Diagnostics

The table below presents the minimum, three quartiles and maximum of the response, the age covariate as well as the various sampling weights.

	QUANTILES				
	0%	25%	50%	75%	100%
Income	70	7200	10800	30773.75	3017069
Age	21	35	43	53	65
d_{CS}	2.25	5.125	8.3583	11.5376	19.6452
d_{SRS}	8.6494	8.6494	8.6494	8.6494	8.6494
$w_{CS}^{pp_1}$	2.0304	5.1366	8.3646	11.2673	22.0418
$w_{SRS}^{pp_1}$	4.5722	7.6263	8.4237	9.4683	13.6398
$w_{CS}^{pp_2}$	2.0372	5.1217	8.3932	11.2335	22.5861
$w_{SRS}^{pp_2}$	5.2461	7.6554	8.3725	9.3764	14.5943
$w_{CS}^{ph_1}$	1.917	5.2027	8.459	11.3375	22.2046
$w_{SRS}^{ph_1}$	4.117	7.6898	8.4628	9.5899	14.0656
$w_{CS}^{ph_2}$	1.9338	5.2069	8.4812	11.3188	22.2665
$w_{SRS}^{ph_2}$	4.9726	7.6627	8.4197	9.4955	15.2401

Table 8.2.2: Quantiles of Variables in IES Regression

It is clear that the theoretical sampling weights, with subscript "CS", have larger variation than the alternative sampling weights and the largest variation is observed once the design weights have been benchmarked, i.e. $w_{CS}^{pp_1}$, $w_{CS}^{pp_2}$, $w_{CS}^{ph_1}$ and $w_{CS}^{ph_2}$. Hence when considering the outlier diagnostics one should determine whether the observation is flagged due to it being outlying in comparison to the other observations, or due to the size of its associated sampling weight.

The figures presented now are bubble plots of the OLS versus SWLS leverages for the person- and household-level auxiliary variables, exponential distance function, benchmarked weights, $w_{CS}^{ph_2}$ and $w_{SRS}^{ph_2}$. The unbenchmarked design weights are not included since survey statisticians are known to use benchmarked sampling weights as their final sampling weights. The trimming methods presented here, in comparison to the untrimmed results, are the 1.5IQR, Hill and M3. All three of these trimming methods are the new methods introduced and defined in this thesis.

Each bubble plot presents the OLS leverages on the x-axis with cut-off denoted by the vertical line, and the SWLS leverages on the y-axis with cutoff denoted by the horizontal line. The size of the bubble is proportional to the weight used. The two cut-off lines divide each plot into four sections. The top left section contains all observations flagged only by SWLS and the bottom right section contains all observations flagged only by OLS.

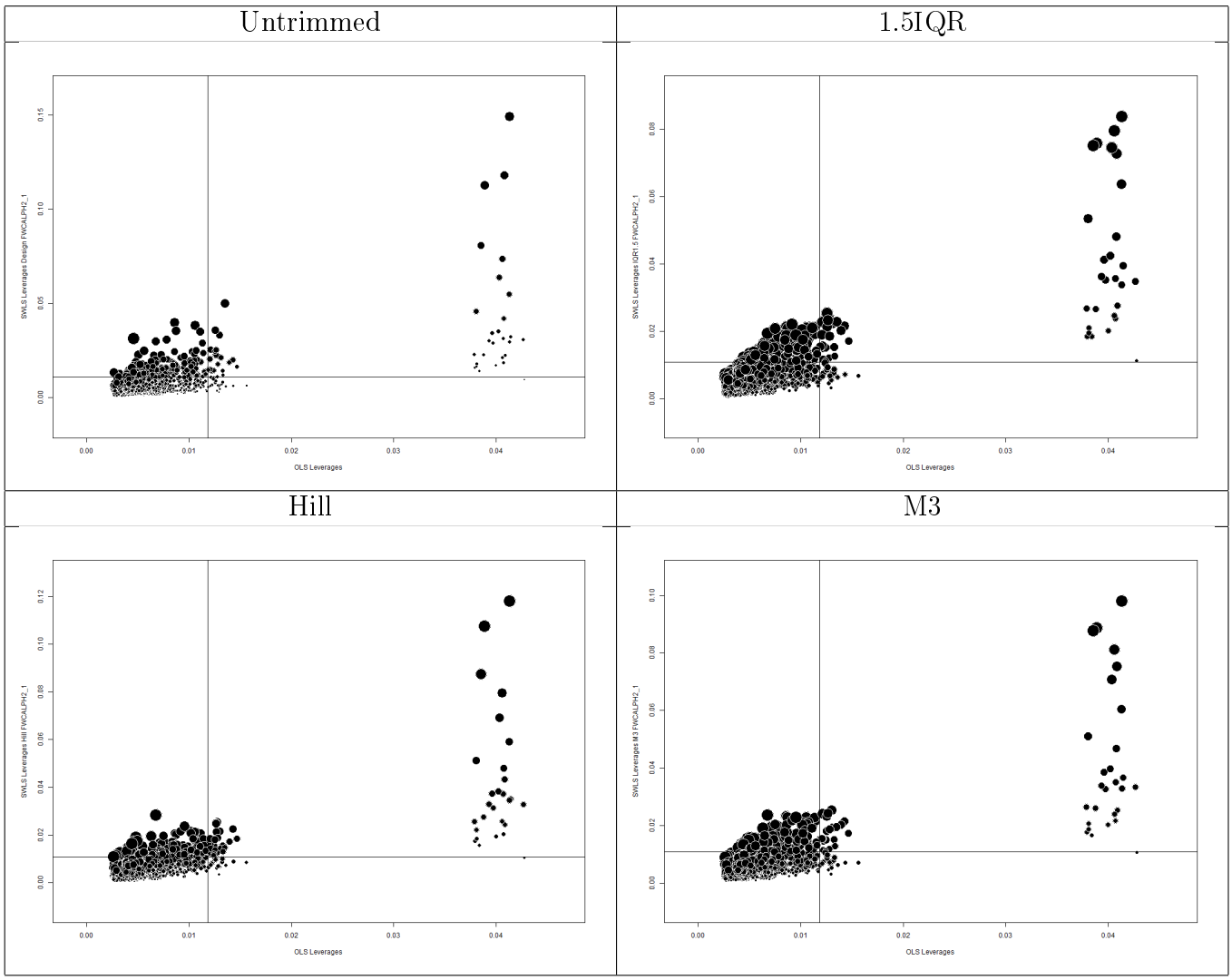


Figure 8.2.3: Bubble plots of OLS versus SWLS Leverages: $w_{CS}^{ph_2}$

The plot of the OLS leverages versus the untrimmed SWLS leverages shows that the OLS leverages identify observations as outliers (bottom right block) that are not identified by the SWLS leverages (top left block). Furthermore, those observations identified by the OLS leverages have smaller sampling weights than those identified by the SWLS leverages. Hence, SWLS leverages identify more observations with large weights as outliers while OLS leverages identify more observations with smaller weights. Now, consider the plots based on the trimmed weights. Although a similar pattern is observed as from the untrimmed plot, the bubbles of the observations flagged by the SWLS leverages are now more uniform in size than previously. The same can be said of the OLS outliers.

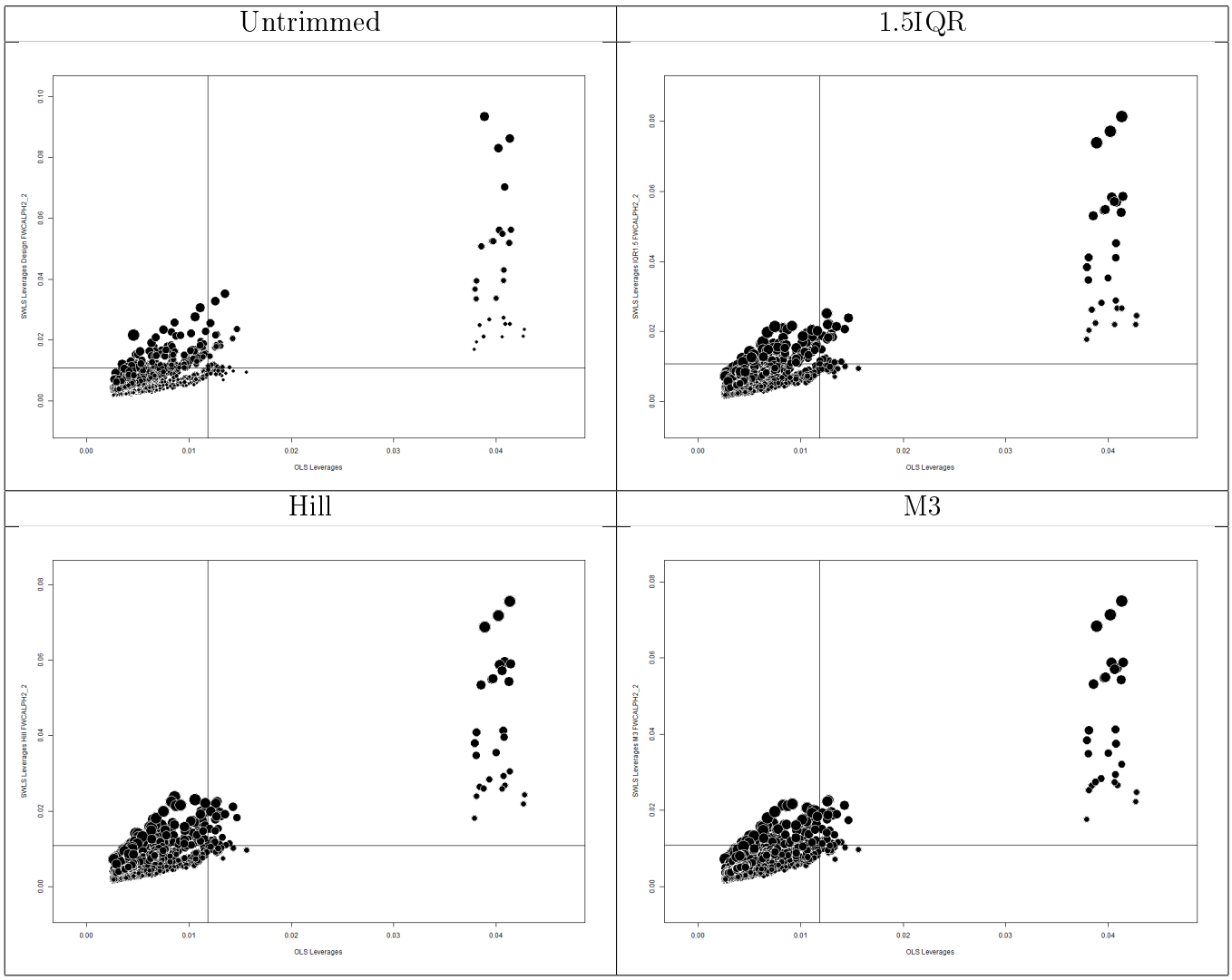


Figure 8.2.4: Bubble plots of OLS versus SWLS Leverages: $w_{SRS}^{ph_2}$

In figure 8.2.4 the observations flagged by SWLS using the untrimmed weights, ph_{SRS}^2, are more similar to the pattern observed in figure 8.2.3. However, the bubbles appear "grouped" according to size and the bubble sizes of the observations flagged only by SWLS, top left corner, are slightly more uniform than at the same point in figure 8.2.3. The application of the trimming methods increase the uniformity of the bubble sizes to such an extent that some observations identified only by OLS leverages before, are now moved to the block that contains only the SWLS flagged observations. Since the remainder of the outlier diagnostic plots based on the "SRS" sampling weights remain similar to those presented here, these figures will not be presented in the thesis document. However, all figures are available for perusal.

The figures presented next are the bubble plots of the DFBetas outlier diagnostic of the age predictor. The OLS and SWLS diagnostics are presented on the same axes as before with their respective lower and upper cut-offs.

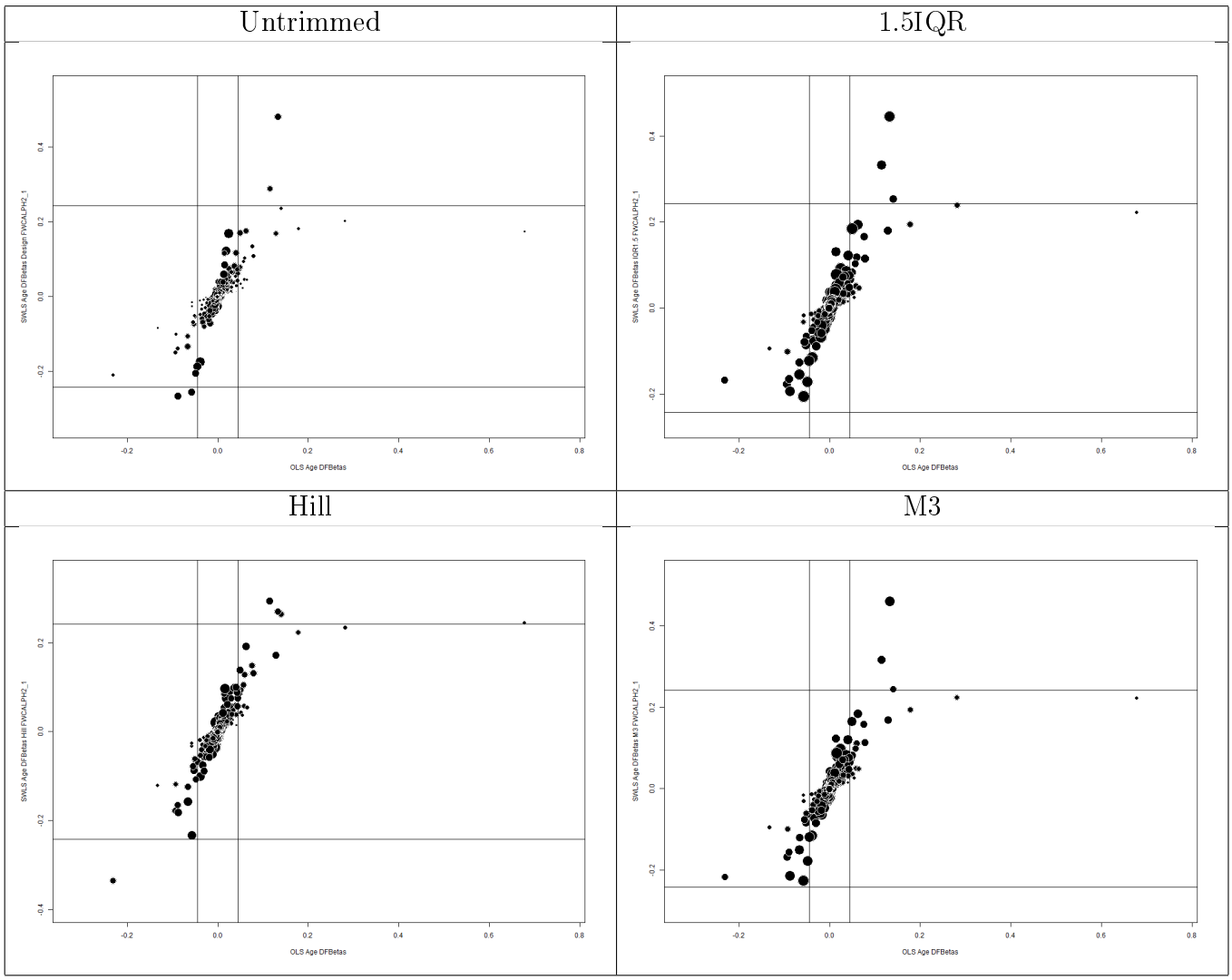


Figure 8.2.5: Bubble plots of OLS versus SWLS Age DFBetas: $w_{CS}^{ph_2}$

Firstly it is observed that the SWLS DFBetas has flagged no observations that differ from those flagged by the OLS diagnostic. However, the OLS diagnostic flagged a few observations not identified by the SWLS diagnostic and it is clear that the flagged observations have small sampling weights. The trimming methods cause the bubble sizes to be more similar and some of the observations flagged by both OLS and SWLS become "unflagged" when using the trimmed weights.

The final group of figures present the DFFits diagnostic in bubble plots of the OLS versus the SWLS diagnostic. The OLS and SWLS diagnostics, along with their respective lower and upper cut-offs, are presented as before.

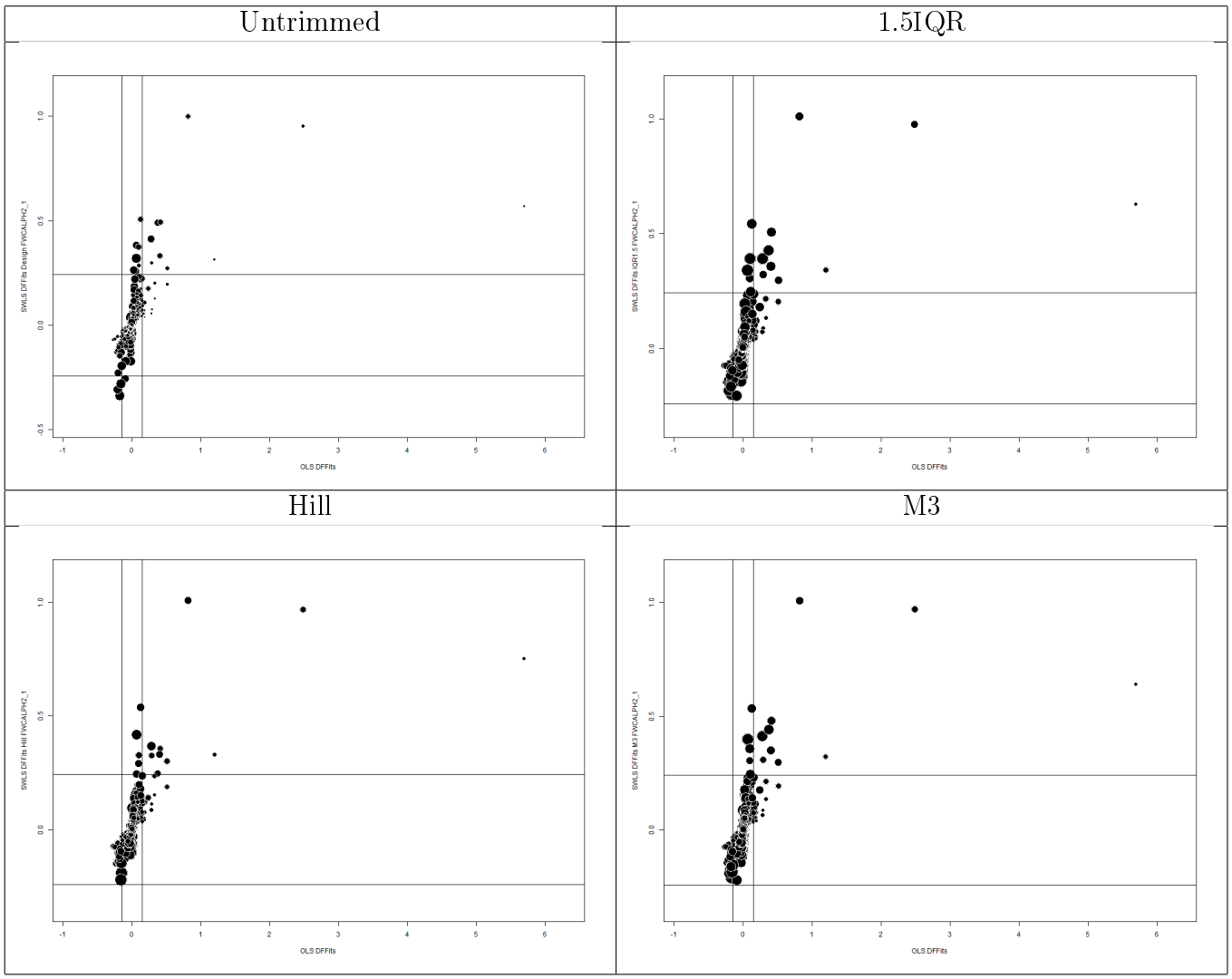


Figure 8.2.6: Bubble plots of OLS versus SWLS DFFits: $w_{CS}^{ph_2}$

According the the untrimmed DFFits bubble plot, there are observations flagged only by the SWLS diagnostic and observations flagged only by the OLS diagnostic. Those flagged by SWLS typically have larger weights than those flagged by OLS. Notice how the trimming of the sampling weight not only slightly decreases the variability in the bubble sizes, it also changes the SWLS outlier status of some observations.

8.2.4 Summary and Conclusions

The previous sections presented the various model evaluation diagnostics which include the coefficient of multiple determination, the model prediction error and the model outlier diagnostics. Each of the diagnostics were compared according to the type of linear model applied (OLS, WLS or SWLS), whether the diagnostic was obtained using the "CS" or the "SRS" sampling weights, and what the effect of trimming the weights, is.

On average the maximum coefficient of determination of 0.3035 was obtained when the theoretical design weights were included in the linear model. Generally the R^2's obtained when using the "SRS" weights were smaller than the "CS" values, but their standard deviations were in most cases smaller than the "CS" values' standard deviations.

The .632 bootstrap estimated prediction error method performed based on the SWLS model using the theoretical design weights and their associated benchmarked weights. It performed especially well based on the Hill trimmed $w_{CS}^{ph_2}$ sampling weights used in the SWLS model. It was noted that further research needs to be done to determine the effect of the number of samples used in the simulation study on the results.

The table below presents the number of outliers and the weight range of the flagged observations under OLS, WLS and SWLS for each outlier diagnostic under "CS" and "SRS" untrimmed and trimmed person- and household-level benchmarked sampling weights.

			OLS			WLS			SWLS		
			Count	Weight	Range	Count	Weight	Range	Count	Weight	Range
$w_{CS}^{ph_2}$	No Trimming	Leverages	70	3.58	22.27	202	3.93	22.27	238	3.93	22.27
		Std Resid	20	3.81	18.35	40	3.81	19.42	20	3.81	18.35
		DFBetas	36	3.89	18.19	50	3.89	19.67	2	11.29	11.41
		DFFits	43	3.81	18.35	78	3.89	20.20	17	3.89	19.42
		Cook's D	0			0			0		
	1.5IQR	Leverages	70	3.58	22.27	202	3.93	22.27	228	3.93	22.27
		Std Resid	20	3.81	18.35	40	3.81	19.42	19	3.81	18.35
		DFBetas	36	3.89	18.19	50	3.89	19.67	3	11.29	14.35
		DFFits	43	3.81	18.35	78	3.89	20.20	15	3.89	19.42
		Cook's D	0			0			0		
	Hill	Leverages	70	3.58	22.27	202	3.93	22.27	215	3.93	22.27
		Std Resid	20	3.81	18.35	40	3.81	19.42	19	3.81	18.35
		DFBetas	36	3.89	18.19	50	3.89	19.67	4	4.64	14.35
		DFFits	43	3.81	18.35	78	3.89	20.20	15	3.89	19.42
		Cook's D	0			0			0		
	M3	Leverages	70	3.58	22.27	202	3.93	22.27	227	3.93	22.27
		Std Resid	20	3.81	18.35	40	3.81	19.42	19	3.81	18.35
		DFBetas	36	3.89	18.19	50	3.89	19.67	3	11.29	14.35
		DFFits	43	3.81	18.35	78	3.89	20.20	15	3.89	19.42
		Cook's D	0			0			0		
$w_{SRS}^{ph_2}$	No Trimming	Leverages	70	7.69	15.24	156	6.12	15.24	194	6.12	15.24
		Std Resid	20	7.01	12.70	31	7.01	12.70	19	7.01	12.70
		DFBetas	36	7.43	15.24	46	7.43	15.24	4	9.25	11.45
		DFFits	43	7.01	12.70	64	7.01	12.70	11	7.52	12.70
		Cook's D	0			0			0		
	1.5IQR	Leverages	70	7.69	15.24	156	6.12	15.24	194	6.12	15.24
		Std Resid	20	7.01	12.70	31	7.01	12.70	19	7.01	12.70
		DFBetas	36	7.43	15.24	46	7.43	15.24	4	9.25	11.45
		DFFits	43	7.01	12.70	64	7.01	12.70	11	7.52	12.70
		Cook's D	0			0			0		
	Hill	Leverages	70	7.69	15.24	156	6.12	15.24	186	6.12	15.24
		Std Resid	20	7.01	12.70	31	7.01	12.70	19	7.01	12.70
		DFBetas	36	7.43	15.24	46	7.43	15.24	3	9.25	11.38
		DFFits	43	7.01	12.70	64	7.01	12.70	9	7.52	12.70
		Cook's D	0			0			0		
	M3	Leverages	70	7.69	15.24	156	6.12	15.24	185	6.12	15.24
		Std Resid	20	7.01	12.70	31	7.01	12.70	19	7.01	12.70
		DFBetas	36	7.43	15.24	46	7.43	15.24	3	9.25	11.38
		DFFits	43	7.01	12.70	64	7.01	12.70	10	7.52	12.70
		Cook's D	0			0			0		

Table 8.2.3: Number of Outliers Identified and Associated Weight Ranges ($w_{CS}^{ph_2}$ versus $w_{SRS}^{ph_2}$)

It is clear that OLS, WLS and SWLS flagged different numbers of observations at outliers. The leverage diagnostic in all three cases seems to flag the largest number of outliers. Furthermore, the weight ranges confirm that OLS typically flags observations with smaller weights while SWLS flags observations with larger weights. Although it appeared from the bubble plots that the trimming methods decreased the general size variability of the bubbles, the weight ranges of the flagged observations did not change according to the information in the table. Notice how the weight ranges of the "SRS" flagged observations remain constant.

Finally, from the different model evaluation diagnostics the following can be concluded:

1. OLS and SWLS achieve different results;

2. the diagnostics obtained from the "CS" and the "SRS" sampling weights, differ; and

3. the weight trimming methods do not necessarily have a positive influence on the model evaluation diagnostics.

The model evaluation is followed by the model parameter analysis. These results are presented in the next section.

8.3 Model Parameter Analysis

After the fitted model has been evaluated and a decision regarding the use of the model has been made, the attention shifts to the inference concerning the model parameters. This section presents various point and interval estimators along with diagnostics by which the estimators can be evaluated.

8.3.1 Parameter Estimation Diagnostics

Recall the unknown regression parameters, $\{\beta_j\}$, $j = 1, ..., p$, that are estimated by $\left\{\hat{\beta}_j\right\}$, the estimators obtained from the application of a linear model to a sample. Now, consider the jth parameter, β_j, that is estimated by $\hat{\beta}_j$. The performance of the estimator will be investigated using the same diagnostics given in section 6.3. Also here the effectiveness of the bootstrap resampling technique is evaluated by letting the rth replicate sample denote a bootstrap population from which B with-replacement bootstrap samples are selected and bootstrap estimators, $\left\{\hat{\beta}_{r_{b_j}}\right\}$ $b = 1, ..., B$, of β_j are obtained. Only a summary of the results will be presented here, but the complete collection of results and functions, programmed using the R statistical software, are available with the author for perusal. A summary of the findings regarding the regression parameter estimation, are now presented.

- With regards to the "true" bias of the estimator of a regression coefficient, consider the following figure of the "true" bias of the estimator of the regression coefficient of age based on equation (6.3.1):

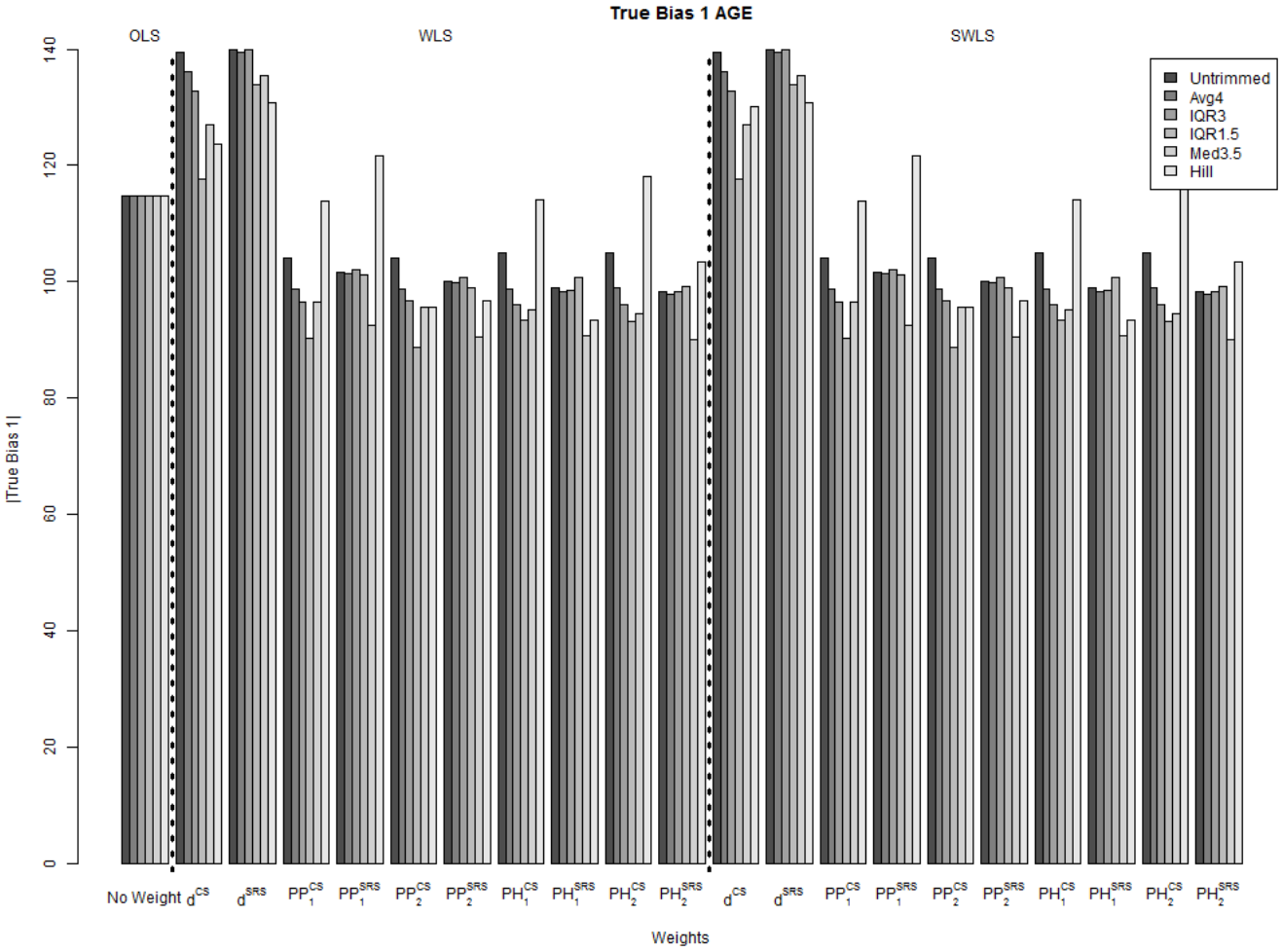


Figure 8.3.1: "True" Bias

On average it was observed that incorporating the sampling weights into the linear model does reduce the bias of an estimator, especially when using the benchmarked sampling weights, i.e. $w_{CS}^{pp_1}$, $w_{SRS}^{pp_1}$,...,$w_{CS}^{ph_2}$ and $w_{SRS}^{ph_2}$. It was also found that the bias under the alternative design weights, d_{SRS}, and their associated benchmarked weights, is larger than for the theoretical weights. Finally, considering the untrimmed versus trimmed weight biases, it was observed that the biases are adjusted downwards by the application of most of the trimming methods.

- Consider the figure of the "true" RMSE of the estimator of the regression coefficient of gender based on equation (6.3.3):

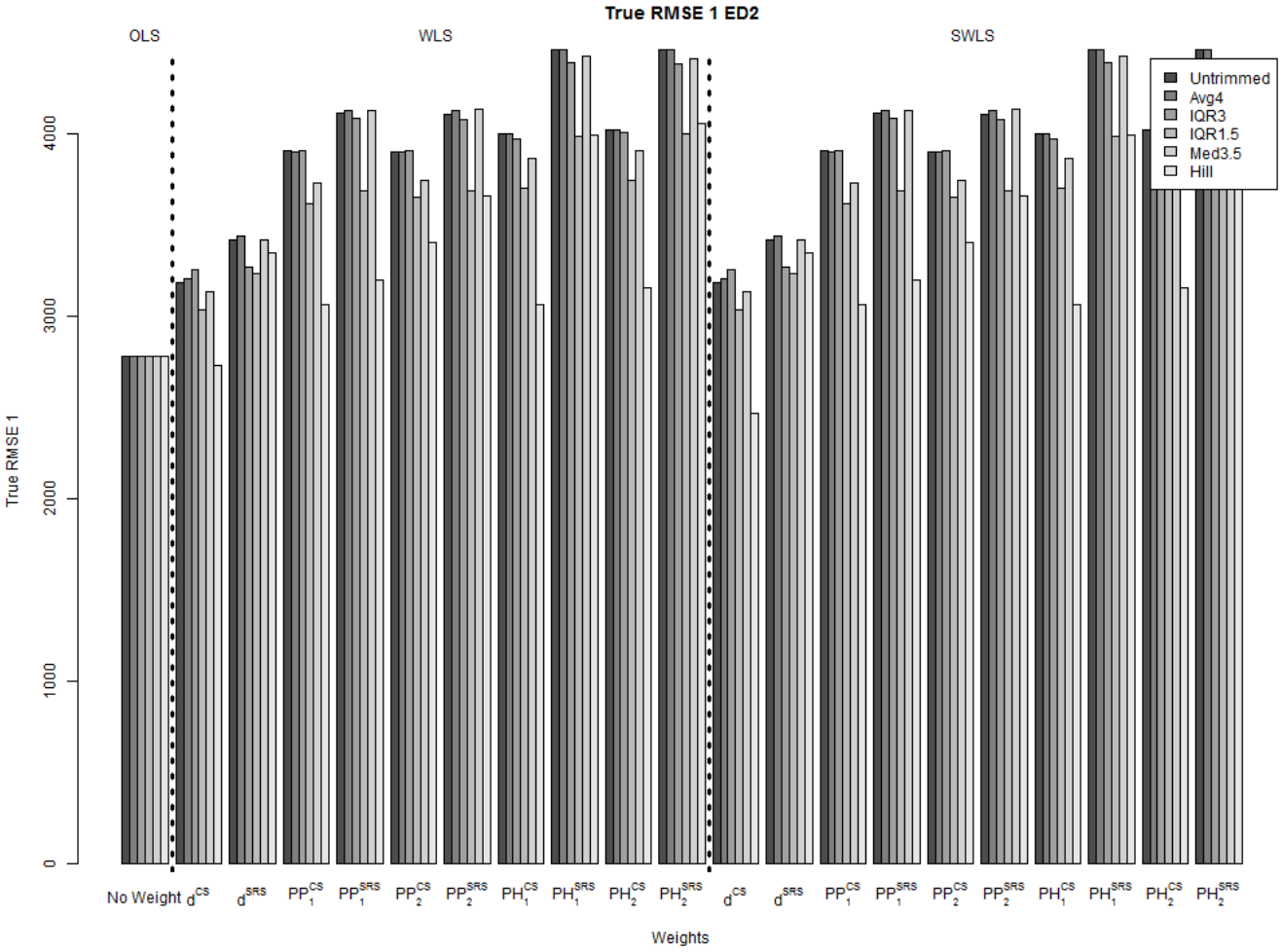


Figure 8.3.2: "True" RMSE

It was clear that the sampling weights mostly decreased the RMSE. This was especially true for the unbenchmarked design weights, but when the trimming methods were applied the RMSE obtained from the benchmarked weights were still smaller than under OLS. Once again the 1.5IQR and Hill trimming methods performed well against the other trimming methods. It was also observed that the described trend was more apparent for the theoretical design weights (d_{CS}) and its benchmarked weights than for the alternative. Considering that the alternative sampling weights increased the bias as well as the RMSE, it can only be concluded that this approach does not have a positive effect on the precision of an estimator. In most cases it can even be concluded that it performs worse than the OLS estimator where the sampling weights are completely ignored.

- Another diagnostic considered was the median absolute deviation (MAD) as defined in 6.3.5 and 6.3.6. From these results it was clear that WLS and SWLS estimators perform better than the OLS estimators when compared to the "truth". The alternative design weights, d_{SRS},

and their benchmarked weights mostly increased the MAD to values that even exceed the OLS values. Considering the theoretical sampling weights, with the exception of the design weights, the trimming methods appear to increase the MAD of the estimator. However, it was found that the Hill trimming method performed well compared to the other methods.

- The gap between the estimated bias and the "truth" is decreased by using the appropriate linear model (SWLS) and the "CS" sampling weights, while in some cases the "SRS" weights' gaps remained larger than that of OLS. The weight trimming methods generally decreased the difference further, especially under the 1.5IQR and Hill trimming methods. Of importance is the difference between the wrong (OLS, WLS) and the correct (SWLS) application of the bootstrap method. The SWLS difference, especially when using theoretical (d_{CS}) benchmarked weights, was the smallest among all the considered differences. Consider figure 8.3.3 for an example of this:

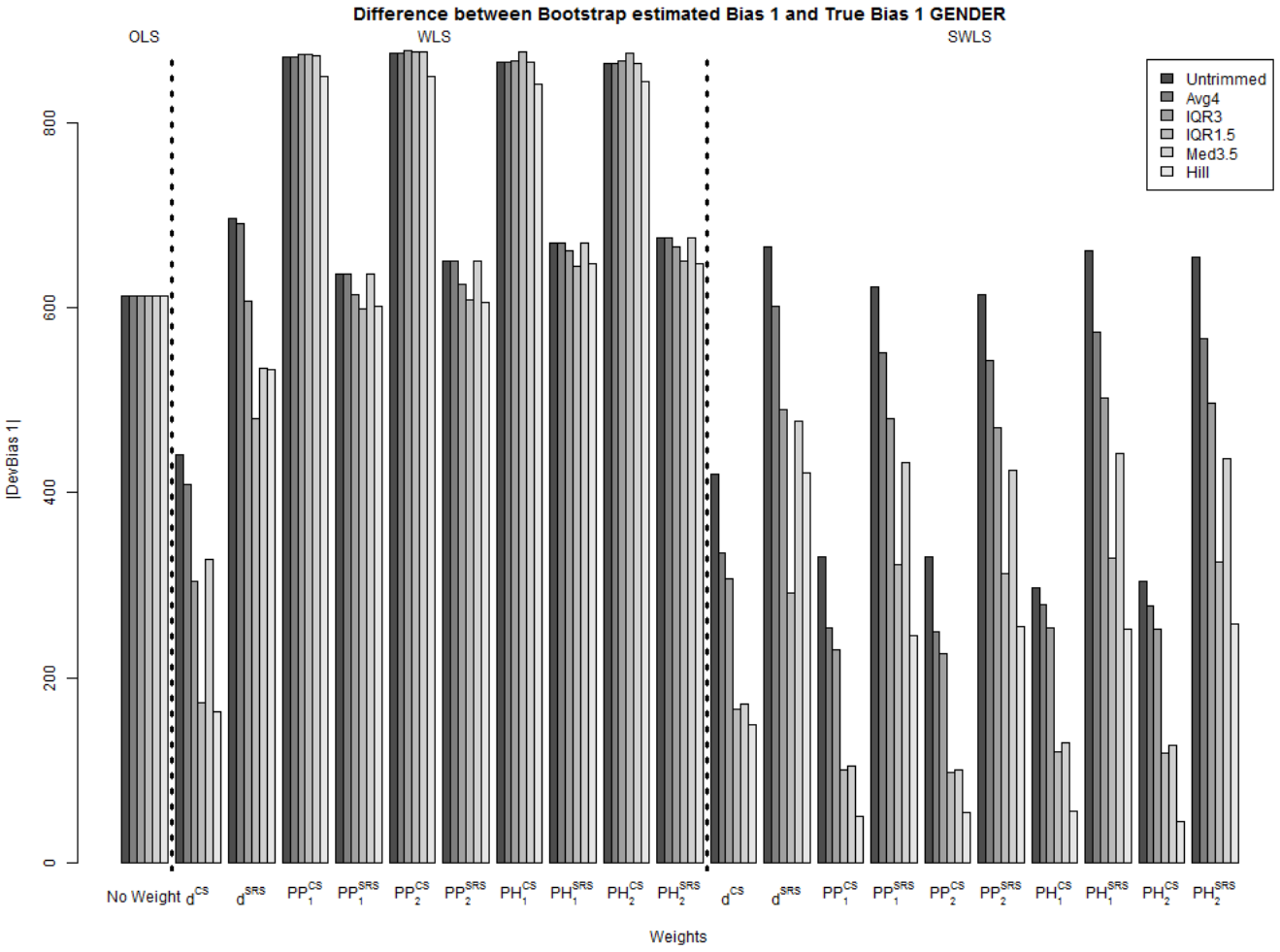


Figure 8.3.3: Difference between Bootstrap Estimated Bias and "True" Bias

- Concerning the difference between the "true" RMSE and the bootstrap estimated RMSE it

was seen that the benchmarked "CS" sampling weights decreased the RMSE difference after being trimmed by the 1.5IQR or Hill methods. Once again the results based on the "SRS" sampling weights were even worse than the OLS and WLS applications. Consider figure 8.3.4 as an example of this.

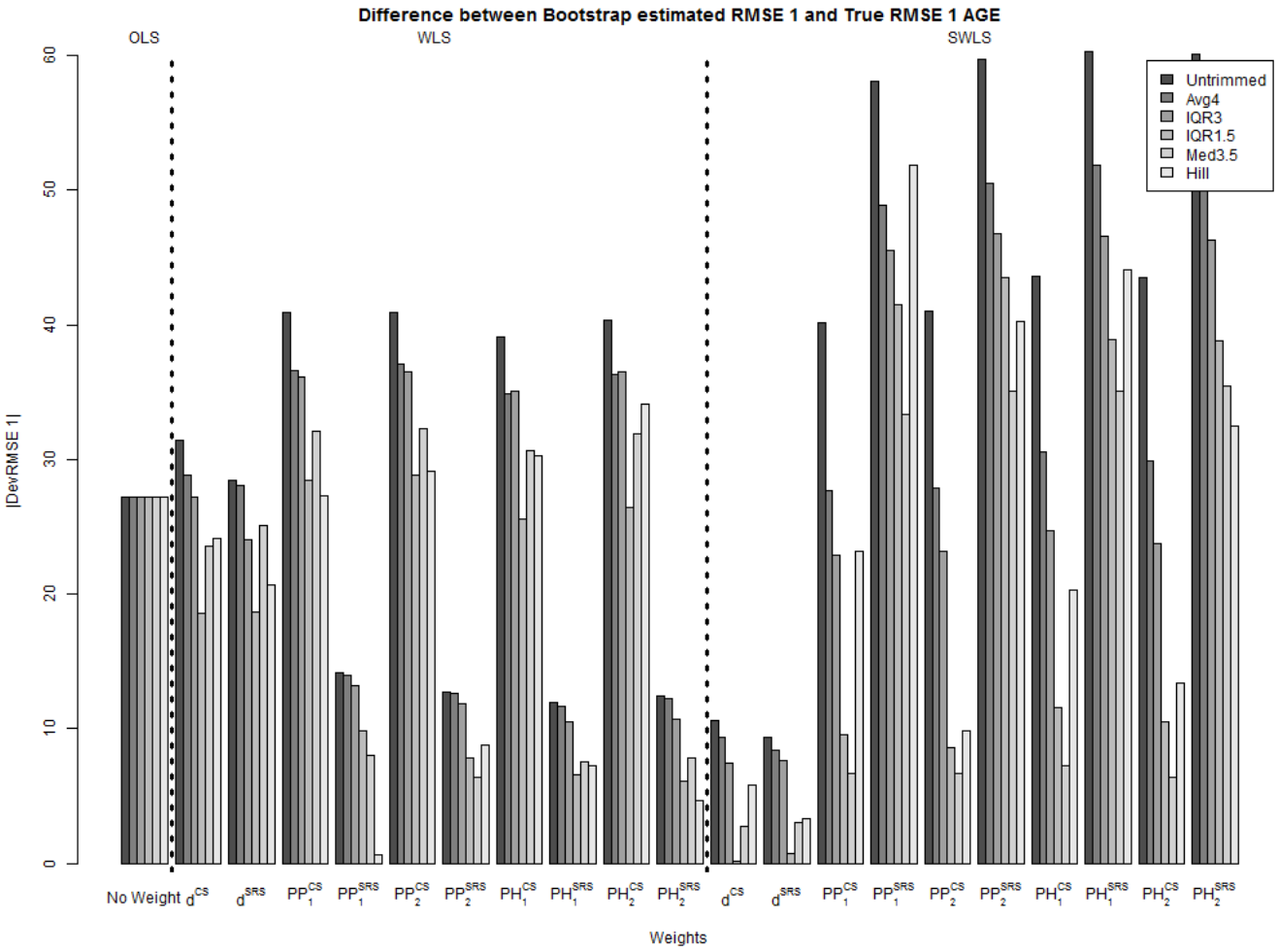


Figure 8.3.4: Difference between Bootstrap Estimated RMSE and "True" RMSE

- The relative bias of the model estimated variance under SWLS was found to be smaller than the OLS relative biases, moreover the d_{CS} and especially its benchmarked weights showed a definite smaller relative bias in comparison to OLS. The d_{SRS} weights and associated benchmarked weights increased the relative bias of the model estimated variances beyond the OLS relative biases, even after the application of the weight trimming methods. The "CS" estimated model variances improved further with the application of the trimming methods. The Hill method performed especially well in this regard.

- The bootstrap estimated variances performed better in terms of relative bias than the model estimated variances. Furthermore, the SWLS relative biases under bootstrap variance esti-

mation were smaller than the OLS and WLS relative biases. In some cases, namely "ph^1_{CS}" and "ph^2_{CS}, the Hill trimming further improved the relative bias of the bootstrap estimated variance. Also, the relative biases of the variances under the application of the theoretical weights were again smaller than those of the alternative weights.

The next section presents the diagnostics of the various confidence interval estimators of the regression parameters.

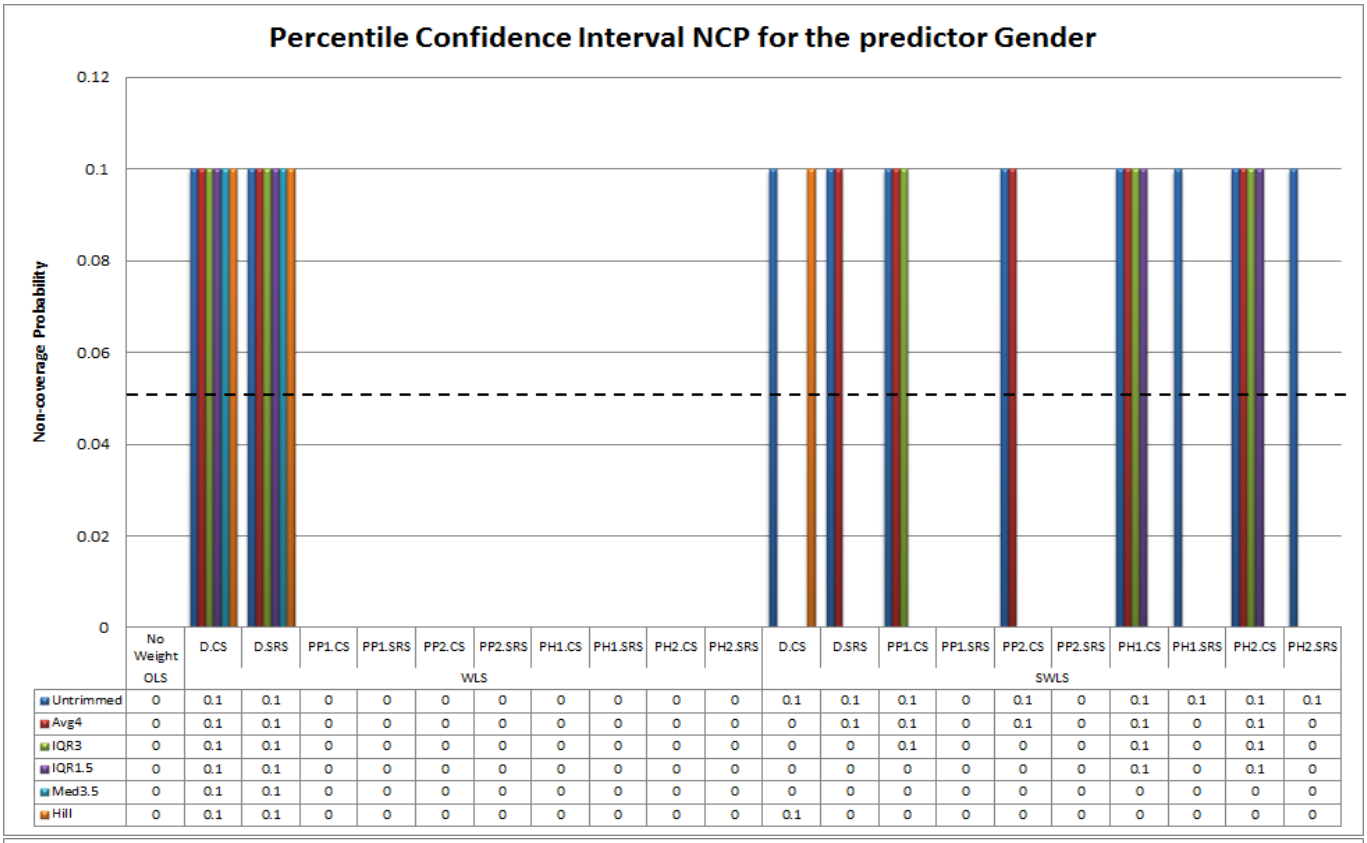
8.3.2 Confidence Interval Diagnostics

Consider the jth parameter, β_j, that is estimated by $\hat{\beta}_j$. This section is concerned with the various confidence intervals that can be constructed for the jth regression parameters, viz. standard (asymptotic) based on the model estimated variance, standard (asymptotic) based on the bootstrap estimated variance, percentile interval, bootstrap-t interval with second level bootstrap estimated variance, bootstrap-t interval with second level jackknife estimated variance, BCa interval with jackknife estimated acceleration, and BCa interval with bootstrap estimated acceleration.

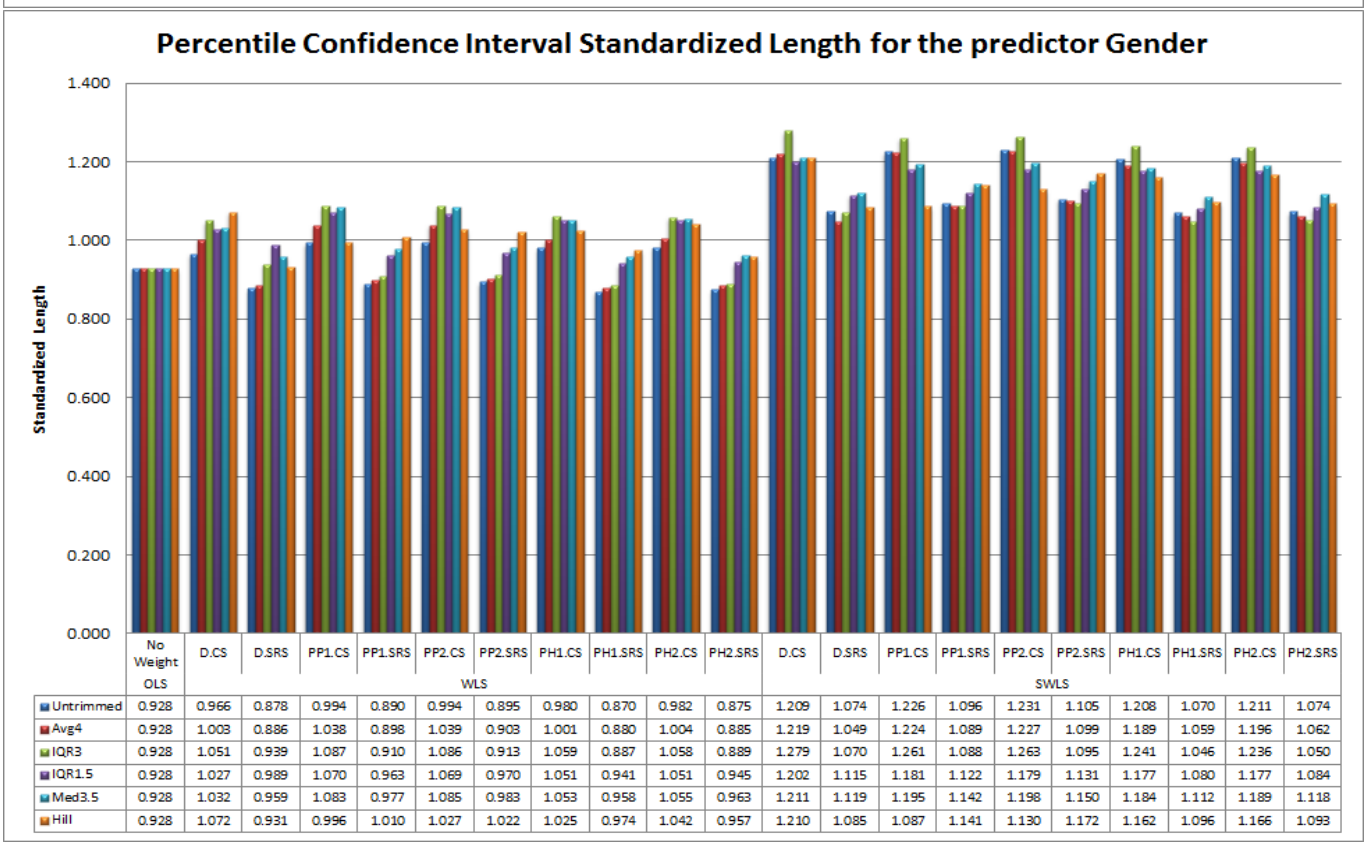
The confidence intervals will be assessed through the consideration of their respective non-coverage probabilities (NCP), lengths and standardized lengths, as discussed in section 6.3. Here too only a summary of the findings will be given, but the entire collection of results and the functions programmed in R are available with the author for perusal. Now consider a summary of the findings:

- Both standard (asymptotic) intervals, viz. using the model estimated variance or the bootstrap estimated variance, did not work well.

- Regarding the percentile interval, SWLS outperformed OLS and WLS, especially based on the benchmarked theoretical design weights, i.e. $w^{pp_1}_{CS}$, $w^{pp_2}_{CS}$, $w^{ph_1}_{CS}$ and $w^{ph_2}_{CS}$. It should be mentioned that the untrimmed and trimmed weights presented similar results. Consider figure 8.3.5 as an example of these conclusions:



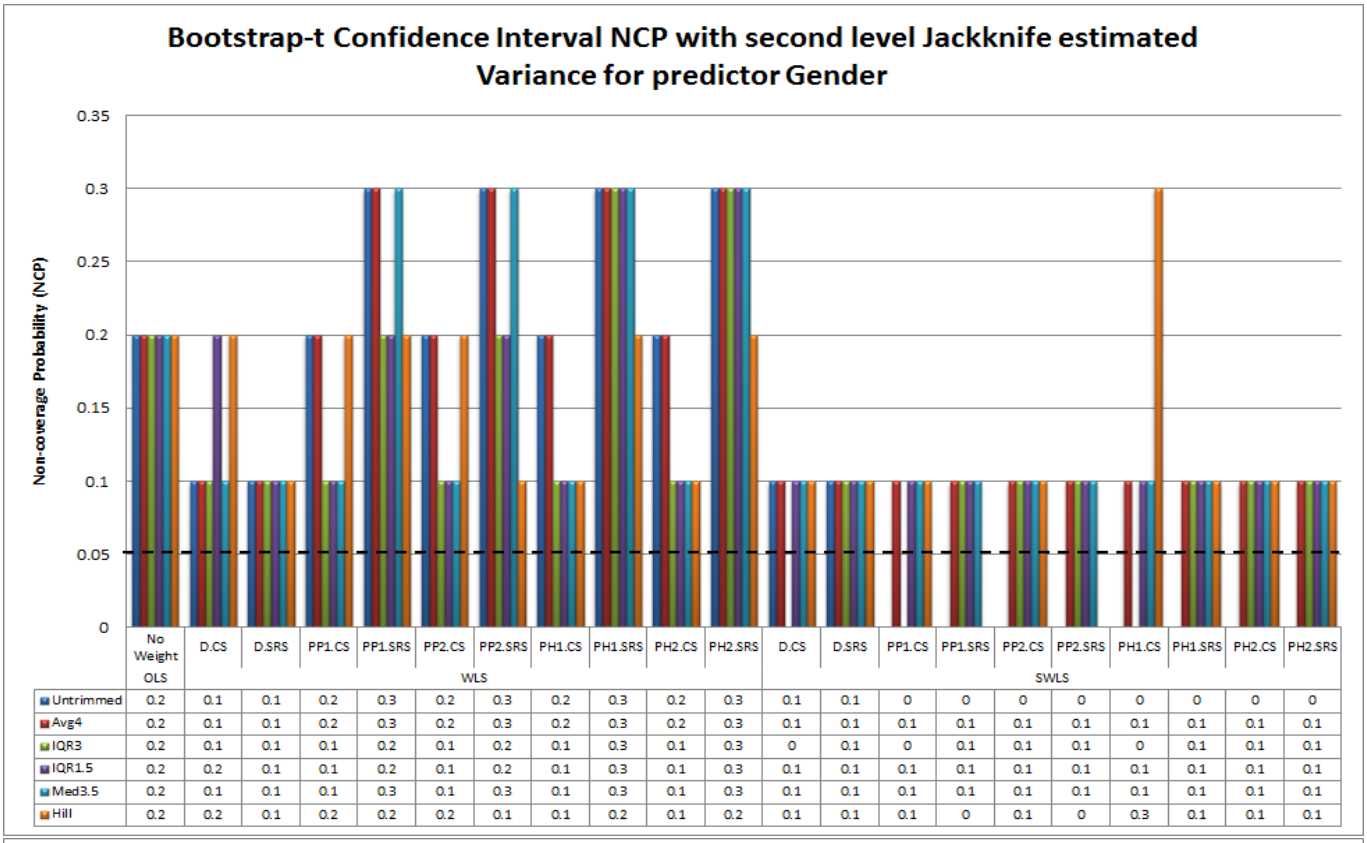
	No Weight	D.CS	D.SRS	PP1.CS	PP1.SRS	PP2.CS	PP2.SRS	PH1.CS	PH1.SRS	PH2.CS	PH2.SRS	D.CS	D.SRS	PP1.CS	PP1.SRS	PP2.CS	PP2.SRS	PH1.CS	PH1.SRS	PH2.CS	PH2.SRS
	OLS	WLS										SWLS									
Untrimmed	0	0.1	0.1	0	0	0	0	0	0	0	0	0.1	0.1	0.1	0	0.1	0	0.1	0.1	0.1	0.1
Avg4	0	0.1	0.1	0	0	0	0	0	0	0	0	0	0.1	0.1	0	0.1	0	0.1	0	0.1	0
IQR3	0	0.1	0.1	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0.1	0	0.1	0
IQR1.5	0	0.1	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0	0.1	0
Med3.5	0	0.1	0.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hill	0	0.1	0.1	0	0	0	0	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0



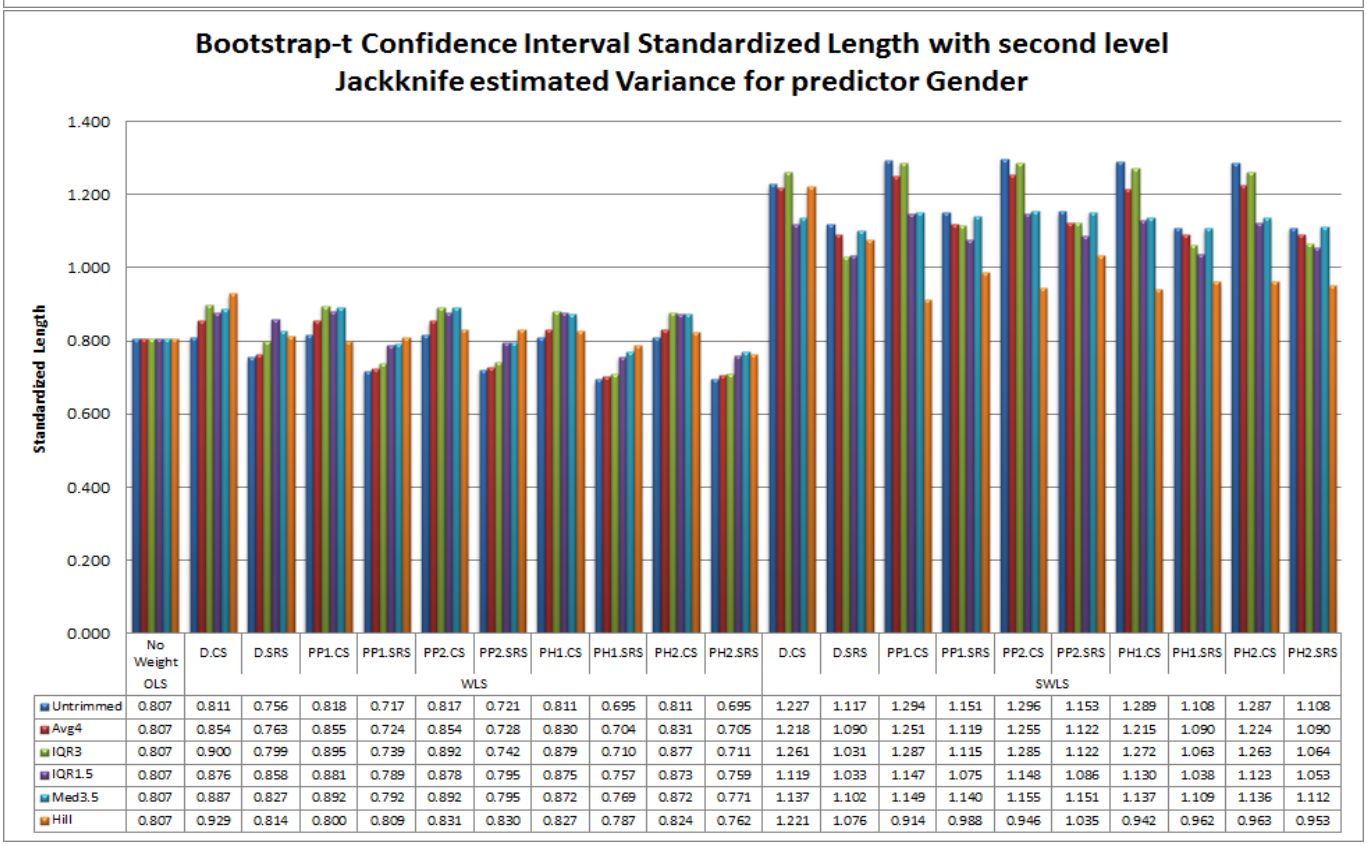
	No Weight	D.CS	D.SRS	PP1.CS	PP1.SRS	PP2.CS	PP2.SRS	PH1.CS	PH1.SRS	PH2.CS	PH2.SRS	D.CS	D.SRS	PP1.CS	PP1.SRS	PP2.CS	PP2.SRS	PH1.CS	PH1.SRS	PH2.CS	PH2.SRS
	OLS	WLS										SWLS									
Untrimmed	0.928	0.966	0.878	0.994	0.890	0.994	0.895	0.980	0.870	0.982	0.875	1.209	1.074	1.226	1.096	1.231	1.105	1.208	1.070	1.211	1.074
Avg4	0.928	1.003	0.886	1.038	0.898	1.039	0.903	1.001	0.880	1.004	0.885	1.219	1.049	1.224	1.089	1.227	1.099	1.189	1.059	1.196	1.062
IQR3	0.928	1.051	0.939	1.087	0.910	1.086	0.913	1.059	0.887	1.058	0.889	1.279	1.070	1.261	1.088	1.263	1.095	1.241	1.046	1.236	1.050
IQR1.5	0.928	1.027	0.989	1.070	0.963	1.069	0.970	1.051	0.941	1.051	0.945	1.202	1.115	1.181	1.122	1.179	1.131	1.177	1.080	1.177	1.084
Med3.5	0.928	1.032	0.959	1.083	0.977	1.085	0.983	1.053	0.958	1.055	0.963	1.211	1.119	1.195	1.142	1.198	1.150	1.184	1.112	1.189	1.118
Hill	0.928	1.072	0.931	0.996	1.010	1.027	1.022	1.025	0.974	1.042	0.957	1.210	1.085	1.087	1.141	1.130	1.172	1.162	1.096	1.166	1.093

Figure 8.3.5: Percentile Confidence Interval NCP and Standardized Length for predictor Gender

- The bootstrap-t interval based on the second-level bootstrap estimated variance performed similarly to the percentile interval, but performed even better in terms of NCP and length when using the second-level jackknife estimated variance. Concerning the untrimmed and trimmed weights the same conclusions could be made as for the percentile interval. See figure 8.3.6.



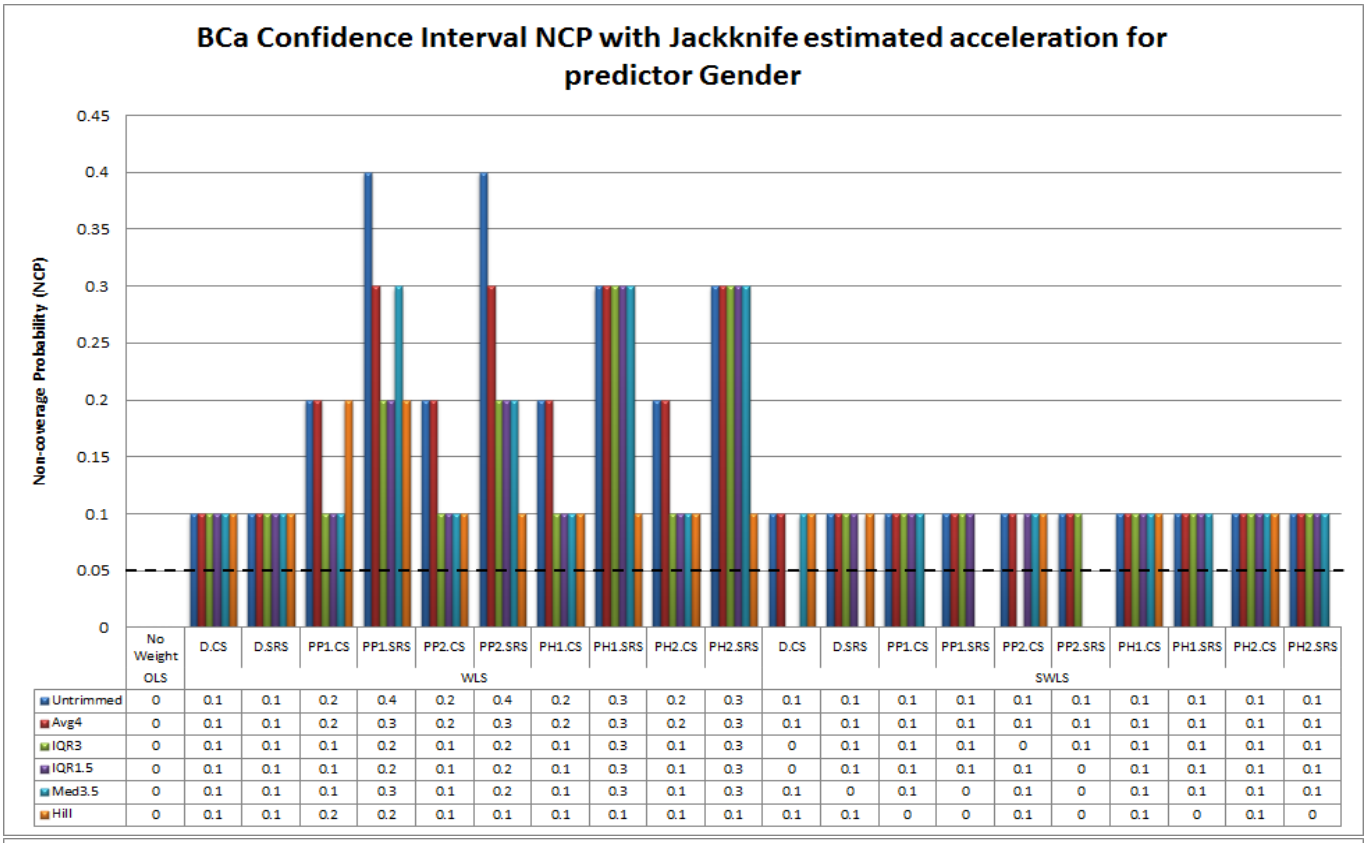
	No Weight	D.CS	D.SRS	PP1.CS	PP1.SRS	PP2.CS	PP2.SRS	PH1.CS	PH1.SRS	PH2.CS	PH2.SRS	D.CS	D.SRS	PP1.CS	PP1.SRS	PP2.CS	PP2.SRS	PH1.CS	PH1.SRS	PH2.CS	PH2.SRS
	OLS	WLS										SWLS									
Untrimmed	0.2	0.1	0.1	0.2	0.3	0.2	0.3	0.2	0.3	0.2	0.3	0.1	0.1	0	0	0	0	0	0	0	0
Avg4	0.2	0.1	0.1	0.2	0.3	0.2	0.3	0.2	0.3	0.2	0.3	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
IQR3	0.2	0.1	0.1	0.1	0.2	0.1	0.2	0.1	0.3	0.1	0.3	0	0.1	0	0.1	0.1	0.1	0	0.1	0.1	0.1
IQR1.5	0.2	0.2	0.1	0.1	0.2	0.1	0.2	0.1	0.3	0.1	0.3	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Med3.5	0.2	0.1	0.1	0.1	0.3	0.1	0.3	0.1	0.3	0.1	0.3	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Hill	0.2	0.2	0.1	0.2	0.2	0.2	0.1	0.1	0.2	0.1	0.2	0.1	0.1	0.1	0	0.1	0	0.3	0.1	0.1	0.1



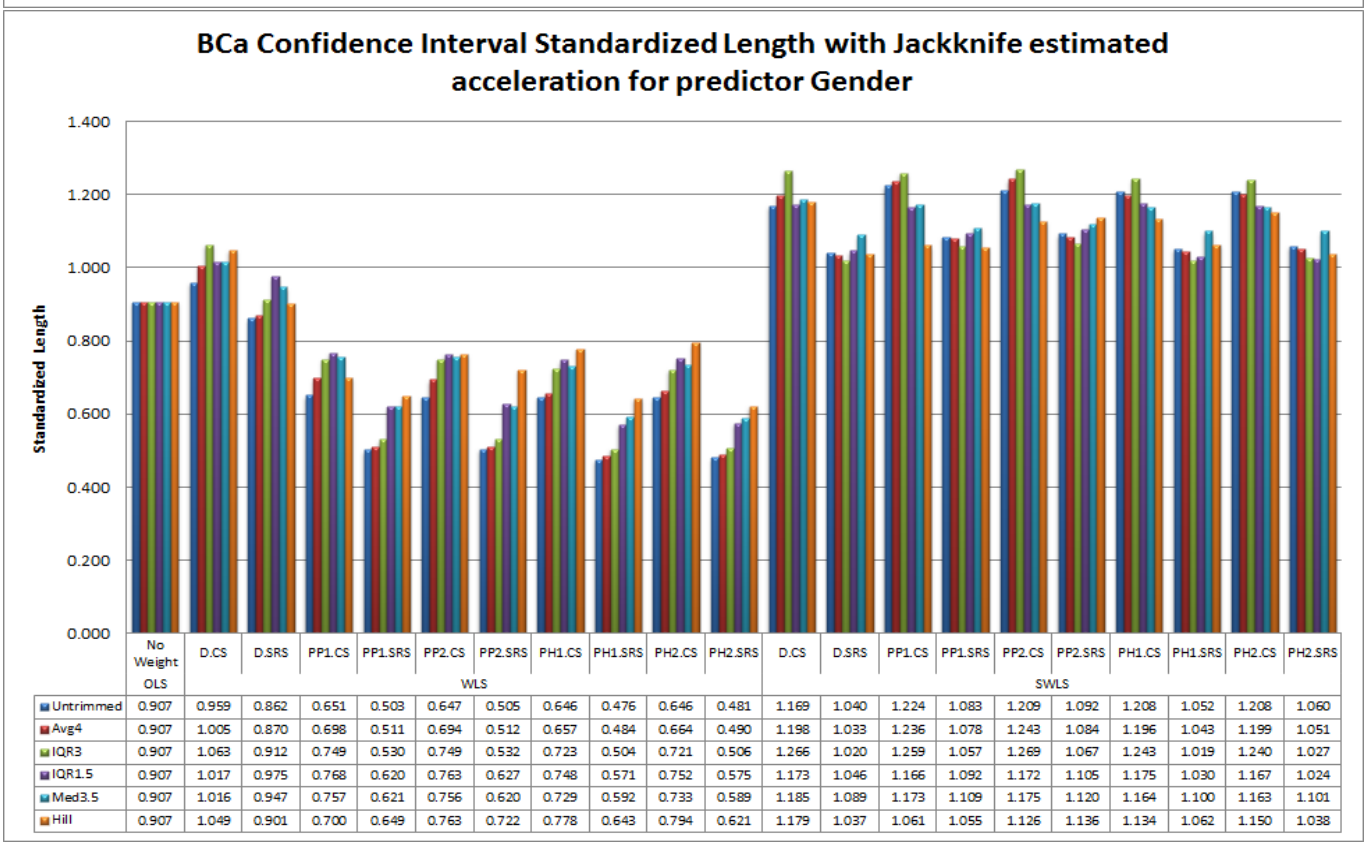
	No Weight	D.CS	D.SRS	PP1.CS	PP1.SRS	PP2.CS	PP2.SRS	PH1.CS	PH1.SRS	PH2.CS	PH2.SRS	D.CS	D.SRS	PP1.CS	PP1.SRS	PP2.CS	PP2.SRS	PH1.CS	PH1.SRS	PH2.CS	PH2.SRS
	OLS	WLS										SWLS									
Untrimmed	0.807	0.811	0.756	0.818	0.717	0.817	0.721	0.811	0.695	0.811	0.695	1.227	1.117	1.294	1.151	1.296	1.153	1.289	1.108	1.287	1.108
Avg4	0.807	0.854	0.763	0.855	0.724	0.854	0.728	0.830	0.704	0.831	0.705	1.218	1.090	1.251	1.119	1.255	1.122	1.215	1.090	1.224	1.090
IQR3	0.807	0.900	0.799	0.895	0.739	0.892	0.742	0.879	0.710	0.877	0.711	1.261	1.031	1.287	1.115	1.285	1.122	1.272	1.063	1.263	1.064
IQR1.5	0.807	0.876	0.858	0.881	0.789	0.878	0.795	0.875	0.757	0.873	0.759	1.119	1.033	1.147	1.075	1.148	1.086	1.130	1.038	1.123	1.053
Med3.5	0.807	0.887	0.827	0.892	0.792	0.892	0.795	0.872	0.769	0.872	0.771	1.137	1.102	1.149	1.140	1.155	1.151	1.137	1.109	1.136	1.112
Hill	0.807	0.929	0.814	0.800	0.809	0.831	0.830	0.827	0.787	0.824	0.762	1.221	1.076	0.914	0.988	0.946	1.035	0.942	0.962	0.963	0.953

Figure 8.3.6: Bootstrap$-t$ Confidence Interval NCP and Standardized Length for predictor Gender

- The BCa intervals, using the jackknife estimated acceleration as well as the bootstrap estimated acceleration, performed very similar with regards to their respective NCP's and lengths. In fact, the BCa results were quite comparable to the bootstrap-t results. Here too it was found that the benchmarked theoretical design weights achieved NCP's close to the desired level of 0.05 while not increasing the interval lengths too much. It should be mentioned that although the achieved NCP's based on the trimmed weights were similar to the NCP's of the untrimmed weights, these NCP's were associated with decreased interval lengths. Thus, the trimmed benchmarked theoretical design weights performed best, especially after the application of the Hill trimming method. See figure 8.3.7.



	No Weight	D.CS	D.SRS	PP1.CS	PP1.SRS	PP2.CS	PP2.SRS	PH1.CS	PH1.SRS	PH2.CS	PH2.SRS	D.CS	D.SRS	PP1.CS	PP1.SRS	PP2.CS	PP2.SRS	PH1.CS	PH1.SRS	PH2.CS	PH2.SRS
	OLS	WLS										SWLS									
Untrimmed	0	0.1	0.1	0.2	0.4	0.2	0.4	0.2	0.3	0.2	0.3	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Avg4	0	0.1	0.1	0.2	0.3	0.2	0.3	0.2	0.3	0.2	0.3	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
IQR3	0	0.1	0.1	0.1	0.2	0.1	0.2	0.1	0.3	0.1	0.3	0	0.1	0.1	0.1	0	0.1	0.1	0.1	0.1	0.1
IQR1.5	0	0.1	0.1	0.1	0.2	0.1	0.2	0.1	0.3	0.1	0.3	0	0.1	0.1	0.1	0.1	0	0.1	0.1	0.1	0.1
Med3.5	0	0.1	0.1	0.1	0.3	0.1	0.2	0.1	0.3	0.1	0.3	0.1	0	0.1	0	0.1	0	0.1	0.1	0.1	0.1
Hill	0	0.1	0.1	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0	0	0.1	0	0.1	0	0.1	0



	No Weight	D.CS	D.SRS	PP1.CS	PP1.SRS	PP2.CS	PP2.SRS	PH1.CS	PH1.SRS	PH2.CS	PH2.SRS	D.CS	D.SRS	PP1.CS	PP1.SRS	PP2.CS	PP2.SRS	PH1.CS	PH1.SRS	PH2.CS	PH2.SRS
	OLS	WLS										SWLS									
Untrimmed	0.907	0.959	0.862	0.651	0.503	0.647	0.505	0.646	0.476	0.646	0.481	1.169	1.040	1.224	1.083	1.209	1.092	1.208	1.052	1.208	1.060
Avg4	0.907	1.005	0.870	0.698	0.511	0.694	0.512	0.657	0.484	0.664	0.490	1.198	1.033	1.236	1.078	1.243	1.084	1.196	1.043	1.199	1.051
IQR3	0.907	1.063	0.912	0.749	0.530	0.749	0.532	0.723	0.504	0.721	0.506	1.266	1.020	1.259	1.057	1.269	1.067	1.243	1.019	1.240	1.027
IQR1.5	0.907	1.017	0.975	0.768	0.620	0.763	0.627	0.748	0.571	0.752	0.575	1.173	1.046	1.166	1.092	1.172	1.105	1.175	1.030	1.167	1.024
Med3.5	0.907	1.016	0.947	0.757	0.621	0.756	0.620	0.729	0.592	0.733	0.589	1.185	1.089	1.173	1.109	1.175	1.120	1.164	1.100	1.163	1.101
Hill	0.907	1.049	0.901	0.700	0.649	0.763	0.722	0.778	0.643	0.794	0.621	1.179	1.037	1.061	1.055	1.126	1.136	1.134	1.062	1.150	1.038

Figure 8.3.7: BCa Confidence Interval NCP and Standardized Length for predictor Gender

8.3.3 Summary and Conclusions

In this section the point and interval estimation of the regression parameters were presented and summaries of the results were given. To conclude the section, consider the main findings below.

1. The correct application of linear modeling to CS data, i.e. SWLS, differed from OLS and WLS in most of the figures. The correct use of sampling weights had, mostly, a positive influence on the diagnostic measures in terms of the minimization of the diagnostic measures.

2. Throughout it was observed that the estimator diagnostics obtained under the "CS" design weights and their associated benchmarked weights, performed better in comparison to the "SRS" sampling weights.

3. The weight trimming methods were also found to, in many figures, improve the diagnostics of the estimators. The 1.5IQR and Hill trimming methods that were newly introduced in this thesis performed well in comparison to some generally used cut-offs. These did not appear to increase the bias of an estimator irrespective of decreasing the MSE of an estimator. In fact, in many cases it appeared as if these trimming methods outperformed the estimators based on the untrimmed weights.

4. The parametric intervals, i.e. standard (asymptotic) intervals, did not perform well in comparison to the non-parametric intervals.

5. Among the non-parametric intervals it was found that the bootstrap-t interval (second-level jackknife estimated variance) and the BCa interval performed best.

6. With regards to the effect of the 1.5IQR and Hill trimming methods on the performance of the intervals, it was observed that the intervals based on these weights achieved similar NCP's as the intervals based on the untrimmed weights, but with decreased interval lengths.

This chapter considered the simulation study outlined in chapter 6, but applied to real-world data in the form of the IES 2005. As in chapter 6, this chapter consisted of two main sections, viz. the model evaluation and model parameter analysis sections. Each section presented a small selection of the results obtained from the IES simulation study followed by a conclusion based on the main findings of the section. The next chapter, namely the final chapter of this thesis, will now summarize the literature presented and simulation results obtained for this thesis. It will conclude with a list of suggested topics for further research.

Chapter 9

Conclusions and Further Research

The statistical analysis of complex sampling (CS) data has received some attention over time, especially in developed countries such as Europe, USA and the United Kingdom. The regression analysis of CS data, especially in developing countries, required more work. Discussions with some survey statisticians brought to light that research about the linear modeling of CS data has not really progressed beyond the point of fitting a model to CS data, especially in developing countries such as South Africa and the greater African continent. This being said, most of the documented literature suggested that the modeling of a discrete dependent variable, especially in the developing countries, received the majority of the attention as opposed to the modeling of a continuous response. As such, the linear modeling of a continuous dependent variable was identified as a possible research topic.

The main difference between simple random sampling (SRS) data and CS data, is the design, which includes clustering, according to which the complex sample is selected. The design is represented in the design weight which is defined as the inverse of the inclusion probability of an observation. The design weights could be adjusted for unit non-response and differential non-response and then the final sampling weights are obtained. This weight development process was discussed extensively in this thesis since it is such an important part of complex sampling. However, the differential non-response adjustment of the process can result in extreme sampling weights that increase the variability within the sampling weight distribution. Since sampling weights are included in the inference under complex sampling, this increased variability is carried over to the inference results and thus the precision of the results could be impaired. One solution for this problem is the trimming (or winsorizing) of the sampling weights. Various methods already in use, have been summarized in the thesis. Two new methods, viz. the 1.5IQR and Hill trimming, have been introduced and developed for weight trimming. It has been noted that weight trimming could itself impair the precision of the inference due to inflating the bias of estimators based on the trimmed weights. One of the objectives then was to determine whether trimming the sampling weights in general improved estimation precision.

The main objective of this thesis was to compare the results of an ordinary least squares (OLS) model fitted to CS data, a weighted least squares (WLS) model and a survey-weighted least squares (SWLS) model. The OLS model was included since this model completely ignores the design, and hence the sampling weights, of the CS data. The WLS model was included since some survey researchers in the developing countries have been found to naïvely fit this model to CS data by specifying the sampling weights as the model weight. Hence, it was important to revise the OLS and WLS methodologies and illustrate how and where these differ from the SWLS methodology. The extension of the fitted model to the evaluation of the model brought about further important contributions of this thesis, namely the adjustment of the leave-one-out cross-validation, bootstrap and .632 bootstrap estimation methods of prediction error for application under CS data. This now enables survey statisticians to evaluate how well their models will perform when predicting a future observation.

There are important assumptions that underly linear modeling, but real-world data rarely meet these requirements. Included under linear modeling is the non-parametric bootstrapping pairs linear model. Included under the non-parametric modeling, are the estimation of the model parameters, the estimation of the variances of the estimators of the model parameters, as well as the construction of non-parametric confidence intervals for the model parameters. This part of the thesis forms part of the smaller contributions made since the bootstrap methodology is well-known, even under CS data, but its usefulness in especially the construction of confidence intervals, is not as well-known.

An extensive simulation study was undertaken in this thesis. The simulation study firstly made use of simulated CS data to ensure that the assumptions underlying the linear modeling, are met, and that any differences observed between the output of the different linear models are attributable to the type of linear model. The development of a model to simulate a population from which a complex sample can be selected, is fairly unknown and quite complex. This has been done in this thesis and is considered one of the major contributions of this research.

The same simulation study was then repeated using real-world data in the form of the Income and Expenditure survey (IES), a survey conducted by Statistics South Africa in 2005. The large real-world data set was considered as a surrogate population such that "true" parameter values could be, at the very least, determined approximately. The known values made it possible to do comparisons in terms of effectiveness and accuracy of the developed models used for the estimation and/or prediction. From the simulated populations as well as the surrogate population a number of replicate samples were simulated. These samples each followed a stratified two-stage cluster design which, in the case of the IES samples, was similar to the surrogate population from which they were selected.

From the simulation studies the following conclusions were made:

1. almost all results showed that there is a clear difference between the OLS, WLS and SWLS

output;

2. according to the prediction error estimation methods the SWLS mostly resulted in the smallest estimated prediction errors;

3. the SWLS estimators were mostly the closest to the "true" parameter values in terms of bias and MSE;

4. the non-parametric bootstrap regression models performed well, especially under the IES data;

5. the bootstrap confidence intervals performed better than the standard (asymptotic) interval in terms of non-coverage probability;

6. in general the weight trimming methods appeared to improve the estimation precision; and

7. the 1.5IQR and Hill trimming methods mostly improved the estimation precision even further.

A number of areas for further research were also identified from the this research and the two simulation studies:

- The estimation results showed promise in terms of the application of weight trimming methods to sampling weights with large variability. Elliott (2007; 2008; 2009) has developed weight trimming methods that were based on the Bayes methodology and these will be included as part of further research.

- Another bootstrap regression approach is the so-called bootstrapping residuals method. This method starts by fitting a linear model to the sample data and approximating the residuals from this model. The residuals form a bootstrap population from which with-replacement samples are selected. The re-sampled residuals are then used, along with the original covariates, to calculate a bootstrap response variable. A linear model is fitted to the bootstrap response and the original covariates to obtain a bootstrap estimated regression coefficient. These bootstrap estimated regression coefficients are used to obtain the bootstrap estimated variance of the regression coefficient. This approach could be extended to other inference concerning the regression parameters.

- Cross-validation is widely used for the estimation of prediction error where the data are split into K parts of roughly equal size. $K-1$ parts are used to fit the model while the remaining part is used for testing the fitted model. This is called K-fold cross-validation of which leave-one-out cross-validation, discussed and extended in this thesis to CS data, is a special case where each of the K parts contains a single observation. Note that in the CS case a single observation refers to a single PSU. The prediction error estimation of the SWLS model

performed very well and thus it is important to pursue this further by extending the leave-one-out cross-validation to K-fold cross-validation, the more commonly used cross-validation method of prediction error estimation.

- The prediction error estimation methods under complex sampling were newly developed in this thesis and this in itself, especially with regards to the bootstrap methods, leaves scope for further research. The apparent error, used in both bootstrap methods, was defined here without the sampling weights, but a case could be made for the alternative. This is a debate that will surely need to be considered further.

- Collinearity diagnostics form an important part of the evaluation of the fitted linear model. It is here that one determines the existence of dependencies between covariates which can have a detrimental effect on the estimation precision. Collinearity could make the fitted model appear better than it actually is and covariates could be deemed significant while they aren't, and vice-versa. Work has been done to develop these diagnostics under complex sampling, but the implementation in statistical software is still lacking. This is an area that will receive attention.

- It was mentioned that the simulation of data for complex sampling is a fairly new topic, especially in developing countries. A simulation model has been developed in this thesis, but there is still scope for improvement. A basic two-level model was used here, but this could be extended to more levels. Also, the simulation of auxiliary variables that could be used for both person- and person-household benchmarking of sampling weights, is necessary since the two-level model only simulated person-level auxiliary variables. Furthermore, in terms of spurious observations, the model can be extended to simulate such observations. These are just a few examples of the research that could still be done in this area.

- The number of samples selected from the simulated and surrogate populations as well as the number of bootstrap samples selected from these samples, were limited. Only ten samples were used and two hundred bootstrap samples. The reason for this was the amount of computer time required to conduct the various analyses using these samples. Some of the results showed evidence of the number of samples and re-samples being too small. Thus, further research will see these numbers increased to assess whether there was a significant effect on the results.

- A univariate multiple linear model was investigated in this research, but an area that has been identified for further research, is the multivariate multiple linear model. This sees the extension of the modeling of a single response to the modeling of multiple responses. Once such a model has been developed for complex sampling data, the model evaluation (prediction errors, outlier diagnostics, collinearity diagnostics, etc.) and parameter estimation (variance

estimation, confidence intervals, etc.) will have to be developed as well. There is much scope for further research in this area.

Research done about the statistical analysis of complex sampling data has made some strides in recent years, especially in the developed countries. However, concerning the developing countries, it has been found that much more research is necessary in this field. Up to this point in the developing countries only the tip of the iceberg has been discovered, just imagine what still lies beneath the surface!

Bibliography

Ajayi, O., St. Catherine, E., Carlson, B., Farid, S., Jambwa, M. M., Mishra, U. S., Kordos, J., Turner, A., Yansaneh, I. and Upadhyaya, S. (2005). *Designing Household Survey Samples: Practical Guidelines*. F 98. New York: United Nations.

Asparouhov, T., Muthen, M. and Muthen, B. (2005). "Multivariate Statistical Modeling with Survey Data". In: *Proceedings of the Federal Committee on Statistical Methodology Research Conference*.

Beaumont, J. F. (2008). "A new approach to weighting and inference in sample surveys." In: *Biometrika* 95(3), pp. 539–553.

Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. L. (2004). *Statistics of extremes*. Ed. by Chichester. John Wiley & Sons.

Berning, T. L. (2010). "Improved estimation procedures for a positive extreme value index." PhD thesis. Stellenbosch University.

— (2015). "Quantification of Estimation Instability and its Application to Threshold Selection in Extremes". In: *The South African Statistics Journal* 49(1), pp. 1–25.

Booth, J.G. and Sarkar, S. (1998). "Monte Carlo Approximation of Bootstrap Variances". In: *The American Statistician* 52, pp. 354–357.

Chowdhury, S., Khare, M. and Wolter, K. (2007). "Weight trimming in the National Immunization Survey". In: *American Statistical Association annual meeting proceedings*.

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

Deville, J. C., Särndal, C. E. and Sautory, O. (1993). "Generalized raking procedures in survey sampling". In: *Journal of the American Statistical Association* 88, pp. 1013–1020.

Drees, H., De Haan, L. and Resnick, S. (2000). "How to make a Hill plot". In: *The Annals of Statistics* 28(1), pp. 254–274.

Drees, H. and Kaufmann, E. (1998). "Selecting the optimal sampling fraction in univariate extreme value estimation." In: *Stochastic Processes and their Applications* 75, pp. 149–172.

Efron, B. (1987). "Better bootstrap confidence intervals". In: *Journal of the American Statistical Association* 82(367), pp. 171–185.

Efron, B. and Tibshirani, R. (1986). "Bootstrap methods for standard eerror, confidence intervals, and other measures of statistical accuracy". In: *Statistical Science* 1(1), pp. 54–77.

Efron, B. and Tibshirani, R. (1998). *An introduction to the bootstrap*. Chapman & Hall.

Elliott, M. R. (2007). "Bayesian weight trimming for generalized linear regression models". In: *Survey Methodology* 33(1), pp. 23–34.

— (2008). "Model averaging methods for weight trimming". In: *Journal of Official Statistics* 24(4), pp. 517–540.

— (2009). "Model averaging methods for weight trimming in generalized linear regression models". In: *Journal of Official Statistics* 25(1), pp. 1–20.

Goldstein, H. (2003). *Multilevel statistical models*. third. Hodder Headline Group.

Guillou, A. and Hall, P. (2001). "A diagnostic for selecting the threshold in extreme value analysis." In: *Journal of the Royal Statistical Society* 63, pp. 293–305.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The elements of statistical learning*. Springer.

Heeringa, S.G., West, B.T. and Berglund, P.A. (2010). *Applied survey data analysis*. Taylor and Francis Group.

Hinkley, D. V. (1977). "Jackknifing in unbalanced situations." In: *Technometrics* 19(3), pp. 285–292.

Izrael, D., Battaglia, M. P. and Frankel, M. R. (2009). "Extreme survey weight adjustment as a component of sample balancing". In: *SAS Global Forum* 247, pp. 1–10.

Killip, S., Mahfoud, Z. and Pearce, K. (2004). "What is an intracluster correlation coefficient? Crucial concepts for primary care researchers." In: *Annals of Family Medicine* 2(3), pp. 204–208.

Kim, J. K. and Skinner, C. J. (2013). "Weighting in survey analysis under informative sampling". In: *Biometrika* 100(2), pp. 385–398.

Kirchoff, R. (2010). "Confidence intervals for estimators of welfare indices under complex sampling". MA thesis. Stellenbosch University.

Knight, K. (2000). *Mathematical Statistics*. Chapman & Hall/CRC.

Kovar, J. G., Rao, J. N. K. and Wu, C. F. J. (1988). "Bootstrap and other methods to measure errors in survey estimates". In: *The Canadian Journal of Statistics* 16, pp. 25–45.

Kutner, M. H., Nachtsheim, C. J., Neter, J. and Li, W. (2005). *Applied linear statistical models*. McGraw-Hill/Irwin.

Lehohla, P. (2008). *Income and Expenditure of Housholds 2005/2006: Statistical Release*. Statistics South Africa. Statistics South Africa.

Lemaître, G. and Dufour, J. (1987). "An integrated method for weighting persons and families". In: *Survey Methodology* 13, pp. 199–207.

Li, J. and Valliant, R. (2009). "Survey weighted hat matrix and leverages". In: *Survey Methodology* 35(1), pp. 15–24.

— (2011). "Linear regression influence diagnostics for unclustered survey data". In: *Journal of Official Statistics* 20, pp. 99–119.

Li, J. and Valliant, R. (2015). "Linear Regression Diagnostics in Cluster Samples". In: *Journal of Official Statistics* 31(1). Ed. by Li, J., pp. 61–75.

Liao, D. (2010). "Collinearity Diagnostics for Complex Survey Data". PhD thesis. University of Michigan.

Liao, D. and Valliant, R. (2012). "Variance inflation factors in the analysis of complex survey data". In: *Survey Methodology* 38(1), pp. 53–62.

Lohr, S. (2010). *Sampling: Design and Analysis*. Brooks/Cole.

Lumley, T. (2014). *Analysis of complex survey samples*. URL: `http://r-survey.r-forge.r-project.org/survey/`.

Lumley, T., Diehr, P., Emerson, S. and Chen, L. (2002). "The importance of the normality assumption in large public health data sets". In: *Annual review of public health* 23, pp. 151–169.

Luus, R., Neethling, A. and De Wet, T. (2012). "Effectiveness of weighting and bootstrap in the estimation of welfare indices under complex sampling". In: *The South African Statistical Journal* 46, pp. 85–114.

Miller, R.G. (1974). "An unbalanced jackknife". In: *The Annals of Statistics* 2(5), pp. 880–891.

Molinaro, A. M., Simon, R. and Pfeiffer, R. M. (2005). "Prediction error estimation: a comparison of resampling methods". In: *Bioinformatics* 21.15, pp. 3301–3307.

Neethling, A. (2004). "Calibration and Integrated Weighting in Sample Surveys". PhD thesis. Faculty of Science, University of Witwatersrand.

Neethling, A. and Galpin, J.S. (2006). "Weighting of Household Survey Data: A Comparison of Various Calibration, Integrated and Cosmetic Estimators". In: *The South African Statistical Journal* 40, pp. 123–150.

Neter, J., Wassermann, W. and Kutner, M. H. (1983). *Applied Linear Regression Models*. Richard D. Irwin Inc.

Osborne, J. W. and Waters, E. (2002). "Four assumptions of multiple regression that researchers should always test". In: *Practical Assessment, Research, and Evaluation* 8(2), pp. 1–5.

Pettersson, H. (2005). *The Design of a Master Sampling Frame and Master Sample for Surveys in Developing Countries*. Tech. rep. Statistics Sweden.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H. and Rasbash, J. (1998). "Weighting for unequal selection probabilities in multilevel models". In: *Journal of the Royal Statistical Society* 60(1), pp. 23–40.

Potter, F. J. (1990). "A study of procedures to identify and trim extreme sampling weights". In: *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 225–230.

Rabe-Hesketh, S. and Skrondal, A. (2006). "Multilevel modelling of complex survey data". In: *Journal of the Royal Statistical Society* 169(4), pp. 805–827.

Rao, J. N. K. and Wu, C. F. J. (1987). "Methods for standard errors and confidence intervals from sample survey data". In: *Bulletin of the International Statistical Institute*.

Rao, J. N. K., Wu, C. F. J. and Yue, K. (1992). "Some Recent Work on Resampling Methods for Complex Surveys". In: *Survey Methodology* 18, pp. 209–217.

Rencher, A. (2002). *Methods of Multivariate Analysis*. John Wiley & Sons.

Rust, K. F. and Rao, J. N. K. (1996). "Variance estimation for complex surveys using replication techniques". In: *Statistical Methods in Medical Research* 5, pp. 283–310.

Sahinler, S. and Topuz, D. (2007). "Bootstrap and jackknife resampling algorithms for estimation of regression parameters". In: *Journal of Applied Quantitative Methods* 2(2), pp. 188–199.

Tukey, J. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Valliant, R. (2010). "Linear regression diagnostics for survey data". In: Statistical Society of Canada.

Valliant, R., Jever, J.A. and Keuter, F. (2013). *Practical tools for designing and weighting survey samples*. Springer.

Wu, C. F. J. (1986). "Jackknife, bootstrap and other resampling methods in regression analysis". In: *The Annals of Statistics* 14(4), pp. 1261–1295.