

The Statistical Analysis of Complex Sampling Data

By

Bradley Paulse

A thesis submitted in partial fulfilment of the requirements for the degree of Masters in
Statistical Science at the University of the Western Cape

Department of Statistics and Population Studies
University of the Western Cape



**UNIVERSITY of the
WESTERN CAPE**

Supervisor: Dr. R. Luus

Co-supervisor: Prof. R.J. Blignaut

November 2018

Abstract:

Most standard statistical techniques illustrated in text books assume that the data are collected from a simple random sample (SRS) and hence are independently and identically distributed (*i.i.d.*). In reality, data are often sourced through complex sampling (CS) designs, with a combination of stratification and clustering at different levels of the design. Consequently, the CS data are not *i.i.d.* and sampling weights that are developed over different stages, are calculated and included in the analysis of this data to account for the sampling design. Logistic regression is often employed in the modelling of survey data since the response under investigation typically has a dichotomous outcome. Furthermore, since the logistic regression model has no homogeneity or normality assumptions, it is appealing when modelling a dichotomous response from survey data.

This research considers the comparison of the estimates of the logistic regression model parameters when the CS design is accounted for, i.e. weighting is present, to when the data are modelled using an SRS design, i.e. no weighting. In addition, the standard errors of the estimators will be obtained using three different variance techniques, viz. Taylor series linearization, the jackknife and the bootstrap. The different estimated standard errors will be used in the calculation of the standard (asymptotic) interval which will be compared to the bootstrap percentile interval in terms of the interval coverage probability. A further level of comparison is obtained when using only design weights to those obtained using calibrated and integrated sampling weights. This simulation study is based on the Income and Expenditure Survey (IES) of 2005/2006. The results showed that generally when weighting was used the estimators performed better as opposed to when the design was ignored, i.e. under the assumption of SRS, with the results for the Taylor series linearization being more stable.

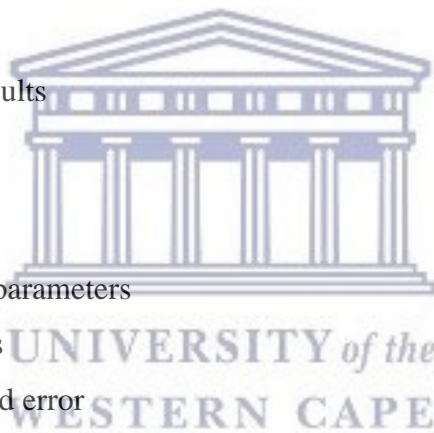
Keywords: complex sampling, inference, weighting, benchmarking, survey data, bootstrap, resampling.

TABLE OF CONTENTS

Chapter 1: Purpose of study and research objectives	1
1.1 Background to the study	1
1.2 Statement of the problem	1
1.3 Purpose and aim of the study	2
1.4 Research questions	3
1.5 Outline of the thesis	3
Chapter 2: Sampling and weighting	5
2.1 Introduction	5
2.2 Probability and non-probability sampling	5
2.2.1 Non-probability sampling methods	6
2.2.1.1 Convenience sampling	6
2.2.1.2 Quota sampling	6
2.2.1.3 Purposive sampling	7
2.2.2 Probability sampling	7
2.2.2.1 Simple random sampling	8
2.2.2.2 Systematic sampling	9
2.2.2.3 Stratified sampling	10
2.2.2.4 Cluster sampling	11
2.2.2.4.1 Cluster sampling with equal probability	12
2.2.2.4.2 Cluster sampling with unequal probability	12
2.3 Complex sampling	14
2.3.1 Design effect	14
2.3.2 Weighting	15
2.3.2.1 Design weight	16

2.3.2.2 Adjusting for non-response	18
2.3.2.2.1 Weighting class adjustments	19
2.3.2.2.2 Inverse of the response rate	20
2.3.2.3 Calibration and integrated weighting	20
2.3.2.4 Integrated weighting techniques	22
2.3.2.4.1 Integrated weights based on person auxiliary variables.	23
2.3.2.4.2 Integrated weights based on person and household auxiliary variables	24
2.4 Conclusion	25
Chapter 3: Logistic regression modelling	26
3.1 Introduction	26
3.2 Model specification and parameter estimation under SRS	26
3.3 Model specification and parameter estimation under CS	30
3.4 Variance estimation	31
3.4.1 Taylor series linearization	32
3.4.2 Resampling methods	35
3.4.2.1 Jackknife repeated replication	36
3.4.2.1.1 JRR under SRS	36
3.4.2.1.2 JRR under CS	37
3.4.2.2 Bootstrap	38
3.4.2.2.1 Bootstrap under SRS	39
3.4.2.2.2 Bootstrap under CS	40
3.5 Confidence intervals for model parameters	42
3.5.1 Standard (asymptotic) confidence interval	42
3.5.2 The bootstrap percentile confidence interval	43
3.6 Conclusion	43
Chapter 4: Research methodology	45
4.1 Introduction	45

4.2 Data collection	45
4.3 Weighting	46
4.4 Response and imputations	47
4.5 Statistical techniques	48
4.5.1 Surrogate population	49
4.5.2 The simulated samples	50
4.5.3 Model and variables	51
4.6 Statistical methodology	55
4.6.1 Assessment of the estimators of the model parameters	55
4.6.2 Assessment of the confidence intervals for the model parameters	56
4.7 Conclusion	58
Chapter 5: Data analysis and results	59
5.1 Introduction	59
5.2 Discussion of results	61
5.2.1 Estimators of model parameters	61
5.2.1.1 The absolute bias	61
5.2.1.2 The mean squared error	68
5.2.2 Confidence intervals for model parameters	76
5.2.2.1 Coverage probability	76
5.2.2.2 Confidence interval length	82
5.3 Conclusion	87
Chapter 6: Conclusion and further research	89
6.1 Introduction	89
6.2 Findings	89
6.5 Further research	90



References	92
Appendices	96
Appendix A: Absolute bias	96
Appendix B: Mean squared error (MSE)	110
Appendix C: Coverage probability	124
Appendix D: Confidence interval length	143



UNIVERSITY *of the*
WESTERN CAPE

LIST OF FIGURES

Figure 1: Mind map of the outline of the thesis.	4
Figure 2: The effect of fitting an ordinary least squares regression to a binary response variable.....	28
Figure 3: Outline of the simulation study.....	60
Figure 4: The absolute bias of the estimator of β_0 under SRS (no weight) and different weighting methods are shown for SAS and R.	62
Figure 5: The absolute bias of the estimator of β_2 under SRS and different weighting methods are shown for SAS and R.	63
Figure 6: The absolute bias of the estimator of β_4 under SRS and different weighting methods are shown for SAS and R.	64
Figure 7: The absolute bias of the estimator of β_5 under SRS and different weighting methods are shown for SAS and R.	65
Figure 8: The absolute bias of the estimator of β_{11} under SRS and different weighting methods are shown for SAS and R.	66
Figure 9: The absolute bias of the estimator of β_{12} under SRS and different weighting methods are shown for SAS and R.	67
Figure 10: The absolute bias of the estimator of β_{20} under SRS and different weighting methods are shown for SAS and R.	68
Figure 11: The MSE of the estimator of β_0 under SRS (no weight) and different weighting methods are shown for SAS and R.	69
Figure 12: The MSE of the estimator of β_2 under SRS and different weighting methods are shown for SAS and R.....	70
Figure 13: The MSE of the estimator of β_4 under SRS and different weighting methods are shown for SAS and R.....	71
Figure 14: The MSE of the estimator of β_5 under SRS and different weighting methods are shown for SAS and R.....	72

Figure 15: The MSE of the estimator of β_{11} under SRS and different weighting methods are shown for SAS and R.....	73
Figure 16: The MSE of the estimator of β_{12} under SRS and different weighting methods are shown for SAS and R.....	74
Figure 17: The MSE of the estimator of β_{20} under SRS and different weighting methods are shown for SAS and R.....	75
Figure 18: The coverage probabilities for β_0 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.....	77
Figure 19: The coverage probabilities for β_4 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.....	79
Figure 20: The coverage probabilities for β_7 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.....	81
Figure 21: The confidence interval lengths for β_0 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.....	83
Figure 22: The confidence interval lengths for β_4 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.....	85
Figure 23: The confidence interval lengths for β_7 under SRS and other weighting methods using TSL, JRR, the bootstrap and for the bootstrap percentile interval are shown for SAS and R.....	86

LIST OF TABLES

Table 1:	Re-grouped education variable.	52
Table 2:	Re-grouped household size variable.	52



UNIVERSITY *of the*
WESTERN CAPE

Declaration

I declare that this thesis is my own work, that the reproduction and publication thereof by the University of the Western Cape will not infringe on any third party rights and that this thesis has not been published previously.

Bradley Paule

November 2018



Signed:

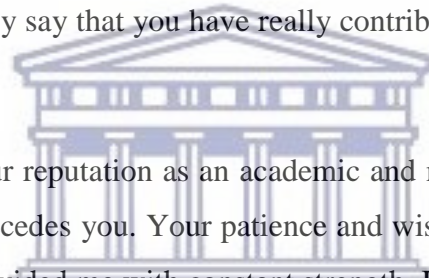


UNIVERSITY *of the*
WESTERN CAPE

ACKNOWLEDGEMENTS

I would like to thank the following:

- I would like to express thanks to God, your mercy and love have been an important component in my success.
- My family, this has been a tough time for you and you have sacrificed a lot. Your support has been tremendous.
- Dr. Retha Luus. Your support, motivation and strength have kept me going over this period. I have learned a lot from you and your hard work and precision is something to strive towards. I can honestly say that you have really contributed to making me a better person.
- Prof. Renette Blignaut. Your reputation as an academic and more importantly, a loving figure in the department precedes you. Your patience and wisdom have been something that has uplifted me and provided me with constant strength. I will be forever thankful.
- Finally, I would like to express thanks and acknowledgement to the National Research foundation in collaboration with South African Statistical Association for providing me with funding for the research.



UNIVERSITY of the
WESTERN CAPE

GLOSSARY

CS:	Complex sampling
Design:	Design weight
EA:	Enumerated area
EPSEM:	Equal probability of selection method
FPL:	Food poverty line
IES:	Income and Expenditure Survey
<i>i. i. d.:</i>	Independent and identically distributed
JRR:	Jackknife repeated replication
<i>Lin_{ph}:</i>	Calibration and integrated weighting using the linear distance method on person auxiliary variables and household
<i>Lin_{pp}:</i>	Calibration and integrated weighting using the linear distance method on person auxiliary variables
MLE:	Maximum likelihood estimator
MOS:	Measure of size
MS:	Master sample
MSE:	Mean squared error
PPS:	Probability proportionate to size
PSU:	Primary sampling unit
<i>RR_{ph}:</i>	Calibration and integrated weighting using the raking ratio method on person and household auxiliary variables and household
<i>RR_{pp}:</i>	Calibration and integrated weighting using the raking ratio method on person auxiliary variables only
SAS:	Statistical analysis system

SRS:	Simple random sampling
SRSWOR:	Simple random sampling without replacement
SRSWR:	Simple random sampling with replacement
SSU	Secondary sampling unit
TSL:	Taylor series linearization
USU:	Ultimate sampling unit
UWC:	University of the Western Cape



UNIVERSITY *of the*
WESTERN CAPE

Chapter 1: Purpose of study and research objectives

1.1 Background to the study

“In spite of its wide range of usefulness, sampling practice has been neglected in the training of statisticians, in the textbooks and treatises, and in the planning and analysis of most experiments and studies. However, like Cinderella, it has risen from neglect to a position of well-deserved importance”, (Stephan, 1948, p. 12). Sampling forms an integral part of statistics. In fact, in order to do proper inference, the sampling design is of utmost importance. The statistics depicted in textbooks are often based on the assumption that the data are from a simple random sample (SRS) when, in reality, most large-scale surveys make use of stratified multistage cluster sampling, or complex sampling (CS), which consists of a combination of different sampling methods (Lumley & Scott, 2015; Heeringa, et al., 2010). According to this sampling method, the observation units are selected by some design that is employed to ensure that the sample selected represents the target population as closely as possible. CS produces data that are not independent and identically distributed (*i.i.d.*), as is the case with an SRS (Lohr, 2010; Heeringa, et al., 2010; Luus, 2016). Instead, the observations have unequal inclusion probabilities associated with them which imply that, should CS data be analysed under the assumption of being *i.i.d.*, all standard errors, confidence intervals and hypothesis tests will be incorrect (Lohr, 2010; Heeringa, et al., 2010; Luus, 2016; Berger & De La Riva Torres, 2016).

1.2 Statement of the problem

CS sampling is an efficient and cost-effective method to collect data and gives more representative samples. As a result, more researchers and analysts are employing CS designs for data collection (Lumley & Scott, 2015). CS data could contain a great number of categorical variables. Researchers often want to establish a multivariate relationship between a response variable that is categorical and explanatory variables which can be a combination of categorical and numerical variables (Kutner, et al., 1996; Heeringa, et al., 2010).

Logistic regression is often employed in the modelling of survey data, since a great number of variables have dichotomous outcomes (Heeringa, et al., 2010; Cheung, 2005; Archer, et al., 2007). Furthermore, since the logistic regression model has no homogeneity or normality assumptions, it is appealing when modelling a dichotomous response from survey data (Archer, et al., 2007; Heeringa, et al., 2010). Estimates and variances of the model parameters may be calculated incorrectly if the design is not accounted for in the inference. Analysis of data obtained from CS needs to be made apparent to ensure the validity of the statistics that are presented. Most researchers are still inclined to use the same techniques under SRS (Lumley & Scott, 2015). This poses the problem of reporting incorrect results and can lead to incorrect conclusions. Therefore, results coming from analyses where the survey design has been ignored must be viewed with caution (Lumley & Scott, 2015; Lumley, 2011).

1.3 Purpose and aim of the study

The objectives of this research are:

1. to illustrate what the major differences in inference results are when ignoring the sampling design as opposed to correctly accounting for it, and the errors that can arise in inference as a result thereof;
2. to show how results obtained using statistical packages SAS and R compare for estimators and the variances of estimators for the logistic regression model when ignoring the sampling design as opposed to correctly accounting for it;
3. to illustrate the precision of standard (asymptotic) confidence intervals obtained under CS using Taylor series linearization (TSL), the jackknife or bootstrap variance estimation for the logistic regression;
4. to show how the bootstrap percentile confidence interval, a non-parametric confidence interval, compares to the standard (asymptotic) confidence interval; and
5. to inform the researchers of the importance of the sampling design when conducting studies, and what statistical methodology to use.

1.4 Research questions

Given the objectives highlighted in the previous section, the following research questions have been identified:

1. How do the estimators of the parameters of the logistic regression model and their estimated variances compare when the sampling design is ignored, i.e. assuming simple random sampling (SRS), as opposed to accounting for the design through CS inference?
2. Is there a difference between estimating the variances of the estimators of the parameters of a logistic regression model when using TSL or employing resampling methods, i.e. the bootstrap and jackknife, for variance estimation?
3. How do the output from the different statistical software compare in the calculation of the variances of the estimators of the parameters of the logistic regression model when using TSL, bootstrap and jackknife in a CS?
4. How do the standard (asymptotic) confidence intervals compare when using TSL or employing resampling methods, i.e. the bootstrap and jackknife, for variance estimation for the logistic regression?
5. How does the bootstrap percentile confidence interval compare to the standard (asymptotic) interval when the sampling design is ignored, i.e. assuming SRS, as opposed to accounting for the design through CS inference?

1.5 Outline of the thesis

Figure 1 is a mind map to outline the major concepts that will be discussed in the thesis. The mind map illustrates how the important concepts are interconnected and the importance of these concepts to provide the basis to answer the research questions.

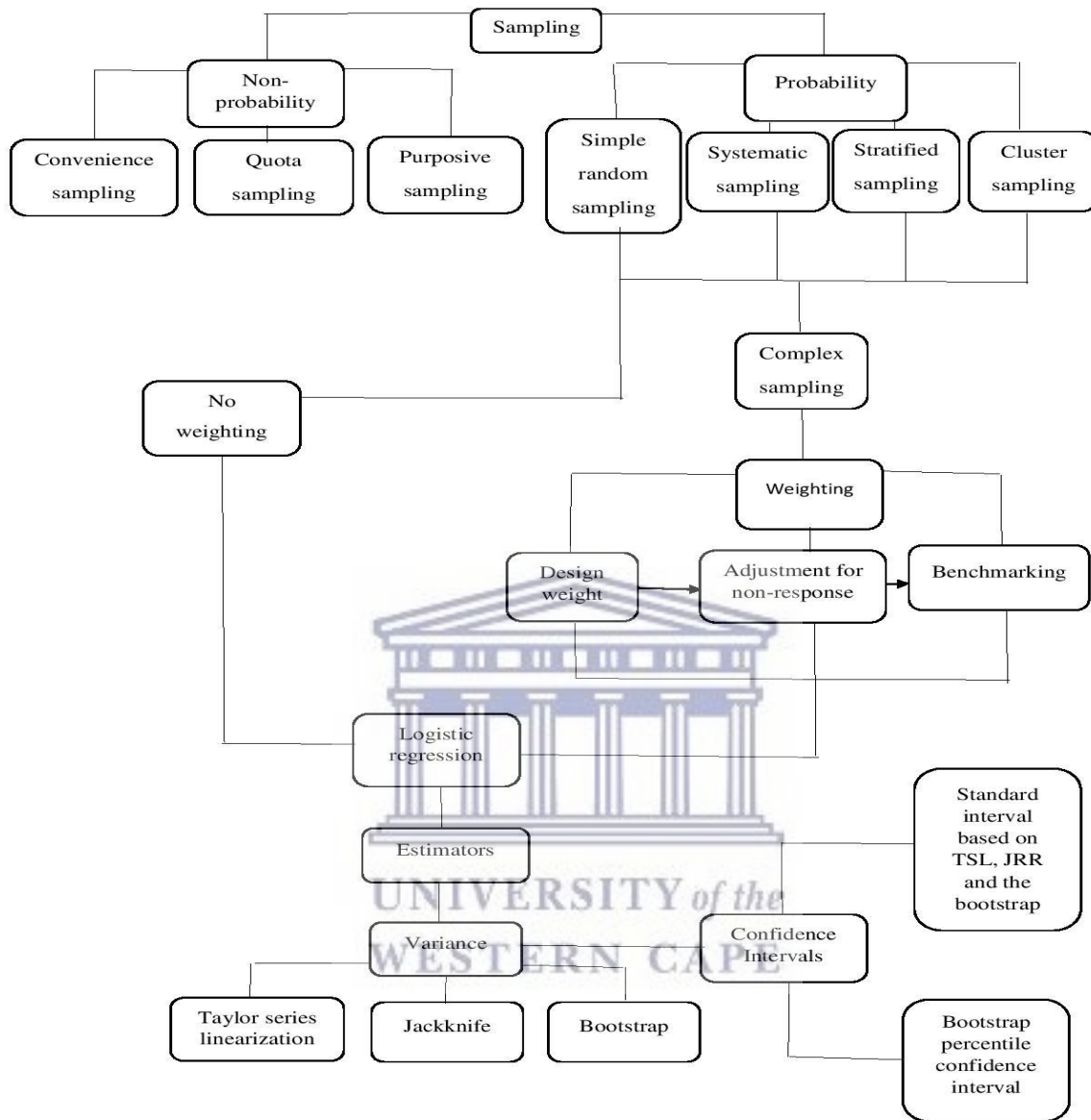


Figure 1: Mind map of the outline of the thesis.

In the next chapter the difference between probability sampling and non-probability sampling will be explained. The chapter will further explore different sampling designs and the impact weighting has to account for the difference in design which is vital for inference.

Chapter 2: Sampling and weighting

2.1 Introduction

Sampling is an integral part of statistics and how and which methods used to gather data is imperative to ascertain which type of analysis should be used. Sampling methods can essentially be grouped into two main categories, i.e. probability and non-probability sampling methods. In this chapter the differences between probability and non-probability sampling methods are explained and a selection of the different sampling techniques within each category is discussed. The chapter further explains the combination of these sampling methods, i.e. complex sampling (CS), and the weighting mechanism used to account for unequal probability of selection, non-response and differential non-response that occur under this sampling technique.

2.2 Probability and non-probability sampling

Consider a finite population U of size N , and suppose a subset of U , of size n , is selected. In probability sampling, the subset is selected such that the elements $\{1, 2, 3, \dots, n\}$ each have a known probability of selection (Yamane, 1967; Lumley, 2011). Non-probability sampling is a collection of sampling techniques that each result in subsets of units for which these probabilities cannot be ascertained, or distributed equitably, and often result in probabilities that are zero (Yamane, 1967; Tansey, 2007). Probability sampling methods include simple random sampling (SRS), stratified sampling, cluster sampling and systematic sampling, and these methods will be discussed here. Complex sampling, which is defined as stratified multistage cluster sampling forms part of the probability sampling methods, and will be discussed as well. Examples of non-probability sampling methods that will be discussed next are convenience sampling, quota sampling and purposive sampling.

2.2.1 Non-probability sampling methods

As mentioned above there are various non-probability sampling methods of which a selection is discussed here. It should be noted that due to the nature of non-probability sampling it is very difficult to generalise from the sample to the greater population and therefore, when data from a non-probability sample are used in research, the results should be viewed in hindsight as having limited scope to establish external validity (Tansey, 2007; Cheung, 2005).

2.2.1.1 Convenience sampling

In this non-probability sampling method the sample is selected that is most easily accessible or available until the desired sample size is acquired (Tansey, 2007; Marshall, 1996). There are no strict rules in terms of selection, and it is drawn in whichever manner suits the researcher (Tansey, 2007). The main reason for using this sampling method is that it may be cost and time efficient. However, some of the drawbacks are that the quality of the data will be low and will lack reliability (Marshall, 1996).

2.2.1.2 Quota sampling

In quota sampling the population is divided into subpopulations from which non-probability samples are selected (Lohr, 2010). The primary reason for making use of quota sampling as opposed to convenience sampling is to ensure that the population is allocated in proportion so that each characteristic is depicted in each subdivision (Tansey, 2007). An example of this would be when a researcher wishes to select a sample of 100 students from different faculties in a university. Suppose 10% of the university belongs to the Science faculty, 50% to Arts and 40% to the Law faculty. These faculties make up the university's total population. Then the sample will comprise of 10 students belonging to the Science faculty, 50 students to the Arts faculty and 40 students to the Law faculty. Note that quota sampling bears an odd resemblance to stratified sampling, discussed in Section 2.2.2.3, but with the exception that, in the subpopulations, probability sampling is not used (Yamane, 1967; Lohr, 2010).

2.2.1.3 Purposive sampling

Purposive sampling is a sampling method in which both the reason for the study as well as the knowledge of the researcher direct the sampling process (Tansey, 2007). In a study the aim is to answer the research question which determines the objectives on which the methodology will be based (Tongco, 2007). A strategy used is to select characteristics that are common to the population you are concerned about under the assumption that errors in judgement in selection will counterbalance each other (Kidder, et al., 1986). However, one drawback is that, when a sample is selected according to an expert's judgement, there is no way to analyse information objectively (Yamane, 1967).

Since non-probability sampling restricts statistical inference, probability sampling needs to be introduced.

2.2.2 Probability sampling

Probability sampling methodology is an essential tool that is vital in order to infer and generalise findings. As opposed to non-probability sampling, making use of probability sampling methods lead to the sample units having inclusion probabilities (the probability that an element is in the sample) that are known (Marshall, 1996; Heeringa, et al., 2010; Lumley, 2011). This is as a result of using a random selection process to obtain the sample which inhibits the possibility of replacing one sampling unit for the next, thus eliminating personal judgement (Luus, 2016). Since the inclusion probability is known, a frequency distribution of the estimates can be obtained (Cochran, 1977). If a probability sampling design is used, a researcher can make inferences about a population with a relatively small sample (Lohr, 2010). There are different probability sampling methods and estimates of the parameters of interest that can be calculated according to the definition of the probability sampling method used.

2.2.2.1 Simple random sampling

Simple random sampling (SRS) is the most fundamental form of probability sampling and provides the theoretical building blocks for other sampling techniques (Lohr, 2010). An SRS is selected in such a way that every possible subset of n units has the same chance of being selected (Thompson, 2010; Lohr, 2010).

There are two ways of selecting an SRS, namely with replacement or without replacement. In SRS with replacement (SRSWR) an element is selected from population U of size N and then, once drawn, that same element is placed back in the set of U . Selecting an SRSWR affords the opportunity for elements in the sample to be repeated. The probability of drawing the first element from U is $\frac{1}{N}$. Since the size of U remains N after the first element is selected, the probability of drawing the second element is also $\frac{1}{N}$. This probability is the same for the third element and so forth. The process will be repeated until a sample of size n is drawn. Therefore, the inclusion probability is $\frac{1}{N}$ which is the same for all elements in SRSWR.

SRS without replacement (SRSWOR) is usually the preferred way of selecting a sample, since in a finite population, sampling the same population element more than once provides no additional information (Lohr, 2010; Cheung, 2005). In SRSWOR there are n distinct elements selected from population U such that every possible combination has an equal chance of being the chosen sample (Luus, 2016; Cochran, 1977). Since this is SRSWOR, the probability that the first unit is drawn is $\frac{n}{N}$, the probability that one of the remaining $(n-1)$ is drawn is $\frac{n-1}{N-1}$, etc. Therefore, the probability that n sampling units are selected in n draws is (Cochran, 1977)

$$\frac{n}{N} \cdot \frac{n-1}{N-1} \cdot \frac{n-2}{N-2} \cdots \frac{1}{N-n+1} = \frac{n!(N-n)!}{N!} = \frac{1}{\binom{N}{n}}.$$

There are $\binom{N}{n}$ possible subsets of size n that can be selected from population U and, as a result, the probability of being the selected sample is $\frac{1}{\binom{N}{n}}$ (Lohr, 2010). Suppose unit j of population U is in the sample, then the other $n-1$ sampling units need to be chosen from the remaining $N-1$ units left in the population. There are $n-1$ combination $N-1$ possible samples

that can be selected, or $\binom{N-1}{n-1}$. Let the inclusion probability of the j^{th} unit be denoted by π_j . It follows that (Lohr, 2010)

$$\pi_j = \frac{\text{number of samples including unit } j}{\text{number of possible samples}} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}.$$

Note that when the population is large, SRSWR and SRSWOR are indistinguishable since N is so large that the probability of selecting unit j from U is very small (Lohr, 2010). SRS is relatively simplistic to employ. However, there is no guarantee that the sample selected is a representative sample and it requires a complete up-to-date sampling frame which is a list or specification of elements in the population from which a sample may be selected (Luus, 2016; Lohr, 2010).

SRS, which is *i.i.d.*, is often the sampling design depicted in statistical theory and most statistical methods assume that the data come from an SRS.

2.2.2.2 Systematic sampling

Systematic sampling is used as an alternative sampling method to SRS if there exists no list of the population or if that list is random (Lohr, 2010). In simple terms, the selection of a systematic sample requires the selection of every k^{th} element in the population in which the first element selected is random (Madow, 1946).

Consider a sample of size n from population U of size N . One method of selecting a systematic sample is to partition the population into groups and then randomly select a unit from each partition (Yamane, 1967). Let k be defined as the selection interval length where k is equal to $\frac{N}{n}$, if this results in an integer. Otherwise k is selected as the next integer following $\frac{N}{n}$. The systematic sampling method begins by finding a random integer between 1 and k , say W , which is the first unit to be included in the sample. The next integer is found by adding a width of k to the first integer, i.e. $W+k$, which becomes the second unit in the sample. This process is repeated until the desired sample size is obtained. It should be noted that systematic sampling forms part of probability sampling as long as it uses a random starting point (Lohr, 2010).

The difference between systematic sampling and SRS is that in systematic sampling all the subsets of size n do not have the same probability of being selected (Luus, 2016). However, if the population is in random order the sampling method is much like SRS (Luus, 2016). Also, systematic sampling is a form of cluster sampling, which is discussed in Section 2.2.2.4. To illustrate this, suppose a population $U = \{1,2,3,4,5,6,7,8,9\}$ from which a sample of size 3 must be drawn. For this case $k = \frac{9}{3} = 3$. To select a systematic sample, one has to select a number at random from 1 to 3 which implies that one must draw that element and every third element thereafter. Thus, the population contains three clusters $\{1,4,7\}$, $\{2,5,8\}$ and $\{3,6,9\}$ and by simply selecting an SRS gives a sample of one cluster (Lohr, 2010).

2.2.2.3 Stratified sampling

The word “stratum” is the Latin word for “layer”. By this sampling method the population U is divided into H distinct subgroups, called strata, such that each population unit belongs to only one stratum (Lohr, 2010; Sitter, 1992). Stratification partitions the population in such a manner that the strata are homogenous which ensures that the variance within a stratum is as small as possible while the between-strata variance is as large as possible (Luus, 2016; Thompson, 2010; Lumley, 2004). This results in estimators with smaller standard errors and estimators with better precision in comparison to SRS (Luus, 2016; Heeringa, et al., 2010; Lohr, 2010).

Consider a population of H strata and let N_h denote the population size of the h^{th} stratum, $h=1,2, \dots,H$. Hence, $N_1 + N_2 + N_3 + \dots + N_H = N$ where N is the size of the population (Lohr, 2010). In stratified sampling, the simplest form of sampling is to take an SRS per stratum. Hence, from stratum h select an SRS of size n_h , $h=1,2, \dots,H$, and then the total sample size, denoted by n , is the sum of the stratum sample sizes, i.e. $n_1 + n_2 + n_3 + \dots + n_H = n$. Consider stratum h and suppose an SRS of size n_h is selected. The inclusion probability of the j^{th} unit in the h^{th} stratum is then given by $\pi_{hj} = \frac{n_h}{N_h}$.

One possible reason for making use of a stratified sample as opposed to an SRS is that, with an SRS there is a possibility of obtaining estimators that are unfavourable. For example,

suppose a sample of different race groups is to be selected using an SRS design. There is a possibility of obtaining a sample of only one race group whereas, with a stratified sample, this is prevented by dividing the population into strata according to race and then selecting an SRS from each stratum. Another reason is to ensure that the sample is representative of the population. Suppose there are more crabs than lobsters in a particular pond. One would divide the strata so that this disproportionality is reflected in the selected sample. Thirdly, stratified sampling can result in lower cost and could be more convenient (Lohr, 2010). Lastly, since the subgroups are independent, different probability sampling methods can be used within strata. Consequently, samples are selected without increasing the selection bias, and inferences can be done on individual strata (Luus, 2016).

2.2.2.4 Cluster sampling

Cluster sampling, on surface level, has a resemblance to stratification since individual elements in the population are grouped into N subgroups. However, in cluster sampling the N subgroups, or clusters, form the population and a sample of n clusters is selected by some probability sampling method (Lohr, 2010). These clusters are referred to as primary sampling units (PSUs). The PSUs consist of subunits called secondary sampling units (SSUs). Suppose each PSU consists of N_j subunits, $j = 1, \dots, N$. It follows that $N_0 = \sum_{j=1}^N N_j$ is the total number of units in the population (Madow, 1946; Lohr, 2010). Moreover, in a cluster sample there is a strong correlation between observational units in the same cluster. This results in the amount of information contained in a cluster sample to be less than that of an independent SRS of the same size (Heeringa, et al., 2010). Therefore, a cluster sample has less precision as opposed to an SRS of equal size (Heeringa, et al., 2010; Lohr, 2010).

There are two types of cluster sampling, i.e. one-stage cluster sampling and two-stage cluster sampling. In a one-stage cluster sample, either all or none of the elements that comprise a particular PSU is in the sample (Lohr, 2010). To illustrate this, suppose a one-stage cluster sample is designed in which there are N PSUs in the population from which n PSUs are selected by SRS (other sampling designs can be used to select the PSUs). Since this is a one-stage cluster sample, it follows that if PSU j is selected then all the elements in PSU j , which

is equal to N_j , are in the sample. In a two-stage cluster sample, once a PSU is selected a subsample of SSUs is selected from the PSU for the sample. Suppose that a sample of n PSUs is selected and that the j^{th} PSU, for $j=1,2,3,\dots,n$, is in that sample. Now suppose the j^{th} PSU has N_j SSUs from which a subsample of n_j is selected (Lohr, 2010). The selection probability resulting from this cluster sample selection can be equal or unequal across the elements. The next section will explore this in more detail.

2.2.2.4.1 Cluster sampling with equal probability

Cluster sampling with equal probability implies that each sampling unit has the same probability of being selected (Cochran, 1977; Lohr, 2010; Heeringa, et al., 2010). Consider the population of N PSUs. The sample of n PSUs must be chosen in such a way that every PSU has the same probability of being in the sample. The two sampling methods discussed in Section 2.2.2.1 and Section 2.2.2.2, i.e. SRS and systematic sampling, both result in an equal probability of selection method (EPSEM) (Lohr, 2010). As mentioned in Section 2.2.2.1, in an SRS all the elements have the same inclusion probability. Therefore, each cluster or PSU has the same probability of being selected. In a one-stage cluster sample EPSEM occurs when the PSUs are selected by SRS. Since this is a one-stage cluster sample, all the subunits or SSUs are automatically in the sample. To select a two-stage cluster sample with EPSEM the SSUs must be selected by an SRS given that the PSUs were selected by the same sampling design. However, cluster sampling with equal probability is often not feasible in reality and therefore unequal probability sampling needs to be introduced (Lohr, 2010).

2.2.2.4.2 Cluster sampling with unequal probability

In the previous section cluster sampling using an equal probability of selection method (EPSEM) was discussed. This design is simple to implement and easy to explain. However, cluster sampling with EPSEM can result in large variances, can be inefficient and can lead to a greater survey cost (Lohr, 2010). Instead, PSUs are sampled with unequal probability which results in better efficiency and lower variances.

Suppose schools in a district are sampled to determine if students are in need of pens. A cluster sample would be essential to manage cost. Suppose an EPSEM two-stage cluster sample of schools (PSUs) and learners (SSUs) in the district were selected. Using EPSEM, larger schools with a greater number of learners are equally likely to be selected as smaller schools with fewer learners. Moreover, it is expected that the number of pens are proportionate to the number of learners attending that school. This results in a large variance and the survey would be inefficient (Lohr, 2010; Cheung, 2005).

An alternative to the EPSEM cluster sampling method is to select the schools in the district with unequal selection probabilities. Many variables of interest in a PSU are related to the number of elements in a PSU. Suppose there are N schools (PSUs) and school j has N_j students (SSUs), $j = 1, \dots, N$, with a total number of $N_0 = \sum_{j=1}^N N_j$ students. Let $\pi_{i/j}$ denote the selection probability of student i from classroom j . The probability of selecting student i from classroom j on the first draw is (Lohr, 2010)

$$\pi_{i/j} = \frac{N_j}{N_0}.$$

Students belonging to classes with a greater selection probability are more likely to be selected in the sample (Lohr, 2010; Heeringa, et al., 2010). This is an example of probability proportionate to size (pps) sampling. The inclusion probability for a two-stage cluster sample with unequal probability of selection for SSU i of PSU j is given by

$$\pi_{ij} = \pi_j \times \pi_{i/j},$$

where π_j is the probability that PSU j is in the sample, and $\pi_{i/j}$ is the probability that SSU i is in the sample given that PSU j is in the sample.

Cluster sampling with unequal probability is not a form of selection bias which is present in non-probability sampling discussed in Section 2.2.1, since non-probability samples do not have a known probability with which they are sampled and cannot necessarily be estimated. Therefore, survey makers cannot account for unequal probability in the form of weighting (Lohr, 2010).

2.3 Complex sampling

Most large-scale surveys are constructed using complex sampling (CS) designs, i.e. stratified multistage cluster sampling designs (Thomas & Heck, 2001; Walker & Young, 2003). Survey designers implement stratification to improve the efficiency of the sample. Also, certain sample elements occur in natural clusters. It would be more efficient to use cluster sampling which reduces travel cost and improves interview efficiency (Heeringa, et al., 2010). Moreover, pps sampling of the population elements may be implemented to improve the sample sizes for subpopulations of special interest (Lohr, 2010; Heeringa, et al., 2010).

The process by which a CS is selected starts by dividing the population into mutually exclusive strata. As noted in Section 2.2.2.3 a stratified sample makes the sample more representative of the population. The next step is to divide the stratum into PSUs, which are predetermined (Luus, 2016). When dividing the stratum into PSUs it is important to ensure that at least two PSUs can be selected per stratum. This is such that variances of estimators of parameters of interest can be calculated (Luus, 2016; Lohr, 2010). The PSUs can be further divided into SSUs. This process can continue until the population units of interest, i.e. ultimate sampling units (USUs), are reached (Luus, 2016). Survey designers often use complex sampling design features to optimise the variance/cost ratio or to meet precision targets for the subpopulations of the survey population (Heeringa, et al., 2010). The precision of the CS estimators as opposed to that of an SRS is termed the design effect and will be discussed next.

2.3.1 Design effect

As mentioned in Section 2.2.2.3 and Section 2.2.2.4 stratification generally yields more precise estimators, while clustering yields less precise estimators of parameters of interest in comparison to an SRS (Heeringa, et al., 2010; Luus, 2016; Lohr, 2010). Therefore, since CS is a combination of stratification and clustering, a CS design does not necessarily yield better precision estimators in comparison to SRS.

The effects of stratification and clustering on the standard errors of the estimators in relation to an SRS is termed the design effect (Heeringa, et al., 2010). Consider $\hat{\theta}$ as an estimator of some parameter of interest, θ . The design effect is calculated as follows:

$$D^2(\hat{\theta}) = \frac{V_{CS}(\hat{\theta})}{V_{SRS}(\hat{\theta})}, \quad (1)$$

where $D^2(\hat{\theta})$ denotes the design effect for $\hat{\theta}$, $V_{CS}(\hat{\theta})$ is the variance of the estimator under complex sampling, and $V_{SRS}(\hat{\theta})$ is the variance of the estimator under SRS. In order for the estimators under CS to have the same precision as those under SRS, $V_{CS}(\hat{\theta}) = V_{SRS}(\hat{\theta})$ resulting in $D^2(\hat{\theta}) = 1$. This can be done by increasing the sample size of the CS design (Lohr, 2010).

The design effect can be used to optimise the cost and properties of specific design alternatives or to alter SRS computations under a specific sampling plan (Heeringa, et al., 2010). To have knowledge of the estimated design effects enables one to see to what extent the sampling design used produces efficiency or losses relative to an SRS and to identify extreme clustering or weighting that can affect inferences (Heeringa, et al., 2010).

2.3.2 Weighting

Weighting is used to make the sample an unbiased representation of the survey population. Essentially one can think of a weight as the number of population elements represented by the associated sample observation (Heeringa, et al., 2010; Lohr, 2010; Lumley & Alastair, 2017). Weighting can be used to correct some of the flaws associated with unequal probabilities, non-response and differential non-response (Luus, 2016; Walker & Young, 2003; Neethling & Galpin, 2006). The weighting process starts by determining a design weight, then adjustments are made for non-response and further weighting adjustments are needed for differential non-response discussed in Sections 2.3.2.1 to 2.3.2.4. The development stages of the sampling weight are discussed next.

2.3.2.1 Design weight

In a probability sample, each unit in the population has a known probability of being selected, which is determined by a randomisation method that chooses the particular unit to be included in the sample (Lohr, 2010; Lumley, 2011). Let the inclusion probability of the j^{th} observation be π_j and let w_j denote design weight of the observation. The design weight is defined as the inverse of the inclusion probability, i.e.

$$w_j = \frac{1}{\pi_j}, j=1, 2, \dots, n.$$

It has the property that $\sum_{j \in S} w_j = N$, i.e. the sum of the weights across the sampling units equals the population size (Heeringa, et al., 2010; Lohr, 2010; Luus, 2016).

The design weight depends on the inclusion probabilities which may differ depending on the sampling design. Consider the inclusion probability defined in Section 2.2.2.1 for a SRSWOR. It follows that the design weight associated with a sampling unit selected by SRSWOR is given by $\frac{N}{n}$. Therefore, using the definition of a weight defined in Section 2.3.2, every unit in a SRSWOR represents itself and $\frac{N}{n} - 1$ units that are not sampled but are in the target population. As opposed to SRSWOR, SRSWR elements can be drawn more than once. Thus, the inclusion probability of an element under SRSWR is $\frac{1}{N}$. It follows that the design weight under SRSWR is N .

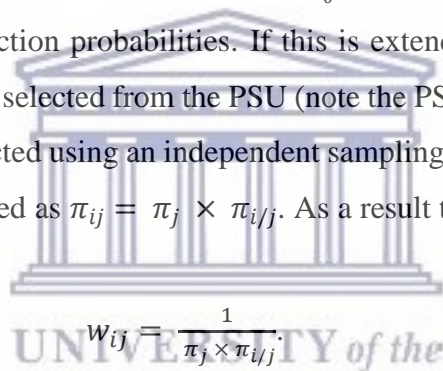
In a stratified sample, sampling units are in distinct subgroups and the inclusion probabilities are calculated per stratum. The inclusion probability of the j^{th} unit in the h^{th} stratum is $\pi_{hj} = \frac{n_h}{N}$ and from the definition of the design weight it follows that $w_{hj} = \frac{N}{n_h}$.

Cluster sampling is a common feature of most CS surveys, and the cluster can be selected by EPSEM or unequal probability. A one-stage cluster sample is a cluster sample that, if a PSU is selected then all the SSUs comprising that PSU are in the sample. If a one-stage cluster sample is selected by EPSEM, this is equivalent to selecting an SRSWR since the PSUs will

form the N elements and n PSUs are to be selected, resulting in each PSU having an inclusion probability of $\frac{n}{N}$. The design weight will be given by $\frac{N}{n}$, the same as that of a SRSWR.

In a two-stage cluster sample with EPSEM, firstly, the n PSUs are selected from the population of N PSUs, then an SRS of SSUs is selected from each of the n selected PSUs. Suppose each PSU consists of N_j SSUs, $j = 1, \dots, N$, and from N_j , n_j SSUs are selected. Then the inclusion probability of SSU i from PSU j is $\frac{nn_j}{NN_j}$. Subsequently, the design weight can be obtained as $\frac{NN_j}{nn_j}$.

In a one-stage cluster sample with pps the SSUs form the basis for the selection probabilities with larger PSUs, i.e. PSUs that have a larger number of SSUs, have a greater chance of being selected. The selection probabilities are given by $\pi_{i/j} = \frac{N_j}{N_0}$. The inclusion probability of PSU j is simply the sum of the selection probabilities. If this is extended to a two-stage cluster sample with pps, then the SSUs selected from the PSU (note the PSUs are selected at the first stage by pps sampling) are selected using an independent sampling method such as SRS. The inclusion probability is calculated as $\pi_{ij} = \pi_j \times \pi_{i/j}$. As a result the design weight for SSU i of PSU j is



$$w_{ij} = \frac{1}{\pi_j \times \pi_{i/j}}$$

Consider now a stratified two-stage cluster sample where a population has been stratified into H strata, each stratum contains N_h PSUs and each PSU contains N_{hj} SSUs, $j = 1, \dots, N_h$, $h = 1, \dots, H$. Consider stratum h . In two-stage cluster sampling, n_h PSUs are selected from stratum h , and n_{hj} SSUs are selected from the j^{th} selected PSU. Since each of the strata is sampled independently, the design weight is calculated within the strata. Let w_{hj} denote the design weight of the j^{th} PSU selected from stratum h and $w_{i/hj}$ the design weight of the i^{th} SSU within PSU j . Then the overall design weight for that sampling unit is

$$w_{hji} = w_{hj} \times w_{i/hj}$$

Note that the sampling weights give no indication of how to calculate the standard errors and therefore inferential statistics using only sampling weights is absolute (Lohr, 2010).

The next step in the weighting process is to adjust the design weight to compensate for non-response, since estimators can be biased and have inaccurate variances when based only on responses (Luus, 2016).

2.3.2.2 Adjusting for non-response

When certain observations in the sampling frame do not respond to the survey it may have an effect on the inference since non-respondents generally differ from those that do respond (Luus, 2016; Lohr, 2010; Cochran, 1977). There are two types of non-response: item non-response, which occurs when answers to certain questions in the survey questionnaire are omitted; and unit non-response, which is observed when the entire sampling unit's information is missing (Luus, 2016; Lohr, 2010; Nations, 2005). Non-response can result in estimation bias specifically when respondents differ significantly from non-respondents (Cheung, 2005). Furthermore, increasing the sample size while not taking into account non-response does not reduce non-response bias. It merely provides more observations that would respond to the survey. In fact, it may worsen non-response since those resources could have been directed to remedy non-response (Lohr, 2010). Lohr (2010) mentions a few remedies for non-response:

1. design the survey so as to minimise non-response to the extent that there is very little to no non-response. This is the best method;
2. take a representative subsample of non-respondents and use that subsample to make inferences on the other non-respondents;
3. use a model to predict values for non-respondents. Weighting class adjustment discussed in Section 2.3.2.2.1 uses a model to adjust for unit non-response. Imputation can be used to adjust for item non-response; or
4. ignore it (not recommended) (Lohr, 2010; Luus, 2016).

To consider the effects of non-response on the sample estimate, suppose there are two strata, i.e. stratum 1 and stratum 2, where stratum 1 is the respondents and stratum 2 the non-respondents. Let N_1 denote the population size of stratum 1 and N_2 the size of stratum 2, where $N_1 + N_2 = N$. Note that there is only information for stratum 1 and suppose the elements

in stratum 1 were selected by SRS. Consider the proportion of respondents and non-respondents given by $P_1 = \frac{N_1}{N}$ and $P_2 = \frac{N_2}{N}$, respectively. The bias of the sample mean can be obtained by

$$\begin{aligned} E(\bar{y}_1) - \bar{Y} &= \bar{Y}_1 - \bar{Y} \\ &= \bar{Y}_1 - (P_1\bar{Y}_1 - P_2\bar{Y}_2) \\ &= P_2(\bar{Y}_1 - \bar{Y}_2), \end{aligned}$$

where \bar{Y} is the population mean, \bar{Y}_1 the population mean for respondents and \bar{Y}_2 the population mean for non-respondents. The bias will be small if the proportion of non-respondents is small or the mean for the population of non-respondents is close to that of the respondents. Since the sample provides no information about \bar{Y}_2 the bias is unknown unless bounds are placed from some source other than the sample information (Cochran, 1977; Lohr, 2010). Therefore, minimising the non-response rate is the only sure way to aid in controlling non-response bias (Lohr, 2010).

2.3.2.2.1 Weighting class adjustments

One way of adjusting the design weights of the respondents is with weighting classes where the weighting classes are formed from variables for which information is known for all the sampling units. The purpose of the adjustment is to make the weights of both the non-respondents and respondents in the same class, similar (Lohr, 2010). Weights of the respondents are increased so that a respondent in the same weighting class as a non-respondent represents the non-respondent's portion as well as their own in the population (Lohr, 2010). Let $\hat{\phi}_c$ denote the response probability for class c , given by

$$\hat{\phi}_c = \frac{\text{sum of the weights for respondents in class } c}{\text{sum of weights for selected sample in class } c}.$$

The sampling weight for each respondent is then multiplied by $1/\hat{\phi}_c$ which is termed the weight factor (Lohr, 2010). The weighting adjustment classes should be formulated as if they are strata (Lohr, 2010). The following conditions would be ideal to eliminate the response bias for estimating means and totals:

1. in class c there is a constant response variable;

2. for every unit in class c the response propensity (the probability that a unit will respond, i.e. $\hat{\phi}_i$) is constant; and
3. the response and the response propensity are uncorrelated.

2.3.2.2.2 Inverse of the response rate

Non-response distorts the results of many surveys, and results from surveys with very low response rates cannot readily be generalized to the greater population. The response rate is simply the number of persons who responded to the survey divided by the number of questionnaires mailed or supplied (Heeringa, et al., 2010; Cochran, 1977).

Let ϕ_i indicate the probability that an element when selected, will respond to the survey. This probability is unknown but assumed to be positive (Lohr, 2010). This probability can be estimated by means of weighting class adjustments discussed in Section 2.3.2.2.1. Then the probability that an element is selected for the sample and responds is

$$\pi_i \times \phi_i,$$

and this product is the response rate (Lohr, 2010). The final weight for a respondent is then $\frac{1}{\pi_i \times \hat{\phi}_i}$, the inverse of the response rate, where $\hat{\phi}_i$ is estimated using the formula in Section 2.3.2.2.1. The main reason for applying non-response factors in survey weights is to reduce bias as a result of non-response across sample elements (Lohr, 2010; Heeringa, et al., 2010; Cochran, 1977).

2.3.2.3 Calibration and integrated weighting

In the weighting process the first step is to compute a design weight, then compensation is made for non-response which can be done by the methods discussed in Section 2.3.2.2. However, it is often the case that the attained sample does not represent the population as intended which results in differential non-response (Neethling & Galpin, 2006; Luus, 2016). Differential non-response occurs when one sampled subgroup has a lower response frequency as opposed to other subgroups. Calibration is used to obtain improved estimates by using auxiliary information in the form of totals. These totals are known marginal counts such as

gender or other categorical variables that are used to form new weights, called calibration weights (Luus, 2016; Neethling & Galpin, 2006). The auxiliary information can be obtained from a census or other administrative files (Deville, et al., 1993).

To assist in obtaining the calibration weights, consider the finite population U of size N consisting of M households, where a sample S of size n is selected consisting of m households, drawn with a known probability. Suppose the k^{th} element is selected from U with inclusion probability π_k , where $\pi_k > 0$ (Deville, et al., 1993; Neethling & Galpin, 2006). The inclusion probabilities can be formulated into a $N \times N$ matrix, $\mathbf{\Pi} = diag(\pi_k)$. Furthermore, let w_k denote the design weight of unit k , which has already been adjusted to compensate for unit non-response. The objective of many surveys is to estimate the finite population total,

$$t_y = \sum_{k \in U} y_k = \sum_U y_k, \quad (2)$$

where y_k is the value of the variable of interest, y , of the k^{th} element, and t_y is the finite population total of the variable of interest. An estimator used to estimate t_y in Equation 2 is the Horvitz-Thompson estimator,

$$\hat{t}_y = \sum_{k \in S} w_k y_k. \quad (3)$$

To formulate the set of new weights, viz. calibration weights, auxiliary information must be used. The auxiliary information is in the form of categorical variables for which responses are known for each unit in the population. Assume that there exists J person level auxiliary variables x_1, x_2, \dots, x_J and consider the k^{th} element. Then a J -vector can be defined as $\mathbf{x}_k = (x_{k1}, x_{k2}, \dots, x_{kj}, \dots, x_{kJ})'$ where $k \in U$. Some examples of person auxiliary variables that can be used are gender, age (categorised), race, etc. The corresponding totals for vector \mathbf{x}_k can be obtained, i.e. $\mathbf{t}_x = \sum_U \mathbf{x}_k$, and placed into a vector. Suppose $\mathbf{x}_k = (1, x_{k1}, x_{k2})'$, then the population total vector will consist of $\mathbf{t}_x = (\sum_U 1, \sum_U x_{k1}, \sum_U x_{k2})$ which results in $\mathbf{t}_x = (N, N\bar{x}_{U1}, N\bar{x}_{U2})$. Furthermore, define a new weight c_k , so that

$$\sum_S c_k \mathbf{x}_k = \sum_U \mathbf{x}_k, \quad (4)$$

where c_k contains the calibration weights and is obtained so that the distance between c_k and w_k is as small as possible subject to the constraint $\sum_U \mathbf{x}_k$ (Deville, et al., 1993). Deville and Sarndal (1993) considered the distance function $\sum_S w_k v_k G(c_k, w_k)$ where v_k is a known

positive weight unrelated to w_k . Moreover, if $c_k = w_k$, then $G(1) = 0$. The following equation should be minimised,

$$\sum_S w_k v_k G(c_k, w_k) - \lambda' (\sum_S c_k \mathbf{x}_k - \sum_U \mathbf{x}_k) = 0, \quad (5)$$

where $\lambda = (\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_j, \dots, \lambda_j)'$ is the Lagrange multiplier vector (Deville, et al., 1993).

Differentiating Equation 5 with respect to c_k yields the following solution,

$$c_k = w_k F(\mathbf{x}'_k \lambda' / v_k), \quad (6)$$

where F is the inverse function of $g(.) = \frac{dG(u)}{du}$, for $u = c_k / w_k$. The λ can be calculated by substituting Equation 6 back into Equation 3. It follows that

$$\sum_S w_k F(\mathbf{x}'_k \lambda' / v_k) \mathbf{x}_k = \sum_U \mathbf{x}_k.$$

The two distance functions that will be used in this thesis are the linear and exponential distance functions. When the function is linear, $G(c_k, w_k) = \frac{w_k v_k}{2} \left(\frac{c_k}{w_k} - 1\right)^2$ is used which results in $F(\mathbf{x}'_k \lambda') = (1 + \mathbf{x}'_k \lambda' / v_k)$. Using Equation 6 the calibrated weights are

$$c_k = w_k (1 + \mathbf{x}'_k \lambda' / v_k).$$

In the exponential method (also known as multiplicative or raking ratio method) the auxiliary variables are expressed in the form of an exponential function (Deville, et al., 1993). The function $G(w_k, c_k) = w_k v_k \left[\frac{c_k}{w_k} \log\left(\frac{c_k}{w_k}\right) - \frac{c_k}{w_k} + 1 \right]$, $\frac{c_k}{w_k} > 0$ is used and yields $F(\mathbf{x}'_k \lambda') = w_k \exp(\mathbf{x}'_k \lambda' / v_k) > 0$. Similarly, the calibration weights are obtained from Equation 6 (Deville, et al., 1993; Neethling & Galpin, 2006). The calibration weights can be used to produce calibration estimates that are more efficient in sample surveys (Neethling & Galpin, 2006).

2.3.2.4 Integrated weighting techniques

The problem associated with calibration techniques at person level is that the person level weights assigned will generally differ from person to person in the same household (Neethling & Galpin, 2006). This results in uncertainty when household characteristics are estimated, since there is no weight that is a representation of the household. Also, the household size is not taken into account nor the fact that persons belonging to the same household should be treated as a cluster (Neethling & Galpin, 2006). Given these shortcomings, integrated

weighting has been developed to achieve one set of weights that overcome these problems. Two integrated weighting methods will be discussed, viz. integrated weighting based on person level auxiliary variables only and integrated weighting based on both person and household auxiliary variables.

2.3.2.4.1 Integrated weights based on person auxiliary variables

Calibration weights assign different weights to different persons in the same household. Integrated weighting per person or both per person and household assigns a single set of weights to the entire household. Consider the finite population U defined in Section 2.3.2.3 consisting of N persons and M households. In Section 2.3.2.3 vector \mathbf{x}'_k was defined. Consider a new matrix \mathbf{X} , with dimensions $N \times J$, the rows of which consist of \mathbf{x}'_k . In other words, row 1 will consist of $\mathbf{x}'_1 = (x_{11}, x_{12}, \dots, x_{1j}, \dots, x_{1J})'$, row 2 of $\mathbf{x}'_2 = (x_{21}, x_{22}, \dots, x_{2j}, \dots, x_{2J})'$, etc. Suppose sample S is selected containing n persons and m households. Matrix \mathbf{X}_S , where the subscript S denotes the sample, is a $n \times J$ matrix. Consider a sample consisting of two households that each contain two and three persons, respectively. Denote household one by (h_1p_1, h_1p_2) and household two by (h_2p_1, h_2p_2, h_2p_3) , therefore, S consist of two households and five persons. Consider the auxiliary variable gender comprising of (M, F) . Then vector $\mathbf{x}'_k = (x_{kM}, x_{kF})'$ for $k \in S$. The matrix \mathbf{X}_S will be a 5×2 matrix in form,

$$\mathbf{X}_S = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \\ x_{41} & x_{42} \\ x_{51} & x_{52} \end{bmatrix}$$

The matrix \mathbf{X}_S can be further adjusted to form a new matrix called \mathbf{Z}_{pp} in which the averages for the auxiliary characteristics are taken. Hence, for members belonging to household h with household size m_h , entries will be defined by $z_{hj} = \frac{a_{hj}}{m_h}$, where $a_{hj} = \sum_{j \in h} x_{hj}$. Consider, the example of two households (h_1p_1, h_1p_2) and (h_2p_1, h_2p_2, h_2p_3) with gender as the auxiliary variable. Suppose in household one both members are females, i.e. $h_1p_1 = F$ and $h_1p_2 = F$ and in household two $h_2p_1 = h_2p_2 = M$ and $h_2p_3 = F$. Then the new entry for $h_1p_1 = \frac{2}{2} = 1$,

i.e. $a_{hj} =$ two females and $m_h =$ two members in the household. The entry for h_1p_2 will be the same. The second household contains two males and one female. The entry for $h_2p_1 = h_2p_2 = \frac{2}{3}$ and $h_2p_3 = \frac{1}{3}$. Therefore matrix \mathbf{Z}_{pp} will be given by

$$\mathbf{Z}_{pp} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ \frac{2}{3} & 0 \\ \frac{2}{3} & 0 \\ 0 & \frac{1}{3} \end{bmatrix}.$$

Interpretation of the weight $\frac{2}{3}$ is, person h_2p_1 belongs to a household where $\frac{2}{3}$ are male. The integrated weighting method can be extended to include both person and household auxiliary variables.

2.3.2.4.2 Integrated weights based on person and household auxiliary variables

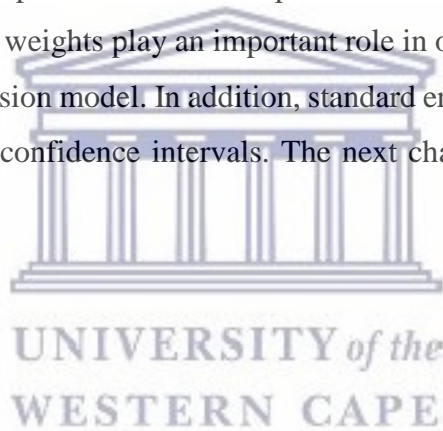
The method described in Section 2.3.2.4.1 is based on person-level auxiliary variables. This can be extended to include both person-level and household-level auxiliary variables. Household auxiliary variables can include province, geographical location, etc. Consider the matrix \mathbf{Z}_{pp} defined in Section 2.3.2.4.1. Now suppose geographical location is added which consists of rural and urban. Then a new matrix can be defined, \mathbf{Z}_{ph} , which has additional columns urban and rural. Consider the two-household example of Section 2.3.2.4.1, and suppose household one lives in a rural area and household two is, urban. Then

$$\mathbf{Z}_{ph} = \begin{bmatrix} 0 & 1 & 0 & \frac{1}{2} \\ 0 & 1 & 0 & \frac{1}{2} \\ \frac{2}{3} & 0 & \frac{1}{3} & 0 \\ \frac{2}{3} & 0 & \frac{1}{3} & 0 \\ 0 & \frac{1}{3} & \frac{1}{3} & 0 \end{bmatrix}.$$

The adjustment of the design weight already compensated for non-response through calibration and integrated weighting and is often referred to as benchmarking. At this stage, the final sampling weight is obtained (Luus, 2016). This completes the development of the final weight.

2.4 Conclusion

In this chapter the differences between probability and non-probability sampling were discussed and the importance of using probability sampling for inferences was highlighted. Furthermore, the different sampling techniques within both probability and non-probability sampling were reviewed. This followed to CS which is a combination of probability sampling techniques. Weighting was introduced and it was shown how it can be used to aid with non-response and differential non-response. The next chapter will discuss the logistic regression for both SRS and CS. In CS the weights play an important role in obtaining estimators of the parameters of the logistic regression model. In addition, standard errors for the estimators can be obtained and subsequently, confidence intervals. The next chapter will explore this for both SRS and CS.



3.1 Introduction

Logistic regression is a statistical technique widely used in the modelling of data where the response has two or more outcomes. Due to the nature of CS, when applying logistic regression modelling to CS data, the model needs to be adjusted to incorporate the CS design through the inclusion of the sampling weights. This chapter explains how the logistic regression model is adapted for use on CS data. It further goes on to discuss different methods of estimating the variances of the estimators of the model parameters under CS, viz. Taylor series linearization, the jackknife, and the bootstrap. These variances, in addition to the standard variance obtained from SRS, will be used to construct standard asymptotic confidence intervals. Furthermore, a non-parametric confidence interval, i.e. the bootstrap percentile interval will be discussed and will be compared to the standard (asymptotic) interval in the analysis.

3.2 Model specification and parameter estimation under SRS

The logistic regression model forms part of the generalised linear models in which the dependent variable follows one of the distributions of the exponential family (Agresti, 2013; O'Connell, 2006). Logistic regression models the odds of an event occurring and estimates the effects of the explanatory variables on those odds (O'Connell, 2006). Furthermore, if the dependent variable has two outcomes, it makes ordinary least squares regression modelling inappropriate (Heeringa, et al., 2010; Kutner, et al., 1996).

To validate this, suppose ordinary least squares regression was used to model a response that is binary. Consider the model

$$Y_i = \mathbf{X}'_i \boldsymbol{\beta} + \varepsilon_i, i = 1, \dots, n, \quad (7)$$

where Y_i is the response that only takes on the values 0 or 1, $\boldsymbol{\beta}$ is the vector $(\beta_0, \beta_1, \dots, \beta_p)$ consisting of $p + 1$ model parameters, \mathbf{X}'_i is the vector $(1, x_1, x_2, \dots, x_p)'$ consisting of p explanatory variables measured for the i th observation, and ε_i is the i th error term. Let $\varepsilon_i =$

$Y_i - \mathbf{X}'_i\boldsymbol{\beta}$. An expression for the error term can be obtained by substituting in the values for Y_i , namely

$$\varepsilon_i = 0 - \mathbf{X}'_i\boldsymbol{\beta}, \quad (8)$$

if $Y_i = 0$, and

$$\varepsilon_i = 1 - \mathbf{X}'_i\boldsymbol{\beta}, \quad (9)$$

if $Y_i = 1$. From the results in Equations 8 and 9 the error term can only take on two values. Therefore, the error term is not normally distributed (Kutner, et al., 1996). Furthermore, an additional problem associated with a response variable with a binary outcome is that the error variances are not constant. To validate this, consider the variance of the model defined in Equation 7 which, by the definition of a variance, yields

$$V(Y_i) = E(Y_i)(1 - E(Y_i)),$$

where $E(Y_i)$ is the expected value of Y_i which is equal to $\mathbf{X}'_i\boldsymbol{\beta}$. The variance of ε_i is the same as that of Y_i (Kutner, et al., 1996). Therefore,

$$V(\varepsilon_i) = E(Y_i)(1 - E(Y_i)),$$

and thus

$$V(\varepsilon_i) = \mathbf{X}'_i\boldsymbol{\beta}(1 - \mathbf{X}'_i\boldsymbol{\beta}). \quad (10)$$

From Equation 10 it is apparent that the variance is dependent on the explanatory variables implying that the error variances will differ at different levels of \mathbf{X} , resulting in the error variance not being constant. Thus, ordinary least squares regression will not be applicable to such data (Agresti, 2013; Kutner, et al., 1996).

In Figure 2, results were obtained from a simulated data set to display a naïve linear regression model for a binary response in the left panel compared to when an S-shaped curve is fitted in the right panel. It shows that a naïve linear regression model does not accurately capture the relationship between the response and explanatory variables and could possibly produce values that are outside the range of 0 and 1. On the other hand the S-shaped curve of the right panel accurately captures the probabilities of 0 and 1 (Kutner, et al., 1996; Heeringa, et al., 2010).

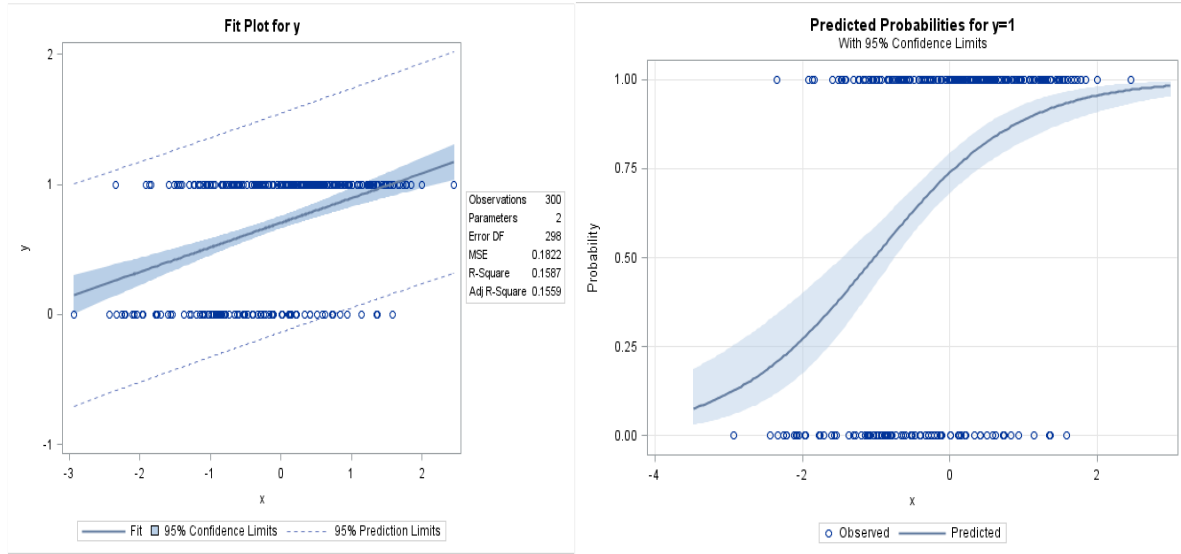


Figure 2: The effect of fitting an ordinary least squares regression to a binary response variable.

The linear regression model in Equation 7 must be altered to compensate for binary outcomes. To address this, a non-linear function must be identified that yields a fitted regression model that is linear in the coefficients for the model covariates and, ideally, when the function is transformed back, the resulting estimated values will fall in the range 0 to 1 (Heeringa, et al., 2010). The above described functions are referred to as link functions and the two commonly used to model binary responses are the logit and probit (Heeringa, et al., 2010; Lohr, 2010; Kutner, et al., 1996).

The link function used to model the logistic regression is the logit link function which transforms the outcome variable to the natural log of the odds (Menard, 2010). Consider Y_i which follows a Bernoulli distribution in which the expected value and variance are given below:

$$E(Y_i) = \pi_i, \tag{11}$$

and

$$V(Y_i) = \pi_i(1 - \pi_i), \tag{12}$$

where π_i is the probability that Y_i equals 1. Using the definition of the link function, the model in Equation 7 can be expressed as (O'Connell, 2006),

$$\text{logit}(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{X}'\boldsymbol{\beta}. \quad (13)$$

Equation 13 can be simplified to

$$\begin{aligned} \exp\left(\ln\left(\frac{\pi_i}{1-\pi_i}\right)\right) &= \exp(\mathbf{X}'\boldsymbol{\beta}) \\ \frac{\pi_i}{1-\pi_i} &= \exp(\mathbf{X}'\boldsymbol{\beta}) \\ \pi_i &= (1-\pi_i) \exp(\mathbf{X}'\boldsymbol{\beta}) \\ \pi_i &= \frac{\exp(\mathbf{X}'\boldsymbol{\beta})}{1+\exp(\mathbf{X}'\boldsymbol{\beta})}. \end{aligned} \quad (14)$$

To obtain the estimators of the model parameters the maximum likelihood function must be derived in order to find the maximum likelihood estimators (MLE) of the parameters (Kutner, et al., 1996; O'Connell, 2006). Since the observations in an SRS are independent and identically distributed (*i.i.d.*), the maximum likelihood function can be obtained as the product of n Bernoulli probability functions (Kutner, et al., 1996),

$$\begin{aligned} g(y_1, y_2, \dots, y_n) &= P(Y_1 = y_1) \times P(Y_2 = y_2) \times \dots \times P(Y_n = y_n) \\ &= \prod_{i=1}^n P(Y_i = y_i) \\ &= \prod_{i=1}^n \pi_i^{Y_i} (1-\pi_i)^{1-Y_i}. \end{aligned}$$

To lessen the computational burden, the logarithm of the maximum likelihood function is found,

$$\begin{aligned} \ln[g(y_1, y_2, \dots, y_n)] &= \ln\left[\prod_{i=1}^n \pi_i^{Y_i} (1-\pi_i)^{1-Y_i}\right] \\ &= \sum_{i=1}^n [Y_i \ln \pi_i + (1-Y_i) \ln(1-\pi_i)] \\ &= \sum_{i=1}^n [Y_i \ln \pi_i + \ln(1-\pi_i) - Y_i \ln(1-\pi_i)] \\ &= \sum_{i=1}^n [Y_i (\ln \pi_i - \ln(1-\pi_i)) + \ln(1-\pi_i)] \\ &= \sum_{i=1}^n Y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^n \ln(1-\pi_i). \end{aligned} \quad (15)$$

Since,

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{X}'\boldsymbol{\beta},$$

Equation 15 becomes

$$\ln[g(Y_1, Y_2, \dots, Y_n)] = \sum_{i=1}^n Y_i(\mathbf{X}'\boldsymbol{\beta}) - \sum_{i=1}^n \ln(1 + \exp(\mathbf{X}'\boldsymbol{\beta})). \quad (16)$$

To obtain the estimates of the parameters, the partial derivatives of $\ln[g(Y_1, Y_2, \dots, Y_n)]$ with respect to the parameters must be obtained and set to zero. However, this is not straightforward and requires iterative estimation procedures such as the Newton-Raphson method or the Fisher Scoring algorithm to obtain the estimators (Kutner, et al., 1996; Heeringa, et al., 2010).

The maximum likelihood function requires that the data be *i.i.d.*. However, this is not the case for CS and this will be discussed in the next section.

3.3 Model specification and parameter estimation under CS

Consider the CS described in Section 2.3.2.1 in which there are H strata where each stratum contains N_h PSUs and each PSU contains N_{hj} SSUs. From each stratum a sample of n_h PSUs is selected followed by n_{hj} SSUs being sampled from each of the selected PSUs, $j = 1, \dots, n_h$ and $h = 1, \dots, H$. Data collected using such a design restricts the normal straightforward use of the maximum likelihood function to obtain estimators of the model parameters for several reasons. The selection probabilities are no longer equal and sampling weights are therefore needed to estimate the finite population values of the logistic regression model parameter. Secondly, in a CS, stratification and clustering result in data that are not independent. The assumption of independence is imperative in order to estimate model parameters and their variances (Archer, et al., 2007; Heeringa, et al., 2010; Lohr, 2010; Lumley & Scott, 2015). Even if the estimators of the parameters were approximately design unbiased, the standard errors would likely be incorrect if the CS involves clustering (Lohr, 2010).

Instead of using the traditional MLE, a pseudo maximum likelihood function is used, i.e. the likelihood function is adapted as if the entire population is used (Archer, et al., 2007; Lohr, 2010; Chambless & Boyle, 1985). As mentioned in Section 2.3.2.1, the design weight has the property of indicating the number of population elements represented by the sample element. By expanding each sample observation by its design weight, a data set of N units is produced. Therefore, the parameter vector \mathbf{B} is introduced which is the MLE of the super population

parameter β which is referred to as the finite population parameter (Binder, 1983; Lohr, 2010). The logistic regression model in Equation (14) can be defined in terms of \mathbf{B} ,

$$\pi = \frac{\exp(\mathbf{X}'\mathbf{B})}{1 + \exp(\mathbf{X}'\mathbf{B})}.$$

The model parameters for the logistic regression of a CS are estimated using the pseudo maximum likelihood function. The Bernoulli probability distribution of the logistic regression model can be expanded using weights,

$$\pi_{hji}^{w_{hji} \times y_{hji}} [1 - \pi_{hji}]^{w_{hji} \times (1 - y_{hji})},$$

where y_{hji} is the binary response variable, π_{hji} is probability that y_{hji} is equal to 1 and w_{hji} is the sampling weight, where $i = 1, \dots, n_{hj}$, $j = 1, \dots, n_h$ and $h = 1, \dots, H$. The pseudo maximum likelihood function is still constructed using the product of the individual contributions, however, the n_h PSUs sampled and n_{hj} SSUs sampled within the given PSU are accounted for thus forming the pseudo maximum likelihood function,

$$l_p(\mathbf{B}) = \prod_{j=1}^{n_h} \prod_{i=1}^{n_{hj}} (\pi_{hji})^{w_{hji} \times y_{hji}} [1 - \pi_{hji}]^{w_{hji} \times (1 - y_{hji})}. \quad (17)$$

The pseudo MLE is similar to the MLE in terms of its functionality except that the pseudo MLE calculates the parameters for the expanded set. Expressed differently, the logistic regression for a CS is being fit to the 'census' data (Archer, et al., 2007; Heeringa, et al., 2010). The estimators are obtained using the iterative estimation procedures discussed in Section 3.2.

Once the estimators are determined the variances and standard errors of the estimators can be obtained. The thesis will discuss three methods to determine the variances under CS, namely Taylor series linearization, the jackknife and bootstrap.

3.4 Variance estimation

To obtain confidence intervals or conduct hypothesis tests, variance estimation is of paramount importance. For statistics based on data collected under the assumption of an SRS, exact expressions for variance estimators in most circumstances can be derived. In a CS design, however, these variance estimators are a bit more intricate and exact formulae can be

cumbersome to obtain (Heeringa, et al., 2010; Lohr, 2010). Several variance estimators exist for a CS design. Of these only the Taylor series linearization, the jackknife and the bootstrap will be discussed in this thesis.

3.4.1 Taylor series linearization

TSL is a method used to approximate complex smooth non-linear functions by simple linear functions of statistics in order to calculate variances, construct confidence intervals and test hypotheses of parameters (Heeringa, et al., 2010; Lohr, 2010; Kolenikov, 2010). TSL is traditionally used when the statistic of interest is a function of moments (Kolenikov, 2010).

Let $\theta = f(T_1, T_2, \dots, T_k)$ be a smooth function of totals T_1, T_2, \dots, T_k , which can be totals of any particular variable of interest, and let $\hat{\theta} = f(t_1, t_2, \dots, t_k)$ be an estimator of θ , where t_1, t_2, \dots, t_k are sample estimates of the corresponding totals (Kolenikov, 2010). Consider a complex sample where $T_l, l=1, 2, \dots, k$ can be estimated by

$$t_l = \sum_{i \in S} w_i y_{il}, \quad (18)$$

where t_l is an estimator of T_l , y_{il} is the response of unit i to item l , and w_i is the sampling weight for unit i . For simplicity the notation has been reduced to only the USU subscript i .

A new variable can be defined for constants a_1, \dots, a_k ,

$$q_i = \sum_{l=1}^k a_l y_{il},$$

such that,

$$\begin{aligned} t_q &= \sum_{i \in S} w_i q_i \\ &= \sum_{i \in S} w_i \sum_{l=1}^k a_l y_{il} \\ &= \sum_{l=1}^k a_l \sum_{i \in S} w_i y_{il} \\ &= \sum_{l=1}^k a_l t_l. \end{aligned}$$

The variance can then be estimated by

$$V(t_q) = V(\sum_{l=1}^k a_l t_l) = \sum_{l=1}^k a_l^2 V(t_l) + 2 \sum_{c=1}^{k-1} \sum_{l=c+1}^k a_l a_c Cov(t_l, t_c), \quad (19)$$

where $V(t_q)$ is the variance of the estimated total for constants a_1, \dots, a_k , $V(t_l)$ is the variance of the estimated total t_l , and $Cov(t_l, t_c)$ is the covariance of t_l and t_c (Lohr, 2010).

Now consider the estimator of the population mean that can be expressed as a weighted combined ratio estimator,

$$\frac{\sum_{i \in S} w_i y_{il}}{\sum_{i \in S} w_i} = \frac{t_l}{N} = \bar{y}. \quad (20)$$

The estimated mean, like many estimators under CS, is a non-linear function of two weighted sample totals. This is true for other estimated quantities too, such as the simple linear and logistic regression coefficients (Heeringa, et al., 2010; Lohr, 2010; Binder, 1983). This is a non-linear statistic and cannot be expressed in the form of Equation 19. To solve the problem of non-linearity of sample estimators, Taylor series expansion is used to approximate the estimates of interest, expressing them as linear combinations of weighted sample totals (Heeringa, et al., 2010).

In order to do so, let

$$\sum_{i \in S} w_i y_{il} = u \text{ and } \sum_{i \in S} w_i = v.$$

From Equation 20 it follows that

$$\bar{y} = \frac{u}{v}.$$

Using Taylor series expansion Equation 20 can be approximated as

$$\begin{aligned} \bar{y}_{TSL} &= \frac{u_0}{v_0} + (u - u_0) \left[\frac{\partial \bar{y}_{TSL}}{\partial u} \right]_{v=v_0, u=u_0} + (v - v_0) \left[\frac{\partial \bar{y}_{TSL}}{\partial v} \right]_{v=v_0, u=u_0} + \text{remainder}, \\ \bar{y}_{TSL} &\approx \frac{u_0}{v_0} + (u - u_0) \left[\frac{\partial \bar{y}_{TSL}}{\partial u} \right]_{v=v_0, u=u_0} + (v - v_0) \left[\frac{\partial \bar{y}_{TSL}}{\partial v} \right]_{v=v_0, u=u_0}, \end{aligned} \quad (21)$$

where u_0 and v_0 are the weighted sample totals which are obtained from the survey data,

$\left[\frac{\partial \bar{y}_{TSL}}{\partial u} \right]_{v=v_0, u=u_0}$ is the derivative of \bar{y} with respect to u evaluated at the expected values of

the sample estimates u_0 and v_0 , and $\left[\frac{\partial \bar{y}_{TSL}}{\partial v} \right]_{v=v_0, u=u_0}$ is the derivative of \bar{y} with respect to v

evaluated at the expected values of the sample estimates u_0 and v_0 .

Note that the quadratic and higher order terms in the full Taylor series expansion are dropped since those terms are assumed inconsequential when the sample sizes are large enough (Woodruff, 1971; Heeringa, et al., 2010; Lohr, 2010). Furthermore, consistent and ideally unbiased estimators are generally used in the place of the expected values of the sample

estimators (Heeringa, et al., 2010). Making use of Equation 19, the variance of the linearized estimator can be calculated. In Equation 21, let

$$\left[\frac{\partial \bar{y}_{TSL}}{\partial u} \right]_{v=v_0, u=u_0} = C \text{ and } \left[\frac{\partial \bar{y}_{TSL}}{\partial v} \right]_{v=v_0, u=u_0} = D.$$

Then Equation 21 reverts to

$$\bar{y}_{TSL} = \frac{u_0}{v_0} + (u - u_0)C + (v - v_0)D,$$

and the variance of \bar{y}_{TSL} can be calculated as

$$\begin{aligned} V(\bar{y}_{TSL}) &= V\left(\frac{u_0}{v_0} + (u - u_0)C + (v - v_0)D\right) \\ &= 0 + C^2 V(u - u_0) + D^2 V(v - v_0) + 2CD \text{cov}(u - u_0, v - v_0) \\ &= C^2 V(u) + D^2 V(v) + 2 CD \text{cov}(u, v). \end{aligned}$$

C and D can be obtained as

$$\left[\frac{\partial \bar{y}_{TSL}}{\partial u} \right]_{v=v_0, u=u_0} = \frac{1}{v_0} \text{ and } \left[\frac{\partial \bar{y}_{TSL}}{\partial v} \right]_{v=v_0, u=u_0} = -\frac{u_0}{v_0^2},$$

which simplifies to

$$V(\bar{y}_{TSL}) = \frac{V(u) + \bar{y}_{TSL} V(v) - 2\bar{y}_{TSL} \text{cov}(u, v)}{v_0^2}.$$

Binder (1983) proposed using a multivariate version of the TSL to calculate the variance of the estimator of a logistic regression model parameter (Binder, 1983; Heeringa, et al., 2010). As discussed in Section 3.3, the estimators of the parameters of the logistic regression model can be obtained from the pseudo MLE defined in Equation 17. Similarly, Equation 17 can be used to obtain a variance-covariance matrix of the logistic regression model parameters. A simplified version of Equation 17 for an observation in stratum h from the j^{th} PSU and the i^{th} SSU is given below,

$$\sum_h \sum_j \sum_i w_{hji} \mathbf{D}'_{hji} [\pi_{hji}(1 - \pi_{hji})]^{-1} (y_{hji} - \pi_{hji}) = \mathbf{0}, \quad (22)$$

where \mathbf{D}_{hji} is a vector of partial derivatives, $\frac{\phi(\pi_{hji}(\mathbf{B}))}{\phi_{B_k}}$, $k = 0, \dots, p$, p is the number of parameters, w_{hji} is the sampling weight and $\pi_{hji}(\mathbf{B})$ is the probability of success of the i^{th} SSU of PSU j from stratum h . Equation 22 reduces to $p + 1$ estimating equations,

$$\sum_h \sum_j \sum_i w_{hji} (y_{hji} - \pi_{hji}(\mathbf{B})) \mathbf{x}'_{hji} = \mathbf{0}. \quad (23)$$

The Newton-Raphson method can be used to obtain the weighted parameter estimates by finding a solution for Equation 23. Using TSL, a sandwich-type variance estimator can be obtained in the form of (Heeringa, et al., 2010)

$$var(\hat{\mathbf{B}}) = (\mathbf{J}^{-1})var[S(\hat{\mathbf{B}})](\mathbf{J}^{-1}),$$

where \mathbf{J} is the matrix of second-order derivatives with respect to \hat{B}_k .

TSL is used in most survey packages under the assumption that the PSUs are sampled with replacement within the strata at the first stage (Kolenikov, 2010). Some advantages of using TSL are that, if the partial derivatives are known, linearization will almost always give the variance estimate of a statistic and, the theory of TSL is well developed (Lohr, 2010). However, it does have some drawbacks. Calculations can be cumbersome when functions are complex. Also, not all statistics yield smooth functions in terms of population totals, and the accuracy depends on the sample size (Lohr, 2010; Kolenikov, 2010).

Although TSL is the default variance estimator in most statistical packages other options are also provided such as the jackknife and the bootstrap. These techniques form part of the resampling methods and will be discussed next.

3.4.2 Resampling methods

Resampling or replication methods, as the names suggest, replicate subsamples of the sampled observations to develop variance estimators for both linear and non-linear statistics (Heeringa, et al., 2010; Wolter, 2007).

Suppose that a sample S is selected by some design and suppose R replicates are obtained from sample S . Consider the r^{th} replicate, $r = 1, 2, \dots, R$. Let the parameter of interest be denoted by θ and let the estimator of θ be denoted by $\hat{\theta}$. Let the estimate of θ obtained from the r^{th} replicate be denoted by $\hat{\theta}^{(r)}$. The variance estimator of $\hat{\theta}$ can generally be defined as (Lohr, 2010)

$$V(\hat{\theta}) = \frac{1}{R(1-R)} \sum_{i=1}^R \{\hat{\theta}^{(r)} - \tilde{\theta}\}^2, \quad (24)$$

where R is the number of replicates; $V(\hat{\theta})$ is the variance estimator; and $\tilde{\theta}$ is a particular measure of central tendency. In the event of the mean of the resampled values, then

$$\tilde{\theta} = \frac{1}{R} \sum_{r=1}^R \hat{\theta}^{(r)}.$$

The resampling methods that will be discussed are the jackknife and the bootstrap. Theoretically, the resampling is done in a way that the sample is re-created for each replicate r . However, in practice this is done by using the sampling weights, i.e. if a particular observation unit is removed for a given replicate, it is simply assigned a weight of zero (Kolenikov, 2010). The weights of the other units need to be increased to ensure that the totals are unbiased for each replicate (Kolenikov, 2010).

3.4.2.1 Jackknife repeated replication

Jackknife repeated replication (JRR) was introduced as a method to reduce bias and can be used for a wide variety of complex designs (Heeringa, et al., 2010; Lohr, 2010; Kolenikov, 2010). The JRR focuses on samples that leave out one observation unit at a time (Efron & Tibshirani, 1994). This thesis will focus on the delete-one jackknife. Firstly, a brief discussion of JRR under SRS will be provided followed by the extension of JRR to the CS case.

3.4.2.1.1 JRR under SRS

Properties of the jackknife for SRSWR and SRSWOR have been extensively investigated and will be briefly explored in this section. Consider the sample S of size n with observations $\{y_1, y_2, y_3, \dots, y_n\}$ in which some parameter θ is estimated by the statistic $\hat{\theta}$. Suppose $S_{(i)}$ denotes the replicate sample in which the i^{th} observation has been removed, i.e. $S_{(i)} = \{y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n\}$. Using the replicate sample $S_{(i)}$ a replicate of $\hat{\theta}$ is obtained, namely $\hat{\theta}_{(i)} = \hat{\theta}(S_{(i)})$. This is repeated until each observation has been deleted once resulting in n replicates of $\hat{\theta}$, i.e. $\{\hat{\theta}_{(i)}\}, i = 1, \dots, n$. Using Equation 24 the JRR estimate of the variance of the estimator of the parameter of interest is calculated as

$$\hat{V}_{JRR}(\hat{\theta}) = \frac{n-1}{n} \sum (\hat{\theta}_{(i)} - \tilde{\theta}_{JRR})^2, \quad (25)$$

where $\tilde{\theta}_{JRR}$ is given by

$$\tilde{\theta}_{JRR} = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{(i)}.$$

The variance produced using JRR includes a factor of $\frac{n-1}{n}$ which is different from $\frac{1}{n-1}$ or $1/n$ which is traditionally used when calculating variances (Efron & Tibshirani, 1994; Lohr, 2010). The factor is derived by considering a special case where $\hat{\theta} = \bar{y}$ and the variance simplifies to

$$\hat{V}_{JRR}(\hat{\theta}) = \frac{1}{n(n-1)} \sum (y_i - \bar{y})^2 = \frac{1}{n} \sum (y_i - \bar{y})^2.$$

The jackknife estimate of the bias is given by

$$\widehat{bias}_{JRR} = (n-1)(\tilde{\theta}_{JRR} - \hat{\theta}).$$

The bias consists of a factor $(n-1)$ which is the same as the factor of the variance given in Equation 25. However, using the special case of letting $\hat{\theta} = \bar{y}$ is not plausible since the sample mean is an unbiased estimator of the population mean. The sample variance can be used instead. Consider

$$\hat{\theta} = \sum (y_i - \bar{y})^2 / n,$$

which has a bias of $-1/n$ times the population variance, and the factor $(n-1)$ in front of $(\tilde{\theta}_{JRR} - \hat{\theta})$ makes \widehat{bias}_{JRR} equal to $-1/n$ times $\sum (y_i - \bar{y})^2 / (n-1)$, the unbiased estimator of the population variance (Efron & Tibshirani, 1994).

When using JRR to estimate the variances of the estimators under CS the sample design needs to be accounted for. This section presented a short description of the JRR under SRS. Since this thesis considers the logistic regression modelling of CS data the next section considers the application of the JRR under CS.

3.4.2.1.2 JRR under CS

In JRR under CS each replicate measures the variance contributed by a single stratum in which case the PSU is removed along with all the observations within that PSU (Kolenikov, 2010; Lohr, 2010; Kish & Frankel, 1974). Deleting one observation at a time will destroy the cluster structure, therefore the entire PSU should be removed (Lohr, 2010). The software that

does the JRR calculations does not actually remove the PSU, but merely assigns a weight of zero to all the cases in that PSU (Heeringa, et al., 2010; Lohr, 2010).

Suppose there are H independent strata and n_h PSUs are chosen from stratum h . Let $\hat{\theta}_{(hj)}$ be the estimator obtained when PSU j of stratum h is deleted. In order to calculate $\hat{\theta}_{(hj)}$ the weights need to be assigned as follows (Lohr, 2010),

$$w_{i(hj)} = \begin{cases} 0, & \text{if observation } i \text{ is in PSU } j \text{ in stratum } h \\ \frac{n_h}{n_h-1} w_{hji}, & \text{if observation } i \text{ is not in PSU } j \text{ but in stratum } h, \\ w_{hji}, & \text{if observation } i \text{ is not in stratum } h \end{cases}$$

where $w_{i(hj)}$ is the adjusted sampling weight, n_h is the number of PSUs in stratum h , and w_{hji} is the original sampling weight of the i th PSU. The jackknife replicate of $\hat{\theta}$ when the (hj) th PSU has been deleted, i.e. $\hat{\theta}_{(hj)}$, is then calculated using the jackknife sampling weights. This procedure is repeated for all PSUs in a stratum and, independently, across all strata (Lohr, 2010; Heeringa, et al., 2010; Kolenikov, 2010). It follows that the jackknife estimator of the variance of $\hat{\theta}$ under CS is given by

$$V_{JRR}(\hat{\theta}) = \sum_{h=1}^H \frac{n_h-1}{n_h} \sum_{j=1}^{n_h} (\hat{\theta}_{(hj)} - \hat{\theta})^2.$$

This is done to ensure that the observational units within a PSU remain together such that the cluster structure remains intact. JRR is applied separately in each stratum at the first stage of sampling (Lohr, 2010).

The jackknife is an “all purpose” method and provides a consistent estimator when the parameter of interest is a smooth function of totals (Kish & Frankel, 1974; Lohr, 2010). However, JRR may require a large number of computations for some sampling designs which can be computationally expensive (Kolenikov, 2010).

3.4.2.2 Bootstrap

Statistics is based on sampling distributions of parameter estimators and test statistics. These distributions can be derived through transformations of random variables or other asymptotic arguments (Kolenikov, 2010). This is not always easy to determine. Bootstrap provides an

alternative to that in which the bootstrap statistics are taken from a distribution that is close to a distribution of an unknown population (Kolenikov, 2010). Using bootstrap statistics, standard errors can be obtained and subsequent confidence intervals can be constructed. This section commences with a summary of the bootstrap under SRS followed by its application to CS data.

3.4.2.2.1 Bootstrap under SRS

Consider an SRS $S=(y_1, y_2, \dots, y_n)$ from an unknown probability distribution F and suppose a parameter of interest θ is to be estimated by $\hat{\theta}$. The accuracy of $\hat{\theta}$ depends on its standard error. If F is unknown, the standard error of $\hat{\theta}$ cannot be readily obtained (Efron & Tibshirani, 1994). The empirical distribution \hat{F} , which assigns a probability of $\frac{1}{n}$ to each element in S , can be used to estimate F . It can be shown that \hat{F} is a sufficient statistic of F , the proof of which is omitted from the scope of this thesis. As a result, the \hat{F} can be used as a basis for obtaining the standard error of $\hat{\theta}$.

Suppose an SRSWR of size n is drawn from S , say $S_1^*=(y_1^*, y_2^*, \dots, y_n^*)$. Corresponding to S_1^* is the replicate of the estimator $\hat{\theta}$, i.e. $\hat{\theta}_1^* = \hat{\theta}(S_1^*)$. Another SRSWR of size n , say S_2^* , can be selected from S and the second replicate of the estimator $\hat{\theta}$, $\hat{\theta}_2^*$, can be obtained similarly to $\hat{\theta}_1^*$. Note that since this is SRSWR, S_1^* and S_2^* can differ. These samples, S_1^* and S_2^* , are referred to as bootstrap samples. This process, i.e. sampling with replacement, can be repeated until all possible samples of S are obtained. All these samples follow the empirical distribution \hat{F} . The estimate of the standard error of $\hat{\theta}$ can be obtained, say $\widehat{se}_{\hat{F}}(\hat{\theta}^*)$, and is referred to as the ideal bootstrap estimator. Note that the ideal bootstrap is a function of the empirical distribution and can be computationally expensive to obtain since it requires all possible samples from S of a certain size (Efron & Tibshirani, 1994).

The bootstrap algorithm is a numerical method to obtain an approximation of $\widehat{se}_{\hat{F}}(\hat{\theta}^*)$. It works by drawing many independent bootstrap samples, calculating replicates of the estimator from each bootstrap sample, and then using these replicates to estimate the

corresponding standard error of $\hat{\theta}$. The result is referred to as the bootstrap estimate of the standard error of $\hat{\theta}$.

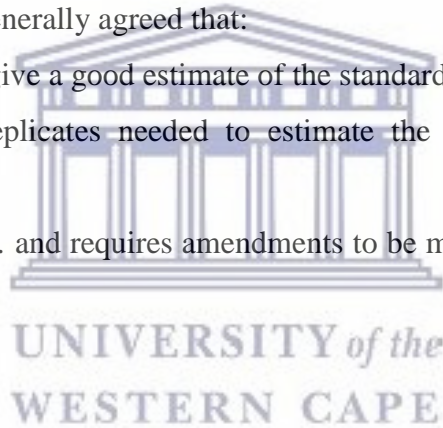
Consider B independent with replacement bootstrap samples, $S_1^*, S_2^*, \dots, S_B^*$, each of size n . A bootstrap replication of $\hat{\theta}$ can be obtained from each bootstrap sample, i.e. $\hat{\theta}_b^*$, for $b = 1, 2, \dots, B$. The resulting bootstrap estimate of the standard error of $\hat{\theta}$ is given by,

$$\widehat{se}_{(B)} = \sqrt{\frac{\sum_{b=1}^B [\hat{\theta}_{(b)}^* - \hat{\theta}^*(\cdot)]^2}{B-1}},$$

where $\hat{\theta}^*(\cdot) = \frac{\sum_{b=1}^B \hat{\theta}_b^*}{B}$. The ideal set up requires $B = \infty$, which results in the “ideal bootstrap estimate” (Efron & Tibshirani, 1994, p. 50). The ideal bootstrap estimate has a smaller standard error as opposed to \widehat{se}_B in an asymptotic sense (Efron & Tibshirani, 1994). However, the ideal bootstrap can be computationally expensive. The bootstrap estimate generally has very little bias. Authors have generally agreed that:

1. $B = 50$ is often enough to give a good estimate of the standard error; and
2. very rarely is $B > 200$ replicates needed to estimate the standard error (Efron & Tibshirani, 1994).

In CS the data is no longer *i.i.d.* and requires amendments to be made to the bootstrap. This is discussed in the next section.



3.4.2.2.2 Bootstrap under CS

The bootstrap can be extended to complex samples in which a bootstrap sample is taken of the PSUs within each stratum (Lohr, 2010; Kolenikov, 2010). Note that, as with the jackknife, observations within the PSU are kept together in the bootstrap iterations (Lohr, 2010).

Consider a complex sample design in which n_h PSUs is selected from stratum h , $h = 1, \dots, H$. Suppose the b^{th} bootstrap sample, for $b = 1, \dots, B$ is taken by selecting an SRSWR of n_h PSUs independently from stratum h . The parameter of interest can be estimated from replicate b by $\hat{\theta}_b^*$; this is repeated B times. The variance of the estimator of the parameter of interest can be calculated by Equation 24. Sitter (1992) highlighted the problem associated with this approach. In the simple case of the sample mean the variance obtained is not an unbiased

estimator and is not consistent (Kolenikov, 2010; Sitter, 1992). This is rectified by the application of a rescaling bootstrap procedure. To construct the b^{th} bootstrap sample a SRSWR of $n_h - 1$ PSUs instead of n_h is taken from the n_h PSUs in stratum h . (Luus, et al., 2010; Kolenikov, 2010). In addition, $n_h - 1$ gives more efficient estimators (Kolenikov, 2010).

Let m_{hj}^* be the number of times PSU j of stratum h appears in the bootstrap sample. Since the PSUs are sampled with replacement, some of the PSUs will appear more than once in the sample and others might not appear at all. Thus the sampling weights of the observations in the bootstrap sample need to be adjusted to compensate for this to ensure that the sum of the sampling weights still equals the population total. The bootstrap sampling weights are then given by

$$w_{hji}^* = w_{hji} \frac{n_h}{n_h - 1} m_{hj}^*, \quad (26)$$

where w_{hji} is the original sampling weight, m_{hj}^* is the number of times the j^{th} PSU appears in the bootstrap sample, and n_h is the number of PSUs that comprises stratum h , $h = 1, \dots, H$. The bootstrap weights can now be used to calculate the bootstrap replicates of $\hat{\theta}$, i.e. $\hat{\theta}^*$. These are then used to calculate the bootstrap variance of $\hat{\theta}$, by firstly using Equation 26 to construct a vector of replicate weights. Let $\hat{\theta}_b^*$ be the estimator of θ calculated in the same way as $\hat{\theta}$, but instead using the weights w_{hji}^* as opposed to w_{hji} . Then using Equation 24 the bootstrap variance for CS can be calculated (Lohr, 2010; Kolenikov, 2010).

The size of B in CS should ideally be selected to be at least as large as the design's degrees of freedom, i.e. $n - H$. Selecting $B < n - H$ does not provide the highest possible rank of the co-variance matrix of the coefficient estimates (Kolenikov, 2010). However, this may not be of concern if $n - H$ is sufficiently large, e.g. exceeds 100 (Kolenikov, 2010).

The bootstrap works well for smooth and non-smooth functions of statistics in general sampling designs (Lohr, 2010). It may, however, be computationally intensive as opposed to the other two variance methods, viz. TSL and JRR. Since different bootstrap samples can be used to compute the variance, the bootstrap variance estimates may differ.

3.5 Confidence intervals for model parameters

Once a sample is selected according to some design, estimators of parameters can be obtained. Moreover, additional information is often desired to make an assessment regarding the accuracy of these estimators. This is often done by constructing confidence intervals. A confidence interval summarises the uncertainty that the true population value lies in a bound placed on the probable error of an estimator from a single sample (Thompson, 2010). Two confidence intervals will be constructed, viz. the standard (asymptotic) confidence interval and the bootstrap percentile confidence interval.

3.5.1 Standard (asymptotic) confidence interval

Consider an estimator $\hat{\theta}$ and suppose the estimator is consistent and asymptotically normal. Let $V(\hat{\theta})$ denote the variance of that estimator. Then the expression

$$\frac{\hat{\theta} - \theta}{\sqrt{V(\hat{\theta})}}, \quad (27)$$

is said to be a pivotal quantity if its distribution does not depend on the parameter θ . For large samples the expression in Equation 27 possesses an approximately normal distribution and can be used to construct an asymptotic confidence interval for the parameter θ (Wackerly, et al., 2008). This assumption is true for TSL, JRR and the bootstrap under certain conditions:

1. the parameter of interest θ can be expressed as a smooth function of totals; and
2. the sample sizes are large: either the number of PSUs is large in each stratum or the survey contains a large number of strata (Binder, 1983; Efron & Tibshirani, 1994; Sitter, 1992; Lohr, 2010).

The generic form for a $100(1-\alpha)\%$ confidence interval for a population parameter, where α is the level of significance, is

$$\hat{\theta} \pm t_{\frac{\alpha}{2}, df} \cdot \widehat{se}(\hat{\theta}), \quad (28)$$

where $t_{\frac{\alpha}{2}, df}$ is the student t -distribution with df degrees freedom under the sampling design and $\widehat{se}(\hat{\theta})$ is an estimate of the standard error of $\hat{\theta}$. Simulation studies suggest that the confidence intervals behave well in practice. These studies also suggest that the TSL and JRR

give similar estimates for variances while the bootstrap gives larger estimates of the variances (Lohr, 2010).

3.5.2 The bootstrap percentile confidence interval

The bootstrap percentile interval is a non-parametric technique used to find percentiles of the bootstrap replicates obtained from the bootstrap samples (Efron & Tibshirani, 1994). In Section 3.4.2.2.1 the empirical distribution function \hat{F} was used to approximate F , from \hat{F} , B bootstrap samples, S^* , were drawn and from each bootstrap sample a bootstrap replicate $\hat{\theta}^*$ was obtained. To generate the percentile interval, the first step is to sort the bootstrap replicates $\hat{\theta}_b^*$, in ascending order, i.e. $\hat{\theta}_{(b)}^*$, $b=1,2,\dots,B$. Then the lower bound of the interval is obtained as the $B \times \frac{\alpha}{2}$ th value of the sorted replicates while the upper bound is obtained by taking the $B \times (1 - \frac{\alpha}{2})$ th value of the sorted replicates (Efron & Tibshirani, 1994; Luus, 2016). The 100 (1 - α)% bootstrap percentile interval is thus given by:

$$[\hat{\theta}_{(B \times \frac{\alpha}{2})}^*, \hat{\theta}_{(B \times (1 - \frac{\alpha}{2}))}^*], \quad (29)$$

where $\hat{\theta}_{(B \times \frac{\alpha}{2})}^*$ and $\hat{\theta}_{(B \times (1 - \frac{\alpha}{2}))}^*$ are the lower and upper bounds of the bootstrap percentile interval, respectively.

The bootstrap percentile interval for CS is exactly the same as defined in Equation 29 with the exception of the weights incorporated in the calculation of the replicates.

3.6 Conclusion

The difference between the estimation of parameters of a logistic model for CS and SRS is important to note. CS data as opposed to data obtained from SRS data are not *i.i.d.* and this has an effect on the MLE of the logistic regression model. This necessary adaptation was noted in Section 3.3 in which a pseudo MLE was obtained. Furthermore, three different variance estimation methods, viz. TSL, JRR and the bootstrap, were discussed including how the different methods are formulated firstly in an SRS setting followed by the CS setting. The variances can be used to construct a standard confidence interval and the bootstrap can be

used to construct a non-parametric confidence interval, namely the bootstrap percentile confidence interval.

The next chapter will provide the methodology to aid in providing solutions to the research questions highlighted. The data sets and sampling design are described and the statistical techniques to compare CS and SRS will be discussed.



UNIVERSITY *of the*
WESTERN CAPE

Chapter 4: Research methodology

4.1 Introduction

In Chapter 1 the problem statement, objectives and research questions were outlined. Chapters 2 and 3 discussed the statistical theory on the major concepts outlined in Chapter 1. In this chapter, the statistical methodology will be outlined and various aspects concerning data collection and data validity will be discussed. In addition, the different statistical methods that will be used to assess the research questions and problem statement, will be highlighted.

4.2 Data collection

The Income and Expenditure Survey (IES) 2005/2006 forms the basis of the simulation study of this thesis. It was based on a newly designed Master Sample (MS) which is used for all surveys conducted by Statistics South Africa (Lehohla, 2008). This newly designed MS was developed from the 2001 population census' enumerated areas (EAs), the smallest geographical areas into which the country is divided for survey purposes (Lehohla, 2008). The MS is designed to focus on all households living in private dwellings and workers living in workers' quarters within the country.

There were 3000 PSUs (note that the EAs were used as the PSUs) in the MS which were divided into four quarterly allocations of 750 each. Within each quarter an SRS of 250 PSUs was selected every month using the updated listings. Then within a selected PSU eight dwelling units were selected systematically (Lehohla, 2008). In total, 2400 dwelling units were covered during the twelve-month period. The survey was conducted from September 2005 to August 2006. The households were sampled and participated for a period of one month after which new sub-samples were taken of households for the new month (Lehohla, 2008). The data for a participant was collected for both the survey month and eleven months prior to the survey being conducted. This information was combined to give an estimated annual figure of expenditure per expenditure item (Lehohla, 2008).

There were three methods used to collect the data. A main questionnaire was used consisting of five sections of which the first covered the household characteristics, the next three sections covered different parts of consumption expenditure and the final section covered household expenditure. For the second method of data collection the participant had a weekly diary in which the daily acquisitions had to be written down. Finally, a summary questionnaire was administered in which the fieldworker had to summarise the total value of each item and transfer it to the appropriate part of the questionnaire (Lehohla, 2008).

4.3 Weighting

The IES data were collected through a complex sampling design, specifically a stratified two-stage cluster sampling design. According to this design, the country was firstly stratified by province after which each of the nine strata was divided into enumerated areas, i.e. PSUs, with each enumerated area consisting of a number of households, i.e. SSUs.

Consider stratum h , $h = 1, \dots, H$, and suppose in stratum h there is N_h PSUs and within each PSU there is N_{hj} SSUs, $j = 1, \dots, N_h$ and $h = 1, \dots, H$. PPS sampling, with number of SSUs the measure of size (MOS), was used to select n_h PSUs from each stratum in the first stage and then systematic sampling was used to select n_{hj} SSUs from each first-stage sampled PSU, $j = 1, \dots, n_h$ and $h = 1, \dots, H$ (Lehohla, 2008). The weighting procedure was applied at two stages (Lehohla, 2008). Consider PSU j selected from stratum h . Let π_{hj} denote the inclusion probability of the j th PSU from stratum h . Then

$$\pi_{hj} = n_h \cdot \frac{N_{hj}}{\sum_j N_{hj}},$$

where N_{hj} is number of SSUs in the selected PSU, $\sum_j N_{hj}$ the total number of SSUs in the stratum, and n_h the number of PSUs sampled from the stratum. Now let $\pi_{i|hj}$ denote the inclusion probability of the i th SSU given that the j th PSU was sampled from stratum h .

Then,

$$\pi_{i|hj} = \frac{n_{hj}}{N_{hj}},$$

where n_{hj} is the number of SSUs sampled from the (hj) th PSU and N_{hj} is the number of SSUs in the (hj) th PSU. The total inclusion probability of the (hji) th SSU is then given by

$$\pi_{hji} = \pi_{hj} \cdot \pi_{i|hj} = n_h \cdot \frac{N_{hj}}{\sum_j N_{hj}} \cdot \frac{n_{hj}}{N_{hj}}.$$

Furthermore, these probabilities were adjusted for non-response and the non-response adjustment factor used was the inverse of the response rate. The response rate is given by

$$r_{hj} = \frac{n_R}{n_T},$$

where n_R is the number of responding SSUs and n_T the total number of households visited (Lehohla, 2008). The design weight adjusted for non-response is thus given by

$$w_{hji} = \frac{1}{\pi_{hji} \times r_{hj}}.$$

The SAS macro CALMAR was used to perform the calibration and integrated weighting whereby w_{hji} was corrected to align with known population totals of certain auxiliary variables for which all information is known. The auxiliary variables are discussed in Section 4.5.2.

4.4 Response and imputations

As discussed in Section 2.3.2.2 there are two types of non-response, namely unit and item non-response. Unit non-response occurs when an entire sampling unit's information is omitted as opposed to item non-response which occurs when certain question responses are omitted (Luus, 2016; Lohr, 2010). Unit non-response is dealt with during weighting while item non-response imputations have to be carried out at different stages. The two stages at which imputations were done on missing data were:

1. imputing for missing diaries; and
2. imputing for item non-response.

Households were required to complete four weekly diaries and a main questionnaire for a period of a month. However, for various reasons, the diaries were not completed for all four weeks. Households that did not diarise their expenditure for a minimum of two weeks were disqualified and treated as non-respondents. This approach was extended to households that

had diaries but no main questionnaire (Lehohla, 2008). Missing values for households with diaries for two or more weeks were imputed. Suppose a household only diarised two weeks of information, then the expenditure for those weeks would be summed together and the total would be divided by two. The result would be used to impute the missing information for the other two weeks. Similarly, if the household had only three weeks of diarised information, then the expenditure would be summed, the total would be divided by three and the result would be used for the fourth week (Lehohla, 2008).

In terms of missing data specifically in which item non-response was present, imputations were done and these items were primarily related to housing. There are three different methods used to measure housing services from owner-occupied dwelling units, namely:

1. interest on loans and mortgage bonds;
2. imputed rent for owner-occupied dwelling units as estimated by respondents; and
3. percentage of the value of the house as an estimate of the rental value of the dwelling unit (Lehohla, 2008).

Essentially, imputations were carried out on missing items according to the following criteria: households that had similar characteristics to the ones missing were identified. Variables such as province, settlement type, type of dwelling unit, value of the house and the number of rooms were used to match households. The average amount for a particular item, as calculated from households of similar characteristics, was used to impute the missing data (Lehohla, 2008).

4.5 Statistical techniques

In order to address the research questions a comparison must be made between the correct implementation of the sampling design in the analyses, CS, and where the sampling design was ignored, SRS. In order to do this the “true” values of the model parameters must be obtained such that the estimates produced by the estimators of the parameters obtained under CS and SRS can be compared to the “truth”. However, the “true” parameter values require knowledge of the population model, which is unknown. Instead, a surrogate population will

be used as a basis for obtaining the “true” parameters. Samples will be selected from the surrogate population to obtain the estimates of the parameters of interest. These estimates will be calculated for both SRS and CS under different weighting methods, i.e. no weighting (None), design weight (Design), linear person-level auxiliary variable weighting (Lin_{pp}), linear person and household-level auxiliary variable weighting (Lin_{ph}), exponential person-level auxiliary variable weighting (RR_{pp}), and exponential person and household-level auxiliary variable weighting (RR_{ph}). The bias and mean squared error (MSE) will be used to determine how close the estimates are to the “true” parameters and will be discussed in Sections 4.6.1. Furthermore, standard confidence intervals will be obtained based on the TSL, JRR and bootstrap estimated variances calculated for no weighting and weighting. In addition, a bootstrap percentile confidence interval will be obtained for the parameters based on estimates obtained from applying no weighting as well as the different weighting methods. The different confidence intervals will be compared based on their coverage probabilities and lengths. These are discussed further in Section 4.6.2.

4.5.1 Surrogate population

The surrogate population that will be used is the Income and Expenditure Survey (IES) 2005/2006. In order to prevent any irregularities, a number of adjustments were made to the IES data set. Firstly, observations having missing data values were removed. Note that although various imputation mechanisms based on sound theory exist to compute those values, it would have presented another level of uncertainty and variability which could affect the inference (Luus, 2016). Furthermore, imputation is not the focus of this thesis. Secondly, only observations with age ranging from 21 to 65 were retained; this was considered a working age. The final adjustment was done on the household expenditure variable. This variable is important as it will be used to construct the response variable. Only household expenditure with positive values were retained. After all the adjustments were made, the surrogate population consisted of 25893 persons. The surrogate population was further adjusted to select only one person per household, namely the oldest person, which was

considered as representative of the head of the household. This resulted in 17541 households which were grouped into 283 PSUs (Luus, 2016).

4.5.2 The simulated samples

Monte Carlo simulation was applied to the surrogate population so that the performance of the different weight-based estimators of the model parameters could be compared. The bootstrap and jackknife methods were then applied to the simulated data with the purpose of obtaining bootstrap and jackknife estimated variances to use in the calculation of the confidence intervals.

The simulation consisted of drawing 100 samples from the surrogate population where each sample followed the same design as the IES 2005/2006, i.e. a stratified two-stage cluster sampling design. The nine provinces of South Africa were used as the strata, with the EAs in each stratum acting as the PSUs and the dwelling units within a PSU as the SSUs. The number of observations in each sample was 2028.

Differential non-response (such as older females being over-represented and younger males being under-represented), as described in Section 2.3.2.3, is often found in practical situations in South Africa. In order to determine this type of non-response error it was simulated in the design of the samples to evaluate the different weighting procedures under non-perfect circumstances. Two sets of auxiliary variables were used, namely person-level auxiliary variables and person and household-level auxiliary variables, to determine which weighting procedure performs best under such circumstances. For the person-level auxiliary variables, indicated by the subscript “pp”: province (9 categories), gender (2 categories), race (4 categories), and age were used. For person and household-level auxiliary variables, indicated by the subscript “ph”: all person-level auxiliary variables, area type (2 categories), and household size (3 categories) (Luus, 2016).

4.5.3 Model and variables

The logistic regression model will be constructed for both the surrogate population and for each of the 100 simulated samples. In order to use logistic regression, as mentioned in Chapter 3, the response variable has to be dichotomous. Since the surrogate population contains household information and the subsequent samples were selected per household, a poverty cut-off value had to be devised that captures household dynamics.

The food poverty line (FPL) is a poverty level used that captures consumption expenditure at household level. This value is rebased to give a value per person. The standard level for the FPL is \$2.34 per person per day. In 2006 this amounted to R16.38 per person per day and R5978.70 per person per year. To convert the value to a per household value an average was determined from the product of the household size and R5978.70. This amounted to R11062.20 household expenditure per year (Lehohla, 2017). Therefore, the response variable based on the food poverty level, Y , is given by

$$Y = \begin{cases} 1, & \text{if household expenditure} \leq R11062.20 \\ 0, & \text{otherwise} \end{cases}$$

The explanatory variables included in the model are age, gender, race, area type, education level, province and household size. The age variable, as mentioned in Section 4.5.1, ranges from 21 to 65. Gender, race, area type, province and household size were coded in such a way that the category with the largest proportion was used as the reference category. In addition, education level, which consists of 28 categories, was re-grouped into 6 smaller categories. Table 1 shows the re-grouping of the education level variable.

Table 1: Re-grouped education variable.

No schooling	Primary	Intermediate	Secondary	Matric	Tertiary
00=No schooling 26="Don't know" NA	01=Grade R 02=Grade 1 03=Grade2 04=Grade3 05=Grade4 06=Grade5 07=Grade6 08=Grade7	09=Grade8 10=Grade9	11=Grade10 12=Grade11 17=Certificate<13 18=Diploma<13	13=Matric 19=certificate=12 20=diploa=12	14=NTC I 15=NTC II 16=NTC III 21=Bachelors 22=Bachelors+Diploma 23=Honours 24=Higher degree 25=Other

Household size, which consisted of 10 categories, was also re-grouped into 3 categories. This is depicted in Table 2 below.

Table 2: Re-grouped household size variable.

Household 1	Household 2	Household 3 and more
one member	two members	three members four members five members six members seven members eight members nine members ten members

A preliminary test was done to determine which category is the largest and this was used as the baseline category for each of the categorical variables. The explanatory variables are defined below:

- Age variable ranges from 21 to 65;

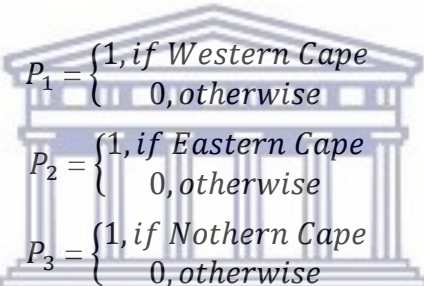
- Gender consists of males (coded 1) and females (coded 2). From the preliminary test it was determined that the males are the largest category. The new variable, G_1 , is defined as:

$$G_1 = \begin{cases} 1, & \text{if Female;} \\ 0, & \text{if Male;} \end{cases}$$

- Area type consists of urban (coded 1) and rural (coded 0). From the preliminary test it was determined that the urban category makes up the greater proportion. The new variable, A_1 , is defined as:

$$A_1 = \begin{cases} 1, & \text{if rural;} \\ 0, & \text{otherwise;} \end{cases}$$

- South Africa is made up of nine provinces. From the preliminary test it was determined that KwaZulu-Natal makes up the largest proportion. The new variables, P_1 to P_8 , are defined as:



$$P_1 = \begin{cases} 1, & \text{if Western Cape} \\ 0, & \text{otherwise} \end{cases}$$

$$P_2 = \begin{cases} 1, & \text{if Eastern Cape} \\ 0, & \text{otherwise} \end{cases}$$

$$P_3 = \begin{cases} 1, & \text{if Northern Cape} \\ 0, & \text{otherwise} \end{cases}$$

$$P_4 = \begin{cases} 1, & \text{if Free State} \\ 0, & \text{otherwise} \end{cases}$$

$$P_5 = \begin{cases} 1, & \text{if North West} \\ 0, & \text{otherwise} \end{cases}$$

$$P_6 = \begin{cases} 1, & \text{if Gauteng} \\ 0, & \text{otherwise} \end{cases}$$

$$P_7 = \begin{cases} 1, & \text{if Mpumalanga} \\ 0, & \text{otherwise} \end{cases}$$

$$P_8 = \begin{cases} 1, & \text{if Limpopo;} \\ 0, & \text{otherwise;} \end{cases}$$

- The race variable consists of four categories, namely Black, Coloured, Asian/Indian and White, coded 1 to 4 respectively. From the preliminary results it was determined that Black is the largest category. The new variables, R_1 to R_3 , are defined as:

$$R_1 = \begin{cases} 1, & \text{if Coloured} \\ 0, & \text{otherwise} \end{cases}$$

$$R_2 = \begin{cases} 1, & \text{if Indian/Asian} \\ 0, & \text{otherwise} \end{cases}$$

$$R_3 = \begin{cases} 1, & \text{if White} \\ 0, & \text{otherwise} \end{cases}$$

- The household variable consists of three categories. From the preliminary results it was determined that a household size of one was the largest category. The new variables, H_1 and H_2 , are defined as:

$$H_1 = \begin{cases} 1, & \text{if two members} \\ 0, & \text{otherwise} \end{cases}$$

$$H_2 = \begin{cases} 1, & \text{if three or more members} \\ 0, & \text{otherwise} \end{cases};$$

- The education level variable consists of six categories. From the preliminary results it was determined that the category “primary school” had the largest proportion. The new variables, E_1 to E_5 , are defined as:

$$E_1 = \begin{cases} 1, & \text{if No schooling} \\ 0, & \text{otherwise} \end{cases}$$

$$E_2 = \begin{cases} 1, & \text{if Intermediate} \\ 0, & \text{otherwise} \end{cases}$$

$$E_3 = \begin{cases} 1, & \text{if High School} \\ 0, & \text{otherwise} \end{cases}$$

$$E_4 = \begin{cases} 1, & \text{if Matric} \\ 0, & \text{otherwise} \end{cases}$$

$$E_5 = \begin{cases} 1, & \text{if Tertiary} \\ 0, & \text{otherwise} \end{cases}$$

Finally, the population logistic regression model is given by

$$\begin{aligned} \text{Povertylev} = & \beta_0 + \beta_1 \text{AGE} + \beta_2 A_1 + \beta_3 G_1 + \beta_4 R_1 + \beta_5 R_2 + \beta_6 R_3 + \beta_7 E_1 + \\ & \beta_8 E_2 + \beta_9 E_3 + \beta_{10} E_4 + \beta_{11} E_5 + \beta_{12} H_1 + \beta_{13} H_2 + \beta_{14} P_1 + \beta_{15} P_2 + \beta_{16} P_3 + \\ & \beta_{17} P_4 + \beta_{18} P_5 + \beta_{19} P_6 + \beta_{20} P_7 + \beta_{21} P_8. \end{aligned} \quad (30)$$

4.6 Statistical methodology

After the data cleaning and variable transformation is done logistic regression models will be obtained for both the surrogate population and the samples. The surrogate population model will represent the “true” parameters to which the sample estimates will be compared. The logistic model defined in Equation 30 will be applied to each sample, either ignoring the sample design (None) or accounting for the design through the inclusion of the different sampling weights. The different weighting procedures that will be used and compared, are: design weighting only (Design), calibration and integrated weighting using the linear distance method on person auxiliary variables (Lin_{pp}), calibration and integrated weighting using the linear distance method on person and household auxiliary variables (Lin_{ph}), calibration and integrated weighting using the raking ratio (exponential) distance method on person auxiliary variables (RR_{pp}), and calibration and integrated weighting using the raking ratio distance method on person and household auxiliary variables (RR_{ph}). The standard errors of the estimators from each of those methods, i.e. None, Design, Lin_{pp} , Lin_{ph} , RR_{pp} and RR_{ph} , will be obtained using TSL, JRR, and the bootstrap. This will be done in SAS and R.

4.6.1 Assessment of the estimators of the model parameters

This section presents the measures that will be used to assess how close the estimators of the model parameters are to the “truth”. Let $\hat{\beta}_i$ denote the estimator of the i th model parameter, β_i , $i = 0, \dots, p$. The estimator will be assessed based on its expected value, bias and mean squared error (MSE). Each of these measures are discussed below.

Consider the r th sample, $r = 1, \dots, R$, and let $\hat{\beta}_{i_r}$ denote the replicate of $\hat{\beta}_i$ obtained when fitting the logistic regression model to the r th sample. The expected value of $\hat{\beta}_i$ is approximated by the average of the R replicates of $\hat{\beta}_i$,

$$\frac{1}{R} \sum_{r=1}^R \hat{\beta}_{i_r}.$$

The closer the expected value of an estimator is to the parameter, the better the estimation method (Wackerly, et al., 2008). The bias of an estimator is defined as the difference between the expected value of the estimator and the parameter, and is approximated by

$$bias(\hat{\beta}_i) = \frac{1}{R} \sum_{r=1}^R \hat{\beta}_{i_r} - \beta_i,$$

where β_i is the “true” value of the model parameter obtained from the surrogate population. If the bias is zero, i.e. $E(\hat{\beta}_i)$ is equal to the “true” value then $\hat{\beta}_i$ is said to be an unbiased estimator. The absolute bias is simply $|\frac{1}{R} \sum_{r=1}^R \hat{\beta}_{i_r} - \beta_i|$.

The mean squared error is defined as the average of the square of the distance between the estimator and its target parameter and is approximated by

$$MSE(\hat{\beta}_i) = \frac{1}{R} \sum_{r=1}^R (\hat{\beta}_{i_r} - \beta_i)^2.$$

There are two aspects to the MSE of an estimator, i.e. the variance of the estimator and the bias. The smaller the MSE the better the estimator. However, if the MSE is large this can be due to a large variance or bias or both. If the estimator is unbiased then the MSE equals the variance (Wackerly, et al., 2008).

The expected value, bias and MSE will be used to assess the performance of the unweighted and different weighted estimators of the logistic regression model parameters.

4.6.2 Assessment of the confidence intervals for the model parameters

Consider β_i , the i^{th} model parameter, estimated by $\hat{\beta}_i$ and let $\{\hat{\beta}_{i_r}\}$ denote the R replicates of $\hat{\beta}_i$ obtained from the R samples. The replicates are used to construct the following 95% confidence intervals for β_i :

1. standard (asymptotic) interval using the TSL estimated variance;
2. standard (asymptotic) interval using the JRR estimated variance;
3. standard (asymptotic) interval using the bootstrap estimated variance; and
4. the bootstrap percentile confidence interval.

In order to determine which of the intervals performs “best” under the different weighting procedures the confidence intervals will be assessed based on their coverage probability and length.

The coverage probability of a confidence interval is defined as the proportion of times that the confidence interval, based on the R replicate samples, contains the parameter β_i . The 95% standard (asymptotic) confidence interval is obtained as:

$$[\hat{\beta}_{i_{LLr}}; \hat{\beta}_{i_{ULr}}] = \hat{\beta}_{i_r} \pm t_{\frac{\alpha}{2}, df} \cdot \widehat{se}(\hat{\beta}_{i_r})$$

where $\widehat{se}(\hat{\beta}_{i_r})$ is the estimated standard error of $\hat{\beta}_{i_r}$ obtained using either TSL, JRR or the bootstrap, $\hat{\beta}_{i_{LLr}}$ is the lower limit of the confidence interval of the r^{th} sample for the i^{th} model parameter, and $\hat{\beta}_{i_{ULr}}$ is the upper limit of the confidence interval of the r^{th} sample for the i^{th} model parameter. This gives $[\hat{\beta}_{i_{LL1}}; \hat{\beta}_{i_{UL1}}], [\hat{\beta}_{i_{LL2}}; \hat{\beta}_{i_{UL2}}], \dots, [\hat{\beta}_{i_{LLR}}; \hat{\beta}_{i_{ULR}}]$. The coverage probability (CP) is then calculated as

$$CP = \frac{\#\{\hat{\beta}_{i_{LLr}} \leq \beta_i \leq \hat{\beta}_{i_{ULr}}\}}{R}$$

The confidence interval for which CP is the closest to 95% is considered the “best”.

Consider the r^{th} sample. B bootstrap samples can be selected from sample r , namely $S_{r_1}^*, S_{r_2}^*, \dots, S_{r_B}^*$. From each bootstrap sample a replicate of $\hat{\beta}_{i_r}$ is obtained, i.e. $\hat{\beta}_{i_{r_1}}^*, \hat{\beta}_{i_{r_2}}^*, \dots, \hat{\beta}_{i_{r_B}}^*$. As discussed in Section 3.5.2 these replicates for sample r are sorted in ascending order, of which the lower bound of the confidence interval is the $B \times \frac{\alpha}{2}th$ value of the sorted replicates and the upper bound of the confidence interval is the $B \times (1 - \frac{\alpha}{2})th$ value of the sorted replicates.

Then a 95% bootstrap percentile confidence interval for the r^{th} sample is given by

$$[\hat{\beta}_{i_{LLr}}^*; \hat{\beta}_{i_{ULr}}^*] = [\hat{\beta}_{i_r^*}^*_{(B \times \frac{\alpha}{2}th)}; \hat{\beta}_{i_r^*}^*_{(B \times (1 - \frac{\alpha}{2})th)}],$$

where $\hat{\beta}_{i_{LLr}}^*$ and $\hat{\beta}_{i_{ULr}}^*$ are the lower and upper bounds for the r^{th} sample for the i^{th} model parameter using the bootstrap percentile confidence interval. Similarly, bounds can be obtained for $r=1, 2, \dots, R$, which gives $[\hat{\beta}_{i_{LL1}}^*; \hat{\beta}_{i_{UL1}}^*], [\hat{\beta}_{i_{LL2}}^*; \hat{\beta}_{i_{UL2}}^*], \dots, [\hat{\beta}_{i_{LLR}}^*; \hat{\beta}_{i_{ULR}}^*]$. The CP for the bootstrap percentile confidence interval is calculated as

$$CP = \frac{\#\{\hat{\beta}_{iLLr}^* \leq \beta_i \leq \hat{\beta}_{iULr}^*\}}{R}$$

The coverage probability will range between zero and one. The closer the coverage probability is to the confidence level, the better the coverage. However, the improved coverage can be due to a large confidence interval length caused by large variances. Therefore, the confidence interval length is considered which is defined as

$$L_{i_r} = \hat{\beta}_{iULr} - \hat{\beta}_{iLLr},$$

where L_{i_r} is the length of the interval for β_i calculated from the r^{th} sample. This gives $L_{i_1}, L_{i_2}, \dots, L_{i_R}$ from which an average length is calculated,

$$\frac{\sum_{i=1}^R L_{i_r}}{R}$$

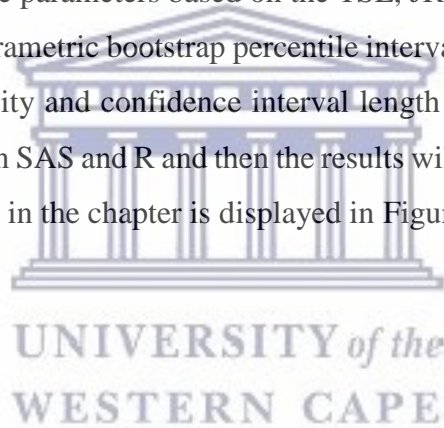
The average length is calculated for each interval type and each weighting approach and will be considered in conjunction with the CP to determine the “best” confidence interval.

4.7 Conclusion

This chapter provided all the tools to assess the research questions outlined in Chapter 1. In this chapter the surrogate population, i.e. the IES 2005/2006 was defined and the response variable coupled with the explanatory variables which are required to build the logistic regression model were constructed and clearly defined. From the surrogate population, samples were selected; each one having a CS design. For each of these samples a logistic regression will be constructed from which estimators will be obtained under SRS and CS and compared to the parameters obtained from the surrogate population. Two such methods were discussed in this chapter, viz. the bias and MSE. Furthermore, variances of these estimators can be obtained. Section 3.4 discussed three variances under CS, i.e. TSL, JRR and the bootstrap. These variances will be used to construct a standard (asymptotic) confidence interval. In addition, a non-parametric confidence interval was discussed, i.e. the bootstrap percentile interval. Two methods were discussed to assess the precision of these confidence intervals, i.e. the coverage probability and the confidence interval length. The next chapter will provide empirical results for the methods mentioned in Chapter 4.

5.1 Introduction

In the previous chapters the building blocks were laid in order to address the research questions outlined in Chapters 1 and 2. Chapter 3 explained the statistical theory and reviewed previous literature while Chapter 4 outlined the statistical methodology and simulated samples needed to build models to compare to the “truth”. As noted in Chapter 4, for each simulated sample a logistic model will be obtained in the form of Equation 30 under the assumption of a SRS and CS in which the Design, Lin_{pp} , Lin_{ph} , RR_{pp} and RR_{ph} weights will be used. From the results the absolute bias and MSE will be calculated and the results will be displayed for a selected number of parameters. In addition, standard (asymptotic) confidence intervals will be obtained for the parameters based on the TSL, JRR and bootstrap estimated variances, including the non-parametric bootstrap percentile interval. The results will be used to obtain the coverage probability and confidence interval length outlined in Section 4.6.2. The analysis will be replicated in SAS and R and then the results will be compared. An outline of the formulation of the results in the chapter is displayed in Figure 3.



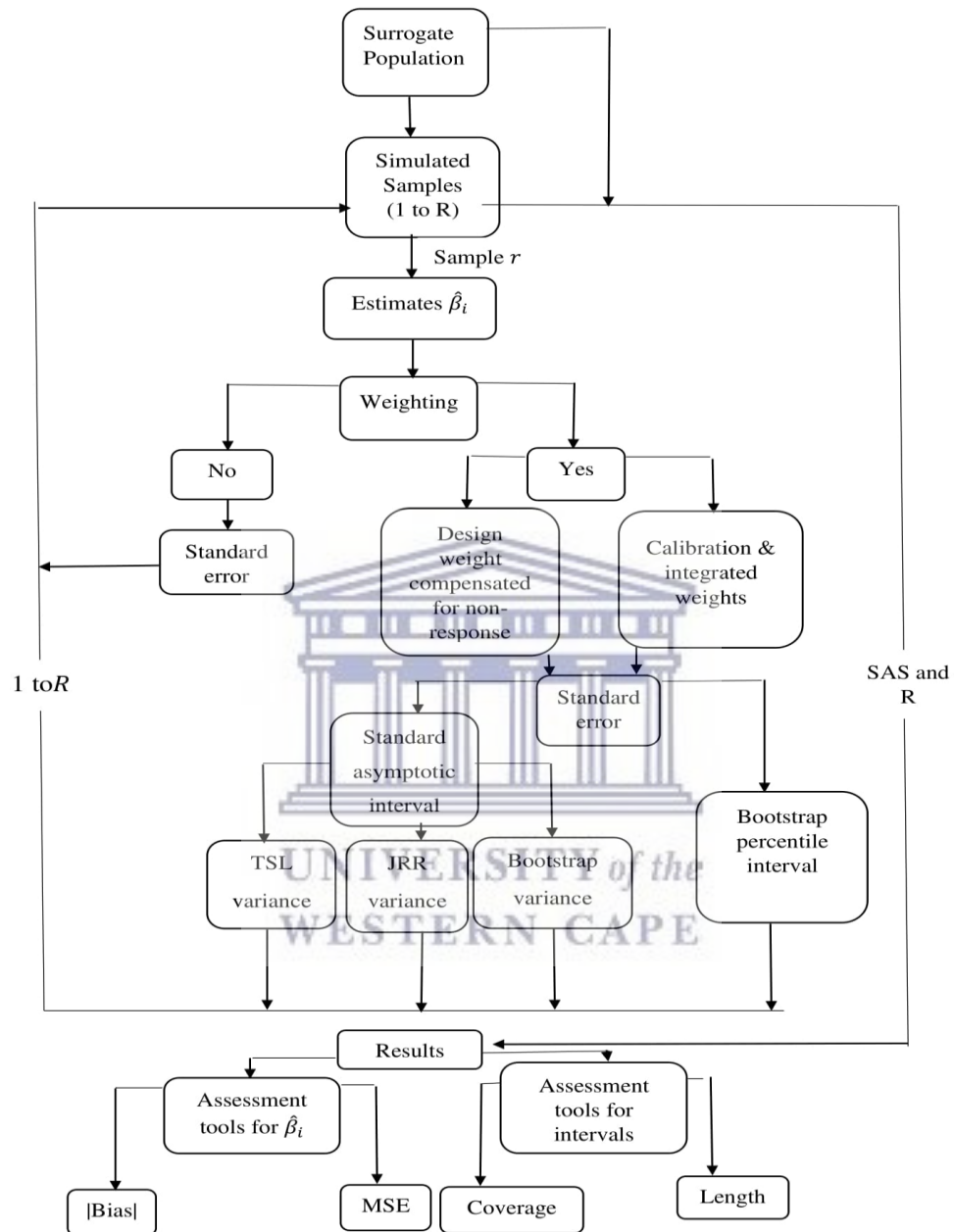


Figure 3: Outline of the simulation study.

5.2 Discussion of results

In this section the results obtained from the simulation study are examined. Only a selection of the results is shown and discussed here. The complete collection of results is presented in Appendix A to D and all programmes used in the simulation study are available from the author at 3315602@myuwc.ac.za.

5.2.1 Estimators of model parameters

As discussed previously, the Income and Expenditure Survey conducted over the period September 2005 until August 2006 forms the surrogate population of this simulation study. This section considers the estimators of the model parameters and their measures of accuracy. Two accuracy measures are discussed and displayed in figures for the estimators, viz. the absolute bias and MSE. Figure 3 gives an outline as to how the results are reported for two statistical packages, namely SAS and R. R is opensource software which was developed as a dialect of the S language, an object-orientated statistical programming language (Seefeld & Linder, 2007; Lumley, 2011). It has a package called “survey” which accommodates CS designs (Lumley, 2011). This was used to obtain the estimates from which the subsequent biases and MSEs were calculated. SAS is a statistical software programme primarily developed for business solutions pertaining to manipulation of data, performance of sophisticated analyses and business intelligence (Simon & Mitterling, 2017). SAS contains “procs” which are used to perform the analyses (Elliot & Woodward, 2010). The “proc survey” was used to incorporate the CS design in the analyses. Similar to R, once the estimates were obtained the absolute biases and MSEs were calculated.

5.2.1.1 The absolute bias

In Section 4.6.1 the absolute bias was discussed as one of the methods to assess how close to the “truth” an estimator is. The estimates were obtained and a selection of the results for the absolute bias are displayed in Figure 4 to Figure 10 for the estimators $\hat{\beta}_0$, $\hat{\beta}_2$, $\hat{\beta}_4$, $\hat{\beta}_5$, $\hat{\beta}_{11}$, $\hat{\beta}_{12}$ and $\hat{\beta}_{20}$. The remainder are included in Appendix A1 to A14. The Figures contain SRS (no

weights), Design, Lin_{pp} , Lin_{ph} , RR_{pp} and RR_{ph} , the weights used under CS. For each method the absolute bias is displayed for R and SAS next to each other.

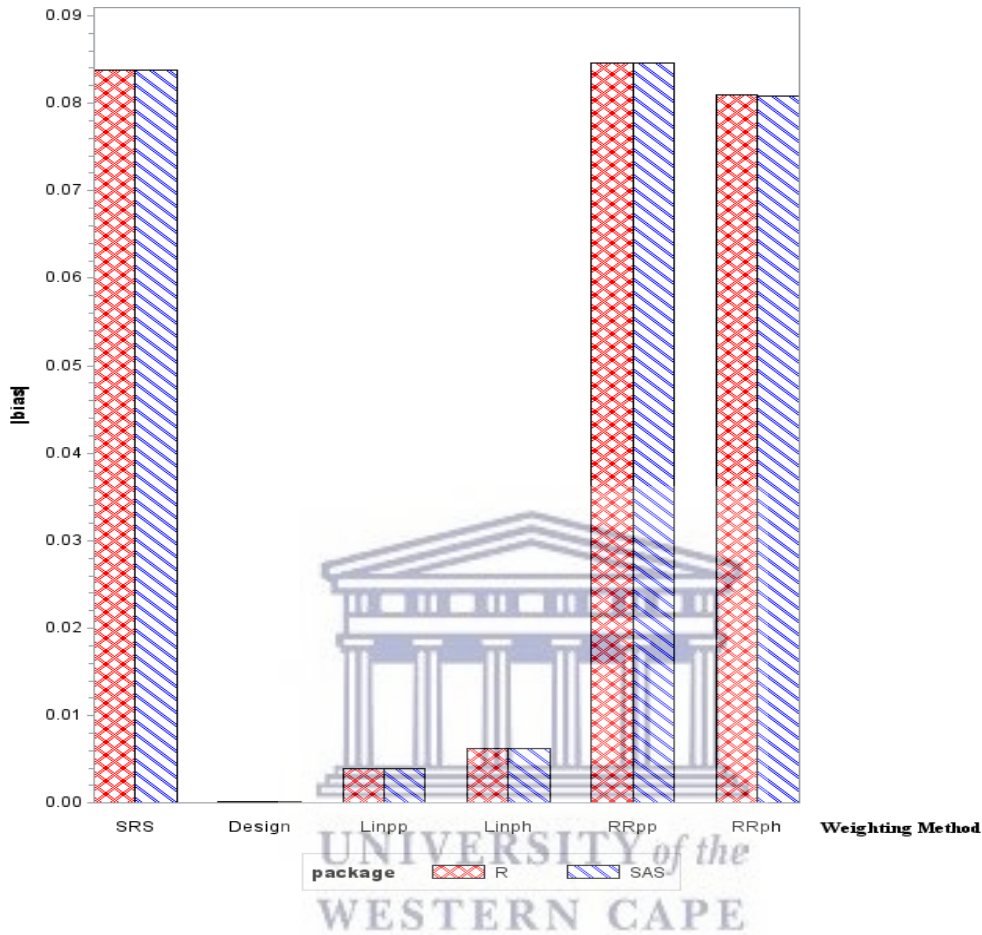


Figure 4: The absolute bias of the estimator of β_0 under SRS (no weight) and different weighting methods are shown for SAS and R.

In Figure 4, it is seen that the absolute bias for the SAS and R output were the same. Estimators based on the design weight showed little to no bias. Similarly, estimators based on Lin_{pp} and Lin_{ph} showed little bias. In contrast greater bias was shown when SRS, RR_{pp} and RR_{ph} were used.

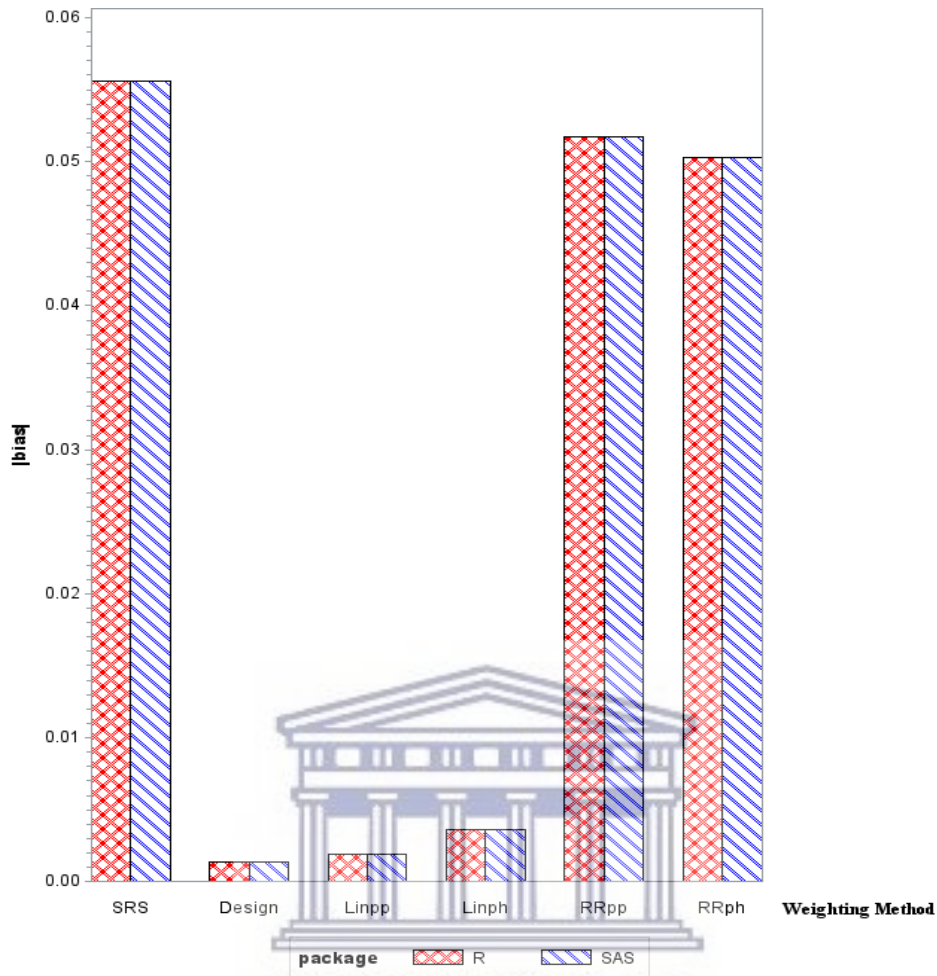


Figure 5: The absolute bias of the estimator of β_2 under SRS and different weighting methods are shown for SAS and R.

Similar to Figure 4, the SAS and R output shown in Figure 5, were exactly the same. Likewise, similar patterns were observed, namely the estimates obtained using the design weight had very little bias in contrast to the estimates obtained from SRS.

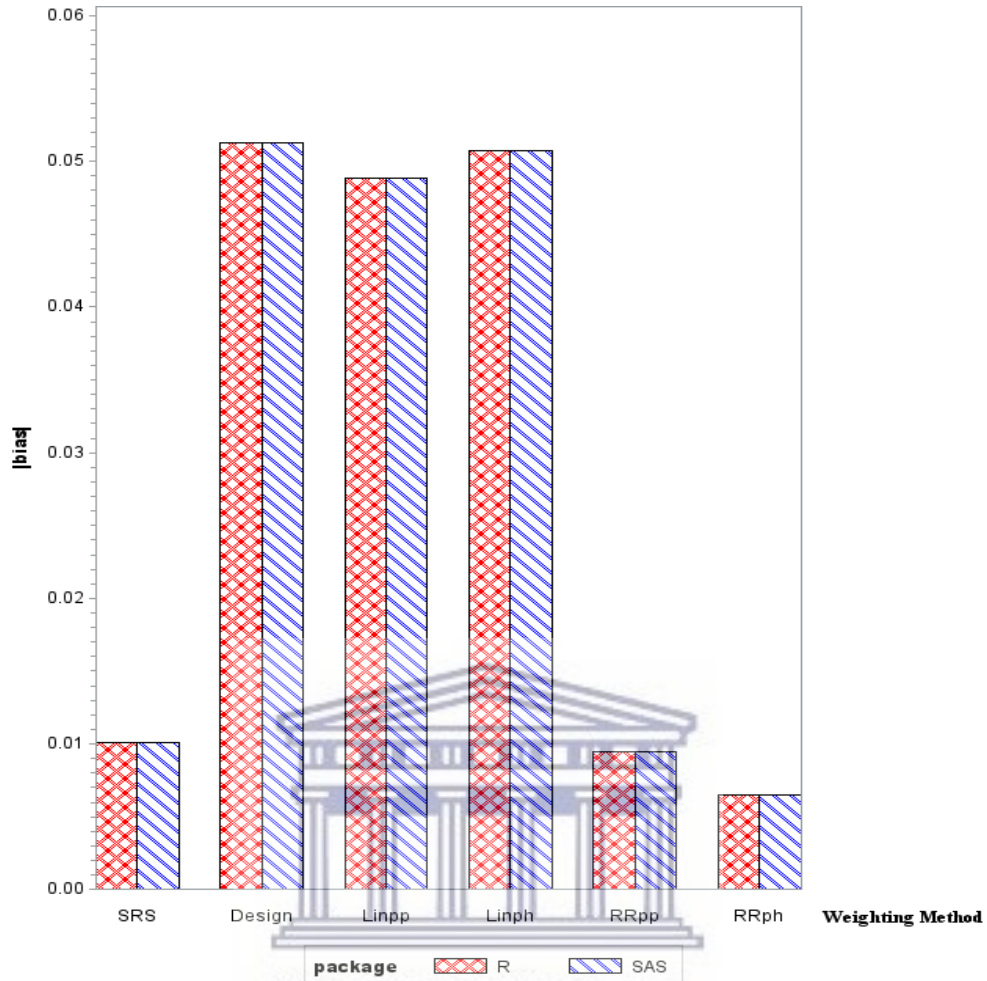


Figure 6: The absolute bias of the estimator of β_4 under SRS and different weighting methods are shown for SAS and R.

In Figure 6, the absolute bias shows a different pattern in comparison to Figure 4 and Figure 5. The estimates obtained from the Design, Lin_{pp} and Lin_{ph} weights show larger absolute bias as opposed to estimates obtained from the other three methods. The bias based on the weight RR_{ph} was the lowest.

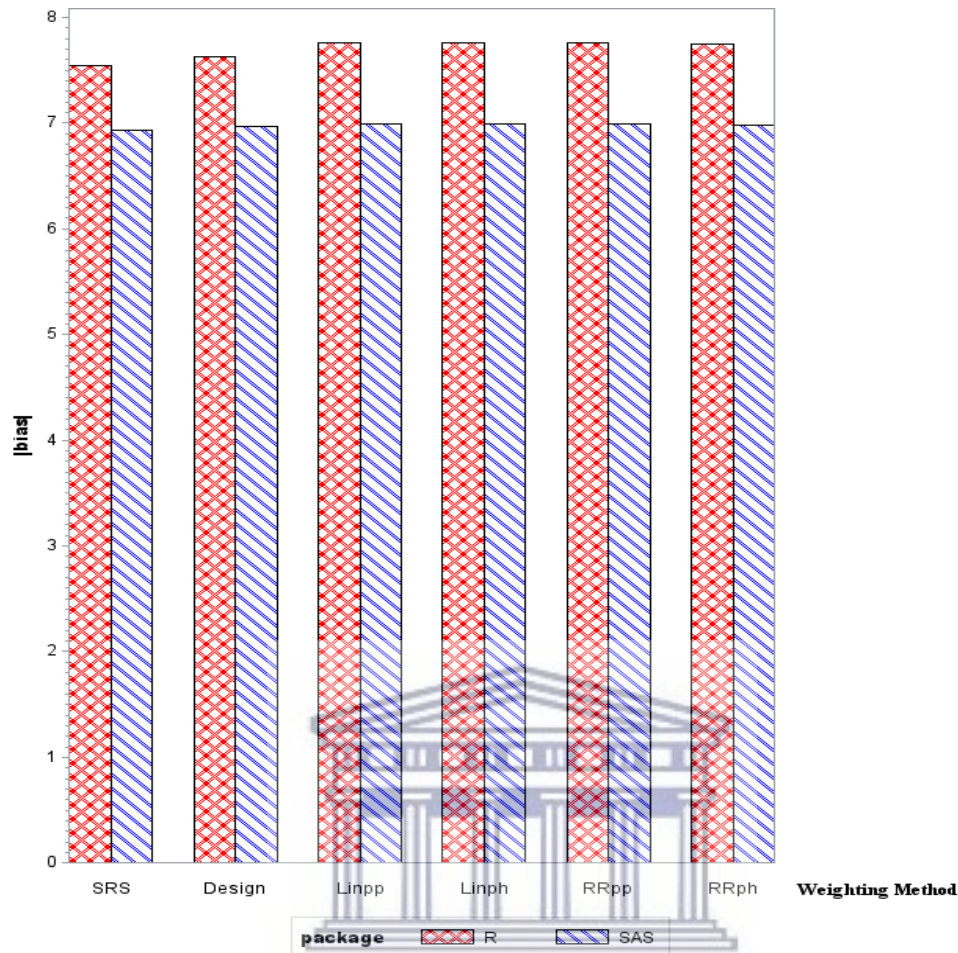


Figure 7: The absolute bias of the estimator of β_5 under SRS and different weighting methods are shown for SAS and R.

In Figure 7, the output from SAS and R of the absolute bias of the estimator of the parameter β_5 , the coefficient of predictor R_2 , differed. The explanatory R_2 represented the Indian or Asian race group. It should be noted that from the preliminary results the frequency of R_2 was small. This resulted in the quasi-separation of data points in some of the samples. When quasi-separation is detected in SAS, the procedure terminates the MLE iteration process and reports the last iteration. In the results window SAS reports that the validity of the model is questionable (SAS Institute, 2017). In R the solution for these estimator's MLE are infinite, however R provides a finite value by falsely converging the iterative procedure (Heinze & Schemper, 2002). The difference between the SAS and R output can be attributed to when

the software terminates the iteration process (Heinze & Schemper, 2002). The SAS output's absolute bias was slightly smaller as opposed to results obtained from R. There was no significant difference across weighting methods from the R output.

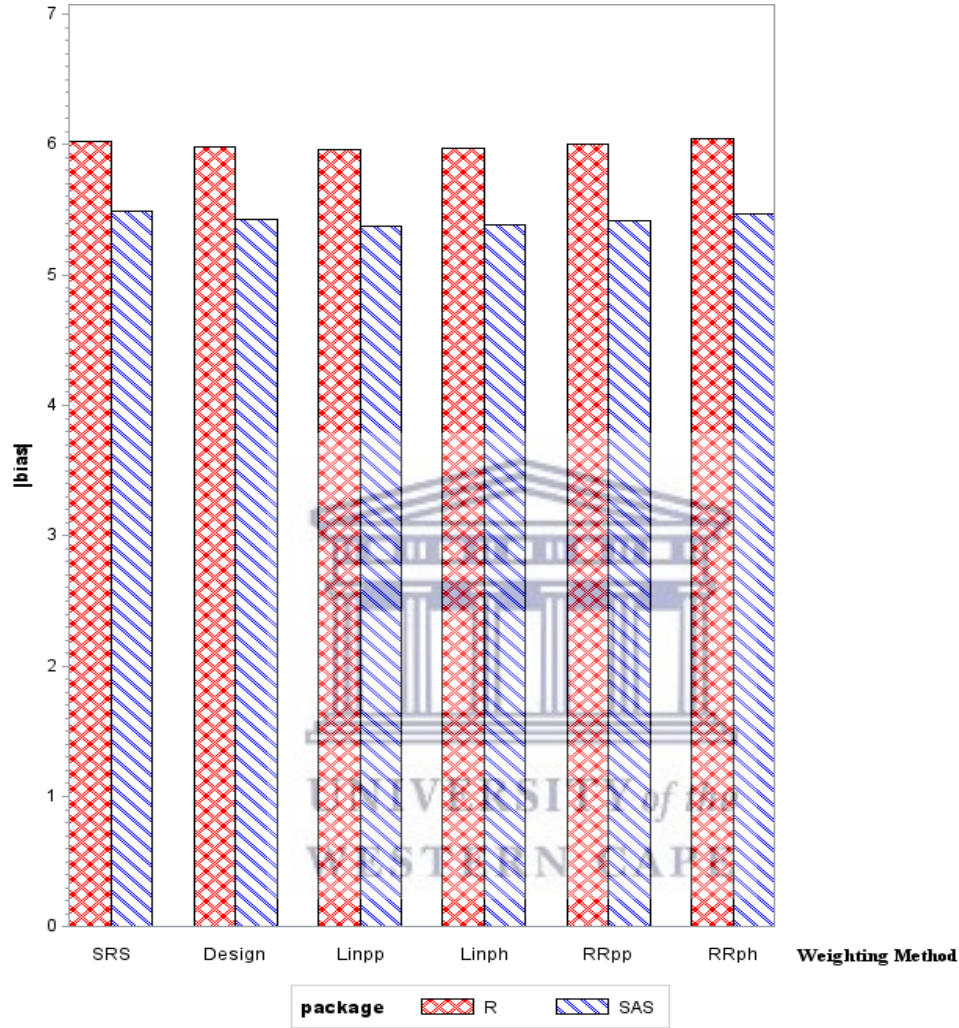


Figure 8: The absolute bias of the estimator of β_{11} under SRS and different weighting methods are shown for SAS and R.

Similar to Figure 7, the output from SAS and R in Figure 8, differed as a result of quasi-separation of data points. The SAS output absolute bias was slightly smaller than output

obtained from R. The estimators based on Lin_{pp} and Lin_{ph} showed smaller absolute bias for both SAS and R.

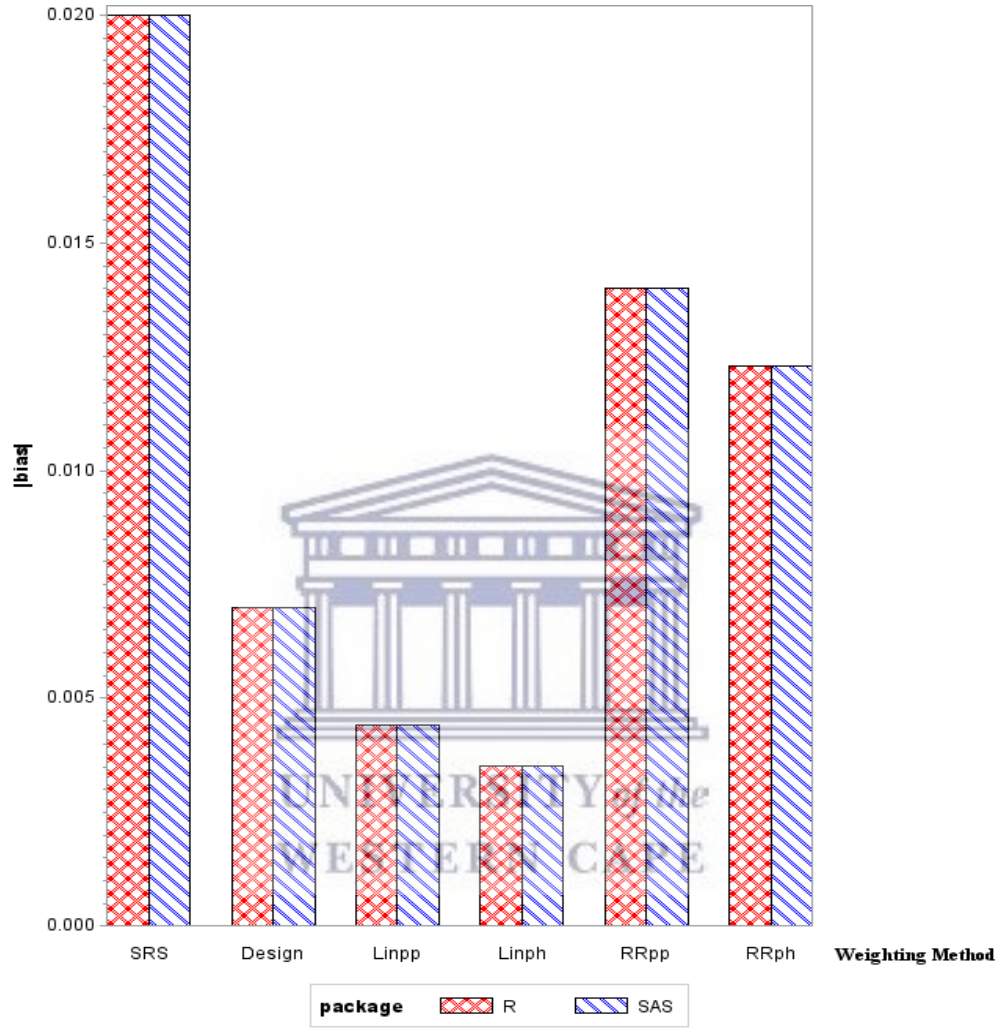


Figure 9: The absolute bias of the estimator of β_{12} under SRS and different weighting methods are shown for SAS and R.

In Figure 9, the estimates obtained under CS achieved absolute biases across the different weighting methods that were lower than those obtained under the assumption of SRS. The

weights Design, Lin_{pp} and Lin_{ph} showed the lowest absolute bias. The output from SAS and R were the same.

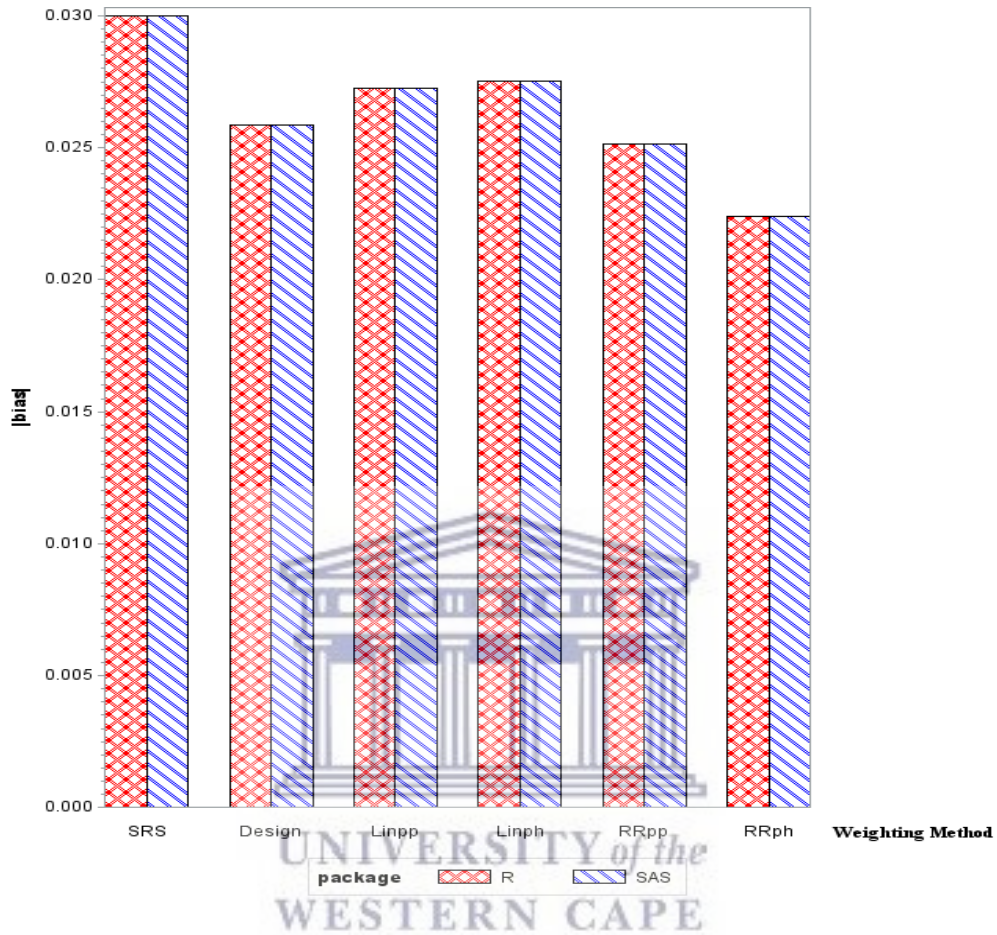


Figure 10: The absolute bias of the estimator of β_{20} under SRS and different weighting methods are shown for SAS and R.

Similar to Figure 9, the absolute bias in Figure 10 from CS was the lowest across weighting methods. The results for SRS had the greatest bias, biases of estimates obtained from weighting methods RR_{pp} and R_{ph} were the lowest.

5.2.1.2 The mean squared error

The MSE was discussed in Section 4.6.1 and will be another measure used to assess how close the selected estimators are to the “truth”. It is comprised of the bias and variance of an

estimator and therefore the results shown in Figure 4 to Figure 10 will comprise some part of the MSE results. The MSE for estimators $\hat{\beta}_0$, $\hat{\beta}_2$, $\hat{\beta}_4$, $\hat{\beta}_5$, $\hat{\beta}_{11}$, $\hat{\beta}_{12}$ and $\hat{\beta}_{20}$ are displayed in Figure 11 to Figure 17 and discussed. The remaining parameters can be found in Appendix B1 to B14.

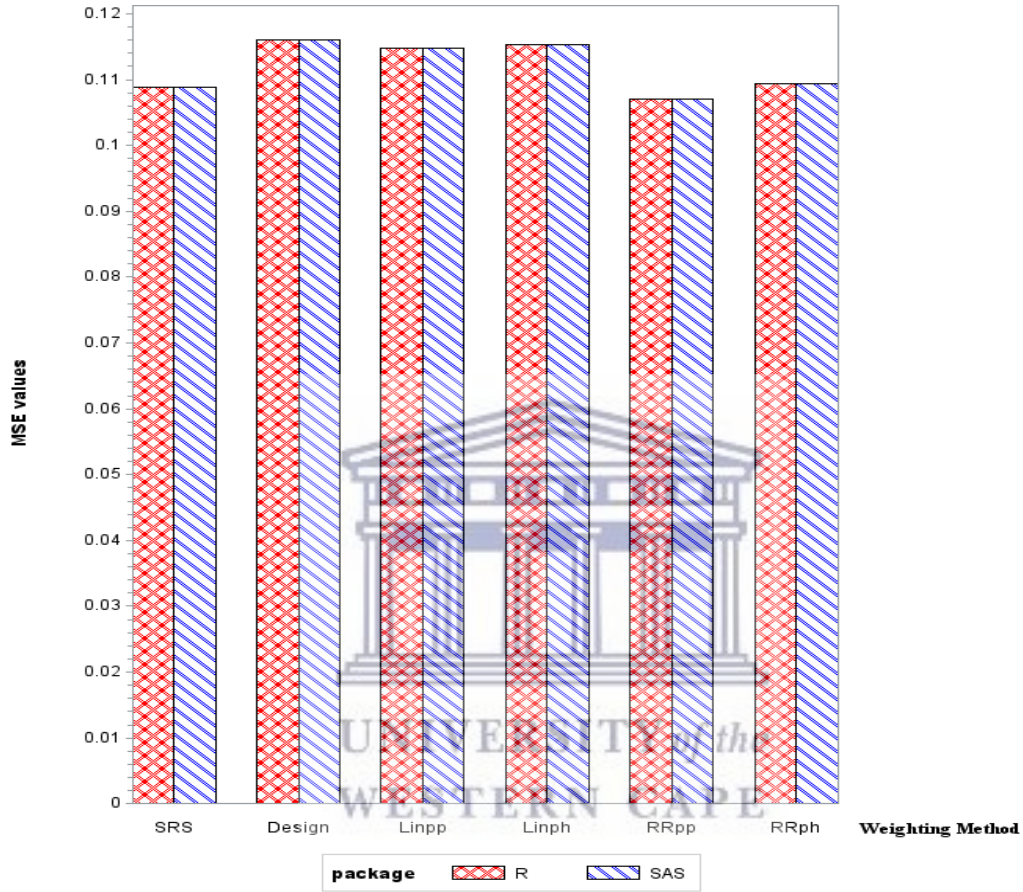


Figure 11: The MSE of the estimator of β_0 under SRS (no weight) and different weighting methods are shown for SAS and R.

In Figure 11, the output for the MSE from SAS and R were the same for the estimator of β_0 . This is consistent with Figure 4. The MSE for estimates obtained from the CS design using the weight RR_{pp} was the lowest. This is closely followed by SRS. These two methods displayed larger absolute bias. This implies that the variance produced by these methods were

the least. Similarly, estimates obtained from the design weight produced the largest MSE with the smallest absolute bias which implies that its variance was large.

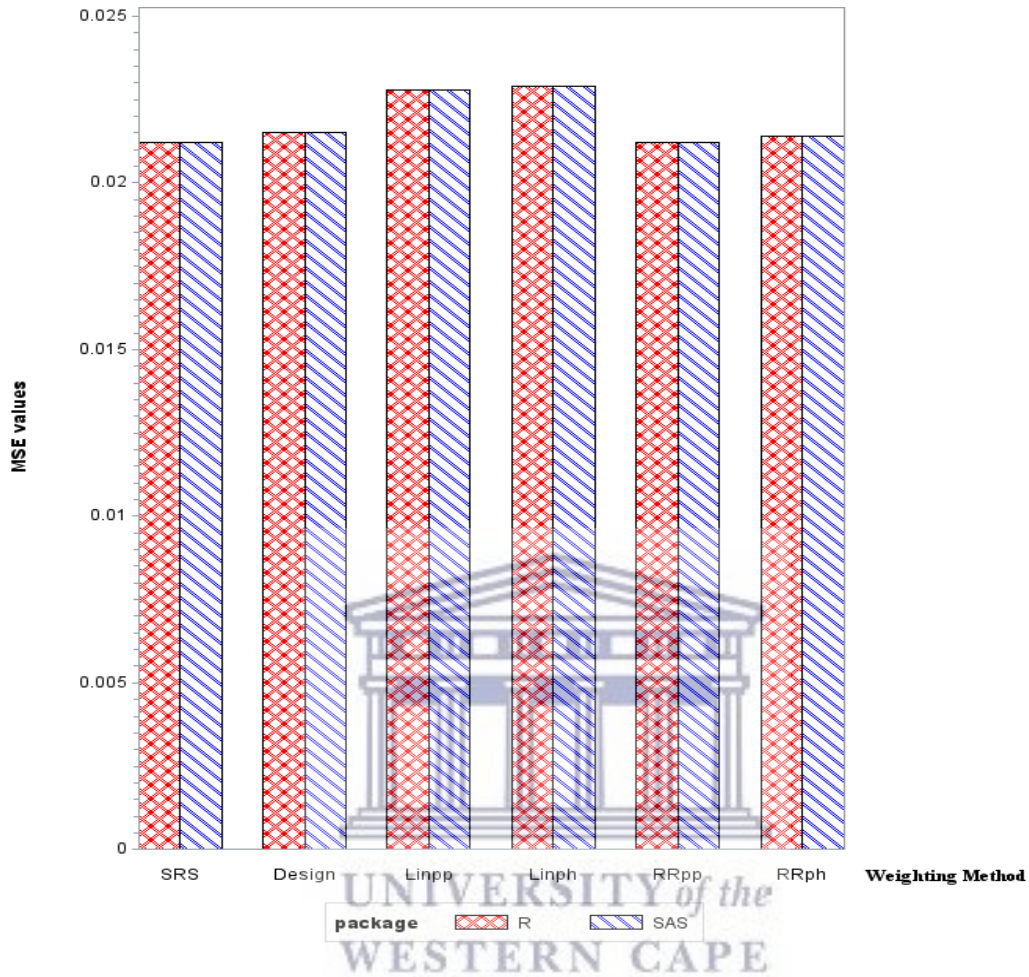


Figure 12: The MSE of the estimator of β_2 under SRS and different weighting methods are shown for SAS and R.

Figure 12, shows similar trends to that of Figure 11, namely that the MSE obtained from RR_{pp} and RR_{ph} were the lowest and the MSE obtained from the Design, Lin_{pp} and Lin_{ph} were the largest. This implies that methods RR_{pp} and RR_{ph} produced low variance. The results for SAS and R were the same.

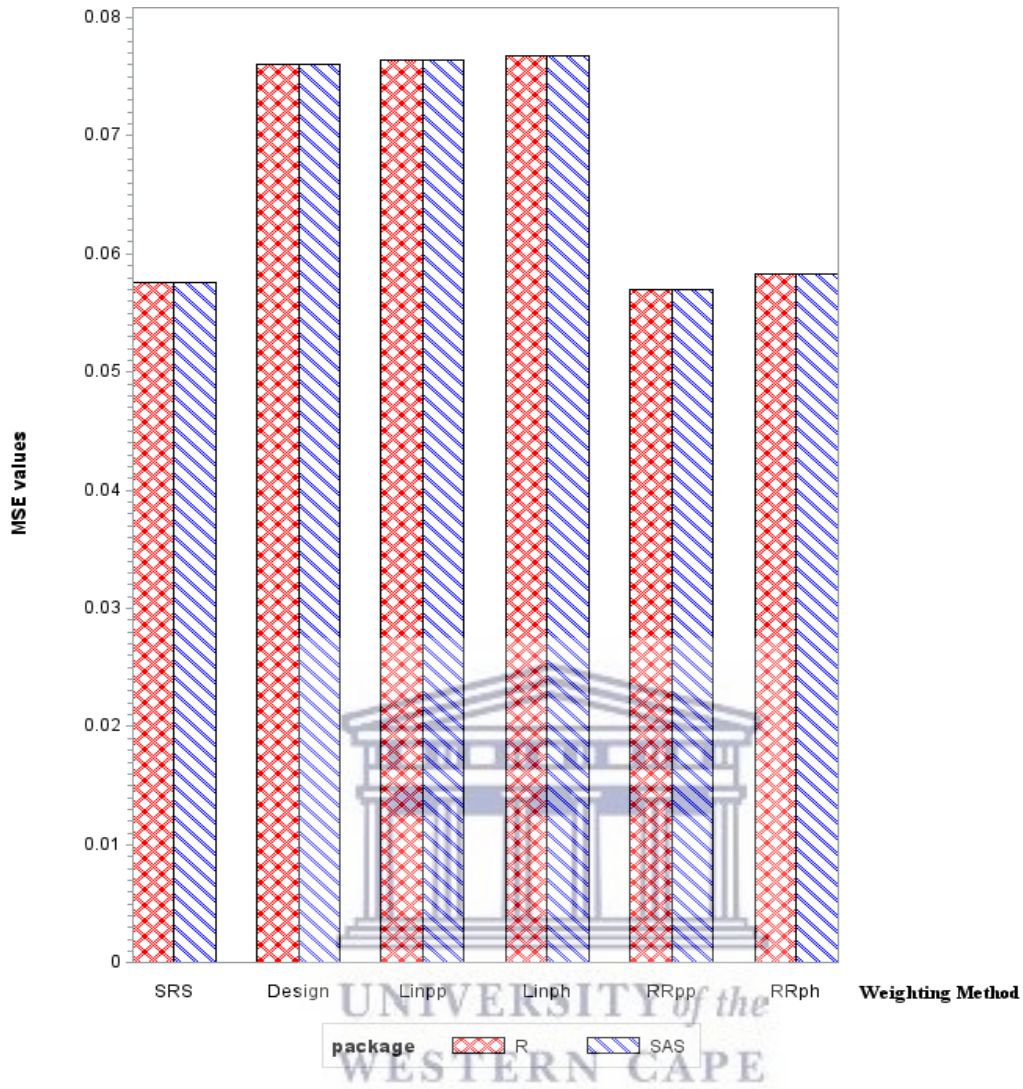


Figure 13: The MSE of the estimator of β_4 under SRS and different weighting methods are shown for SAS and R.

In Figure 6, the absolute bias displayed for Design, Lin_{pp} and Lin_{ph} were the largest. In Figure 13 the MSE based on those methods were also the largest, this implies that these methods produce large variances and large biases. In contrast SRS, RR_{pp} and RR_{ph} produced small absolute bias and lower MSE, with RR_{pp} having the smallest MSE.

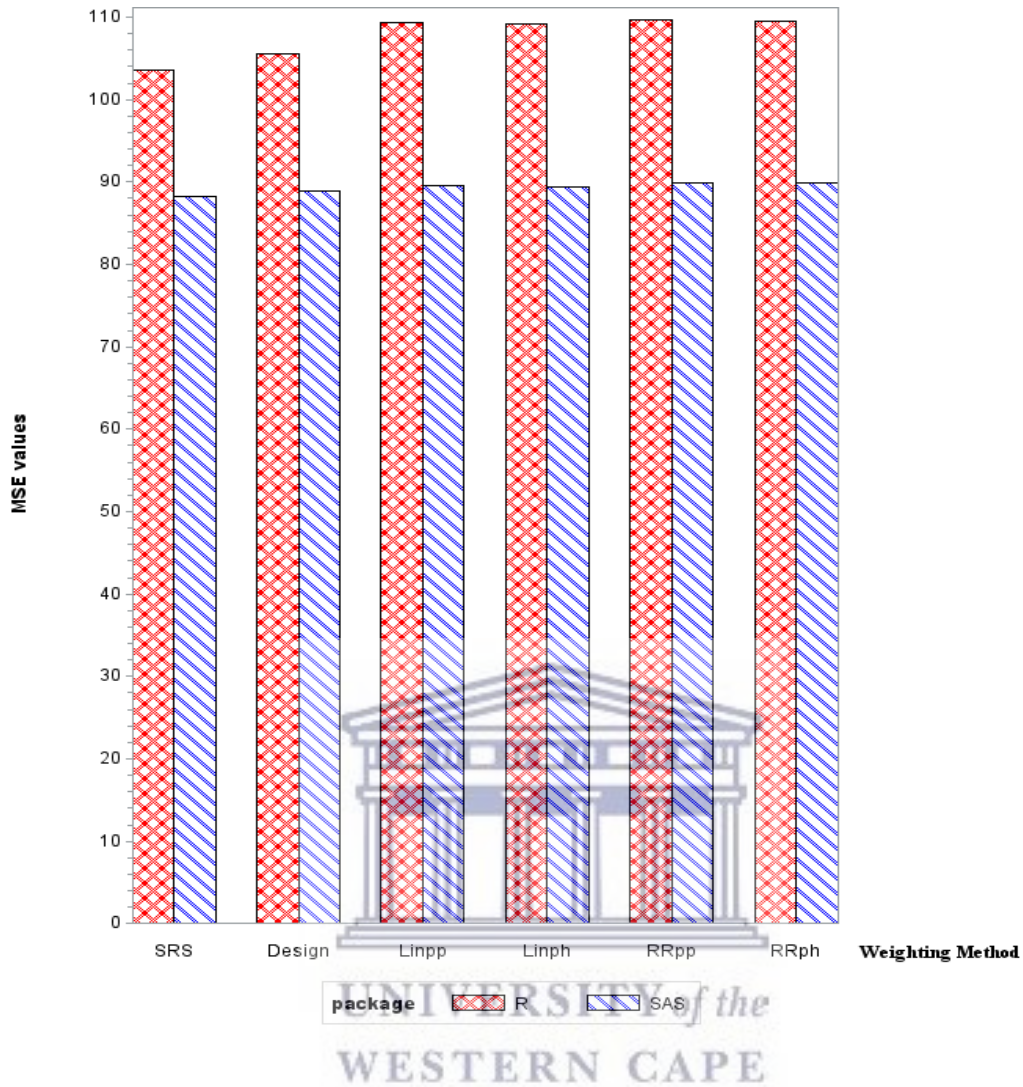


Figure 14: The MSE of the estimator of β_5 under SRS and different weighting methods are shown for SAS and R.

In Figure 14, the output from SAS and R differed which is consistent with Figure 7. The MSE values produced by SAS and R are very different in comparison to the others already discussed. The output obtained from SAS produced lower MSE values across the weights in comparison to R. The estimates obtained from SRS produced lower MSE values from both SAS and R. Once more, these differences were attributed to the quasi-separation of data points, due to small frequencies observed in the explanatory variable R_2 .

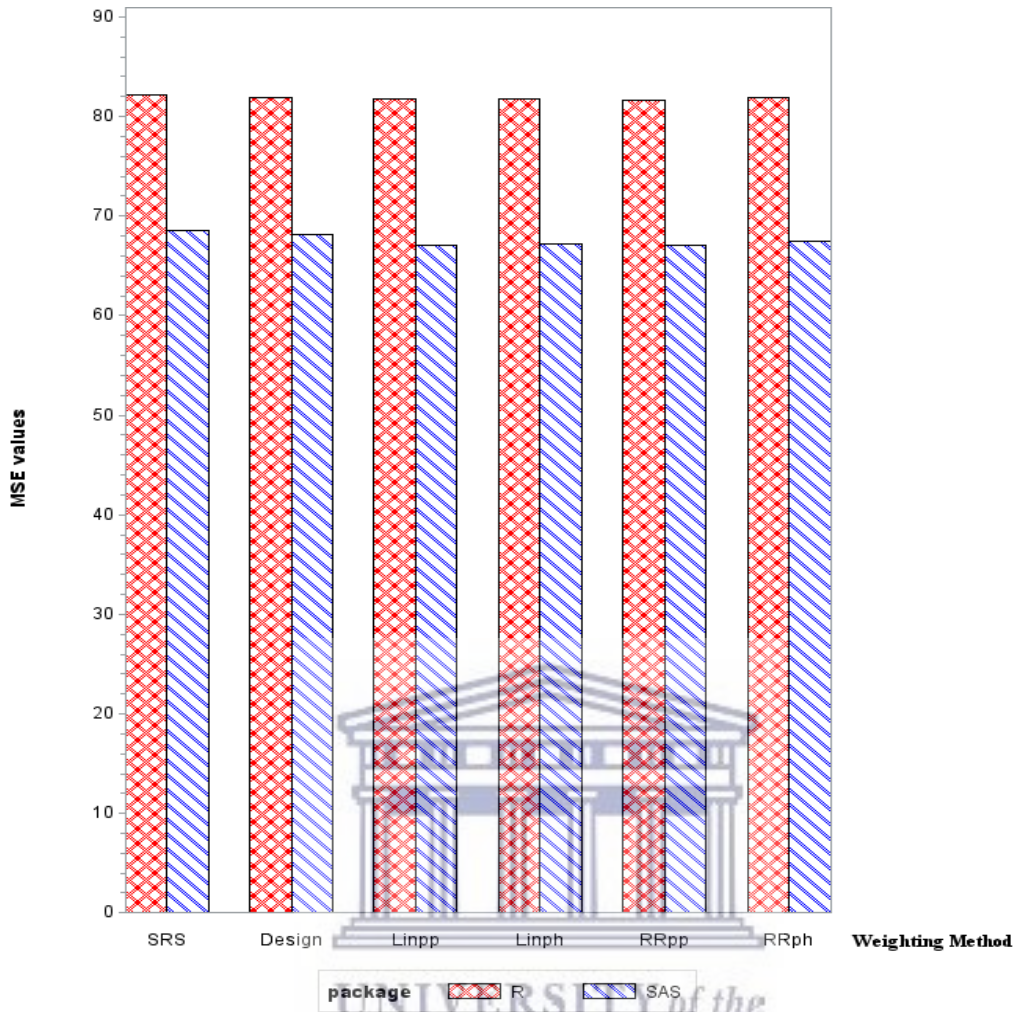


Figure 15: The MSE of the estimator of β_{11} under SRS and different weighting methods are shown for SAS and R.

The SAS and R output differed in Figure 15, for the β_{11} estimator. Once more the SAS output produced lower values across the methods in comparison to those obtained from R. As noted in Figure 8, quasi-separation was present in the variable.

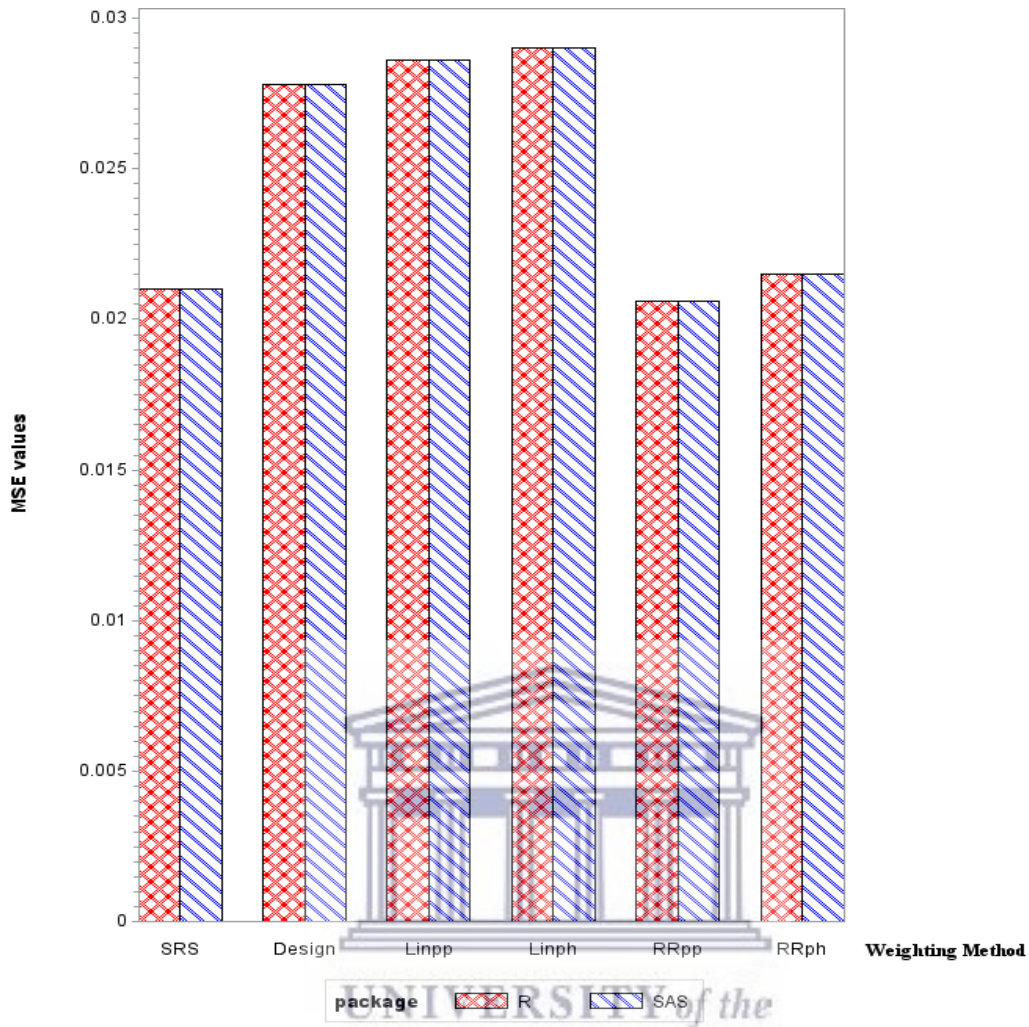


Figure 16: The MSE of the estimator of β_{12} under SRS and different weighting methods are shown for SAS and R.

In Figure 16, the MSE was larger under Design, Lin_{pp} and Lin_{ph} as opposed SRS and RR_{pp} . In Figure 9, the absolute bias under SRS soared which implies that the variance produced for estimates obtained from SRS was very small.

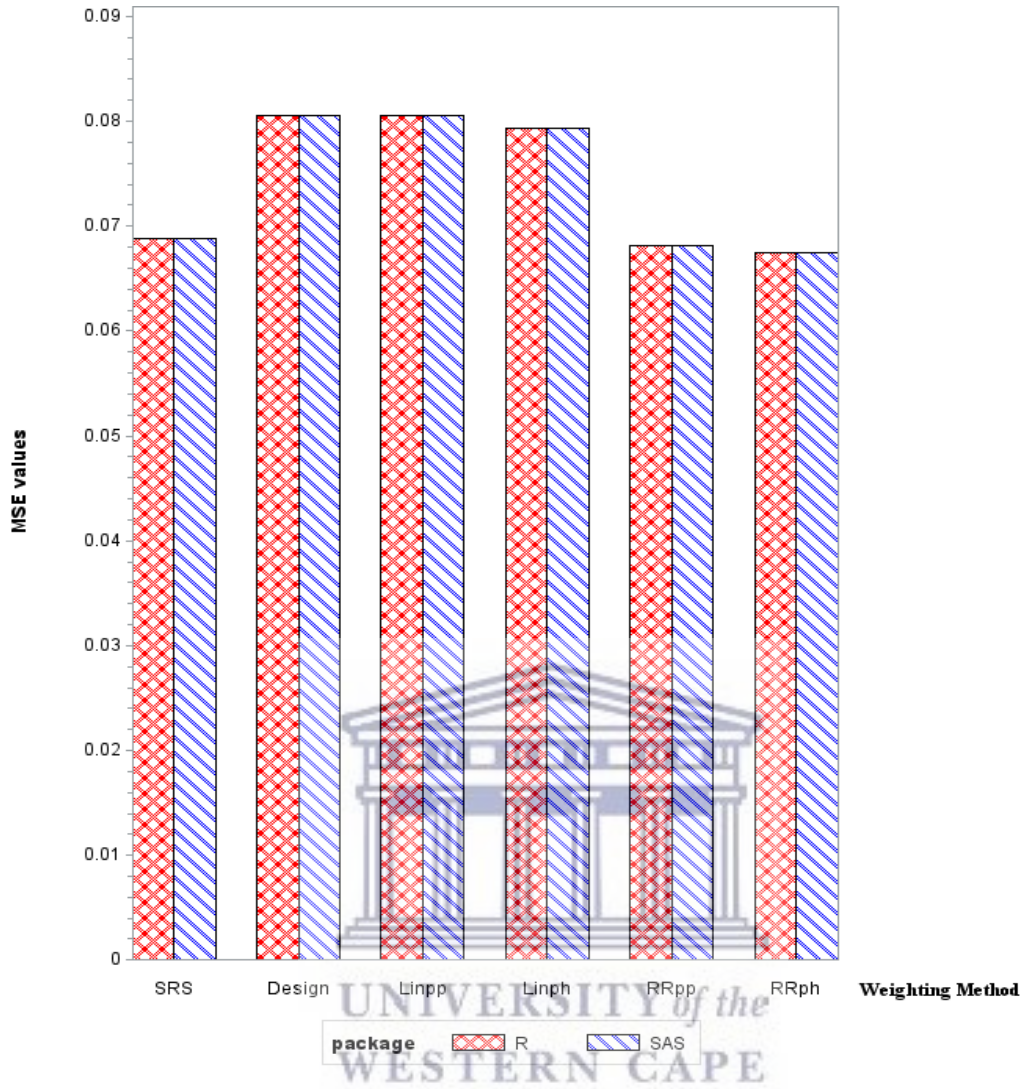


Figure 17: The MSE of the estimator of β_{20} under SRS and different weighting methods are shown for SAS and R.

Figure 17, shows similar trends to those of Figure 16, with RR_{pp} , RR_{ph} and SRS producing small MSE values as opposed to the other three methods. Similarly, these methods produced larger absolute bias values implying that lower variances were observed.

5.2.2 Confidence intervals for model parameters

This section reviews the results obtained for the 95% confidence intervals calculated for the model parameters along with their measures of accuracy. The standard error provides a bound for the model parameter. The coverage probability as defined in Section 4.6.2 provides an indication of the precision of the standard error, and if the model parameter is contained in it. Furthermore, the standard error affects the length of the confidence interval: larger standard errors result in greater lengths and provide less precision. Therefore, the confidence interval lengths are of importance. Three standard (asymptotic) confidence intervals based on TSL, JRR and bootstrap estimated variances, and an additional non-parametric interval, the bootstrap percentile, are obtained. The coverage probability and confidence interval length were subsequently obtained and the output is displayed and discussed.

5.2.2.1 Coverage probability

The standard (asymptotic) confidence interval was formulated in Section 3.5.1 and the results obtained will be displayed and discussed. Three methods were used to obtain the standard errors, viz. TSL, JRR and the bootstrap, that will be used in the calculation of the standard interval. Furthermore, the bootstrap percentile interval is a non-parametric confidence interval obtained from taking the percentiles of the estimates obtained from the bootstrap samples. The bootstrap samples were simulated from the samples discussed in Section 4.5.2. These samples formed the basis from which the estimates were calculated. The coverage probabilities were obtained for both SAS and R and the output is displayed for SRS and CS using the weights Design, Lin_{pp} , Lin_{ph} , RR_{pp} and RR_{ph} . The probability values range from 0 to 1 with the ideal probability being 0.95; the level of significance. A selection of results is displayed in Figure 18 to Figure 20. The remainder is included in Appendix C1 to C19. The top left panel displays the coverage probability for the standard interval based on the TSL estimated variance, the top right panel displays the coverage probability for the standard interval based on the JRR estimated variance, the bottom left panel shows the coverage

probability for the standard interval based on the bootstrap estimated variance and the bottom right panel shows the coverage probability for the bootstrap percentile interval.

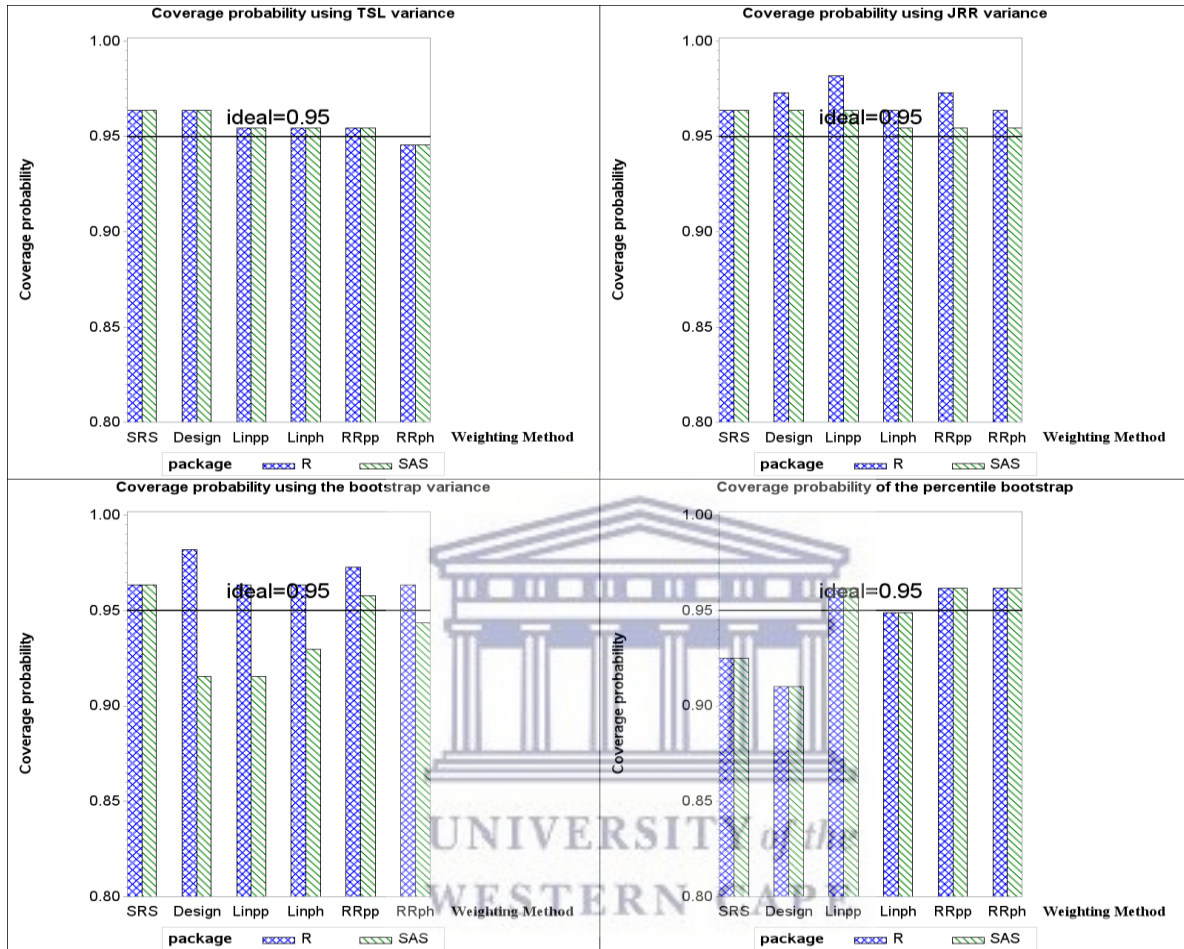


Figure 18: The coverage probabilities for β_0 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

In Figure 18, the R and SAS output for the coverage probability values for the β_0 parameter were the same as when TSL variance estimation was used. The coverage probability of the intervals obtained using the design weight and SRS were the furthest from the level of

significance of 0.95. The coverage probabilities of the remaining intervals were all equidistant from the level of significance.

Figure 18, also shows a difference between the SAS and R outputs for the coverage probabilities of the interval for β_0 in which the JRR estimated variance was used. The differences were attributed to quasi-separation of data points in some of the replicates. The coverage probability for the results from R deviated more from the level of significance as opposed to the SAS results. The results for the estimates obtained from the weights Lin_{ph} , RR_{pp} and RR_{ph} for the SAS output were the closest to the level of significance.

As noted in Section 3.4.2.2.2 the bootstrap variance may differ, therefore the results obtained from SAS and R are different. As shown in Figure 18, R produced larger coverage probabilities than SAS. The weights RR_{pp} and RR_{ph} had better coverage for both SAS and R.

The bootstrap percentile interval's coverage probabilities on the other hand were the same for SAS and R. This is due to the same bootstrap samples being used for both software programs. The coverage probabilities for Lin_{ph} were the closest to the level of significance. This in comparison to the Design weight that deviated the furthest from the level of significance.

Overall, when the TSL estimated variance was used, the results showed more stable coverage probabilities and less deviation from the level of significance across methods was observed. As opposed to the bootstrap that showed greater fluctuations across methods. In terms of the variances produced from the different weighting methods, Lin_{ph} , RR_{pp} and RR_{ph} showed less deviation from the level of significance across methods.

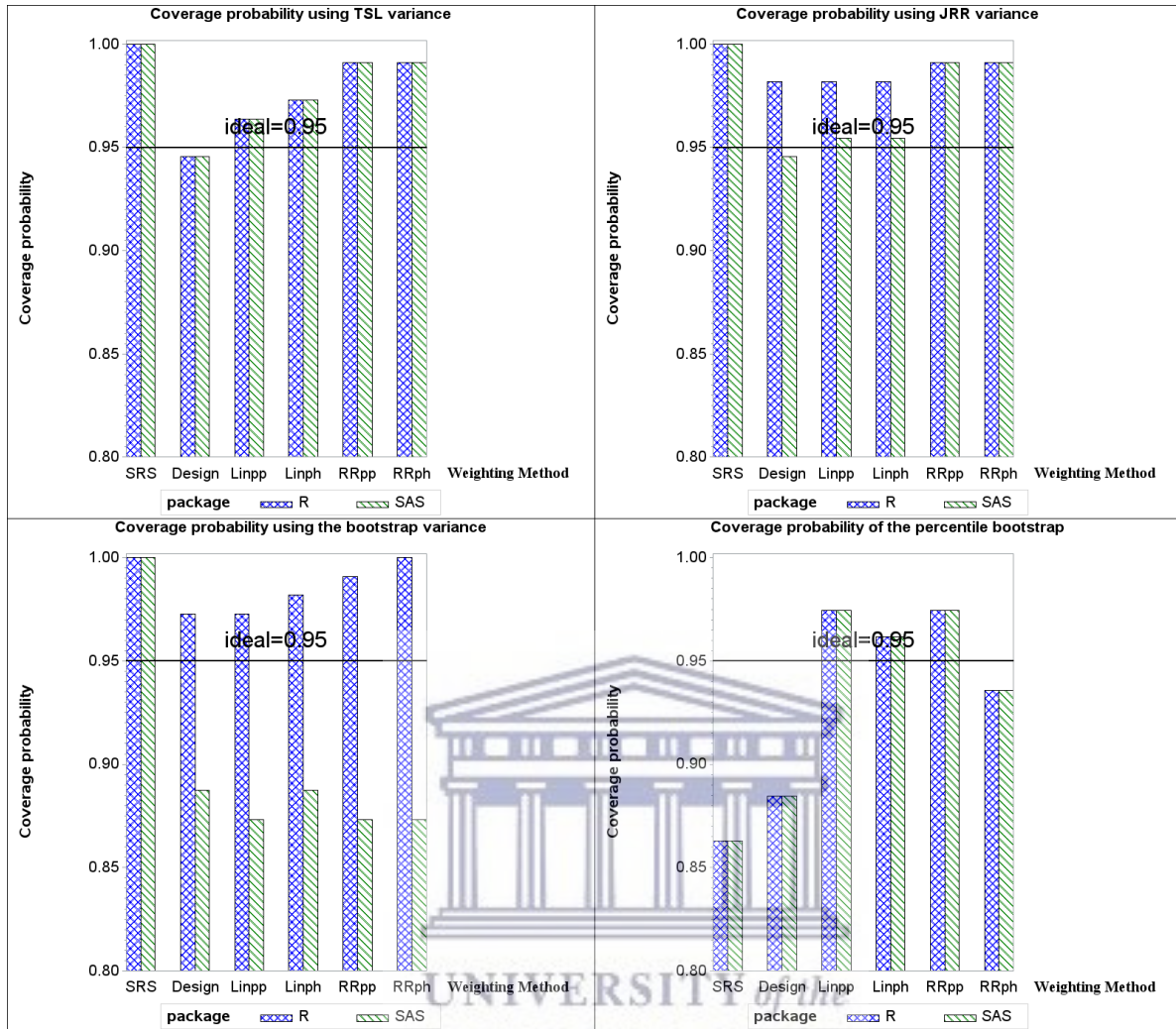


Figure 19: The coverage probabilities for β_4 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

In Figure 19, the results for SAS and R for the coverage probabilities using TSL were the same and better coverage were produced using the Design weight in comparison to coverage probabilities produced for SRS for the parameter β_4 .

The results shown for JRR variance estimation for the Design weight, Lin_{pp} and Lin_{ph} differed for SAS and R. Better coverage probabilities were shown for R across different methods with the Design and Lin_{pp} weights being the closest to the level of significance.

When the bootstrap variance estimation was used the SAS output's coverage probabilities were smaller than when output was obtained from R. R also produced better coverage as opposed to that of SAS, with the Design weight producing the best coverage.

The bootstrap percentile interval's coverage probabilities showed that the coverage probabilities for SRS deviated the furthest from the level of significance. Lin_{ph} and RR_{ph} coverage probabilities were the closest to the level of significance.

In general, the coverage probabilities for β_4 showed contrasting results, once more when the TSL variance estimator was used, better coverage was observed across weighting methods.



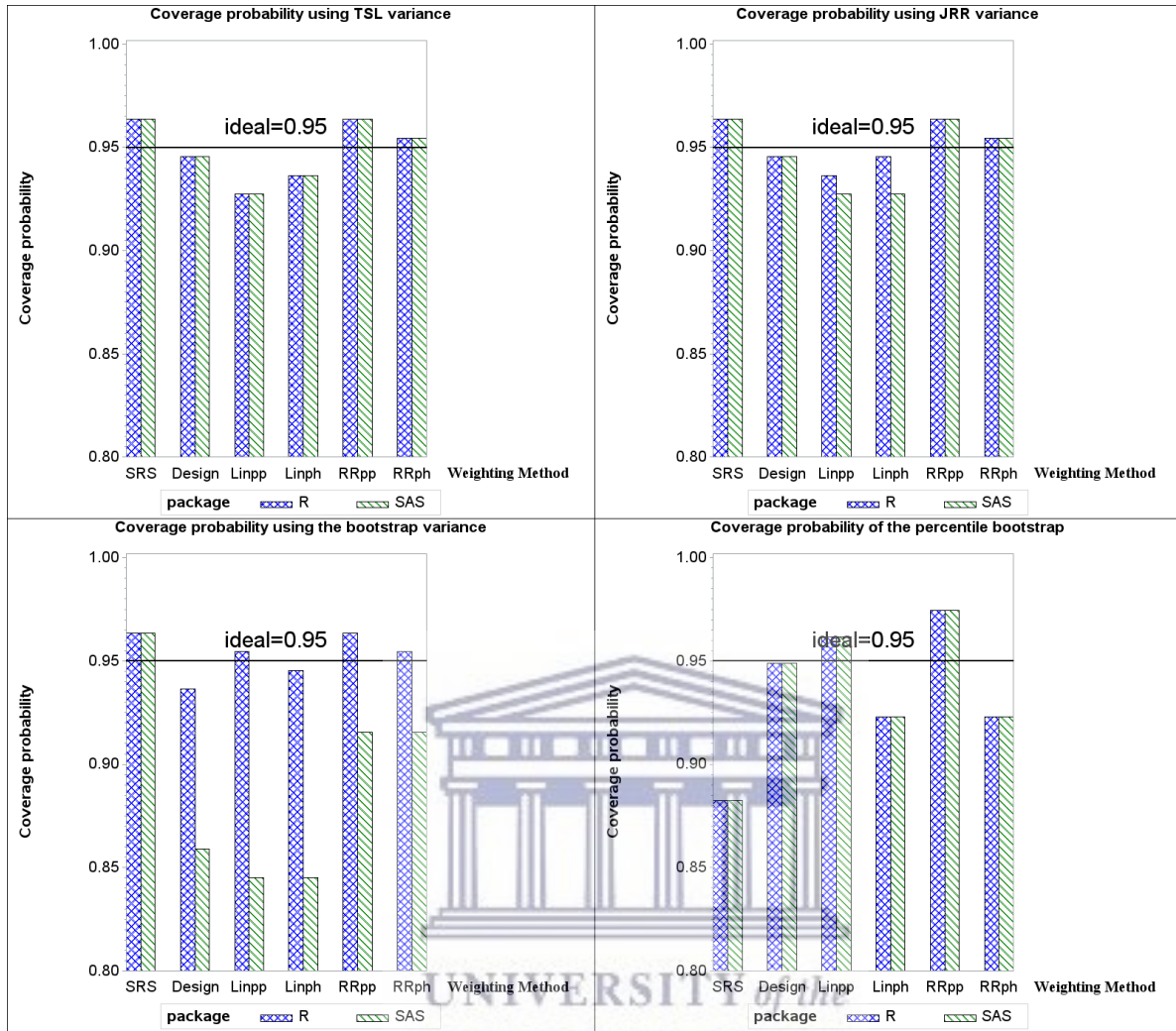


Figure 20: The coverage probabilities for β_7 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

In Figure 20, the TSL estimated variance coverage probabilities for SAS and R were once again the same. The weighting methods Design and RR_{ph} deviated the least from the level of significance with Lin_{pp} deviating the furthest from the level of significance.

The JRR estimated variance coverage probabilities produced slight differences for SAS and R for Lin_{pp} and Lin_{ph} . The coverage probabilities for Design and RR_{ph} deviated the smallest

from the level of significance, closely followed by Lin_{ph} obtained from R. The SAS output displayed for Lin_{pp} and Lin_{ph} deviated the furthest from the level of significance.

Figure 20, also shows contrasting results for SAS and R for the bootstrap estimated variance coverage probabilities. The output obtained from R performed slightly better than that obtained from SAS. The weight RR_{ph} obtained from R was the closest to the level of significance.

The bootstrap percentile interval's coverage probabilities were the same for SAS and R. The weight Design deviated the least from the level of significance.

Overall, for parameter β_7 the weights Design and RR_{ph} deviated the least from the level of significance with output obtained from R generally doing better.

5.2.2.2 Confidence interval length

The confidence interval length was discussed in Section 4.6.2, as noted, improved coverage could be due to large confidence intervals as a result of large variances. Therefore, the confidence interval length provides a scope to validate the coverage probabilities displayed in Figure 18 to Figure 20. The confidence interval length was calculated in both SAS and R. In addition, results were obtained under SRS and CS using the weights Design, Lin_{pp} , Lin_{ph} , RR_{pp} and RR_{ph} . A selection of the results is displayed in Figure 21 to Figure 23. The remainder are in Appendix D1 to D19. Once more the top left panel contains the confidence interval length for the TSL estimated variance, top right for the JRR estimated variance, bottom left the bootstrap estimated variance and bottom right the confidence interval length for the bootstrap percentile interval.

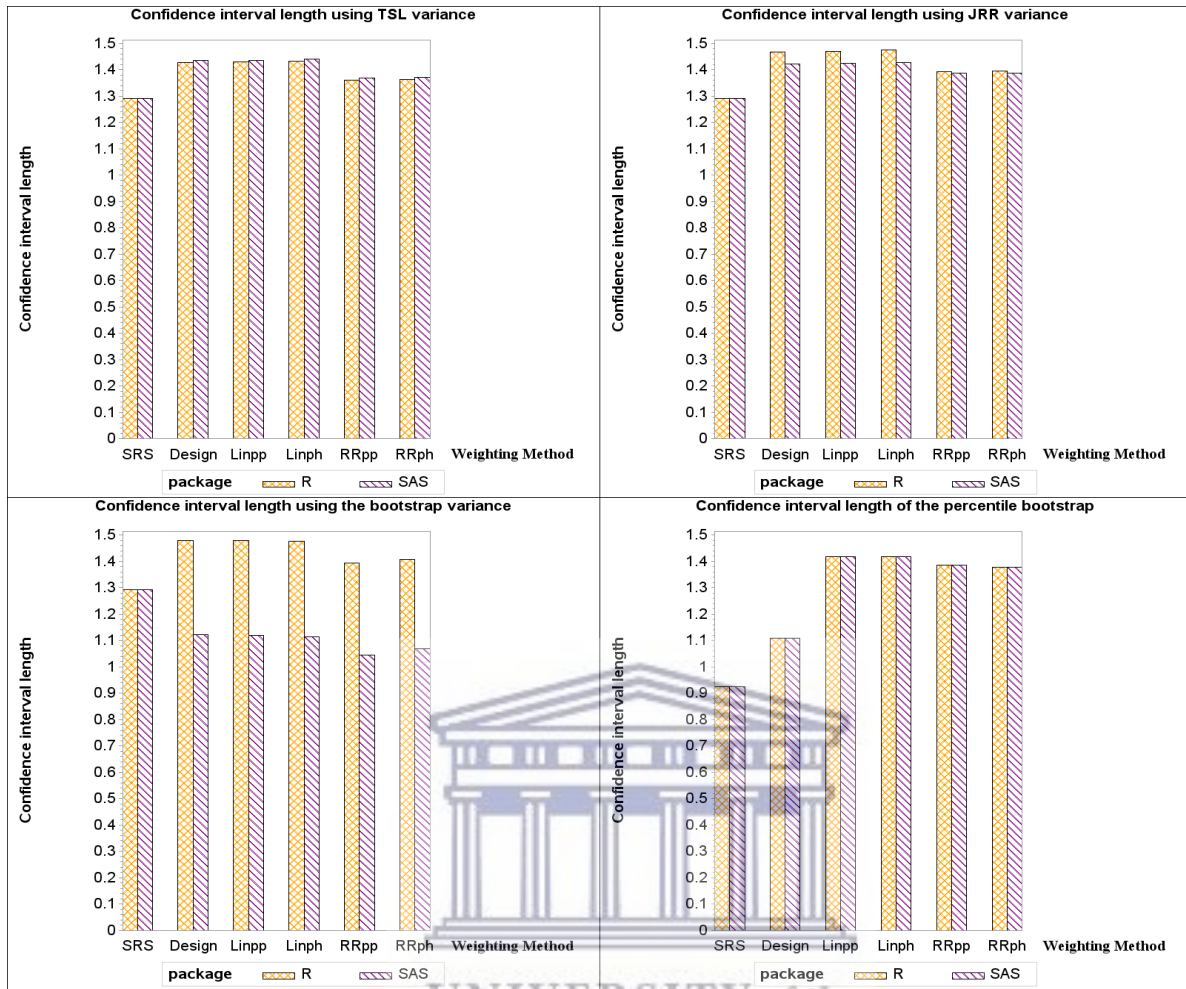


Figure 21: The confidence interval lengths for β_0 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

In Figure 21, the lengths based on the TSL estimated variance do not differ greatly amongst the different methods, with RR_{ph} having smaller lengths. This reinforces the coverage probabilities displayed for RR_{ph} in Figure 18, which were generally closer to the level of significance for β_0 . The values obtained from SAS and R differed slightly which is attributed to decimal differences.

The results displayed for the JRR estimated variance in Figure 21, shows that the confidence interval lengths differed for different weighting methods. Greater lengths were produced for confidence intervals obtained in R as opposed to those obtained in SAS.

The confidence interval lengths for the bootstrap showed results for R were greater than that of SAS. The results obtained from SAS were also lower across weighting methods, indicating more precise standard errors for the bootstrap confidence intervals using SAS.

Lastly, the confidence interval length for the bootstrap percentile interval shows that the confidence interval length for SRS and the design weight were the lowest. In comparison to the other methods which showed lengths very similar to each other. Overall, the confidence interval lengths across variance estimation methods and the bootstrap percentile interval do not differ greatly with lower confidence interval lengths produced under SRS, RR_{pp} and RR_{ph} .



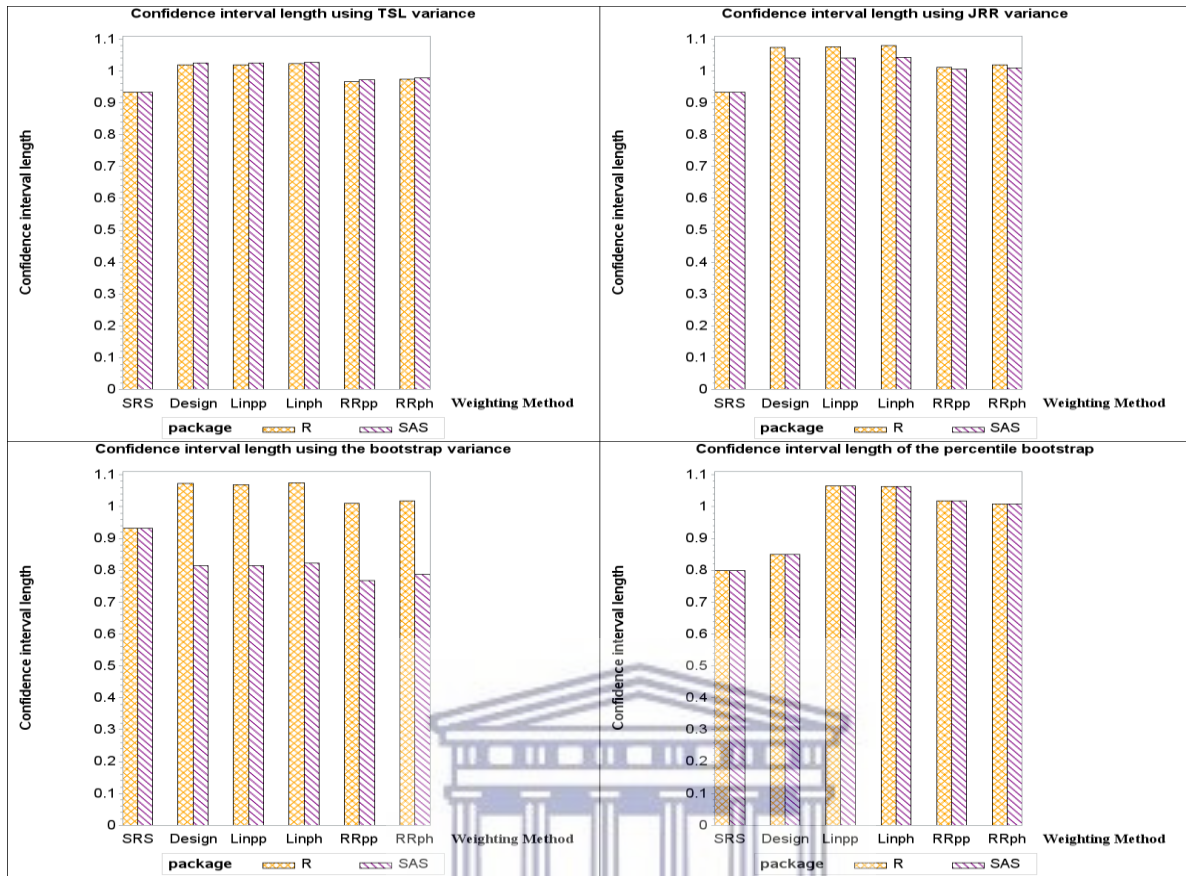


Figure 22: The confidence interval lengths for β_4 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

In Figure 22, shorter confidence interval lengths were displayed using the TSL estimated variance for SRS for β_4 , the other lengths do not differ greatly with slightly smaller lengths produced by RR_{pp} and RR_{ph} .

For the JRR estimated variance larger confidence interval lengths were produced for the β_4 parameter using weights Design, Lin_{pp} and Lin_{ph} . Similar to Figure 19, SAS results produced smaller confidence interval lengths than that of R.

When the bootstrap estimated variance was used better confidence interval lengths were produced in SAS as opposed to R. Lower confidence interval lengths were observed for weights RR_{pp} and RR_{ph} in SAS.

In Figure 22, the bootstrap percentile interval lengths were the lowest for SRS, similar to that observed under SRS for β_0 . The largest confidence interval lengths were produced by Lin_{pp} and Lin_{ph} .

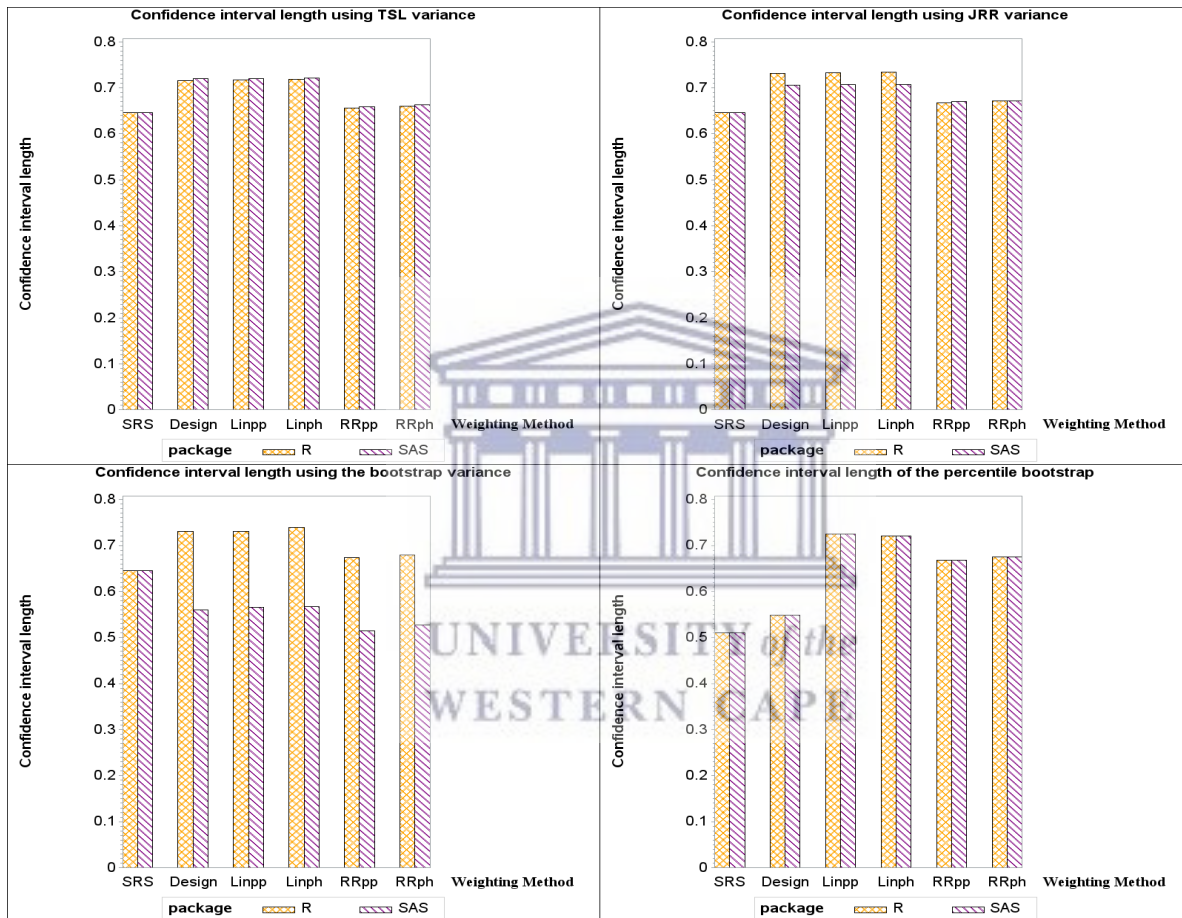


Figure 23: The confidence interval lengths for β_7 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

In Figure 23, the confidence interval lengths do not differ greatly for the TSL estimated variance. As in Figure 22, the confidence interval lengths for RR_{pp} , RR_{ph} and SRS were slightly smaller in comparison to the other weighting methods.

The JRR estimated variance confidence interval lengths showed similar results to that of the TSL estimated variance with RR_{pp} , RR_{ph} and SRS having slightly lower lengths.

The bootstrap estimated variance confidence interval lengths showed that the lengths obtained from R were slightly larger than those obtained from SAS. Lower confidence interval lengths were observed for RR_{pp} and RR_{ph} obtained from SAS.

Lastly the bootstrap percentile confidence once more showed lower lengths for SRS in comparison to the other weighting methods.

5.3 Conclusion

The logistic regression model was discussed in Chapter 3 and in that chapter the standard logistic regression was outlined for *i.i.d.* data and adapted for CS data. It was noted that when the data comes from a CS design in particular, the sampling weights need to be incorporated in the model. This is the “golden standard”. The main purpose of this chapter was to provide empirical results for the study outlined in Chapter 4. The study provides a focal point, using real data (IES 2005/2006) to provide results to aid in answering the research questions and problem statement outlined in Chapter 1. In Chapter 4 methods were discussed to ascertain how close estimators are to the parameters of interest, in particular the MSE and bias. Estimates were obtained when the design was ignored i.e. SRS, and when correctly accounted for. In addition, when the design was correctly accounted for, i.e. CS, five sampling weights were used, viz. Design, Lin_{pp} , Lin_{ph} , RR_{pp} and RR_{ph} . In general, the absolute bias was smaller for CS as opposed to SRS particularly when using the design weight, Lin_{pp} and Lin_{ph} , which showed the smallest bias. The MSE showed mixed results, however, SRS generally showed larger MSE results than sampling weights RR_{pp} and RR_{ph} . The two statistical packages generally showed the same results. In cases where they do differ it was as a result of quasi-separation of data points, where existence of the MLE are questionable.

The variance provides a measure of accuracy for an estimator of a parameter of interest and is used to construct confidence intervals. Three variances were discussed in this thesis, viz. TSL, JRR and the bootstrap. These were used to construct standard (asymptotic) confidence intervals. In addition, a non-parametric confidence interval, i.e. the bootstrap percentile interval, was obtained. In Section 4.6.2, the coverage probability was defined to indicate the proportion of times a parameter is contained in a confidence interval. Also, in Section 4.6.2, the confidence interval length was discussed to ascertain whether good coverage is not due to a larger confidence interval length. Generally, TSL provided better coverage probabilities as opposed to the other three variance methods and smaller lengths. The coverage probabilities for SRS compared to CS showed mixed results, using weighting generally provided better coverage for the different variance methods. Also, when weighting was used the confidence interval lengths were smaller. Generally, RR_{pp} and RR_{ph} provided better coverage and lengths than the other weighting methods. Similarly, the bootstrap percentile provided mixed results with RR_{pp} and RR_{ph} giving better coverage. However, the confidence interval length was better under SRS as opposed to the other methods. Once more, when variables contained low frequencies, the results differed for SAS and R.

Results obtained are generally consistent with literature, with the effects of not correctly accounting for the design apparent. It should be stressed that when SRS appears to perform better than CS, it is an indication of how the results can be presented incorrectly and should not be a basis for ignoring the design. The calibration and integrated weights using the raking ratio distance method presented better overall estimators and their variances provided better coverage and lengths. Furthermore, TSL presented better precision.

Chapter 6: Conclusion and further research

6.1 Introduction

The main objectives of the thesis were outlined in Chapter 1 in which the research questions were presented. The objectives were firstly, to compare results obtained when ignoring the design, i.e. SRS, and correctly accounting for the design. Secondly, to establish which sampling weights provide better estimators when correctly accounting for the design. The sampling weights used were design, Lin_{pp} , Lin_{ph} , RR_{pp} and RR_{ph} . These weights were incorporated in the logistic regression model and estimators were obtained. Also, the thesis aimed to compare CS variances of which three were discussed, viz. TSL, JRR and the bootstrap. These were used to obtain standard (asymptotic) confidence intervals, and comparisons were made between the variances and the standard logistic regression variance i.e. the variance obtained when the design is ignored. In addition, the bootstrap percentile confidence interval was obtained, a non-parametric confidence interval, and results were compared amongst weighting methods. Literature provided a basis for the findings and noted that when the design is ignored, estimates obtained can be incorrect.

6.2 Findings

The surrogate population discussed in Section 4.5.1 was used to obtain the “truth”, and samples were drawn from the surrogate population using a CS design. The estimators obtained from both SRS and CS were compared to the “truth”. The absolute bias was one such method used to compare estimators. In terms of absolute bias, generally, the weighting methods performed better than when no weighting was used. The design weight, Lin_{pp} and Lin_{ph} showed smaller absolute bias. SRS (no weighting) generally showed larger absolute bias. Another measure used to compare estimators was the MSE, which comprises of the bias and the variance. The MSE showed mixed results with RR_{pp} and RR_{ph} having the smallest MSE in general. Since RR_{pp} and RR_{ph} tend to have larger bias in comparison to the design weight it is reasonable to conclude that the variability amongst estimates obtained using RR_{pp}

and RR_{ph} are smaller in comparison to the other methods. Three asymptotic confidence intervals were obtained and the bootstrap percentile confidence interval. Generally, the intervals based on TSL estimated variance produced better coverage and smaller confidence interval lengths in comparison to the other methods. Once more, RR_{pp} and RR_{ph} in general showed better coverage and smaller lengths across variance methods. Similarly, RR_{pp} and RR_{ph} bootstrap percentile confidence intervals showed better coverage probabilities. In terms of the confidence interval length for the bootstrap percentile confidence interval, SRS showed a slightly shorter length. The results for R and SAS were generally the same. However, if quasi-separation of data is present then the results differ. The R results for the bootstrap was generally better, however, results did not differ substantially. In the case when quasi-separation of data points were present SAS estimates out performed R. It should be noted that in the event where SRS appeared to perform better than CS, it should not be considered a basis to ignore the design but should be an indication of how incorrect results can be presented when the sample design is ignored.

6.5 Further research

The following areas for further research were identified from the results of this thesis:

1. How to remedy the quasi-separation of data points, particularly when the data comes from a CS design;
2. Certain weighting methods, in particular the design weight, produces very low absolute bias and large MSE values. Further research can be done concerning why that is the case;
3. Results can be replicated in other software, in particular SPSS, and see how the results compare to those obtained from SAS and R;
4. Model selection criteria can be incorporated and adjusted for CS designs for the logistic regression;
5. AIC and BIC can be assessed for logistic regression for CS; and
6. Multicollinearity can be assessed for the logistic regression for CS.

Researchers across different fields still remain uninformed about CS; even those within the field of statistics. This makes research and education regarding CS imperative and this will provide researchers with better tools to obtain better answers and conclusions for their research questions.



UNIVERSITY *of the*
WESTERN CAPE

References

1. Agresti, A., (2013). *Categorical data analysis*. 3rd ed. New Jersey: John Wiley & Sons, inc.
2. Archer, K. J., Lemeshow, S. & Hosmer, D. W., (2007). Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design. *Computational Statistics & Data Analysis*, 51(9), pp. 4450-4464.
3. Berger, T. G. & De La Riva Torres, O., (2016). Empirical likelihood confidence intervals for complex sampling designs. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 78(2), pp. 319-341.
4. Binder, D. A., (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical review*, Volume 51, pp. 279-292.
5. Chambless, C. E. & Boyle, K. E., (1985). Maximum likelihood methods for complex sample data: logistic regression and discrete proportional hazards models. *Communications in Statistics-Theory and Methods*, 14(6), pp. 1377-1392.
6. Cheung, P., (2005). *Household Sample Surveys in Developing and Transition Countries*. New York: United Nations.
7. Cochran, W. G., (1977). *Sampling Techniques*. 3rd ed. New York: John Wiley & Sons, inc.
8. Deville, J.-C., Sarndal, C.-E. & Sautory, O., (1993). Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*, 88(423), pp. 1013-1020.
9. Efron, B. & Tibshirani, R. J., (1994). *An introduction to Bootstrap*. New York: CRC press.
10. Elliot, A. C. & Woodward, W. A., (2010). *SAS Essentials mastering SAS for research*. 1st ed. San Francisco: John Wiley & Sons, Inc..
11. Heeringa, S. G., West, B. T. & Berglund, P. A., (2010). *Applied Survey Data Analysis*. New York: CRC Press.
12. Heinze, G. & Schemper, M., (2002). A solution to the problem of separation in logistic regression. *Statistics in medicine*, 21(16), pp. 2409-2419.

13. Institute, S., (2008). *SAS/STAT(R) 9.2 User's Guide, Second Edition*. North Carolina: SAS institute.
14. Kidder, L. H., Judd, C. M. & Smith, E. R., (1986). *Research methods in social relations*. 5th ed. s.l.:New York : CBS Pub. Japan Ltd., ©1986.
15. Kish, L. & Frankel, M. R., (1974). Inference from Complex samples. *Journal of the Royal Statistical Society*, 36(1), pp. 1-37.
16. Kolenikov, S., (2010). Resampling variance estimation for complex survey data. *The Stata Journal*, 10(2), p. 165–199.
17. Kutner, M. H., Nachtsheim, C. J., Neter, J. & Li, W., (1996). *Applied Linear Statistical Models*. 5th ed. Chicago: McGraw-Hill Irwin.
18. Lehohla, P., (2008). *Income and Expenditure of Households 2005/2006: Statistical Release*, Pretoria: Statistics.
19. Lehohla, P., (2017). *Poverty trends in South Africa*, Pretoria: Statistical South Africa.
20. Lohr, S. L., (2010). *Sampling: Design and Analysis*. 2nd ed. Boston: Brooks/Cole.
21. Lumley, T., (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9(1), pp. 1-19.
22. Lumley, T., (2011). *Complex Survey A Guide to Analysis Using R*. New Jersey: John Wiley & Sons, Inc.
23. Lumley, T. & Alastair, S., (2017). Fitting Regression Models to Survey Data. *Statistical Science*, 32(2), pp. 265-278.
24. Lumley, T. & Scott, A., (2015). AIC and BIC for modeling with Complex Survey Data. *Journal of Survey Statistics and Methodology*, 3(1), pp. 1-18.
25. Luus, R., (2016). *Statistical Inference of the Multiple Regression Analysis of Complex Survey Data*. s.l.:PhD. thesis, Stellenbosch University.
26. Luus, R., Neethling, A. & de Wet, T., (2010). Effectiveness of weighting and bootstrap in the estimation of welfare indices under complex sampling. *South African Statistical Journal*, 46(1), pp. 85-114.
27. Madow, L. H., (1946). Systematic Sampling and its Relation to Other Sampling Designs. *Journal of the American Statistical Association*, 41(234), pp. 204-217.

28. Marshall, M. N., (1996). Sampling for qualitative research. *Family Practice*, 13(6), pp. 522-526.
29. Menard, S., (2010). *Logistic Regression: From Introductory to Advanced Concepts and Applications*. California: SAGE Publications, Inc.
30. Nations, U., (2005). *Household Sample Surveys in Developing and*. s.l.:United Nations.
31. Neethling, A. & Galpin, J. S., (2006). Weighting of household survey data: a comparison of various calibration, integrated and cosmetic estimators: theory and methods. *South African Statistical Journal*, 40(2), pp. 123-150.
32. O'Connell, A. A., (2006). *Logistic regression models for ordinal response variables*. California: Sage.
33. SAS Institute, (2017). *SAS/STAT 9.4 User's Guide*. Cary, NC: SAS Institute Inc. North Carolina: SAS Institute.
34. Seefeld, K. & Linder, E., (2007). *Statistics Using R with Biological Examples*. s.l.:University of New Hampshire, Durham.
35. Simon, J. & Mitterling, L., (2017). *Macro Language 1: Essentials Course Notes*. s.l.:SAS Institute Inc..
36. Sitter, R. R., (1992). A Resampling Procedure for Complex Survey Data. *Journal of the American Statistical Association*, 87(419), pp. 755-765.
37. Stephan, F. F., (1948). History of the uses of modern sampling. *Journal of the American Statistical Association*, 43(241), pp. 12-39.
38. Tansey, O., (2007). Process Tracing and Elite Interviewing: A Case for Non-probability Sampling. *PS: Political Science & Politics*, 4(40), pp. 765-772.
39. Thomas, S. L. & Heck, R. H., (2001). Analysis of large-scale secondary data in higher education research: Potential Perils Associated with Complex Sampling Designs. *Research in higher education*, 42(5), pp. 517-540.
40. Thompson, S. K., (2010). *Sampling*. Third Edition ed. New Jersey: John Wiley and Sons, inc..
41. Tongco, M. D., (2007). Purposive sampling as a tool for informant selection. *Ethnobotany Research & Applications*, 5(1), pp. 147-158.

42. Wackerly, D. D., Mendenhall, W. & Scheaffer, R. L., (2008). *Mathematical Statistics with Applications*. 7 ed. California: Thomson Learning, Inc.
43. Walker, D. A. & Young, D. Y., (2003). Example Of The Impact Of Weights And Design Effects On Contingency Tables And Chi-Square Analysis. *Journal of Modern Applied Statistical Methods*, 2(2), pp. 425-432.
44. Wolter, K. M., (2007). *Introduction to Variance Estimation*. 2nd ed. New York: Springer.
45. Woodruff, R. S., (1971). A Simple Method for Approximating the Variance of a Complicated Estimate. *Journal of the American Statistical Association*, 66(334), pp. 411-414.
46. Yamane, T., (1967). *Elementary Sampling Theory*. s.l.:Prentice-Hall, Inc., Englewood Cliffs, N.J.



UNIVERSITY *of the*
WESTERN CAPE

Appendices

Appendix A: Absolute bias

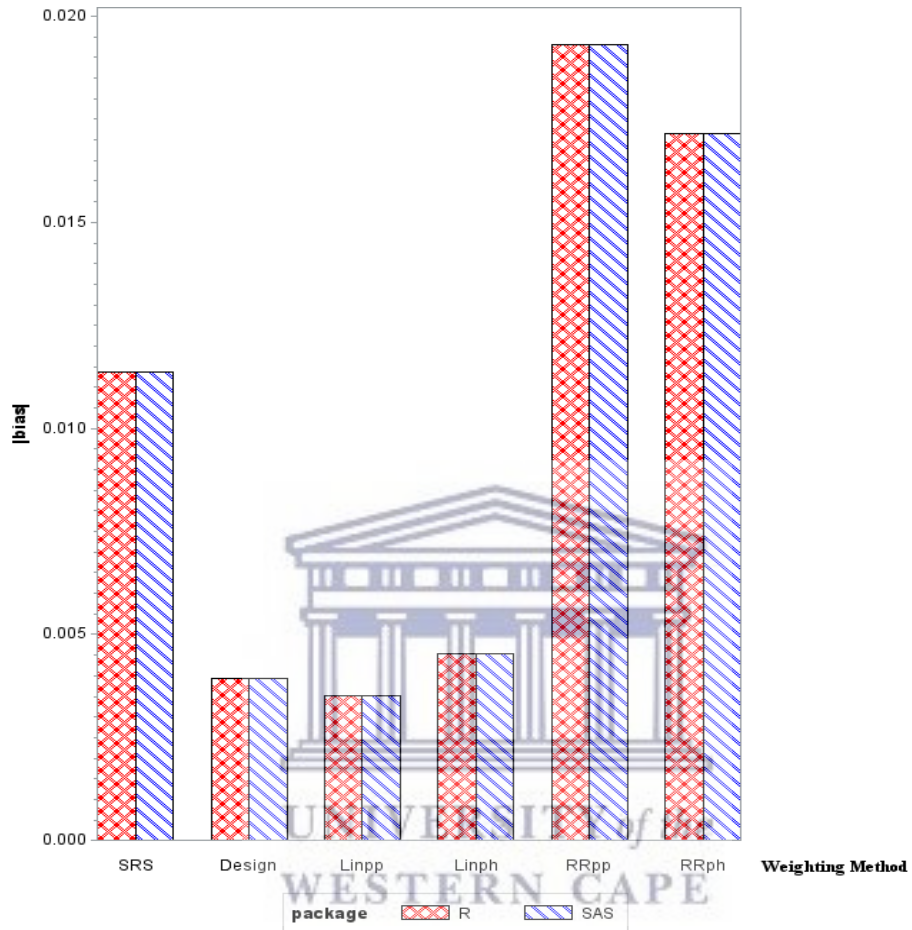


Figure A. 1: The absolute bias of the estimator of β_3 under SRS (no weight) and different weighting methods are shown for SAS and R.

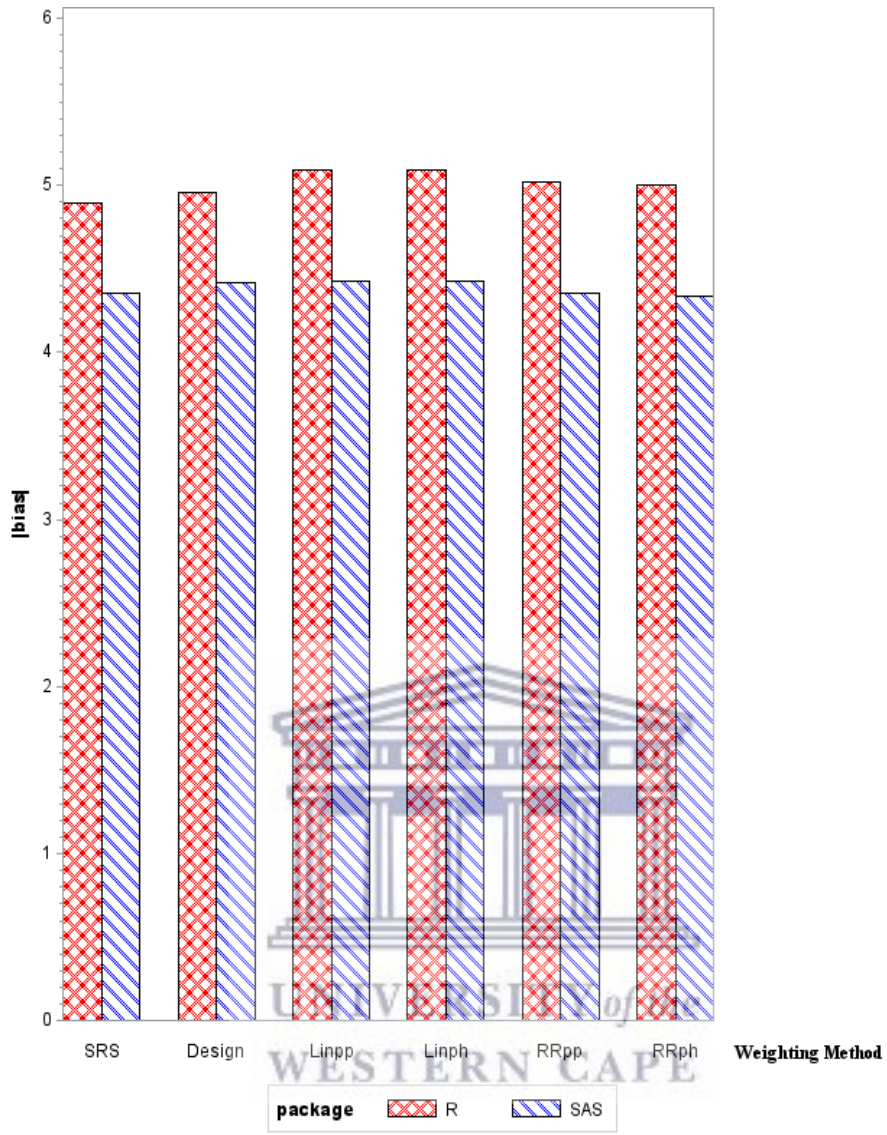


Figure A. 2: The absolute bias of the estimator of β_6 under SRS (no weight) and different weighting methods are shown for SAS and R.

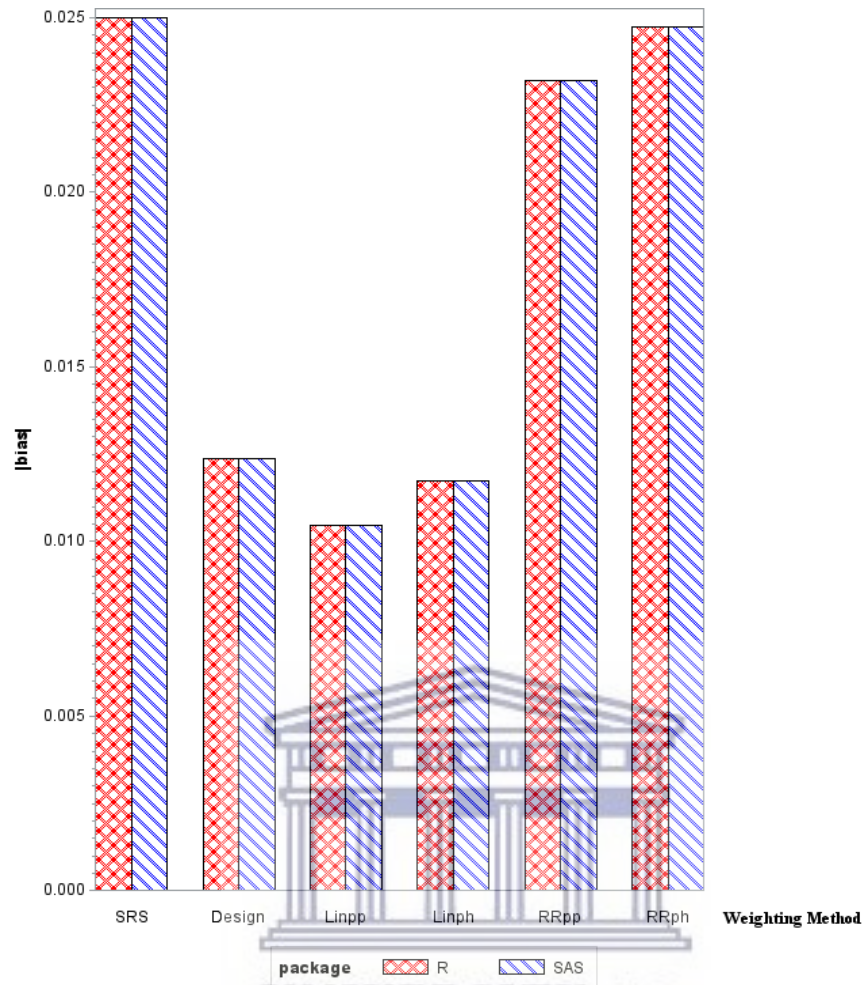


Figure A. 3: The absolute bias of the estimator of β_7 under SRS (no weight) and different weighting methods are shown for SAS and R.

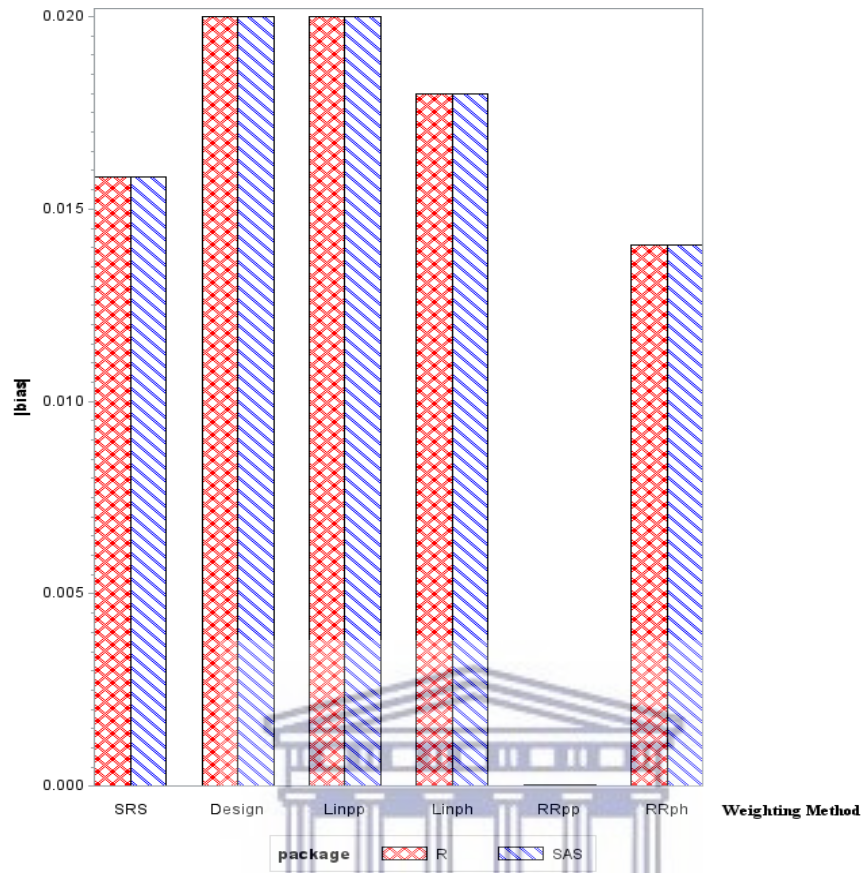


Figure A. 4: The absolute bias of the estimator of β_8 under SRS (no weight) and different weighting methods are shown for SAS and R.

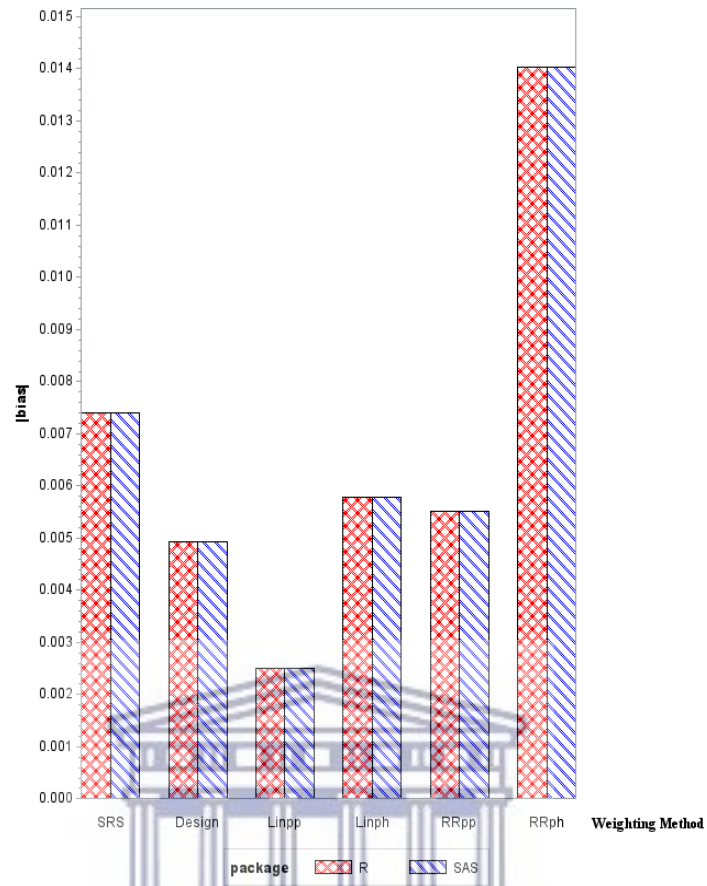


Figure A. 5: The absolute bias of the estimator of β_9 under SRS (no weight) and different weighting methods are shown for SAS and R.

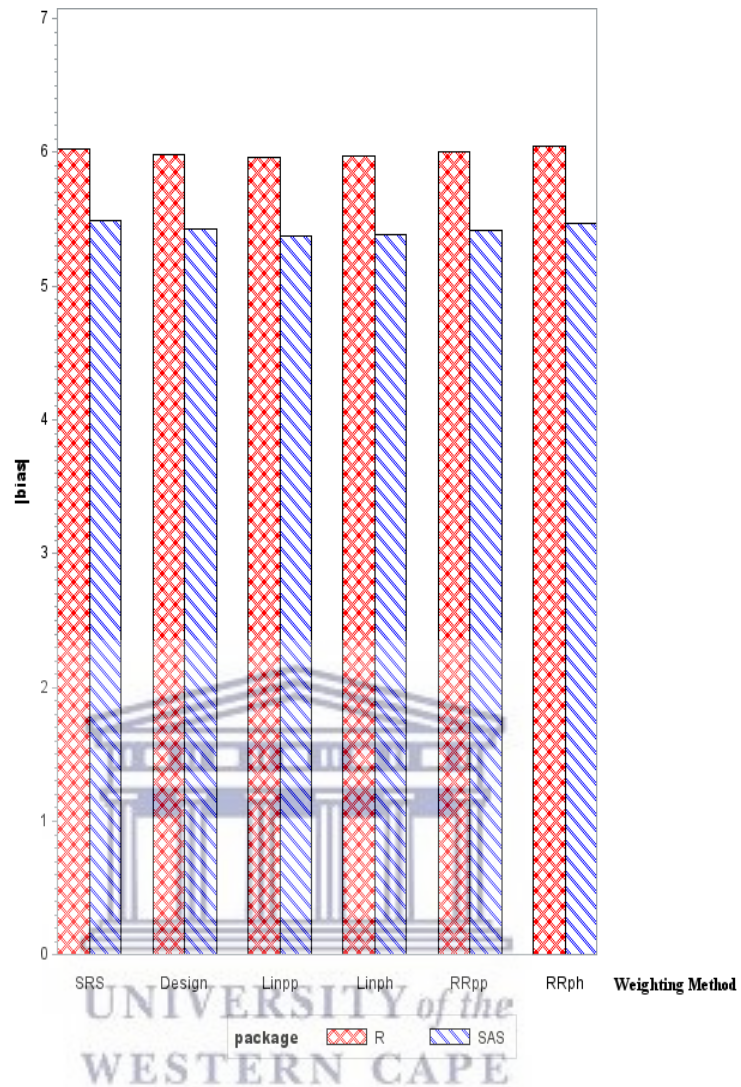


Figure A. 6: The absolute bias of the estimator of β_{10} under SRS (no weight) and different weighting methods are shown for SAS and R.

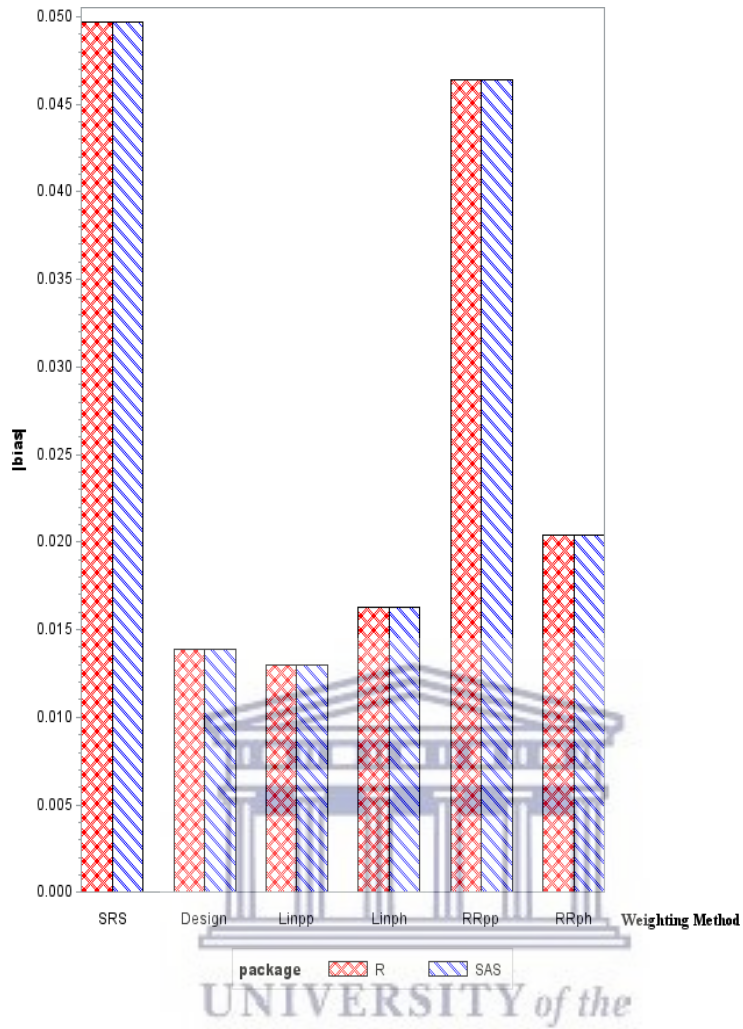


Figure A. 7: The absolute bias of the estimator of β_{13} under SRS (no weight) and different weighting methods are shown for SAS and R.

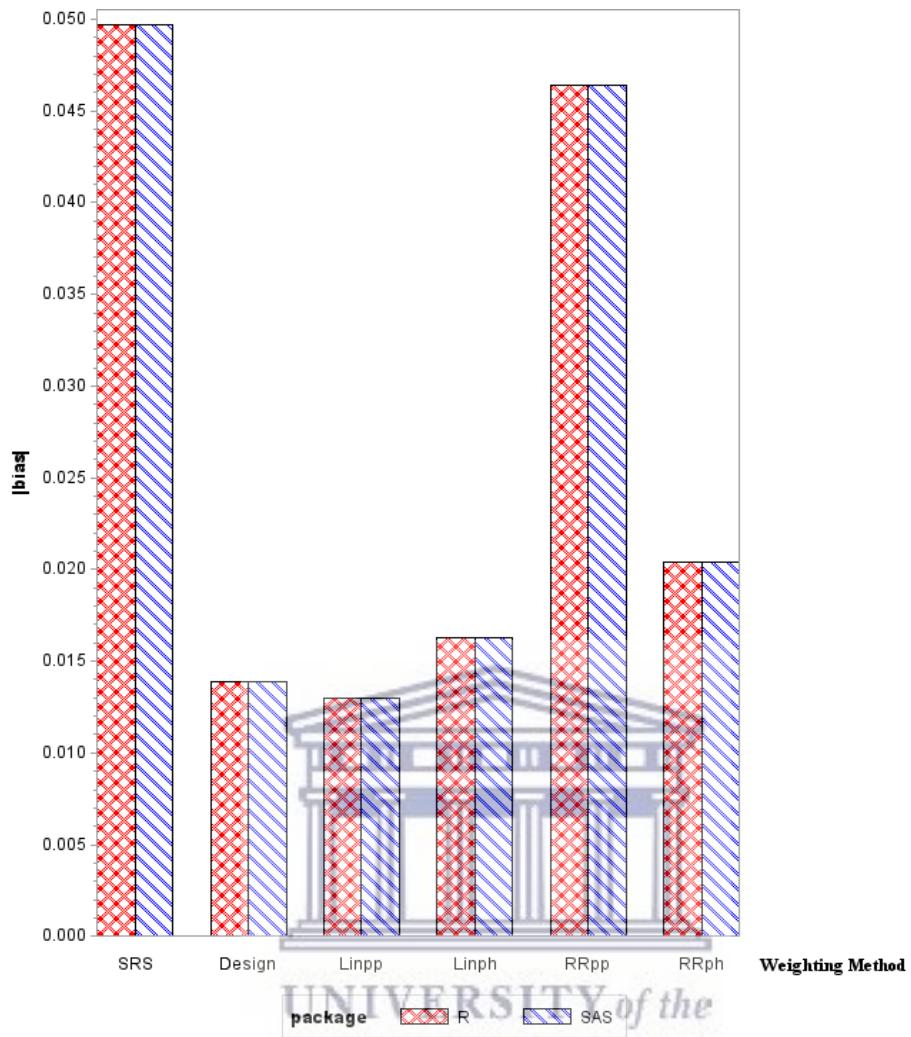


Figure A. 8: The absolute bias of the estimator of β_{14} under SRS (no weight) and different weighting methods are shown for SAS and R.

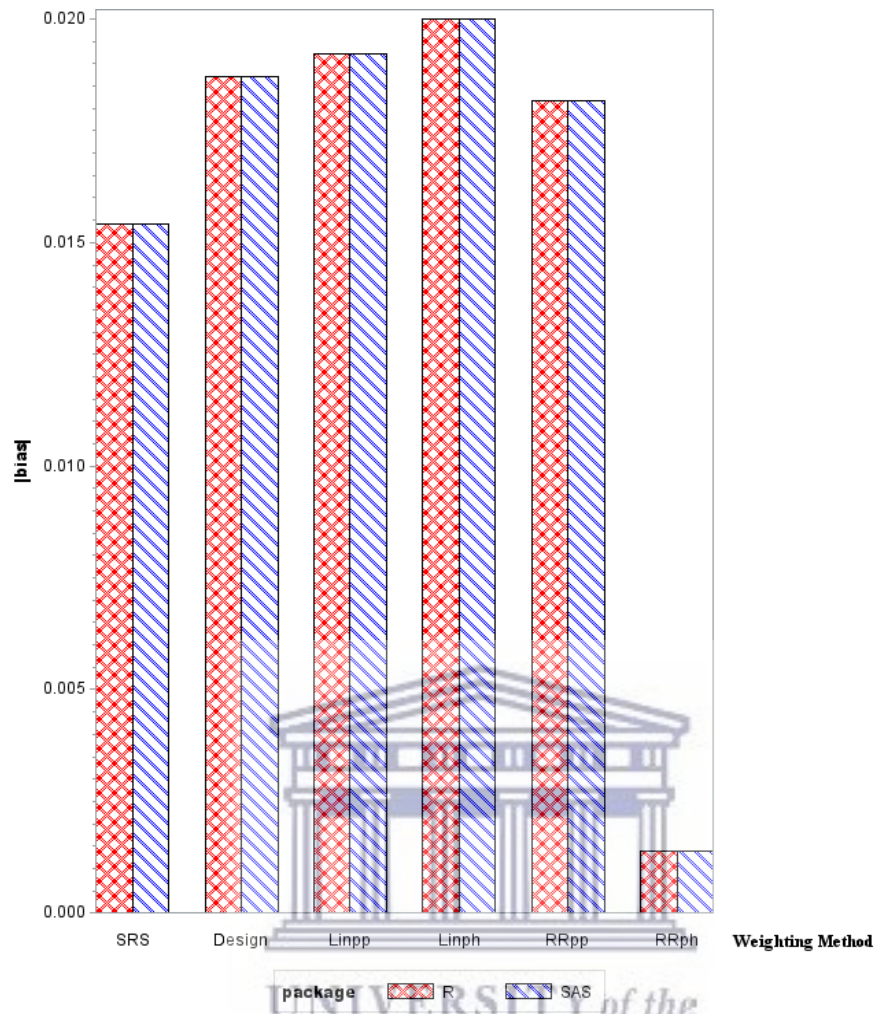


Figure A. 9: The absolute bias of the estimator of β_{15} under SRS (no weight) and different weighting methods are shown for SAS and R.

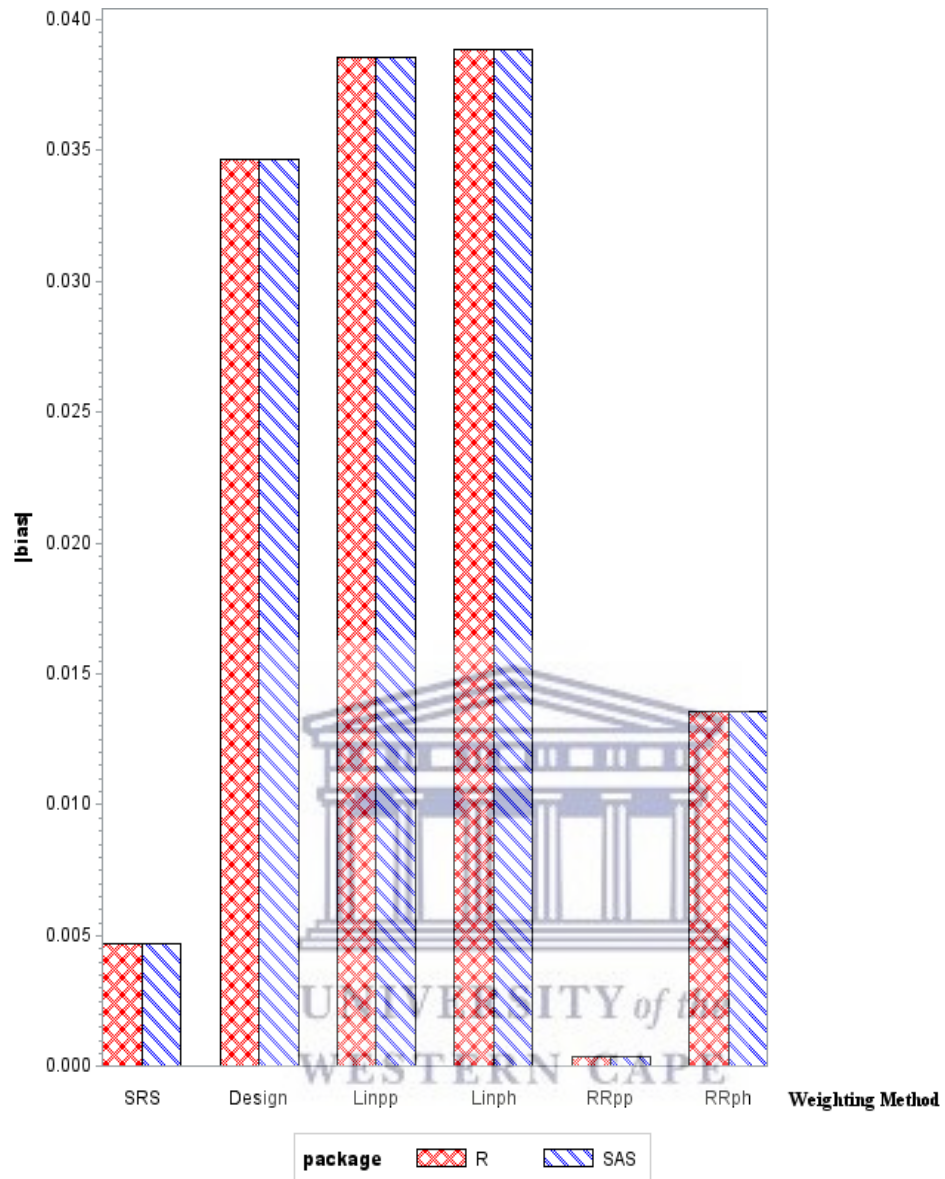


Figure A. 10: The absolute bias of the estimator of β_{16} under SRS (no weight) and different weighting methods are shown for SAS and R.

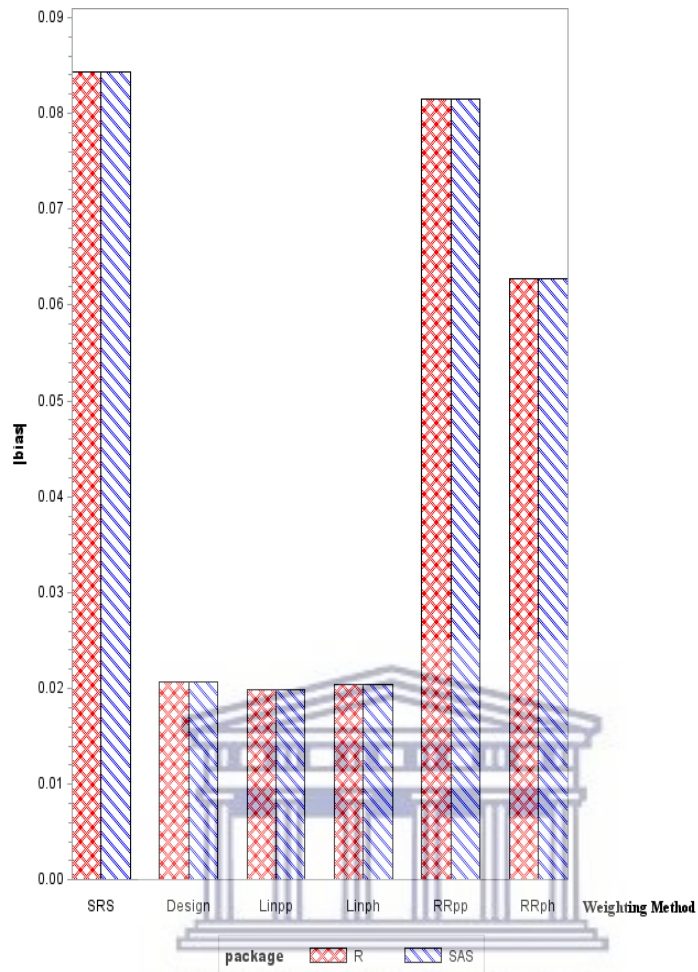


Figure A. 11: The absolute bias of the estimator of β_{17} under SRS (no weight) and different weighting methods are shown for SAS and R.

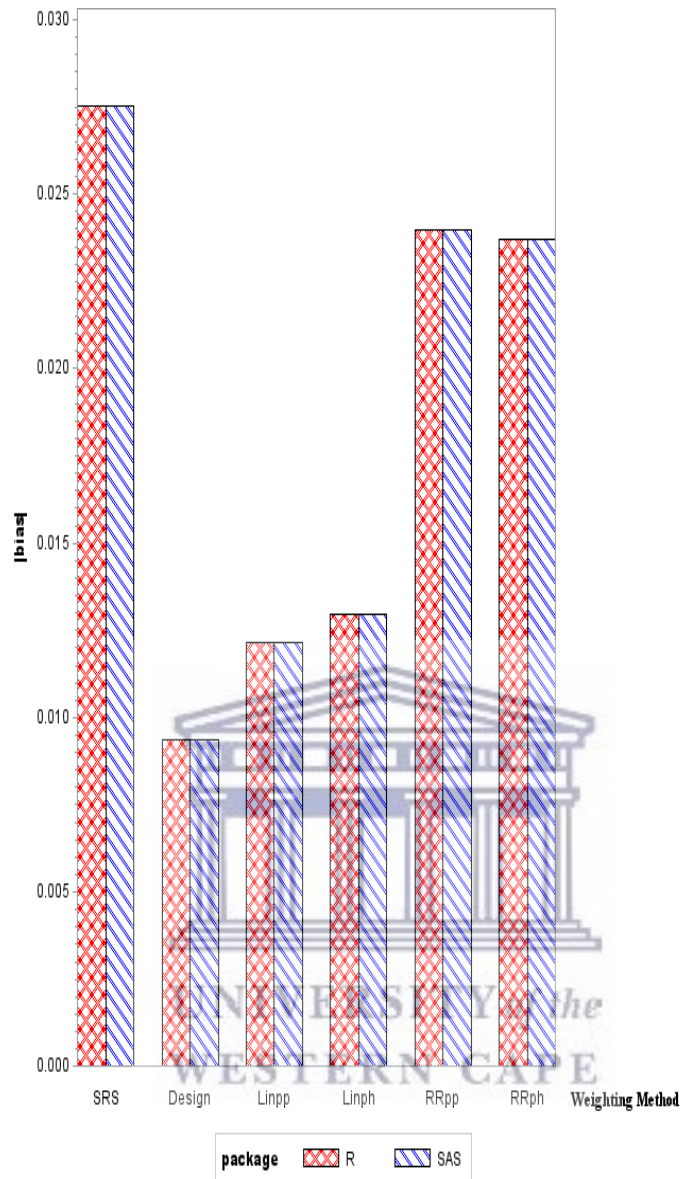


Figure A. 12: The absolute bias of the estimator of β_{18} under SRS (no weight) and different weighting methods are shown for SAS and R.

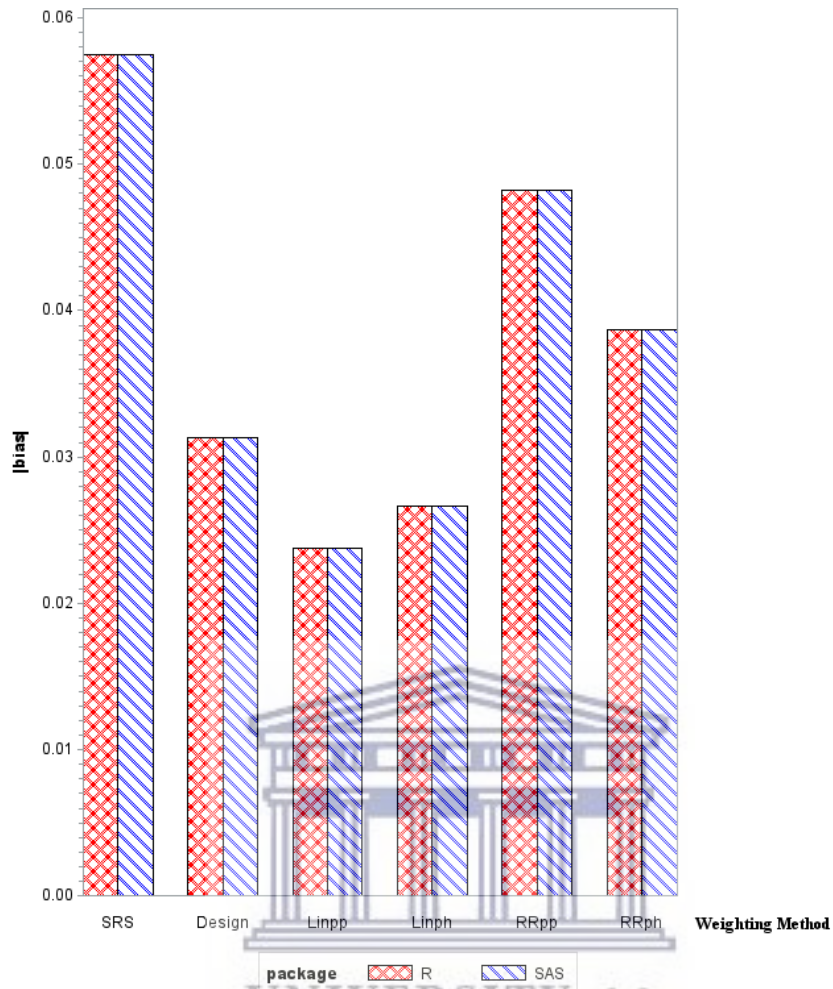


Figure A. 13: The absolute bias of the estimator of β_{19} under SRS (no weight) and different weighting methods are shown for SAS and R.

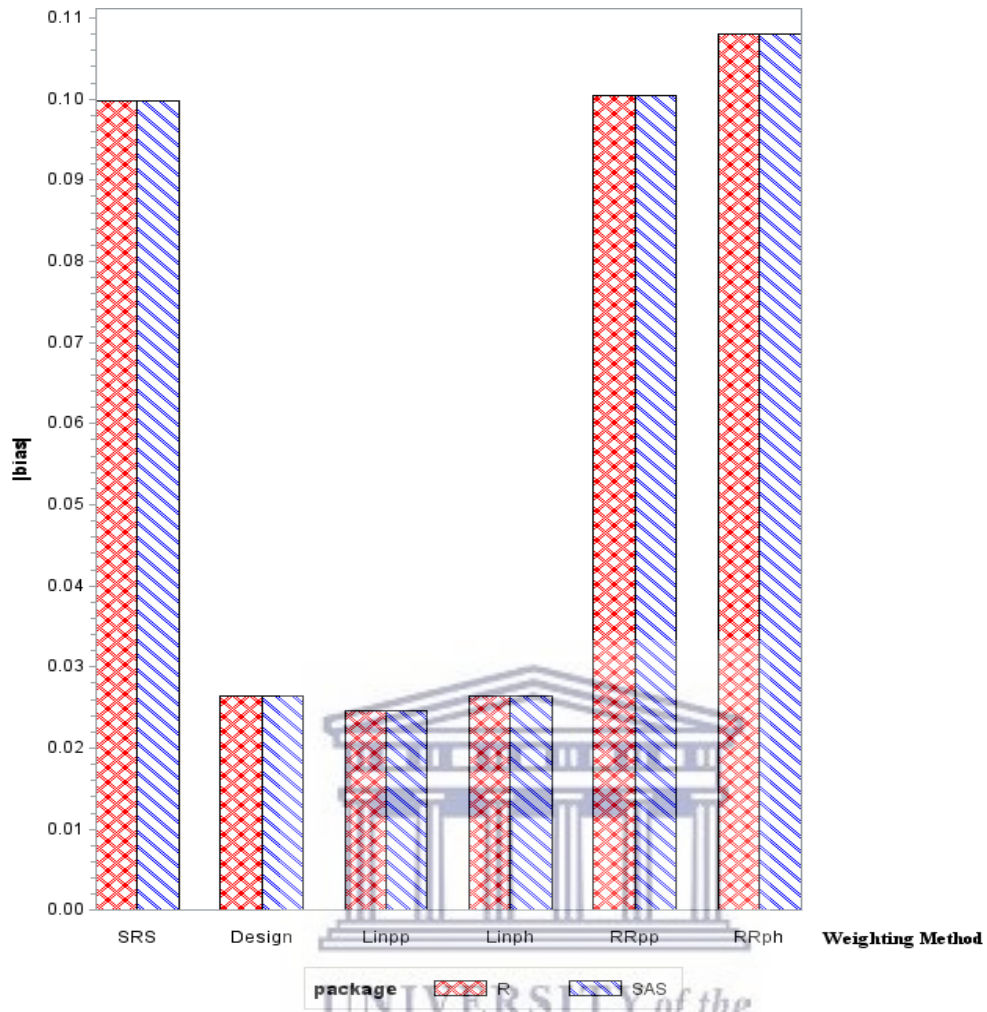


Figure A. 14: The absolute bias of the estimator of β_{21} under SRS (no weight) and different weighting methods are shown for SAS and R.

Appendix B: Mean squared error (MSE)

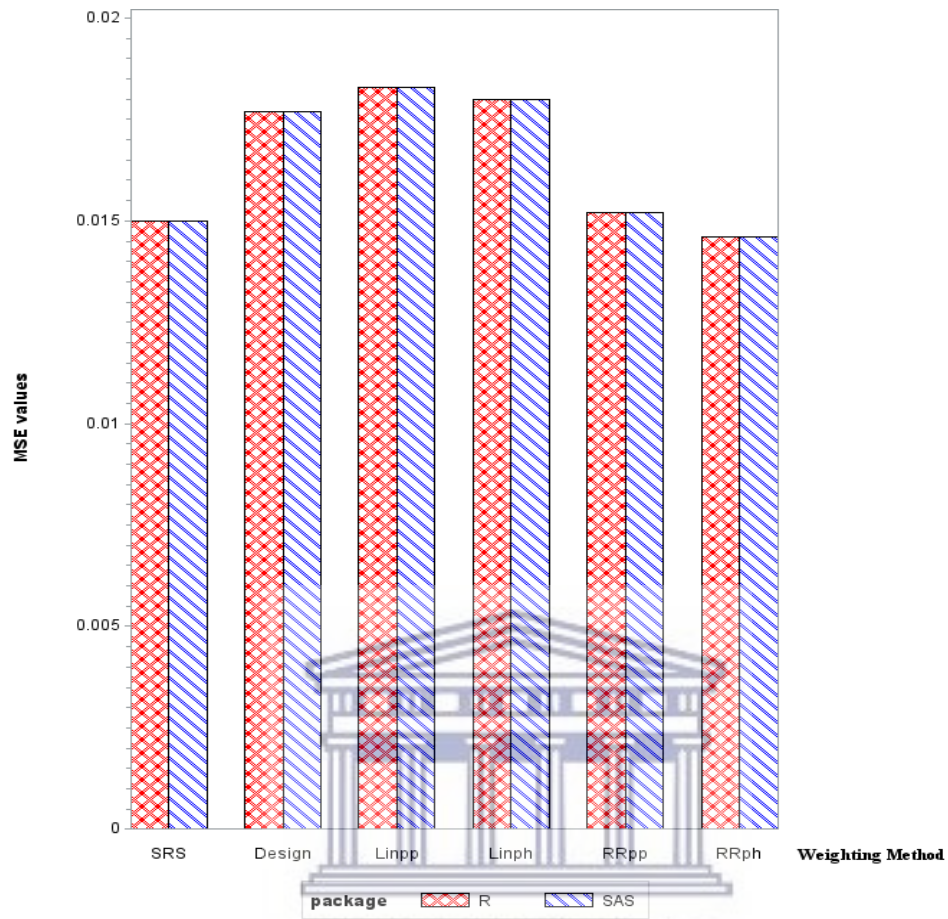


Figure B. 1: The MSE of the estimator of β_3 under SRS (no weight) and different weighting methods are shown for SAS and R.

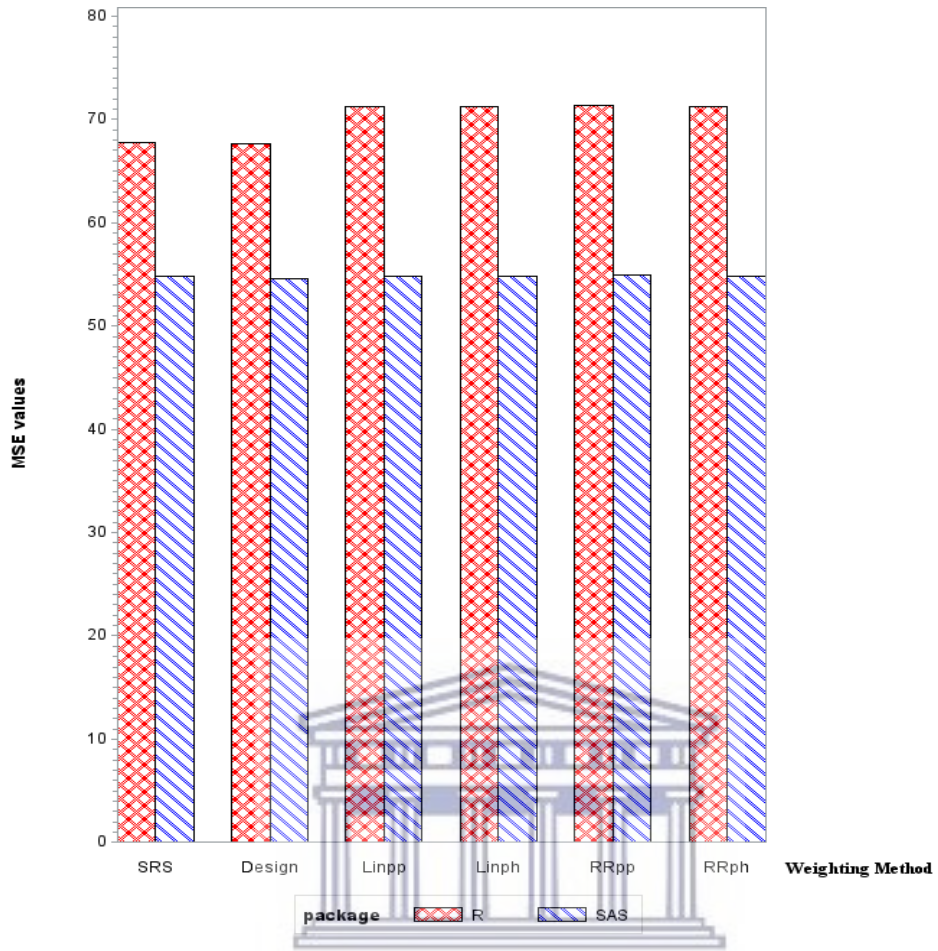


Figure B. 2: The MSE of the estimator of β_6 under SRS (no weight) and different weighting methods are shown for SAS and R.

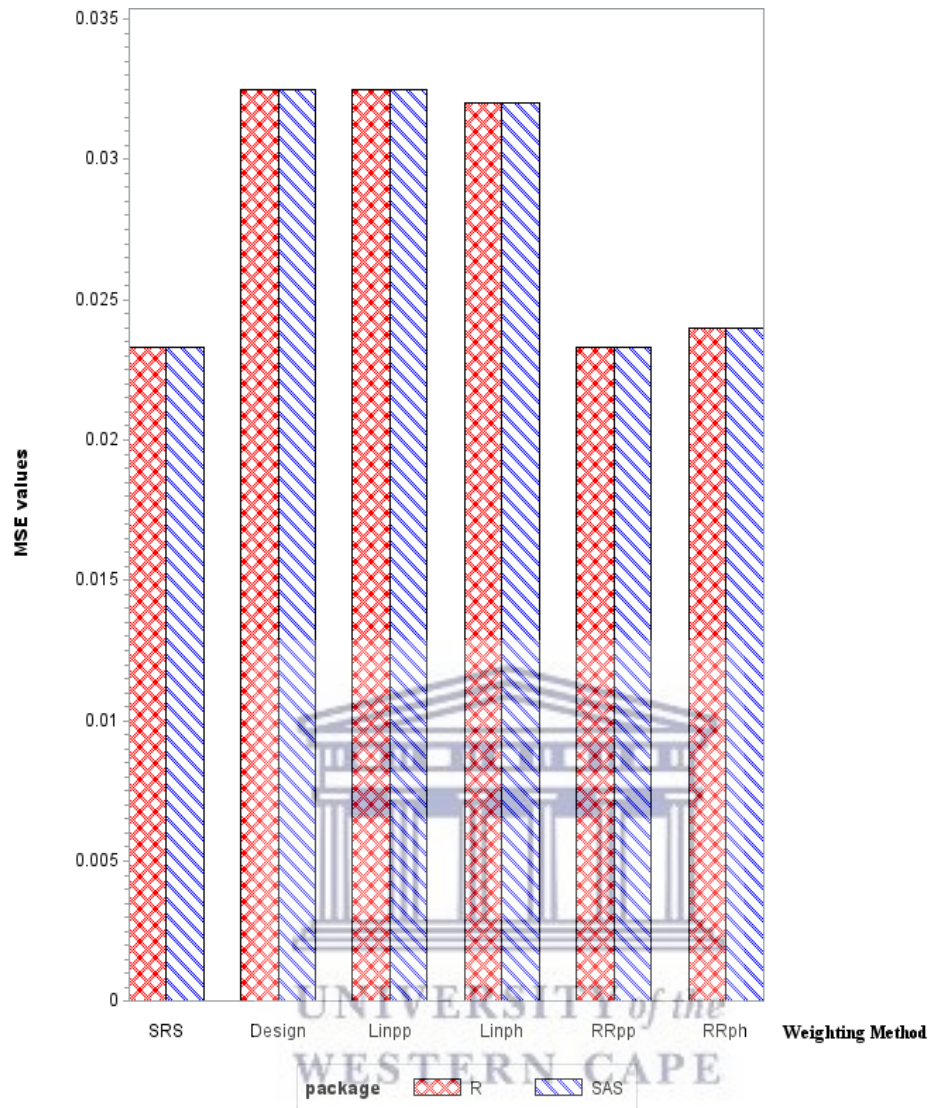


Figure B. 3: The MSE of the estimator of β_7 under SRS (no weight) and different weighting methods are shown for SAS and R.

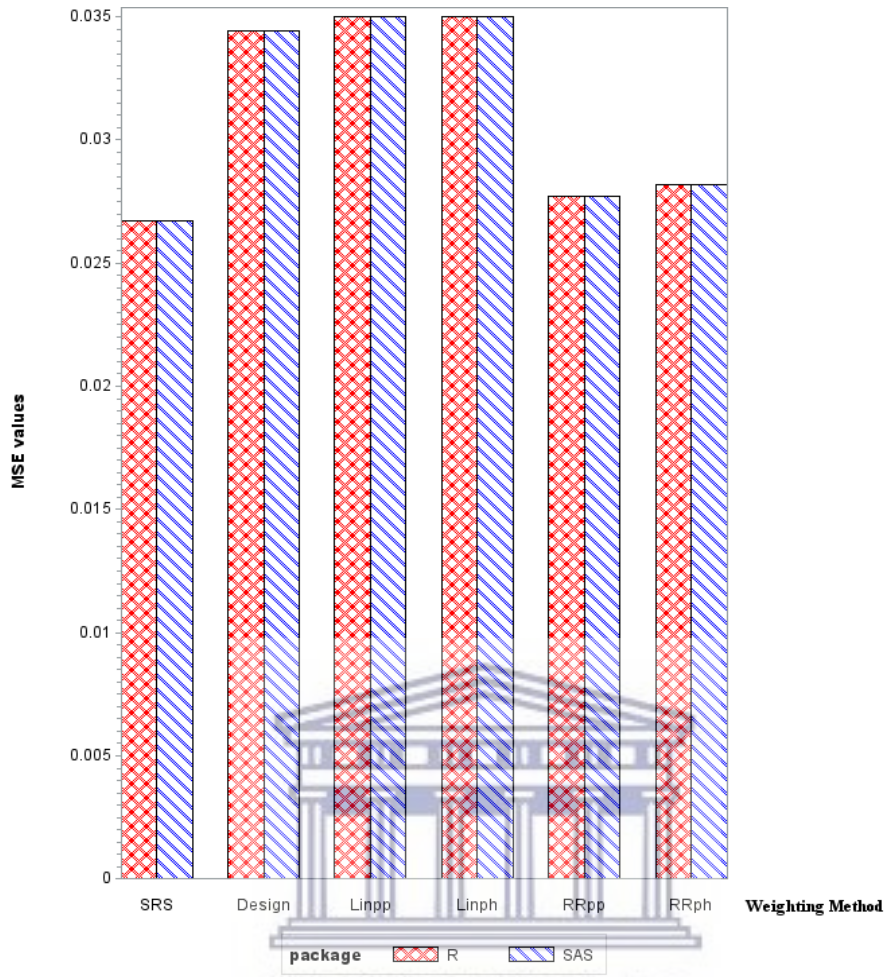


Figure B. 4: The MSE of the estimator of β_g under SRS (no weight) and different weighting methods are shown for SAS and R.

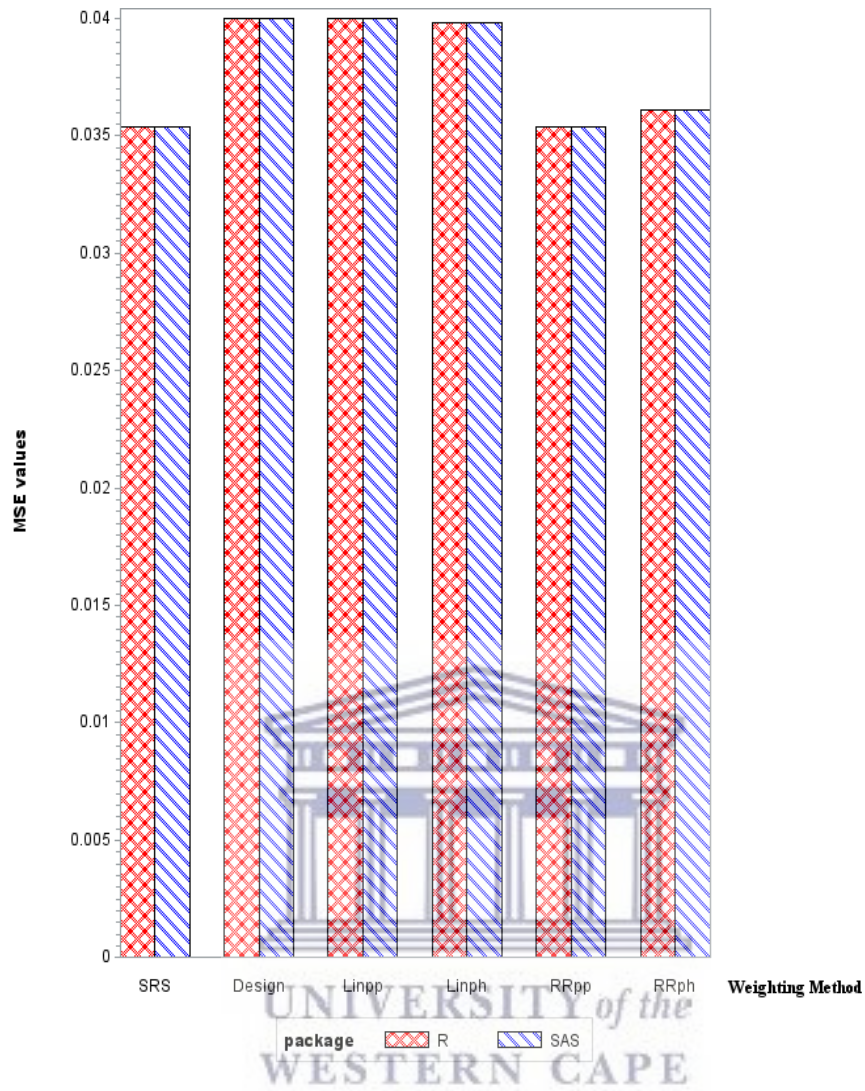


Figure B. 5: The MSE of the estimator of β_9 under SRS (no weight) and different weighting methods are shown for SAS and R.

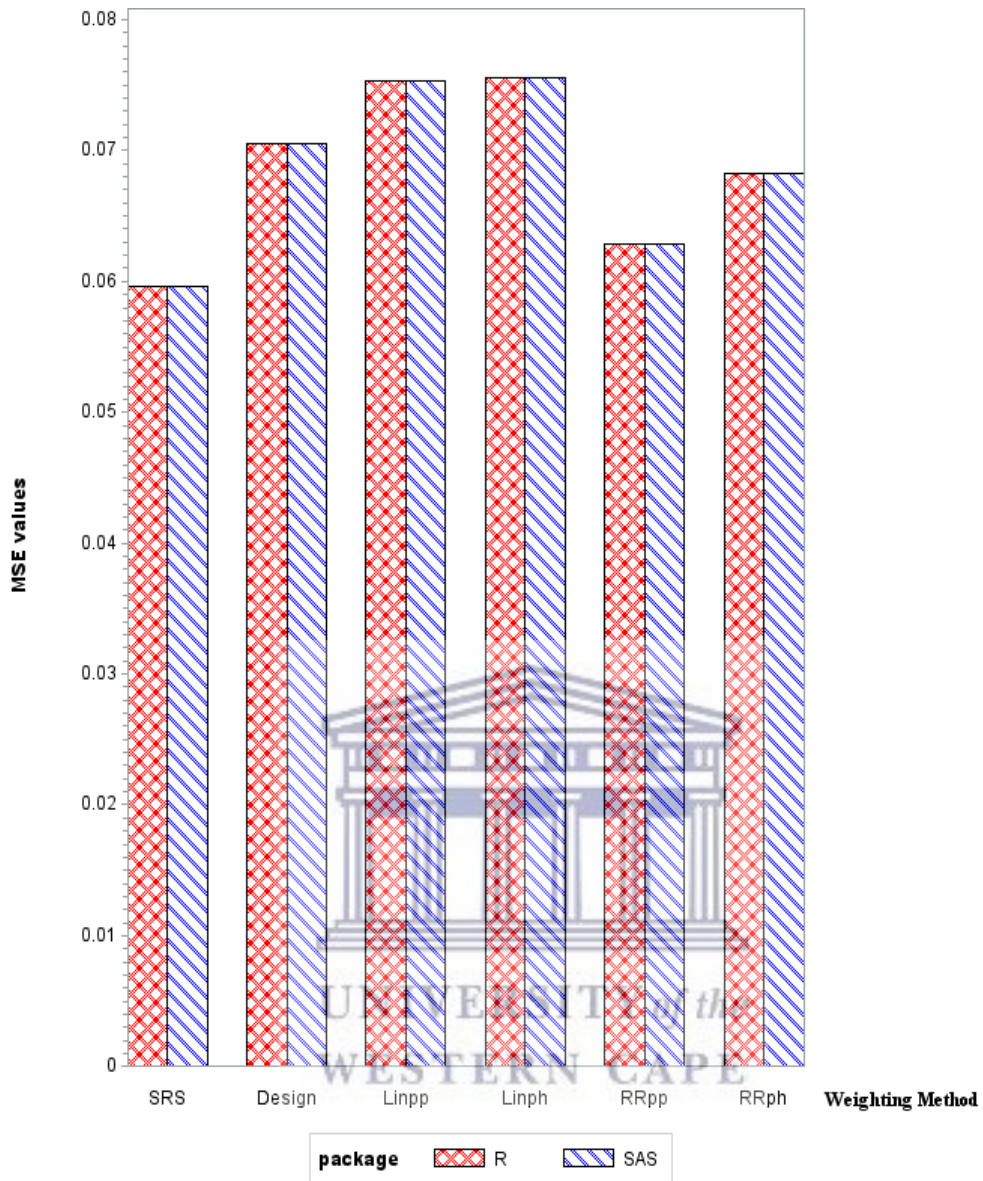


Figure B. 6: The MSE of the estimator of β_{10} under SRS (no weight) and different weighting methods are shown for SAS and R.

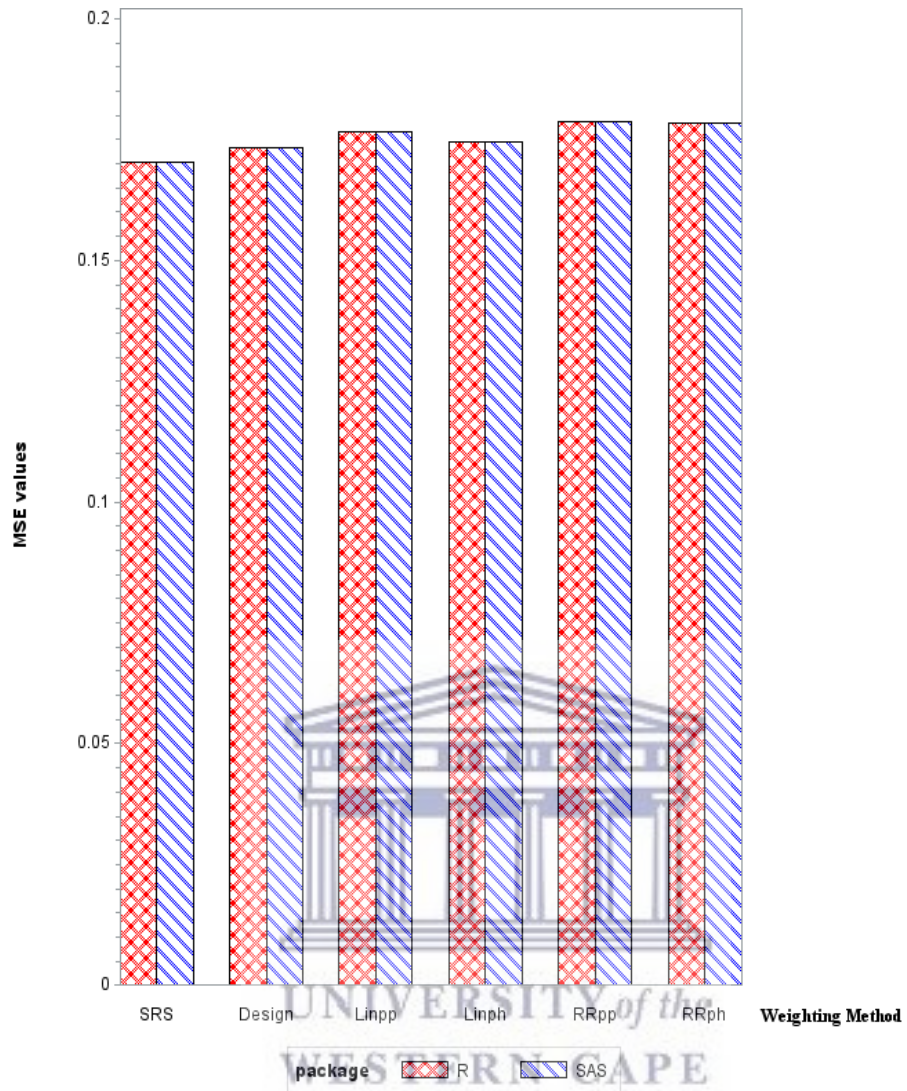


Figure B. 7: The MSE of the estimator of β_{13} under SRS (no weight) and different weighting methods are shown for SAS and R.

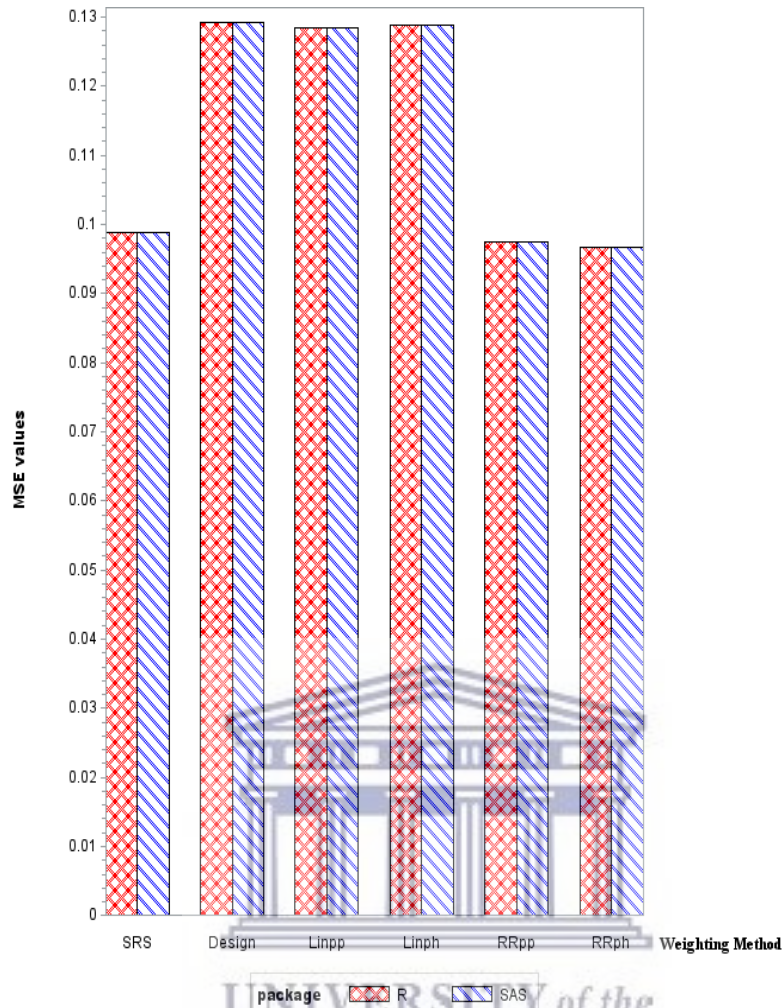


Figure B. 8: The MSE of the estimator of β_{14} under SRS (no weight) and different weighting methods are shown for SAS and R.

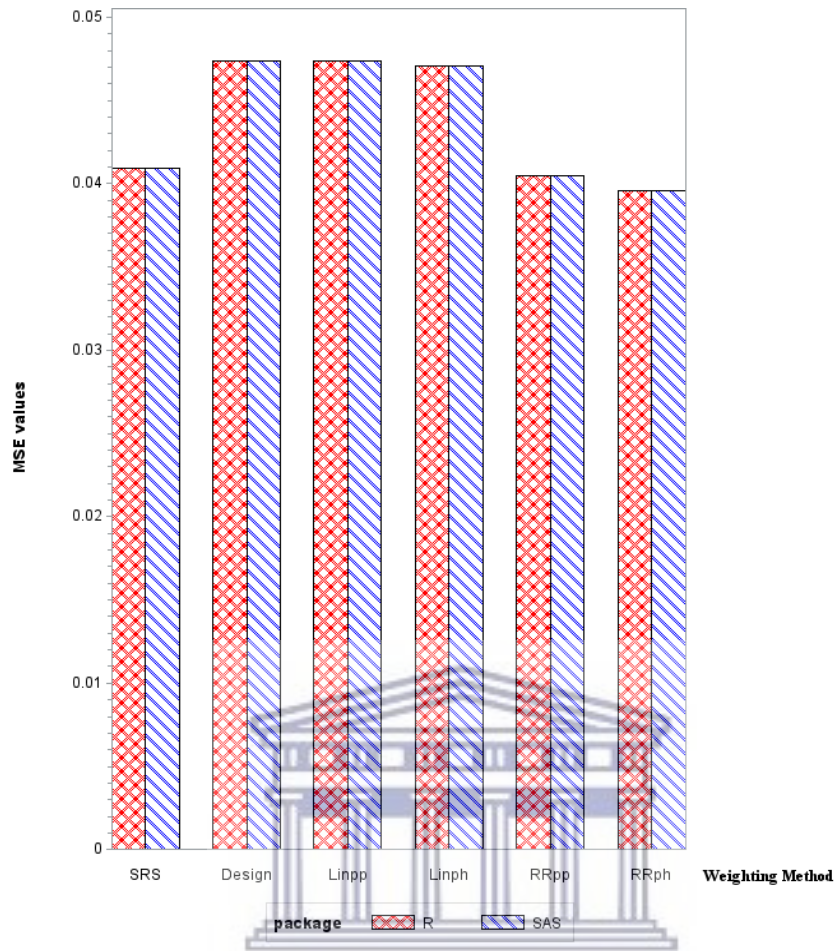


Figure B. 9: The MSE of the estimator of β_{15} under SRS (no weight) and different weighting methods are shown for SAS and R.

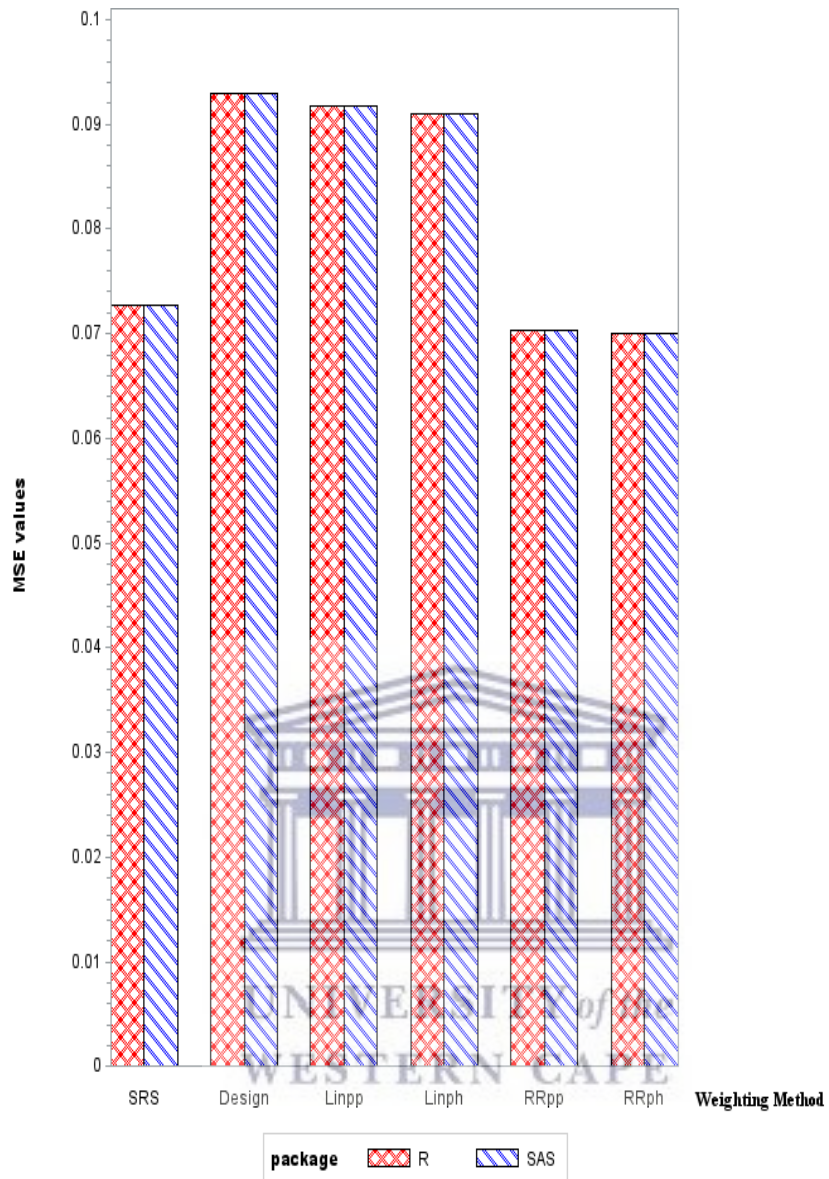


Figure B. 10: The MSE of the estimator of β_{16} under SRS (no weight) and different weighting methods are shown for SAS and R.

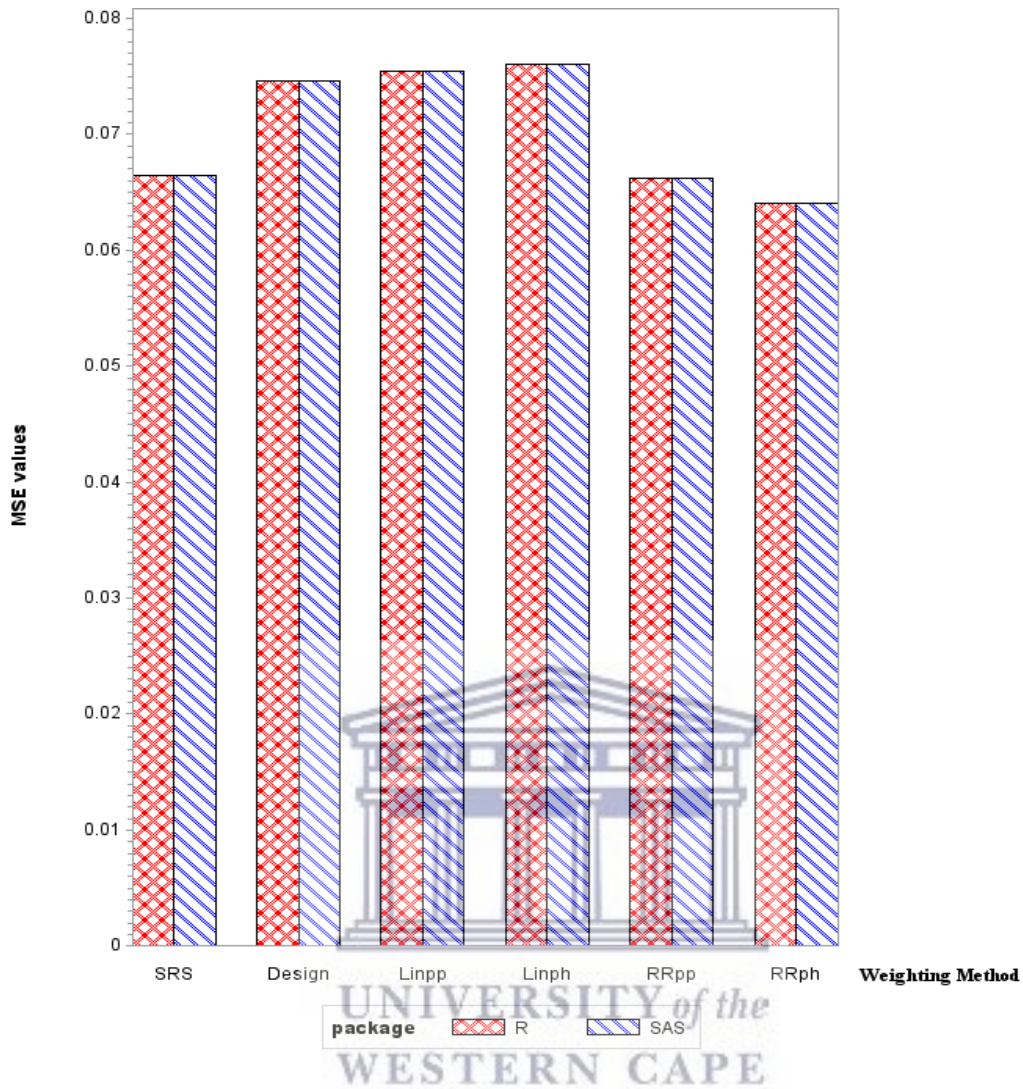


Figure B. 11: The MSE of the estimator of β_{17} under SRS (no weight) and different weighting methods are shown for SAS and R.

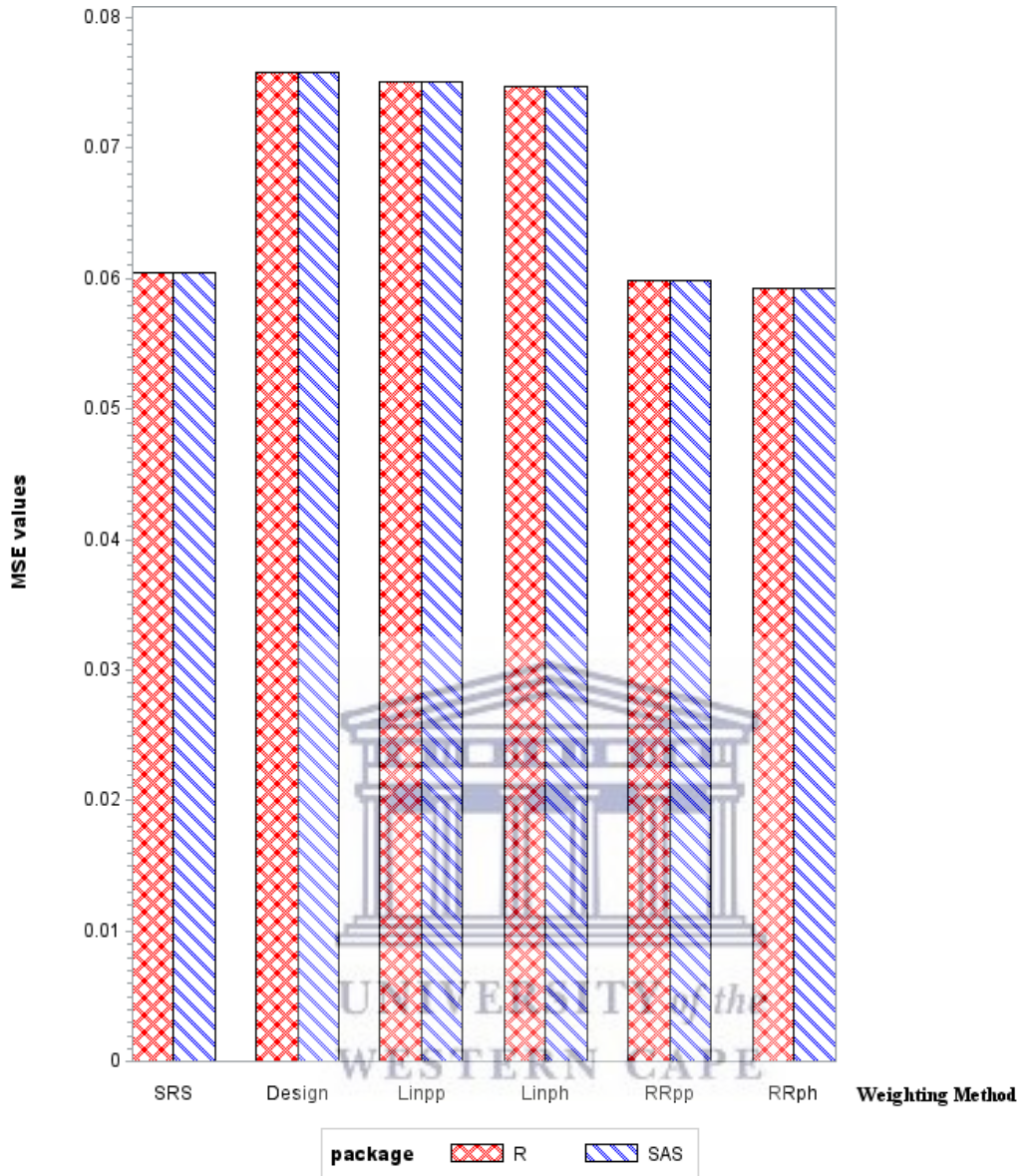


Figure B. 12: The MSE of the estimator of β_{18} under SRS (no weight) and different weighting methods are shown for SAS and R.

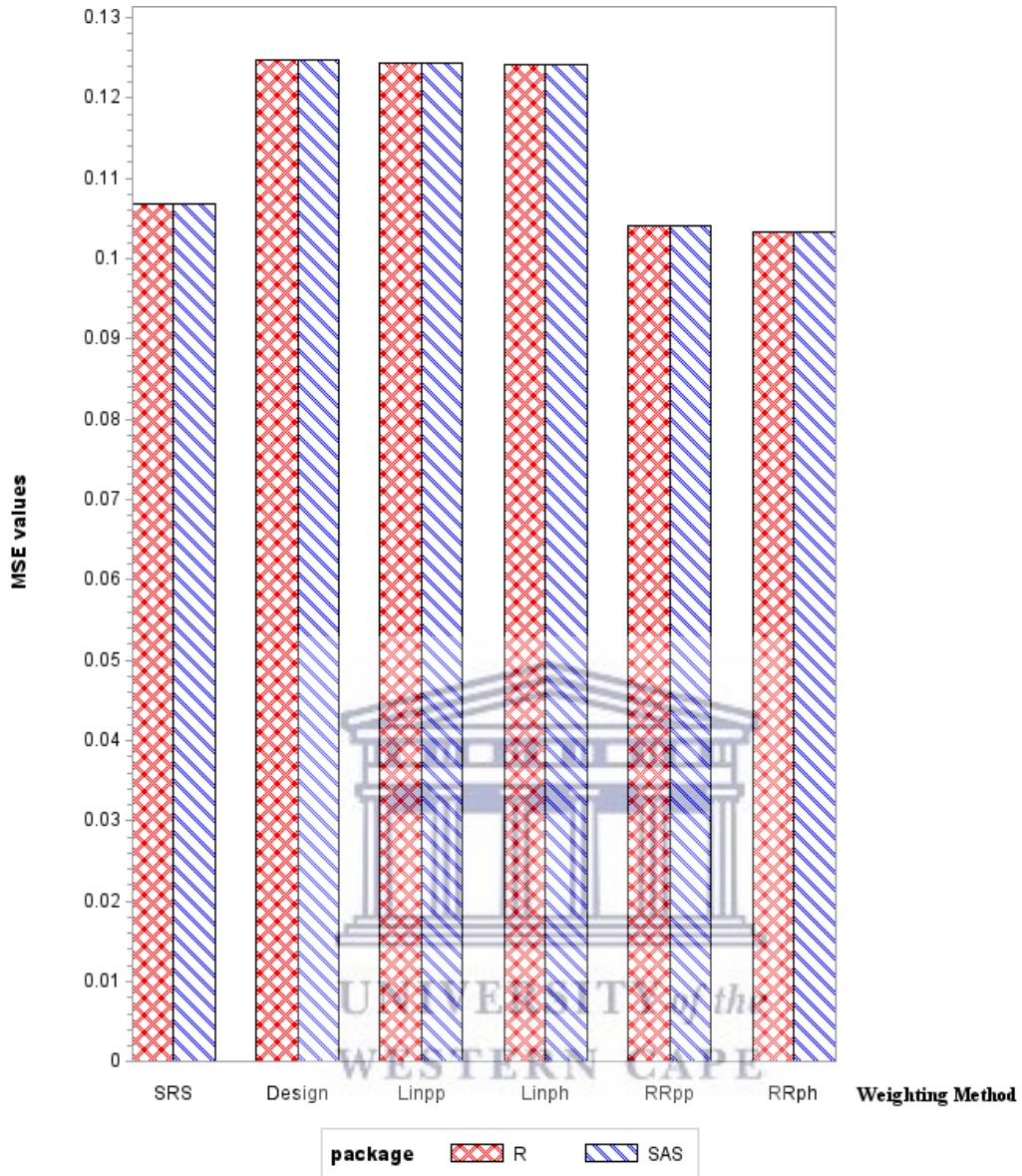


Figure B. 13: The MSE of the estimator of β_{19} under SRS (no weight) and different weighting methods are shown for SAS and R.

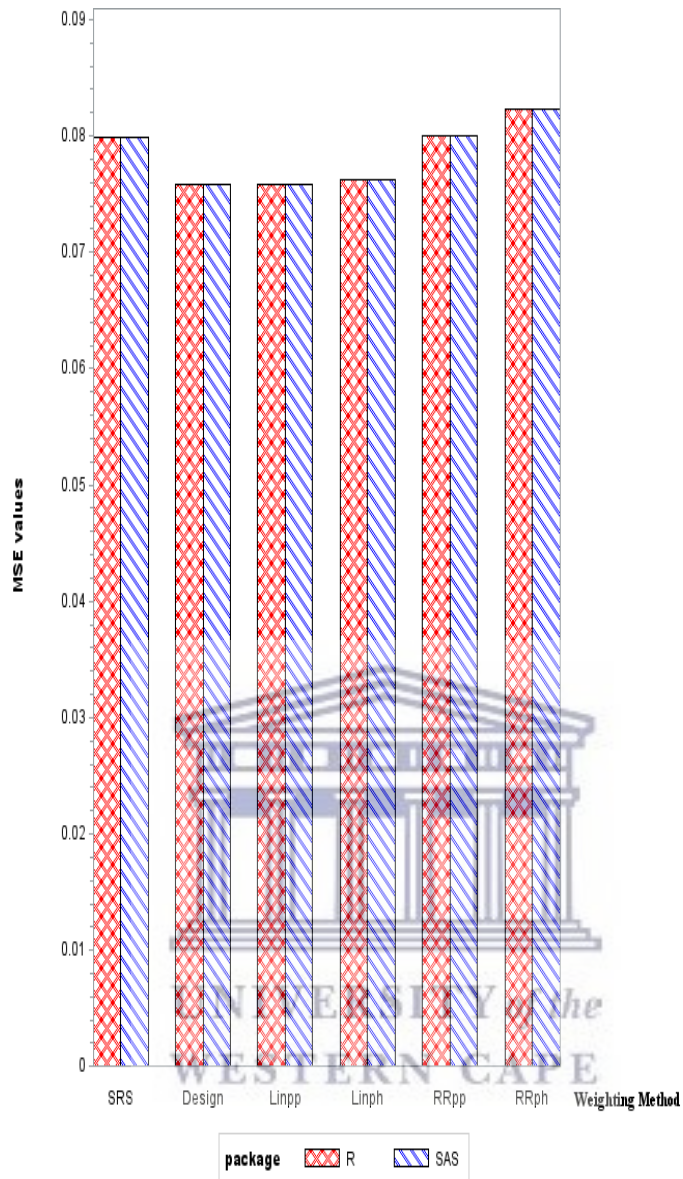


Figure B. 14: The MSE of the estimator of β_{21} under SRS (no weight) and different weighting methods are shown for SAS and R.

Appendix C: Coverage probability

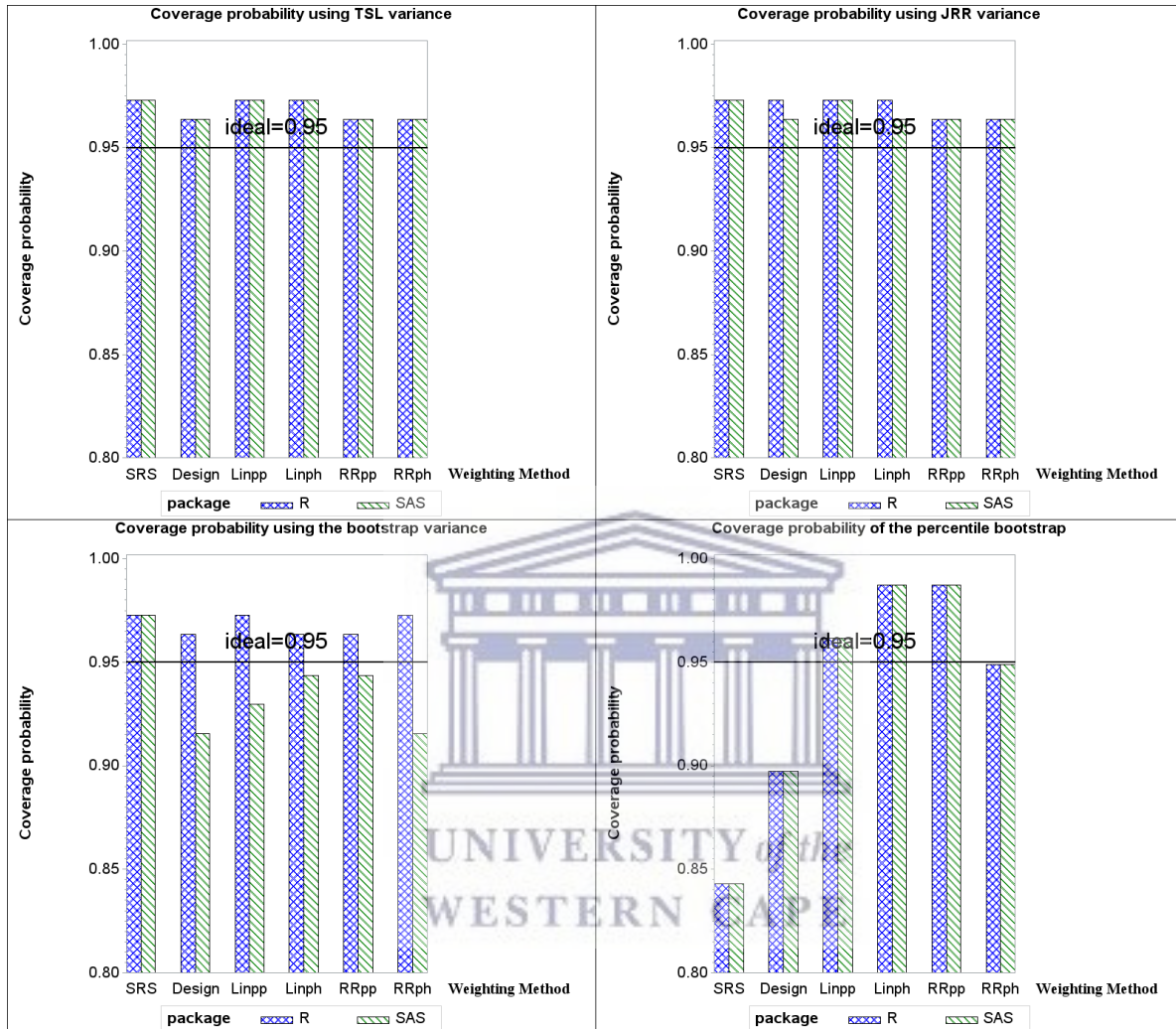


Figure C. 1: The coverage probabilities for β_1 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

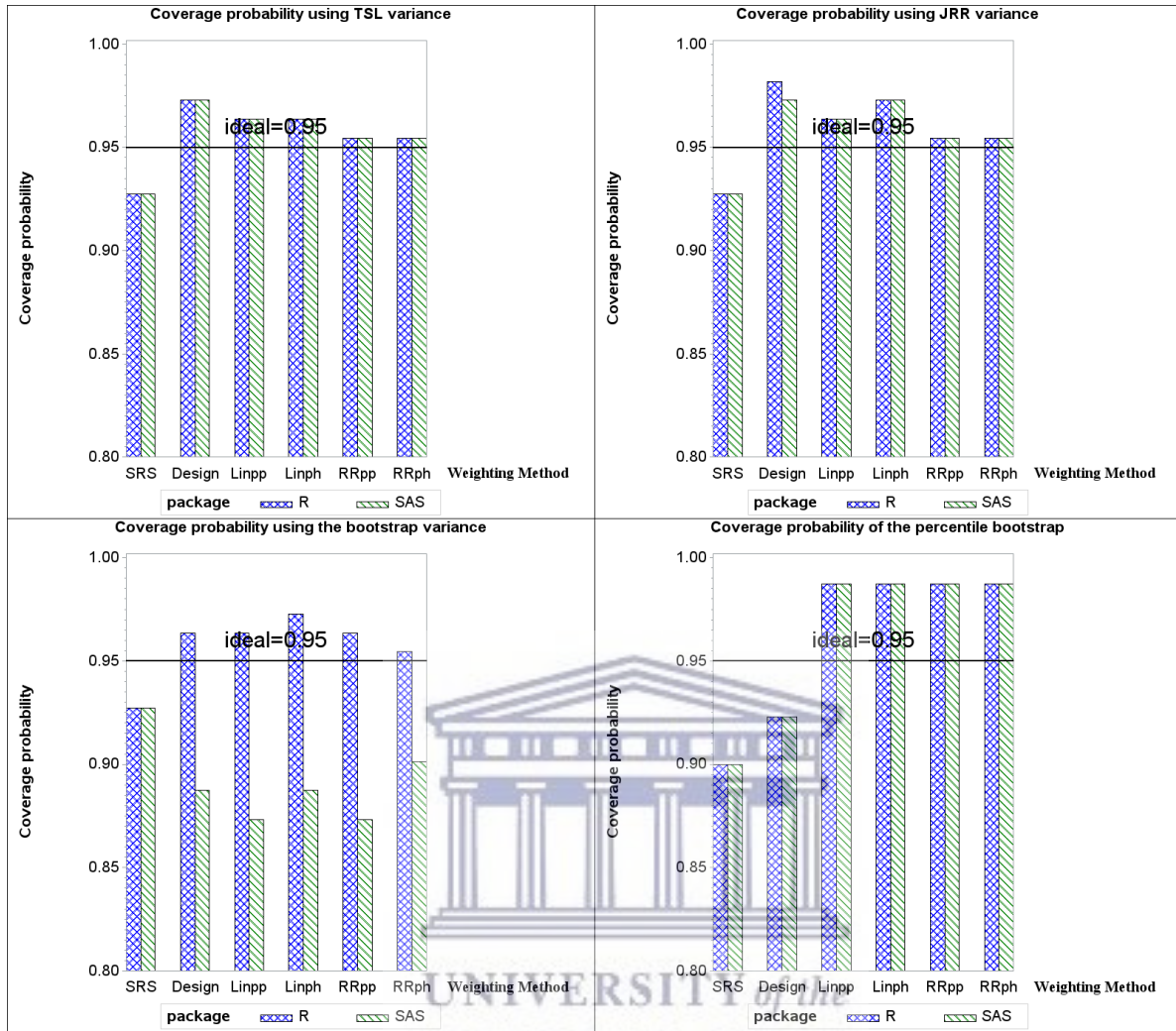


Figure C. 2: The coverage probabilities for β_2 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

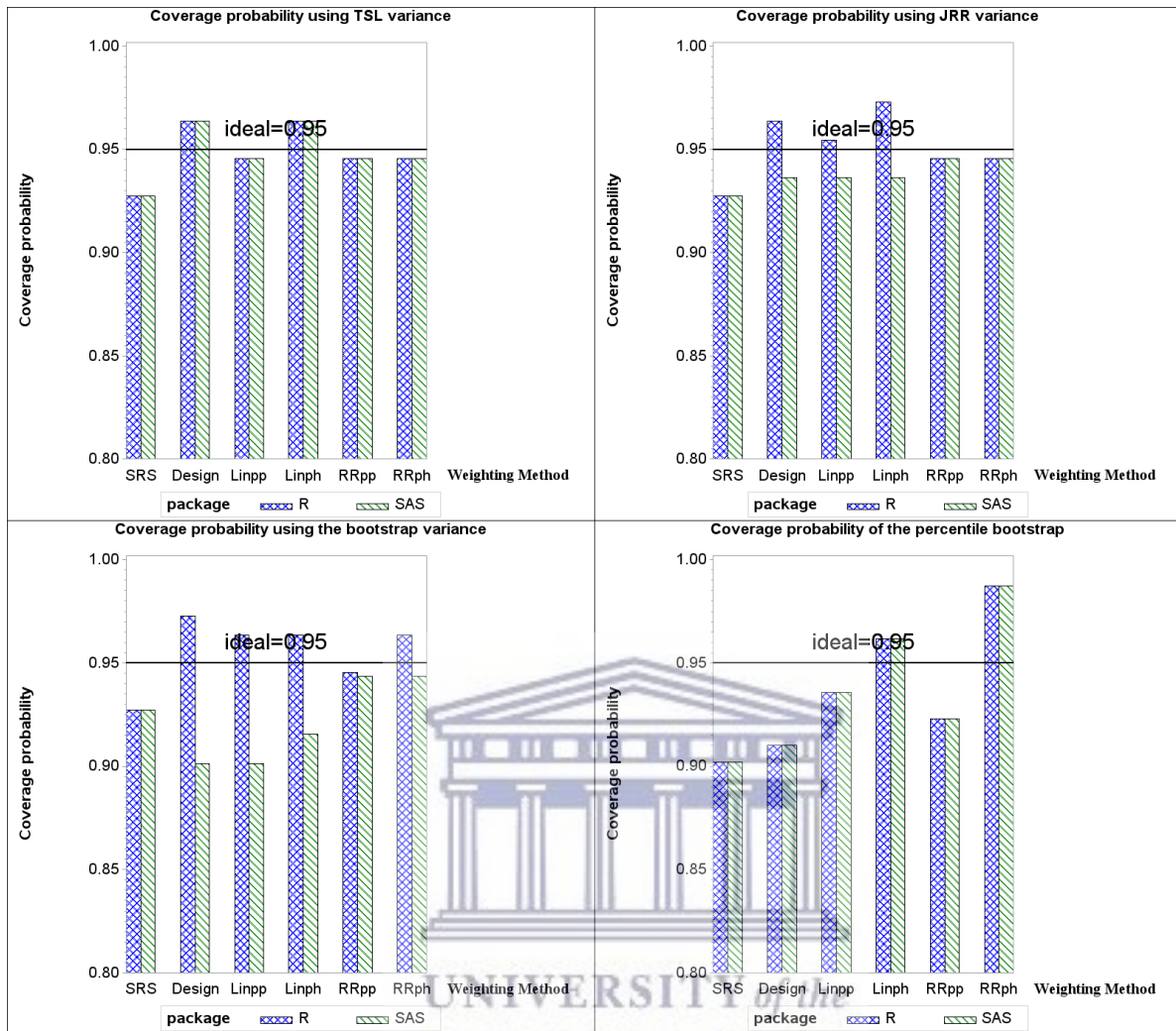


Figure C. 3: The coverage probabilities for β_3 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

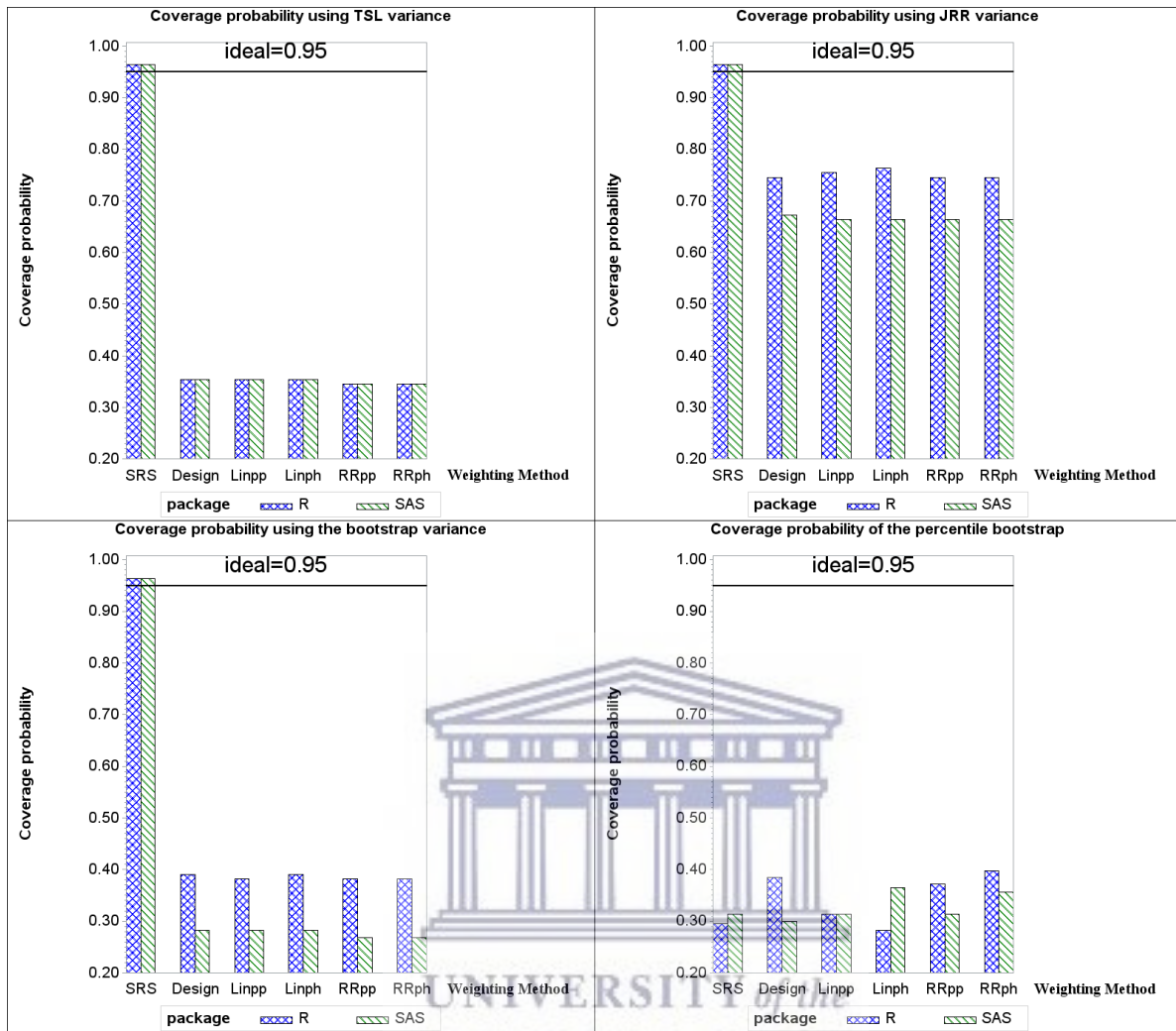


Figure C. 4: The coverage probabilities for β_5 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

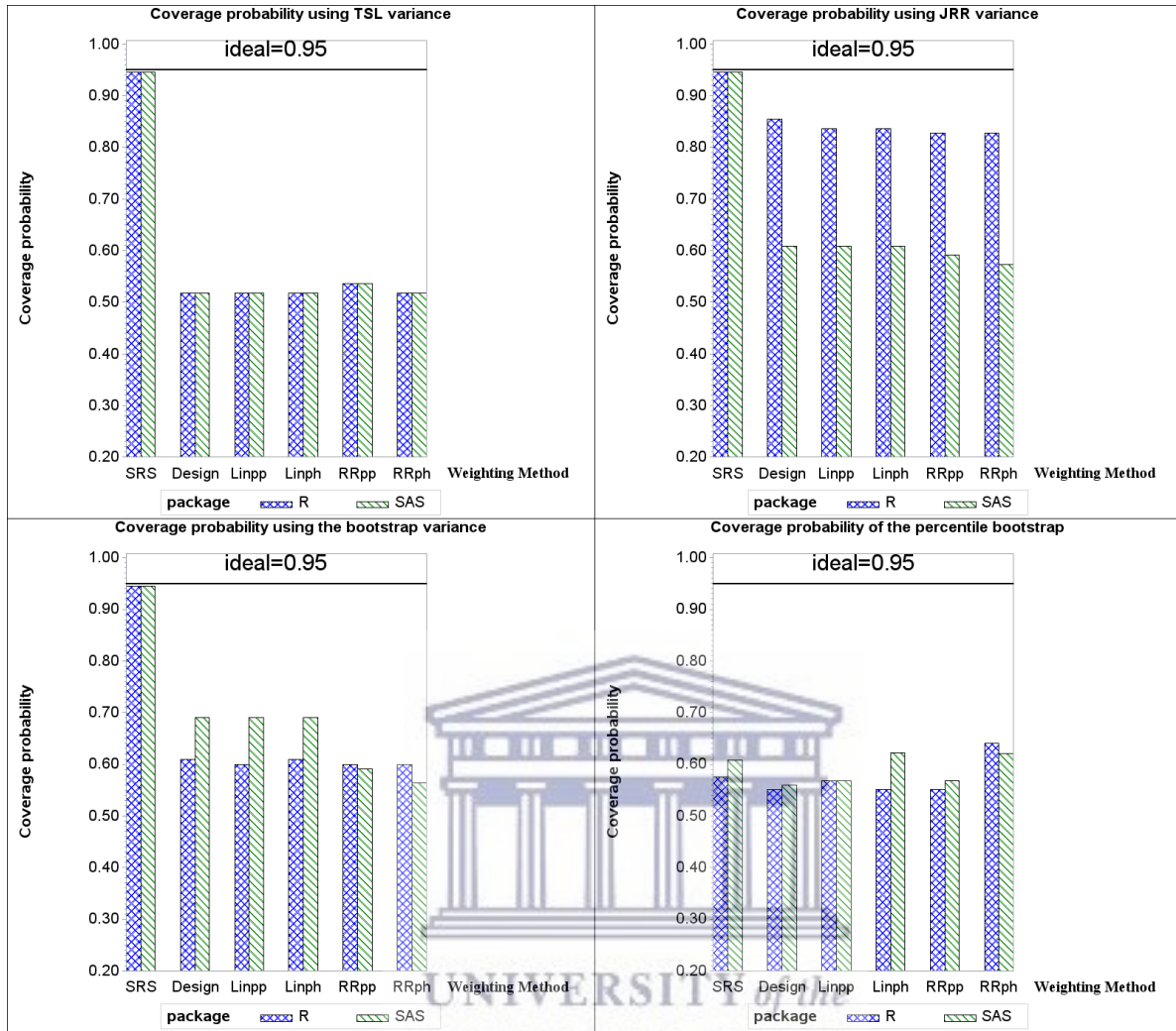


Figure C. 5: The coverage probabilities for β_6 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

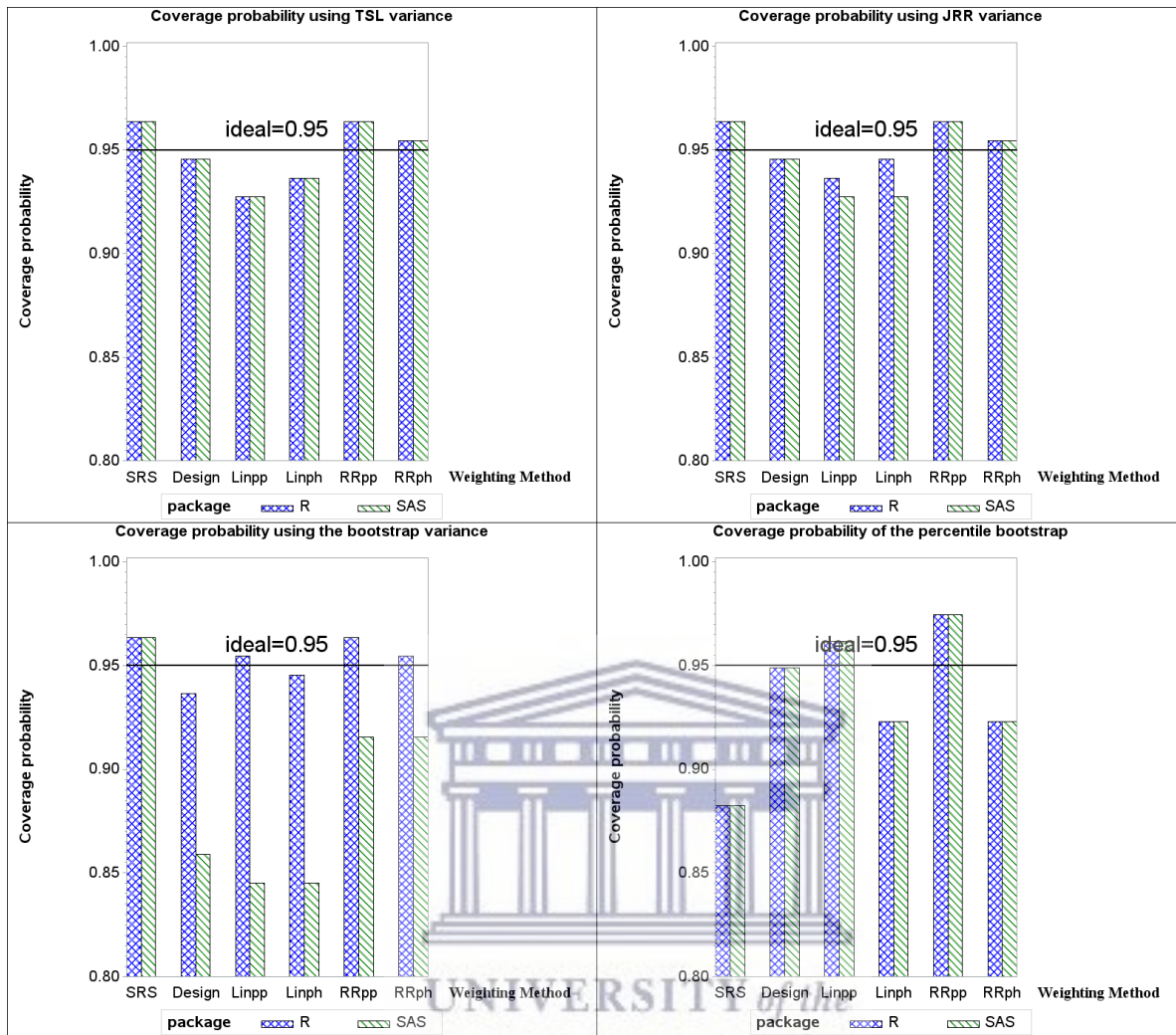


Figure C. 6: The coverage probabilities for β_8 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

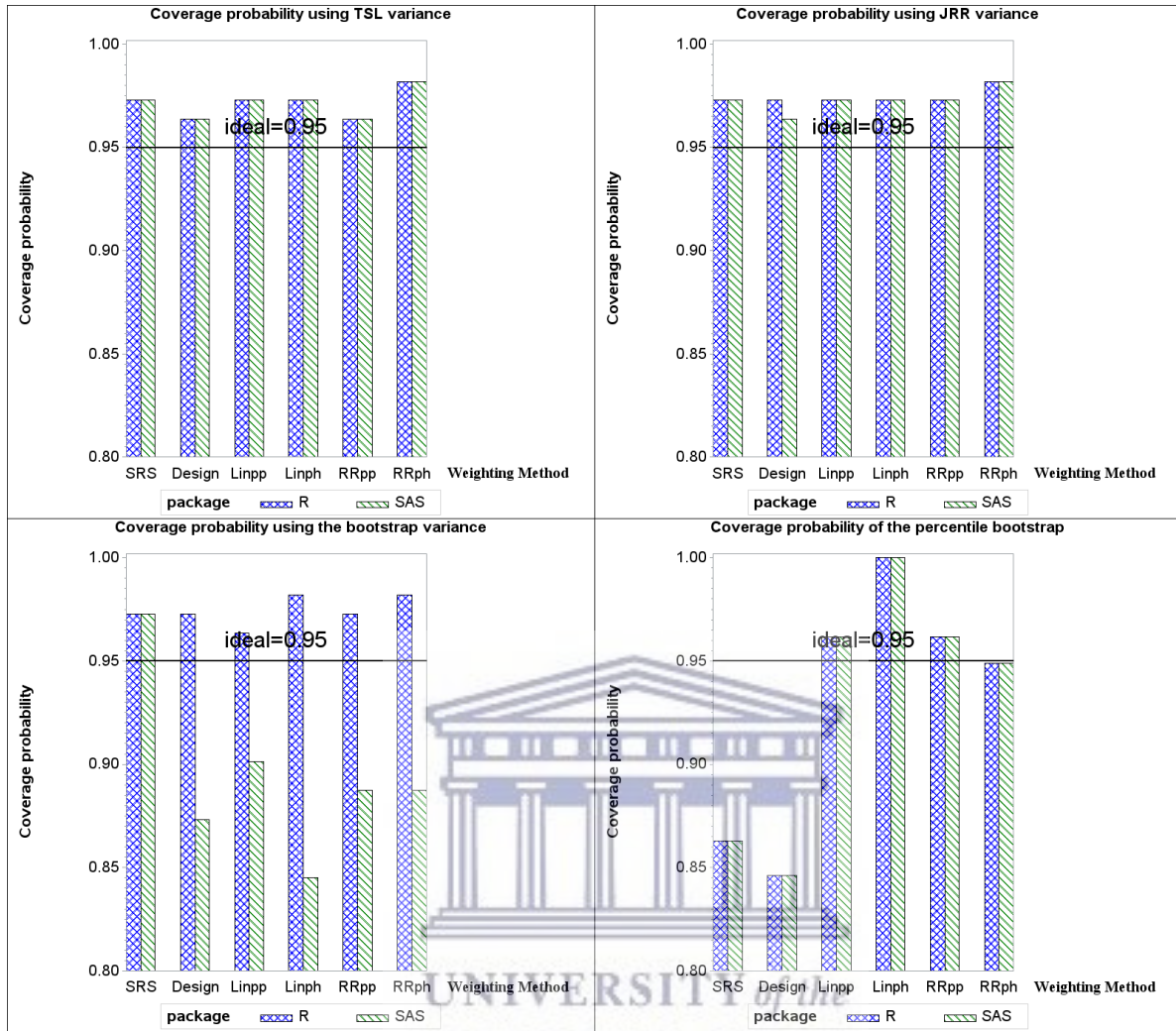


Figure C. 7: The coverage probabilities for β_9 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

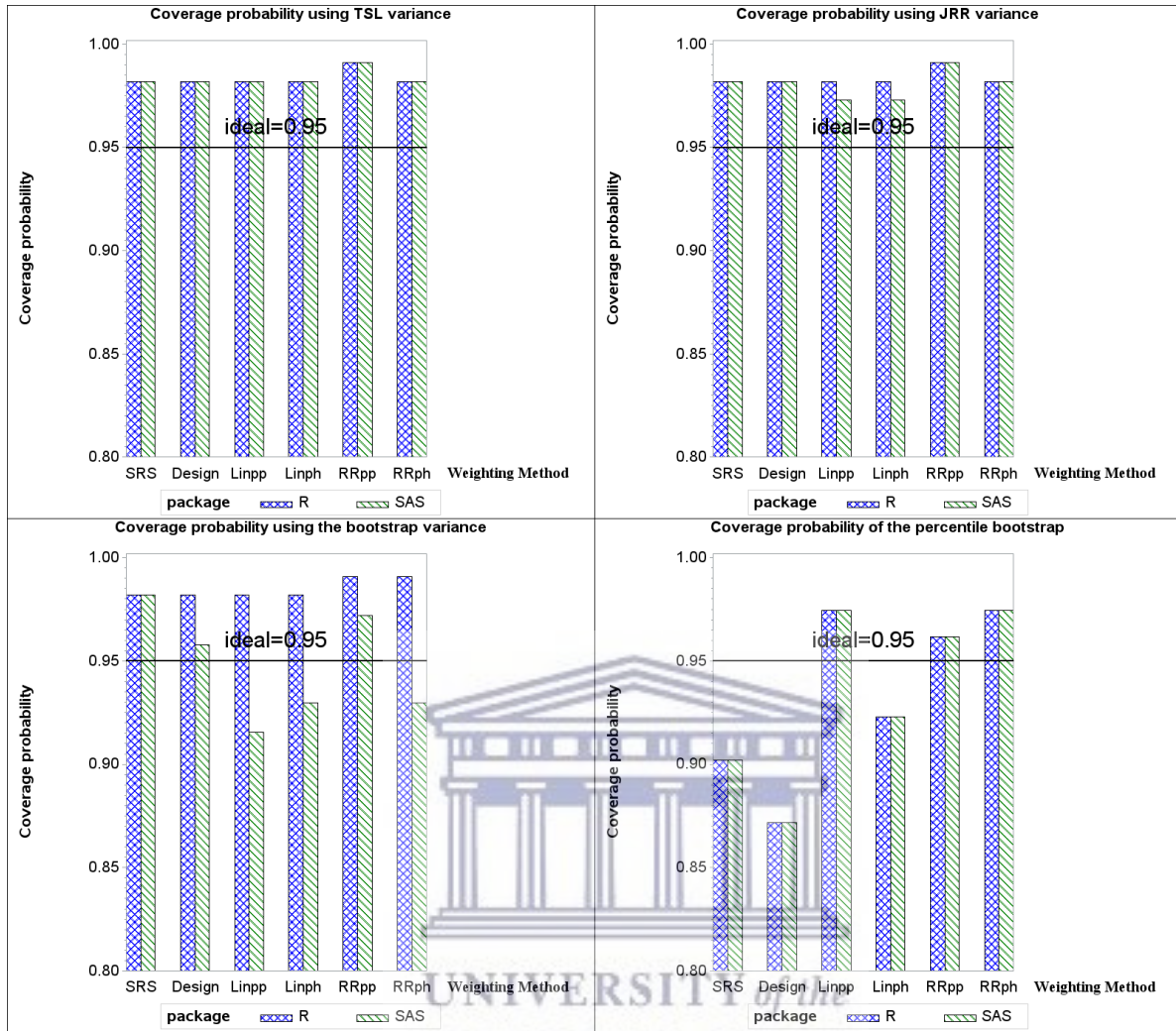


Figure C. 8: The coverage probabilities for β_{10} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

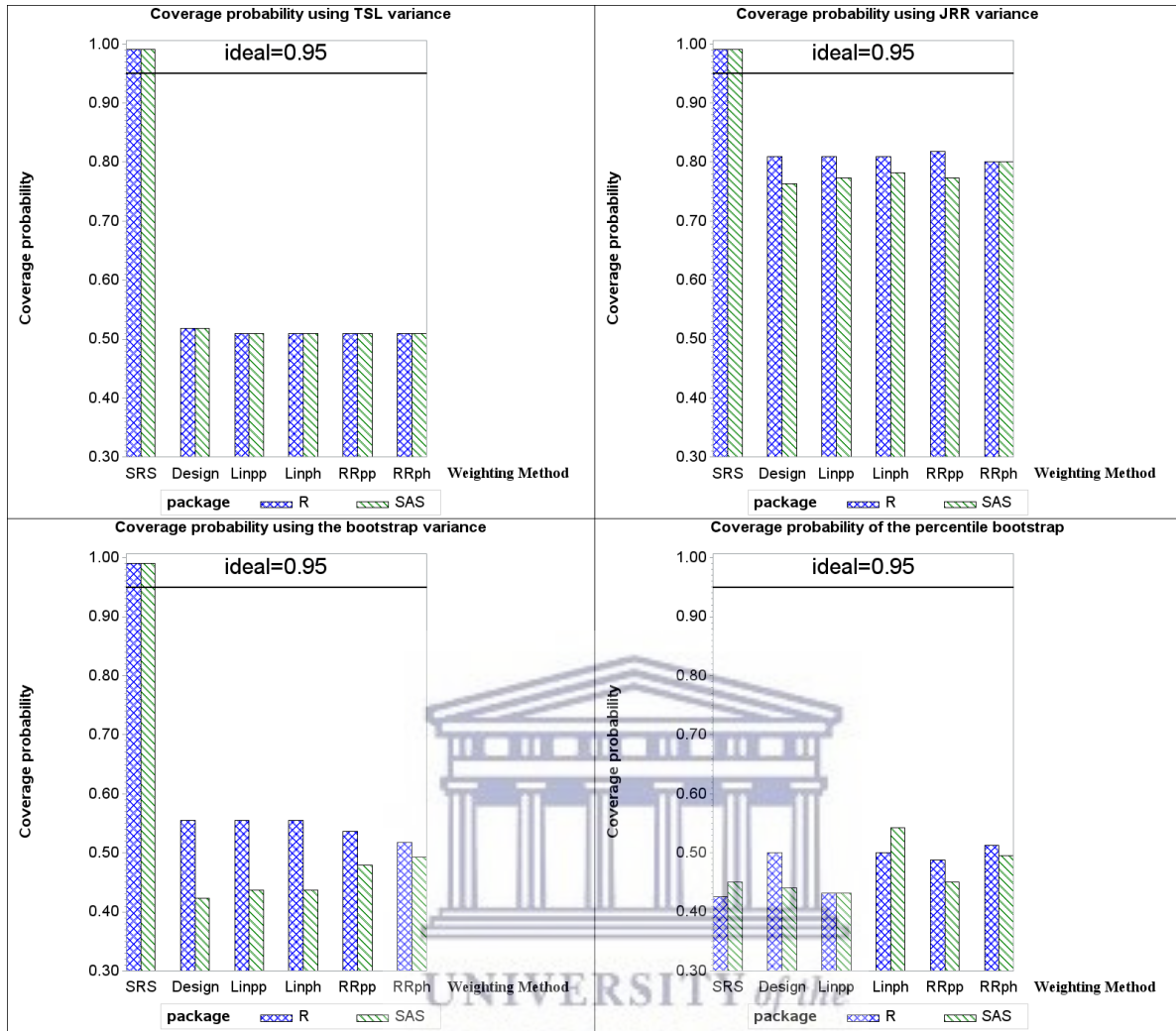


Figure C. 9: The coverage probabilities for β_{11} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

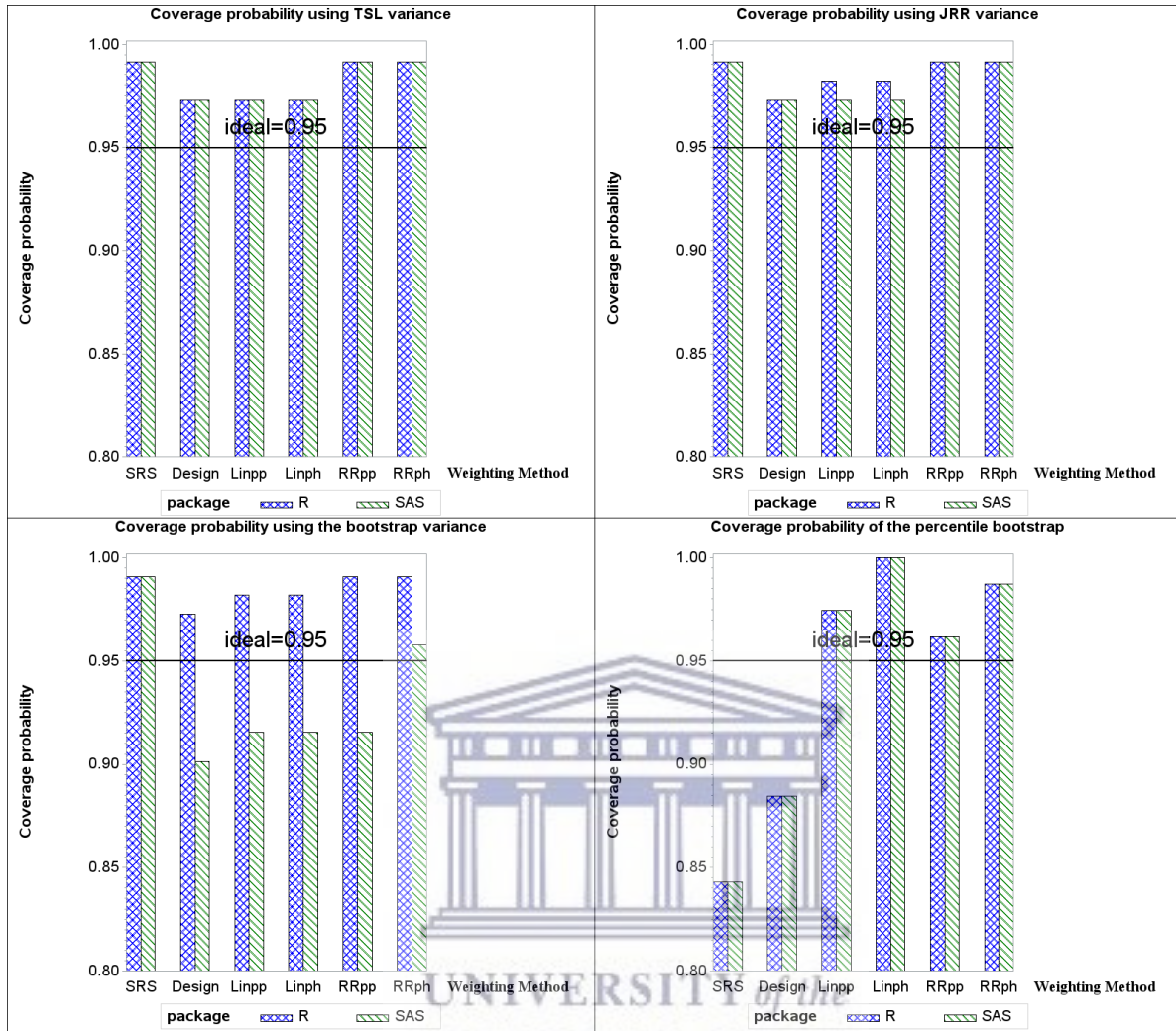


Figure C. 10: The coverage probabilities for β_{12} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

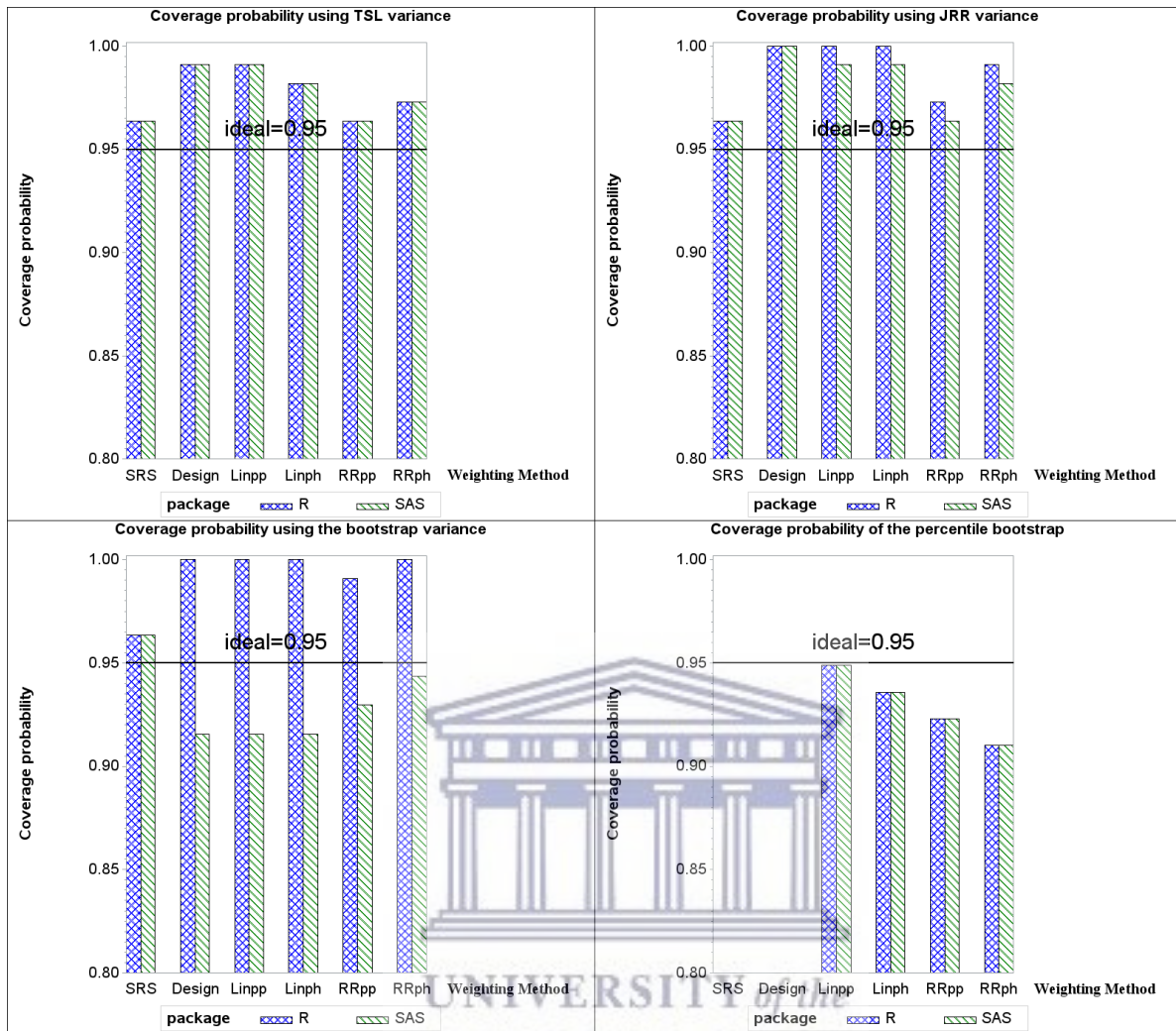


Figure C. 11: The coverage probabilities for β_{13} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

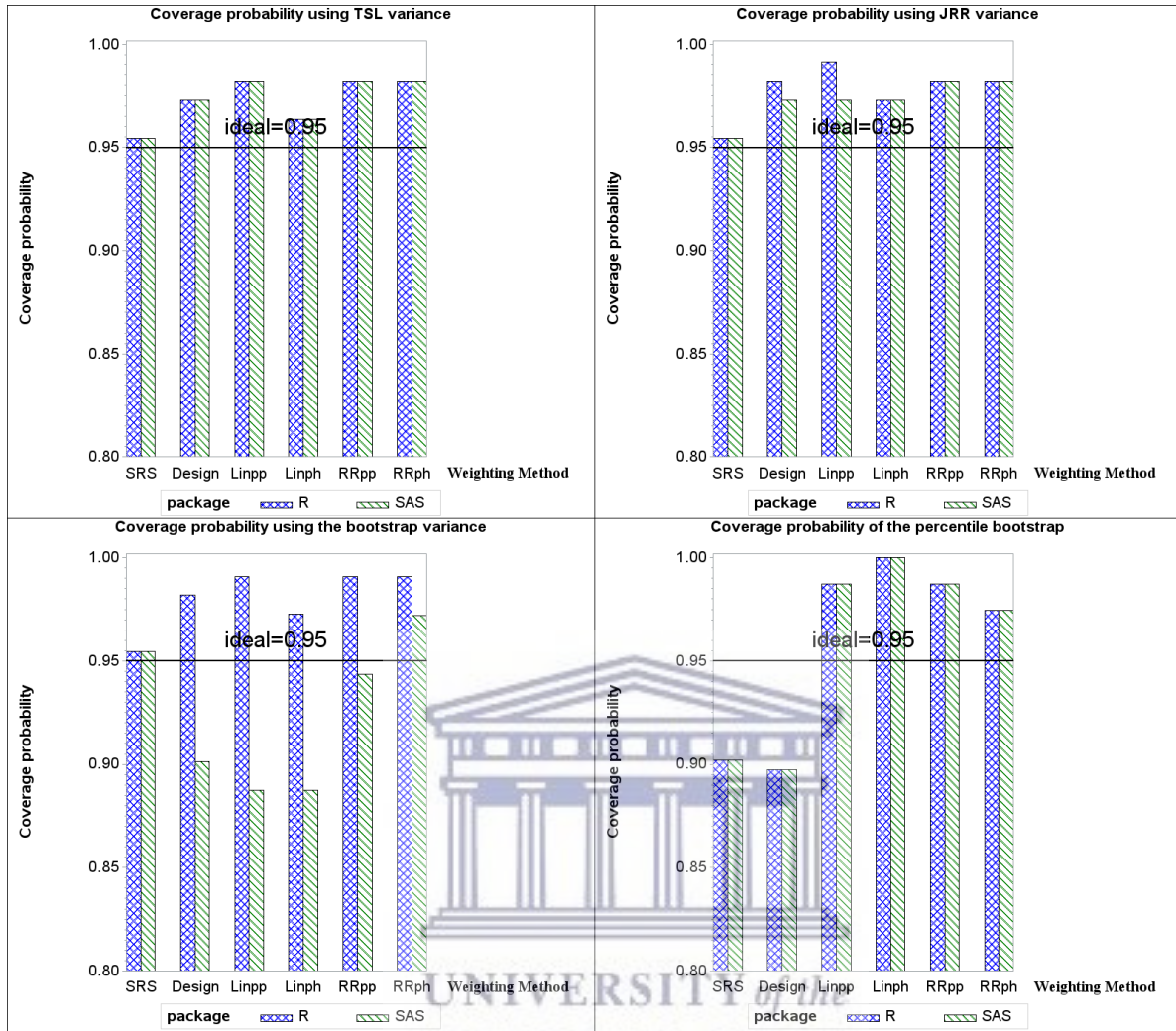


Figure C. 12: The coverage probabilities for β_{14} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

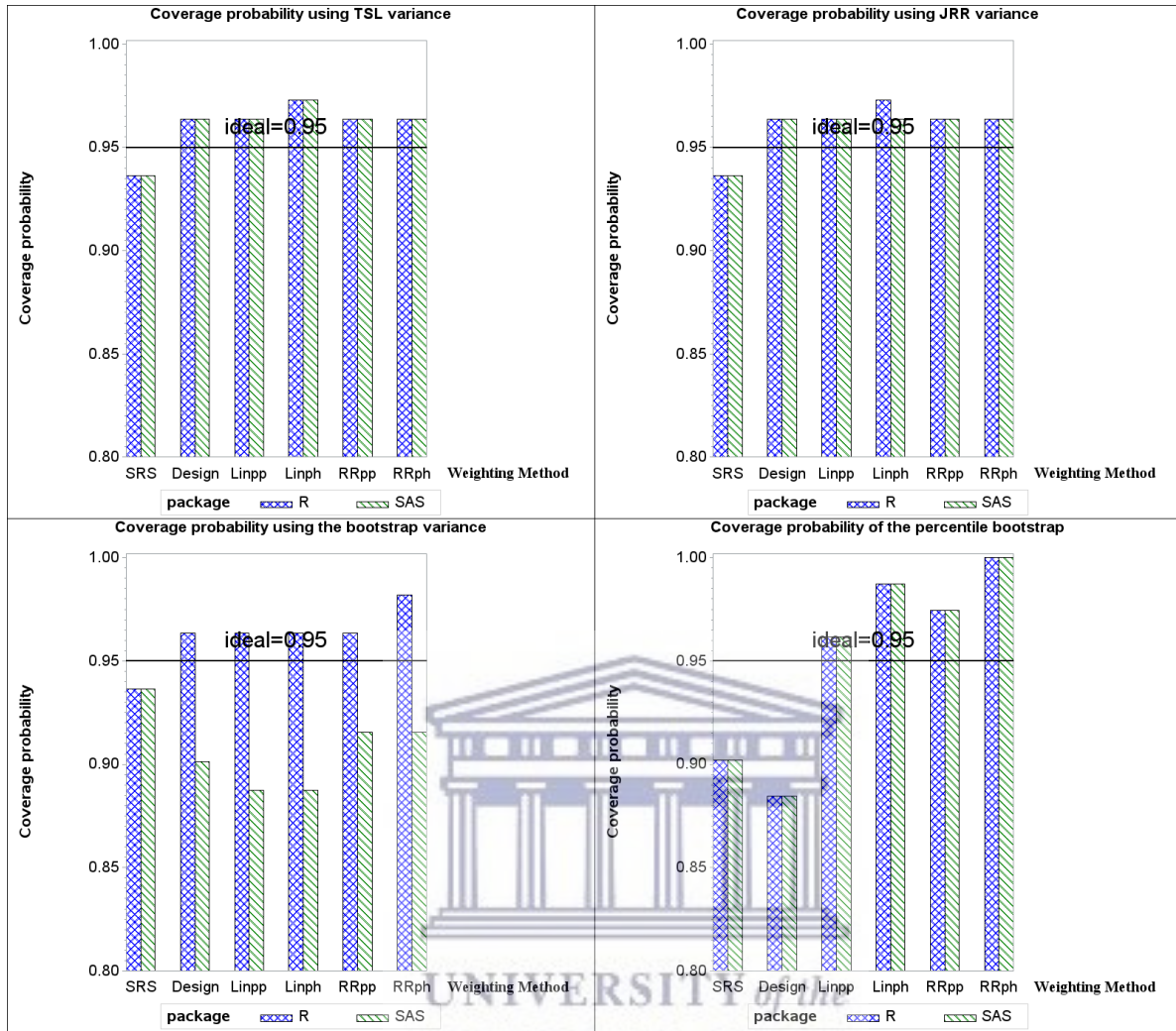


Figure C. 13: The coverage probabilities for β_{16} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

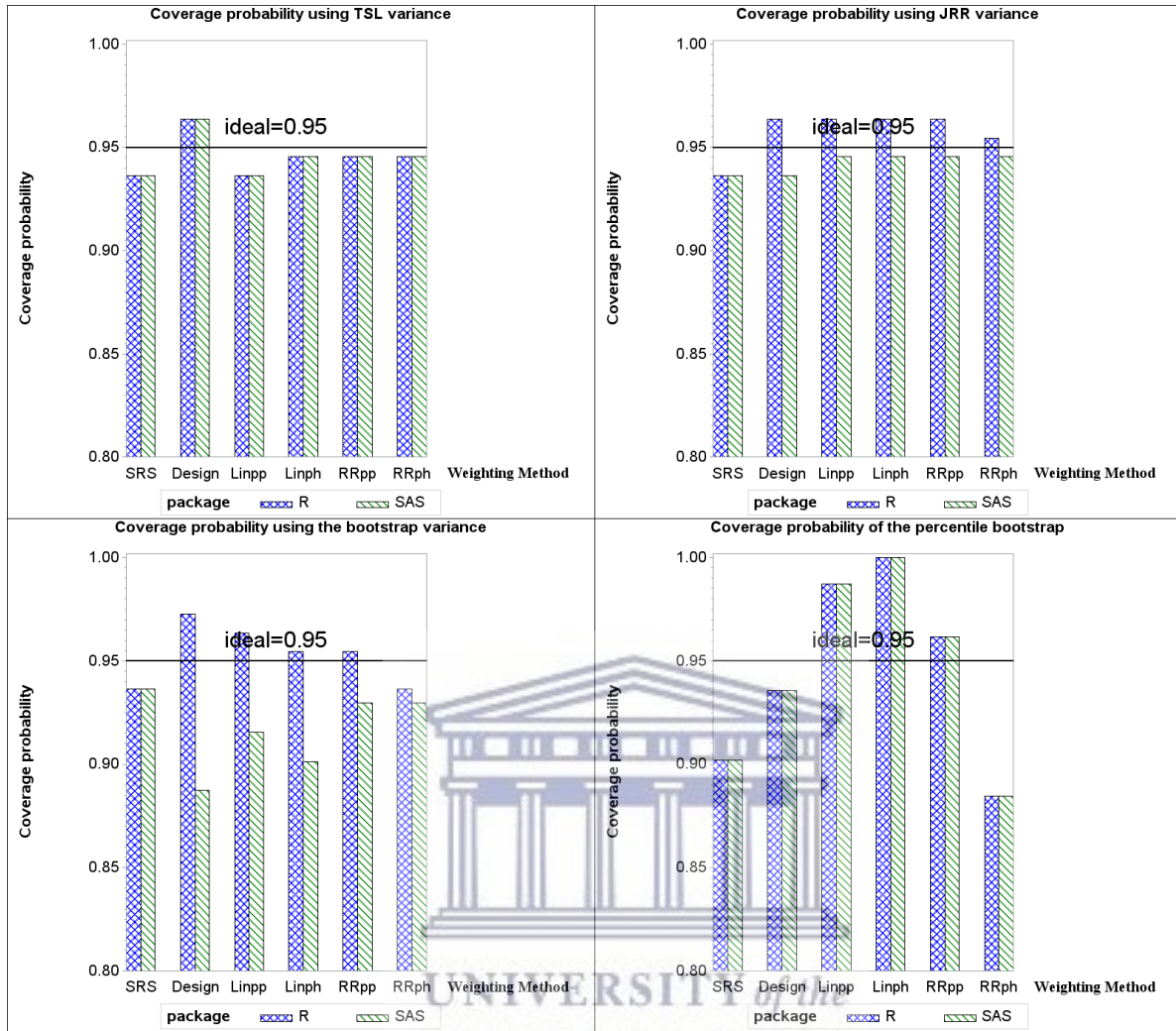


Figure C. 14: The coverage probabilities for β_{16} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

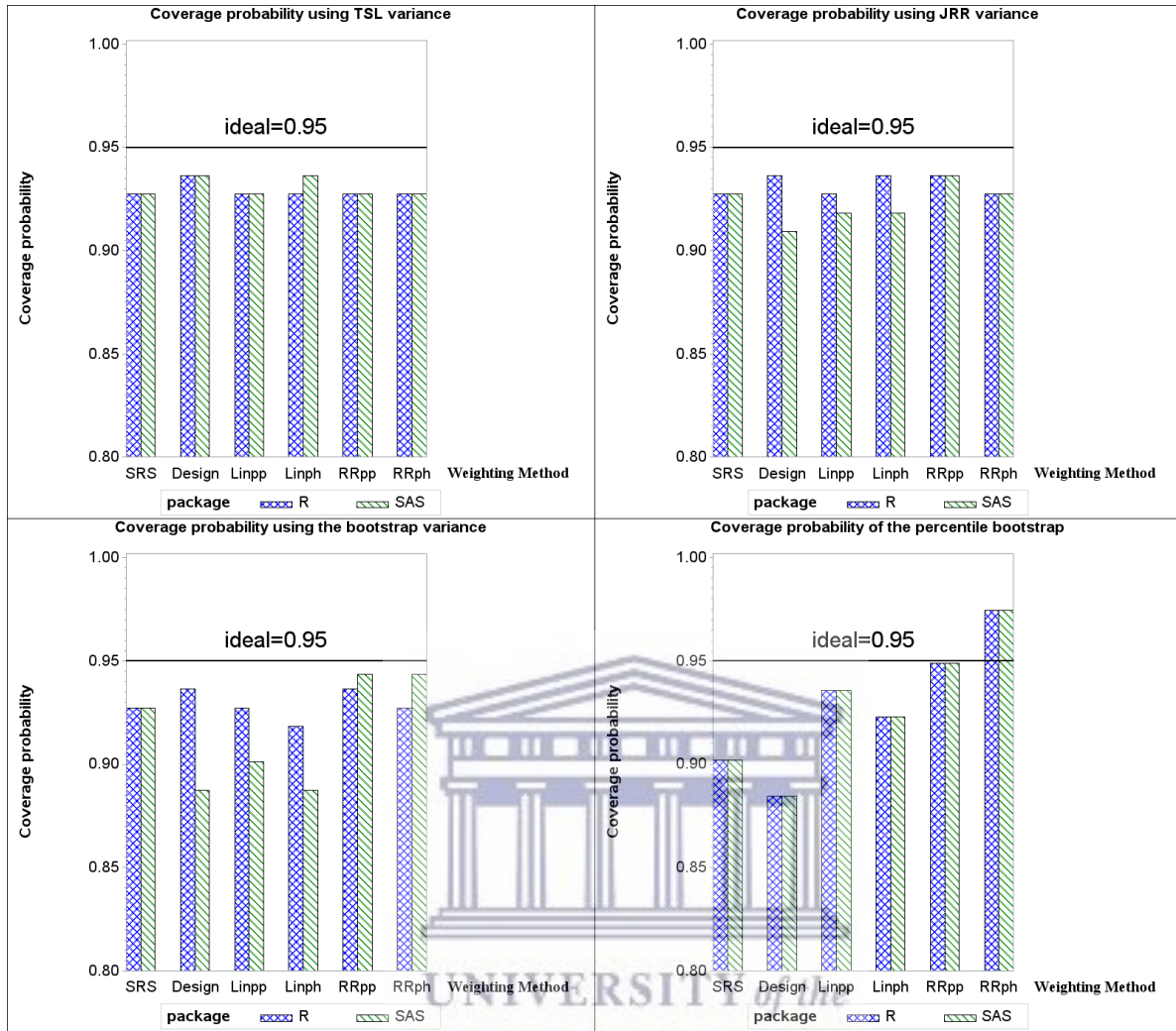


Figure C. 15: The coverage probabilities for β_{17} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

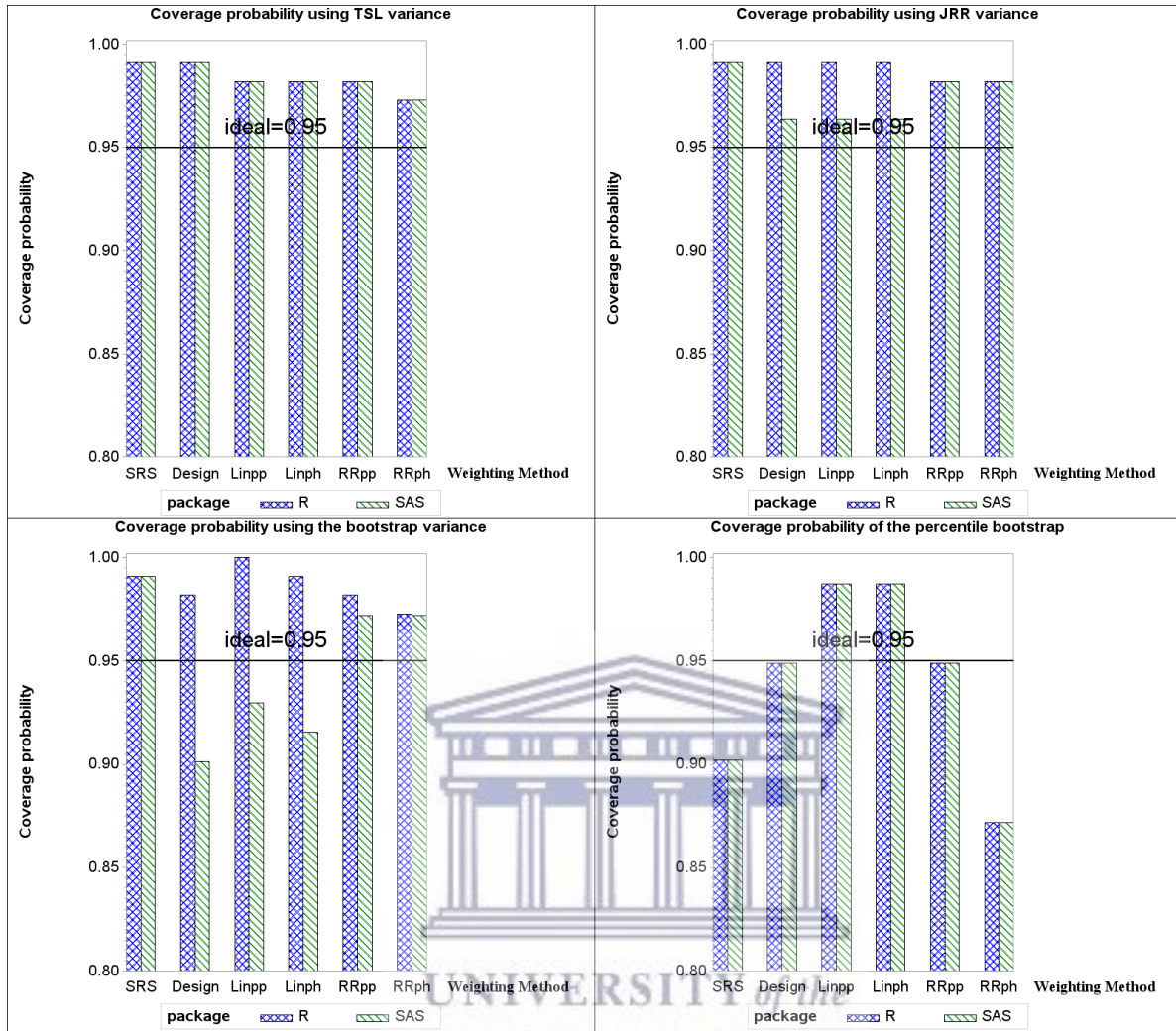


Figure C. 16: The coverage probabilities for β_{18} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

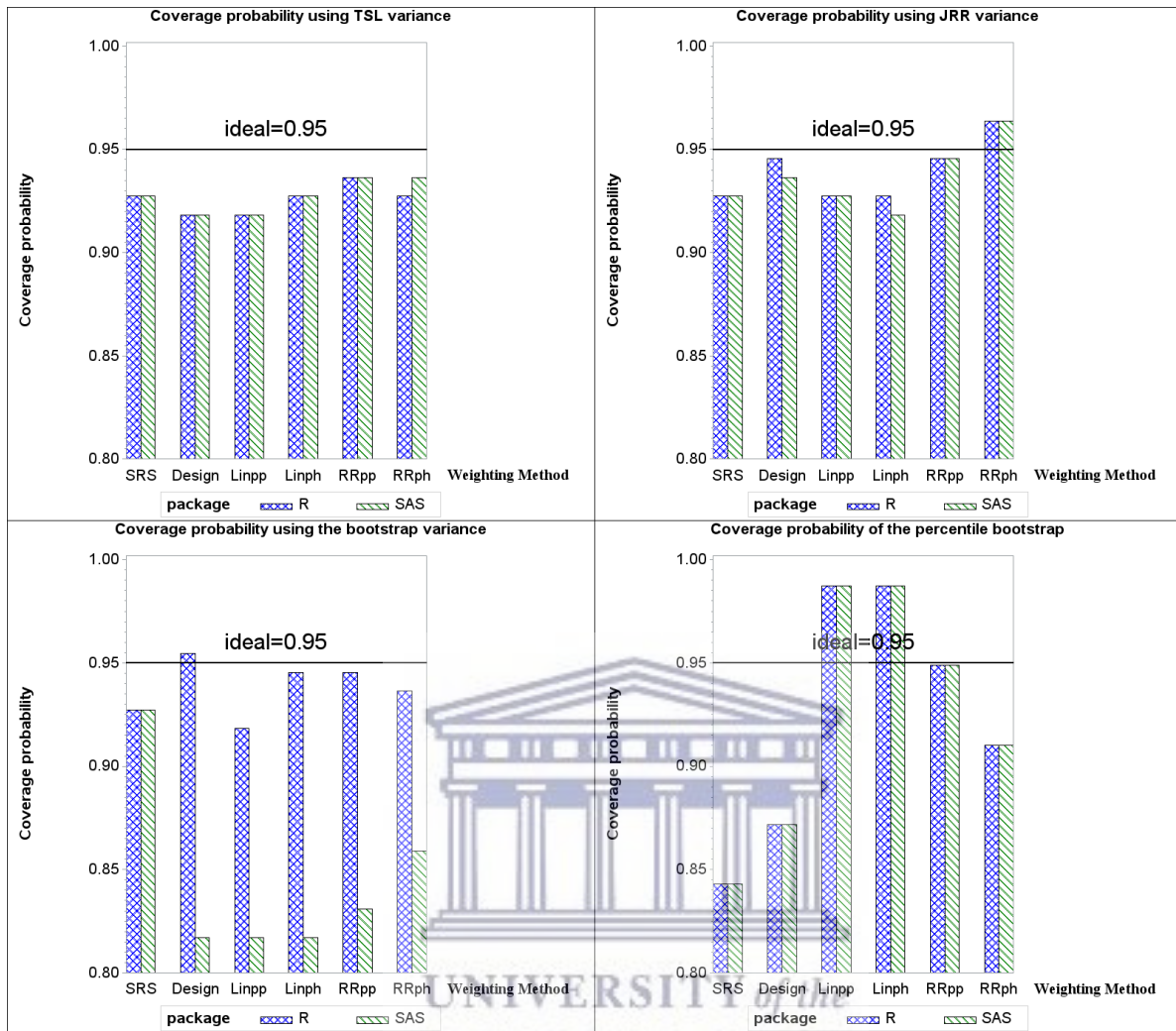


Figure C. 17: The coverage probabilities for β_{19} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

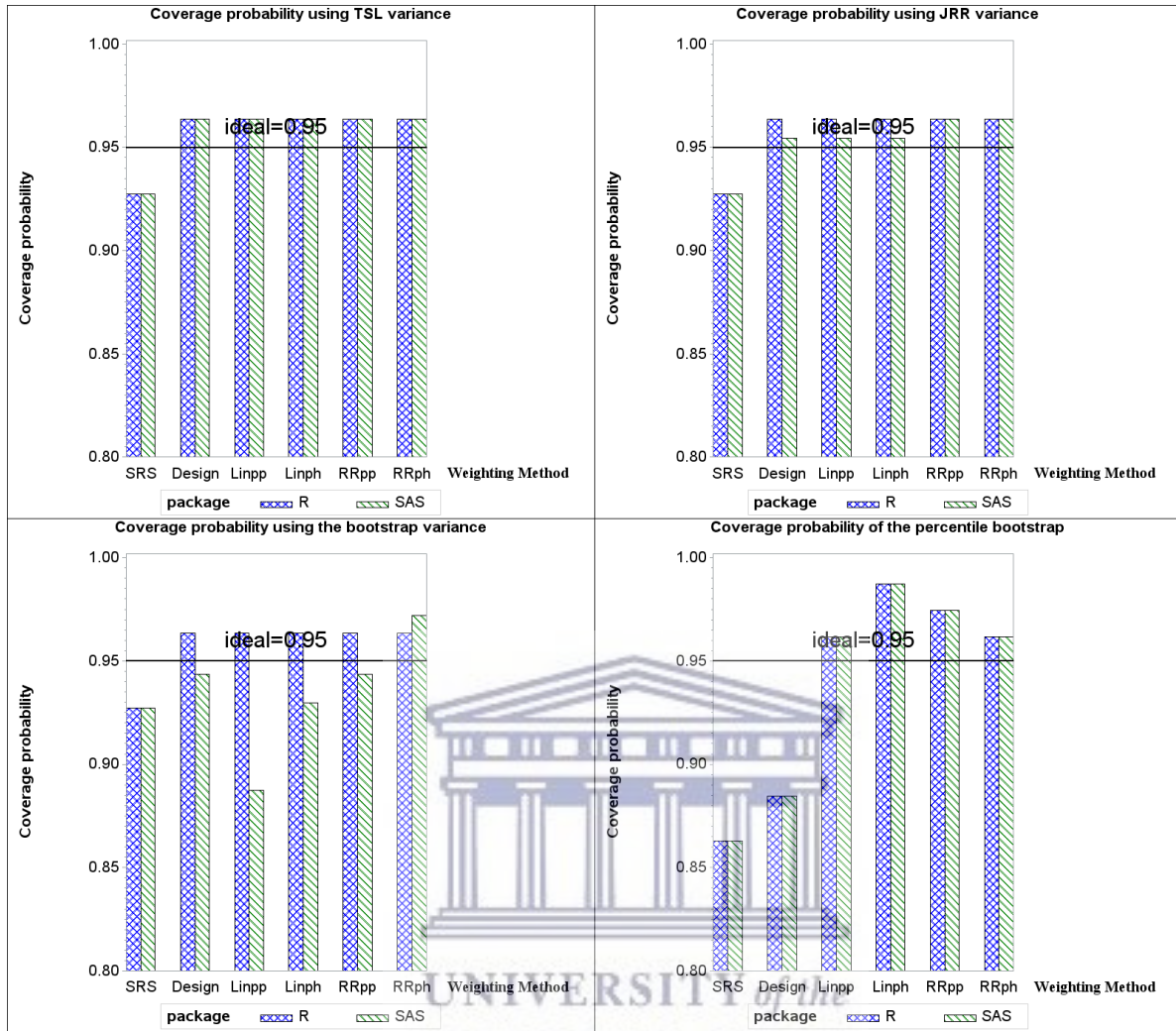


Figure C. 18: The coverage probabilities for β_{20} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

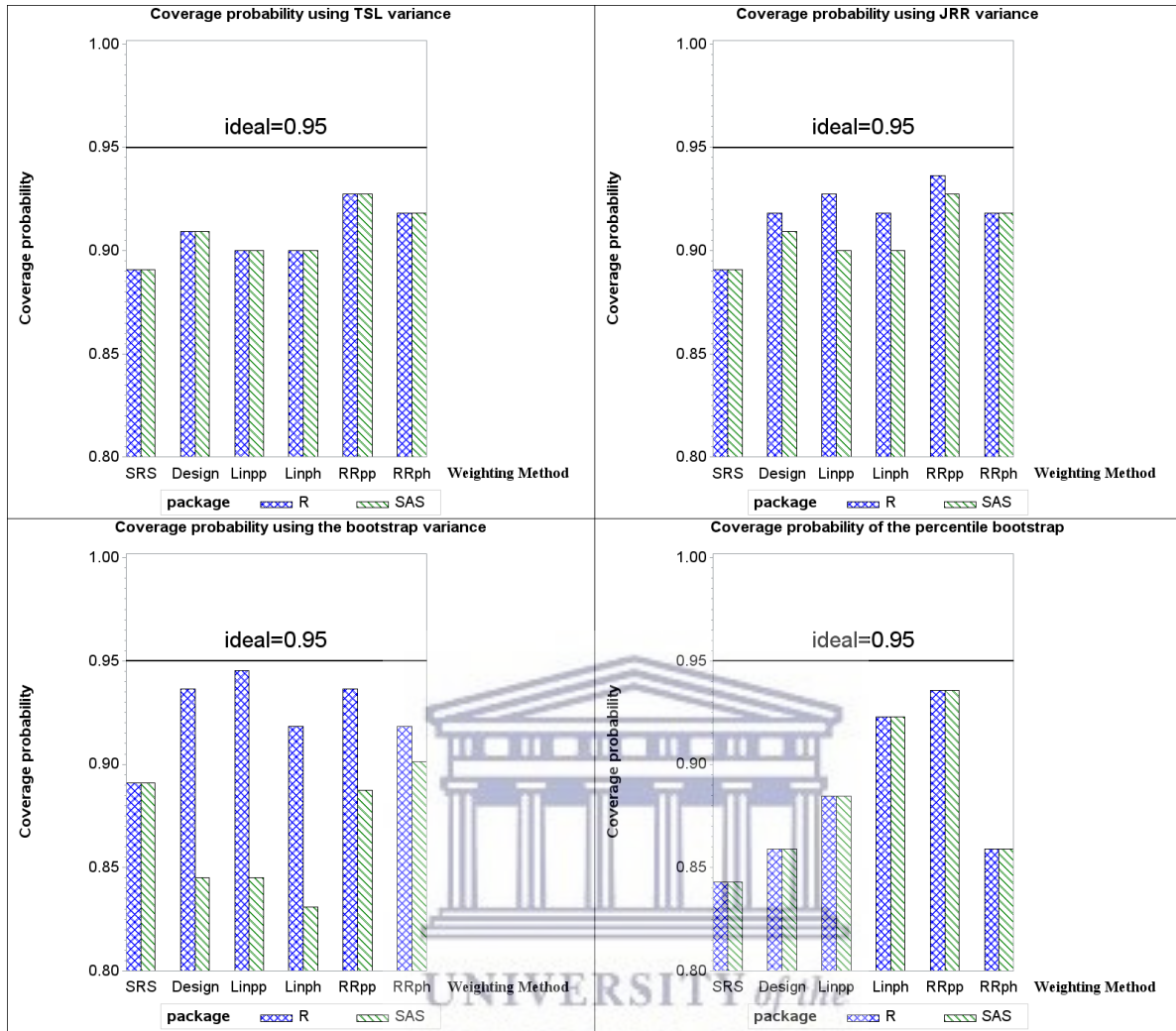


Figure C. 19: The coverage probabilities for β_{21} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

Appendix D: Confidence interval length

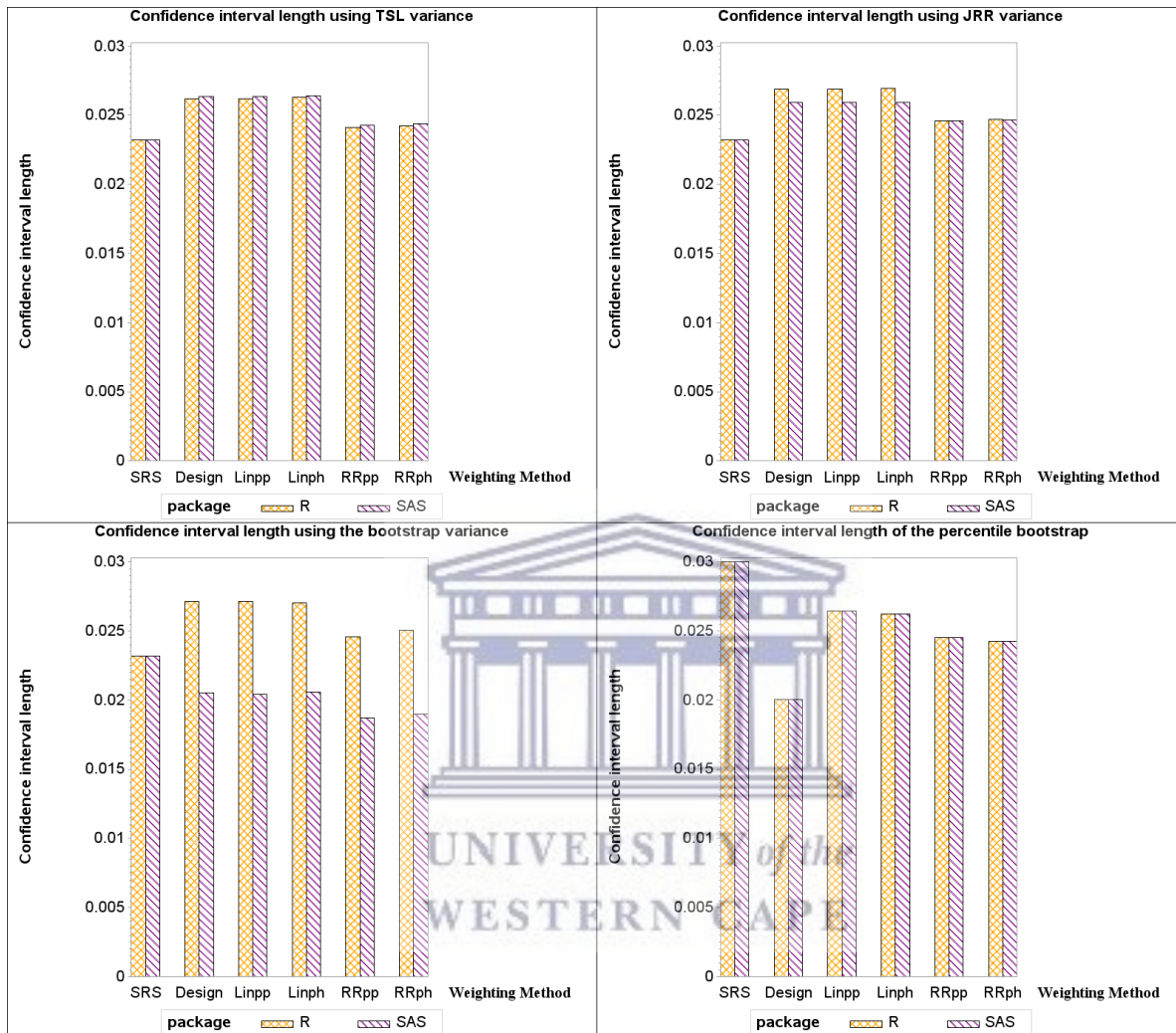


Figure D. 1: The confidence interval lengths for β_1 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

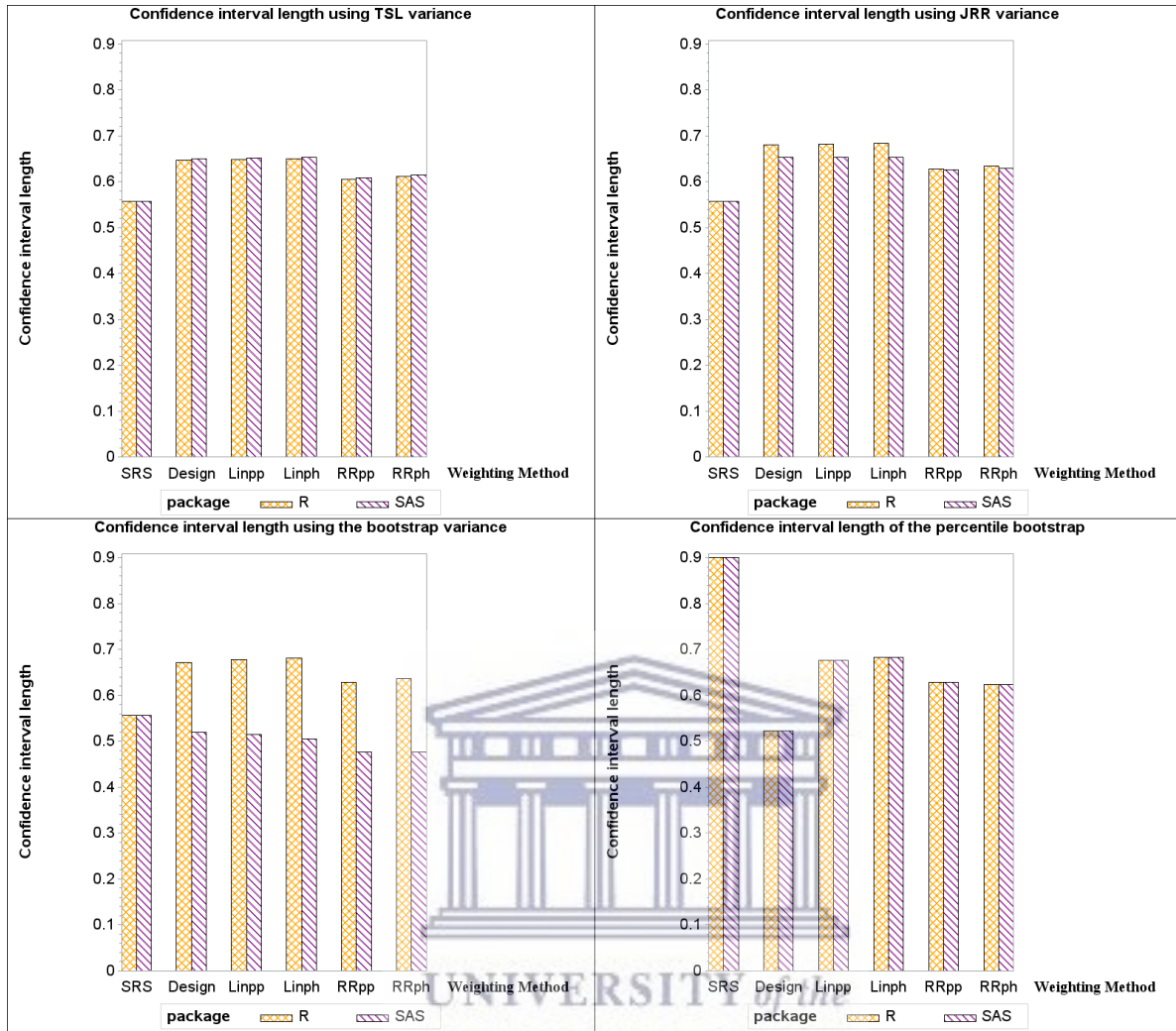


Figure D. 2: The confidence interval lengths for β_2 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

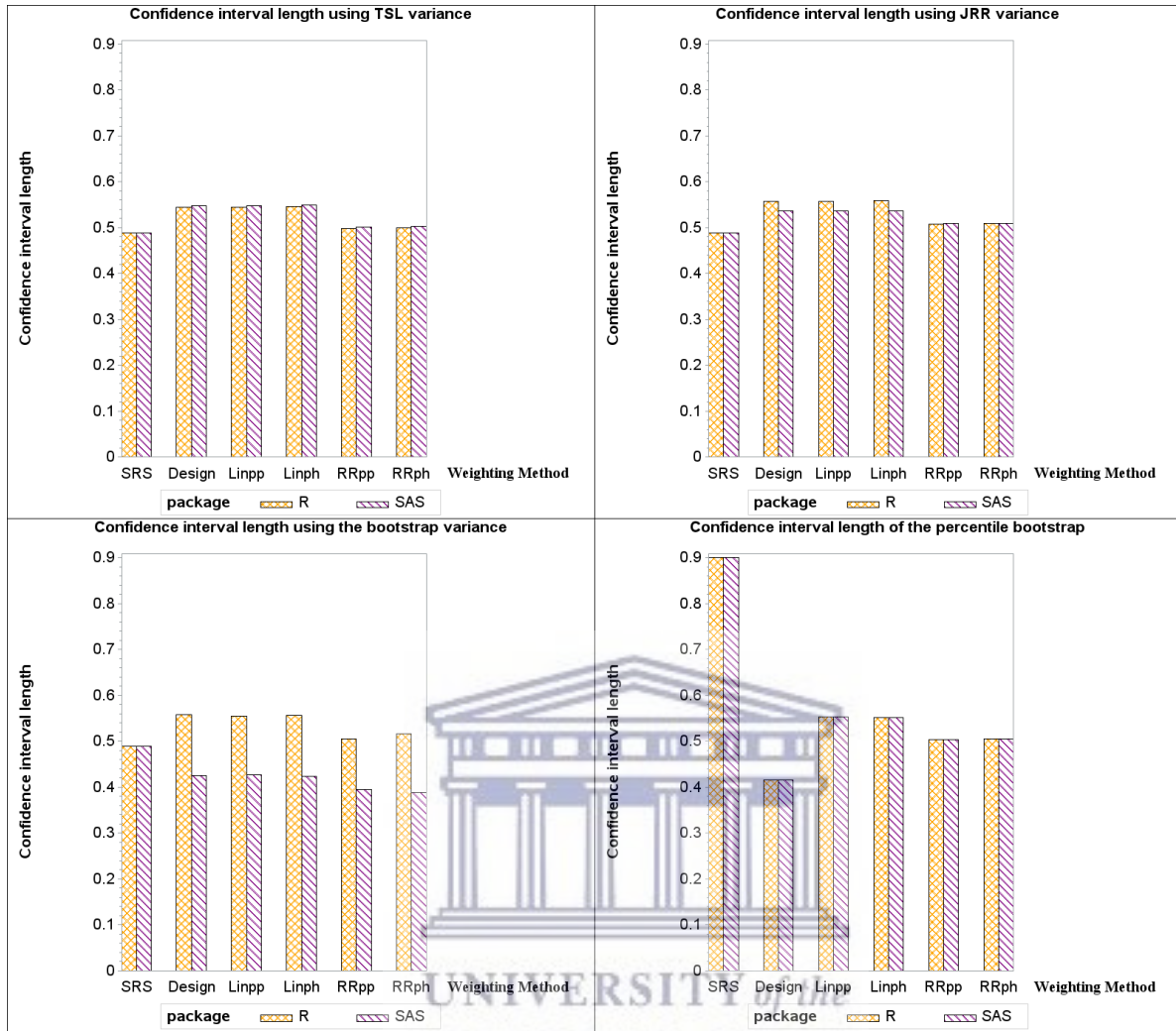


Figure D. 3: The confidence interval lengths for β_3 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

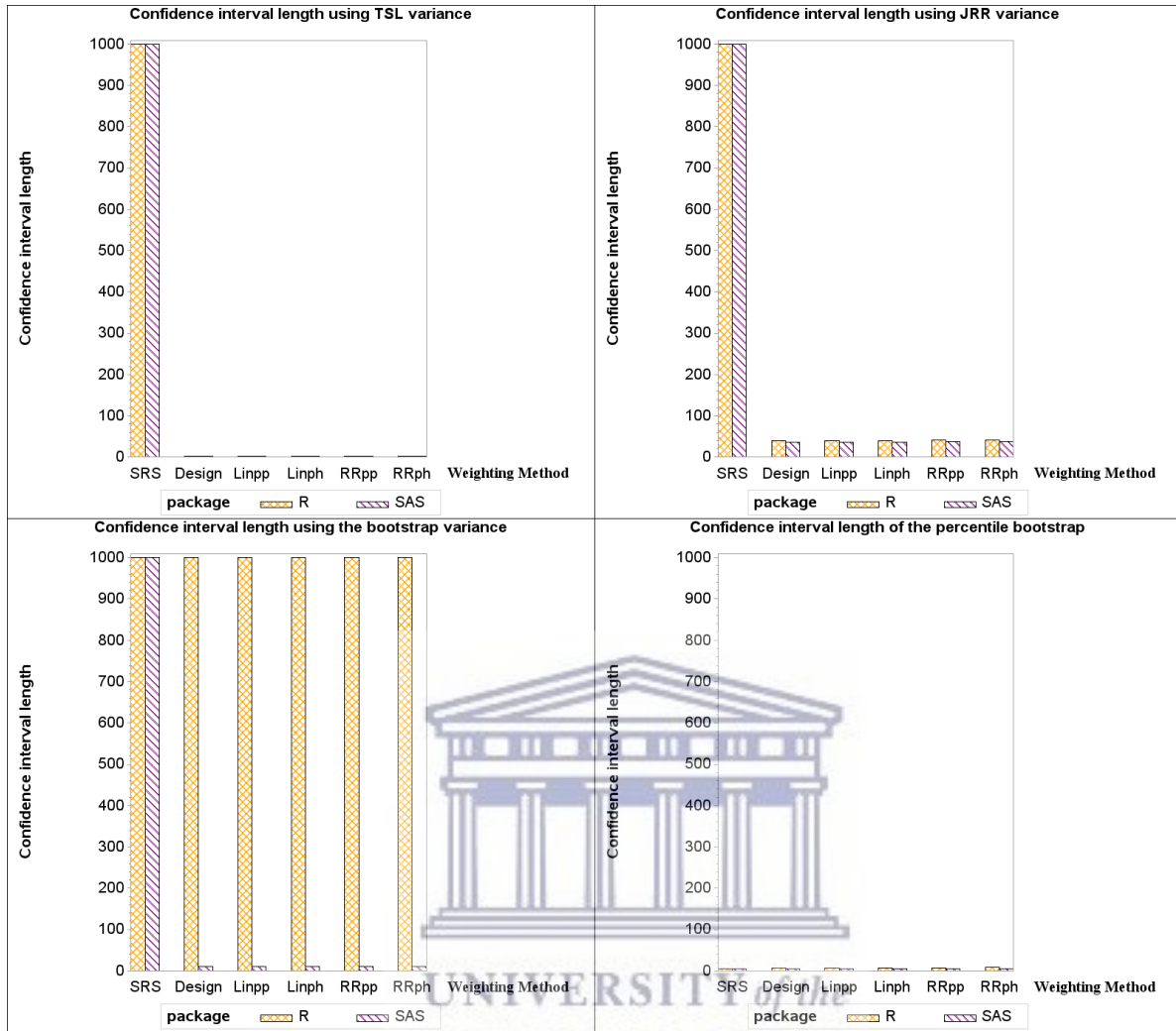


Figure D. 4: The confidence interval lengths for β_5 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

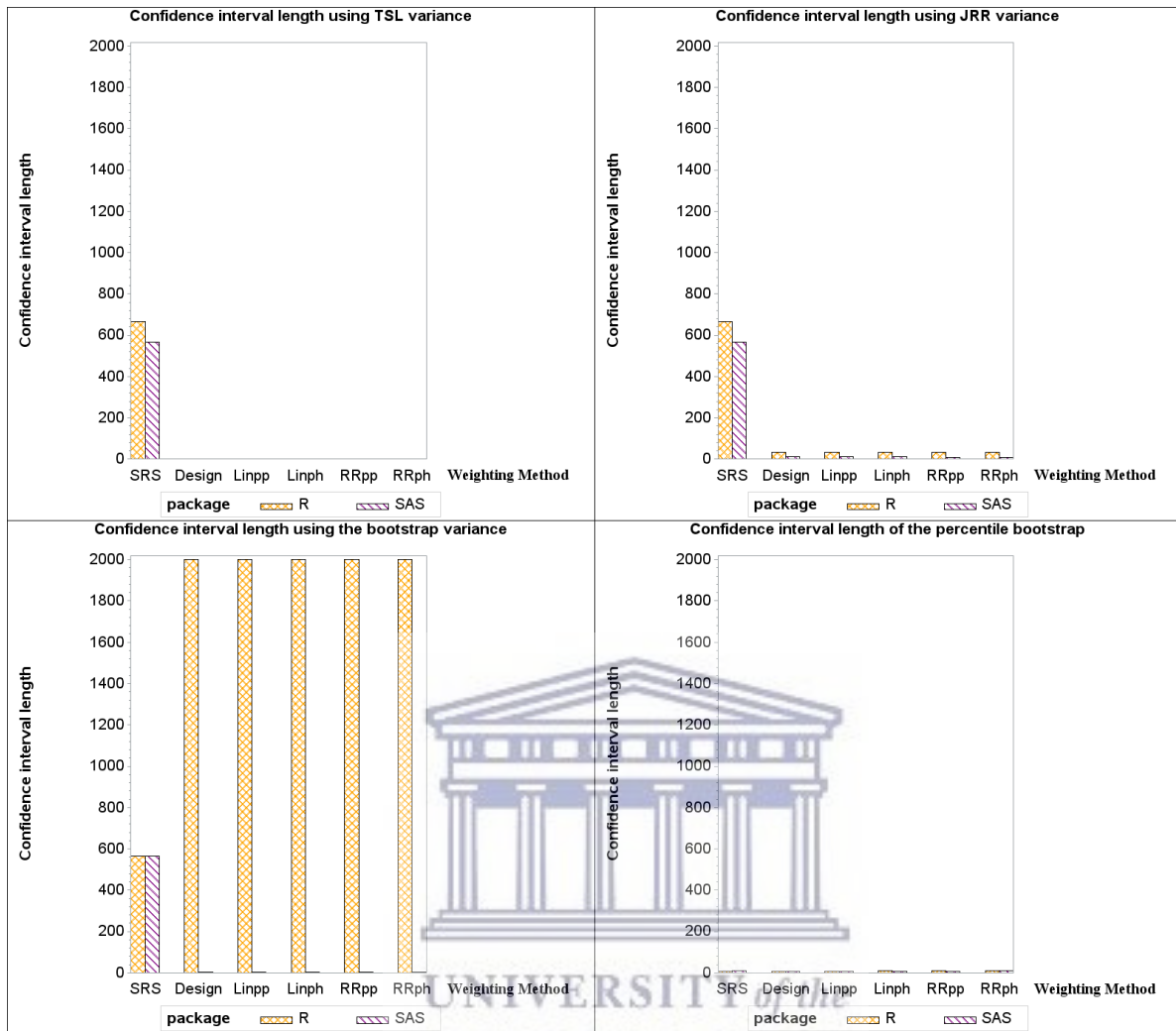


Figure D. 5: The confidence interval lengths for β_6 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

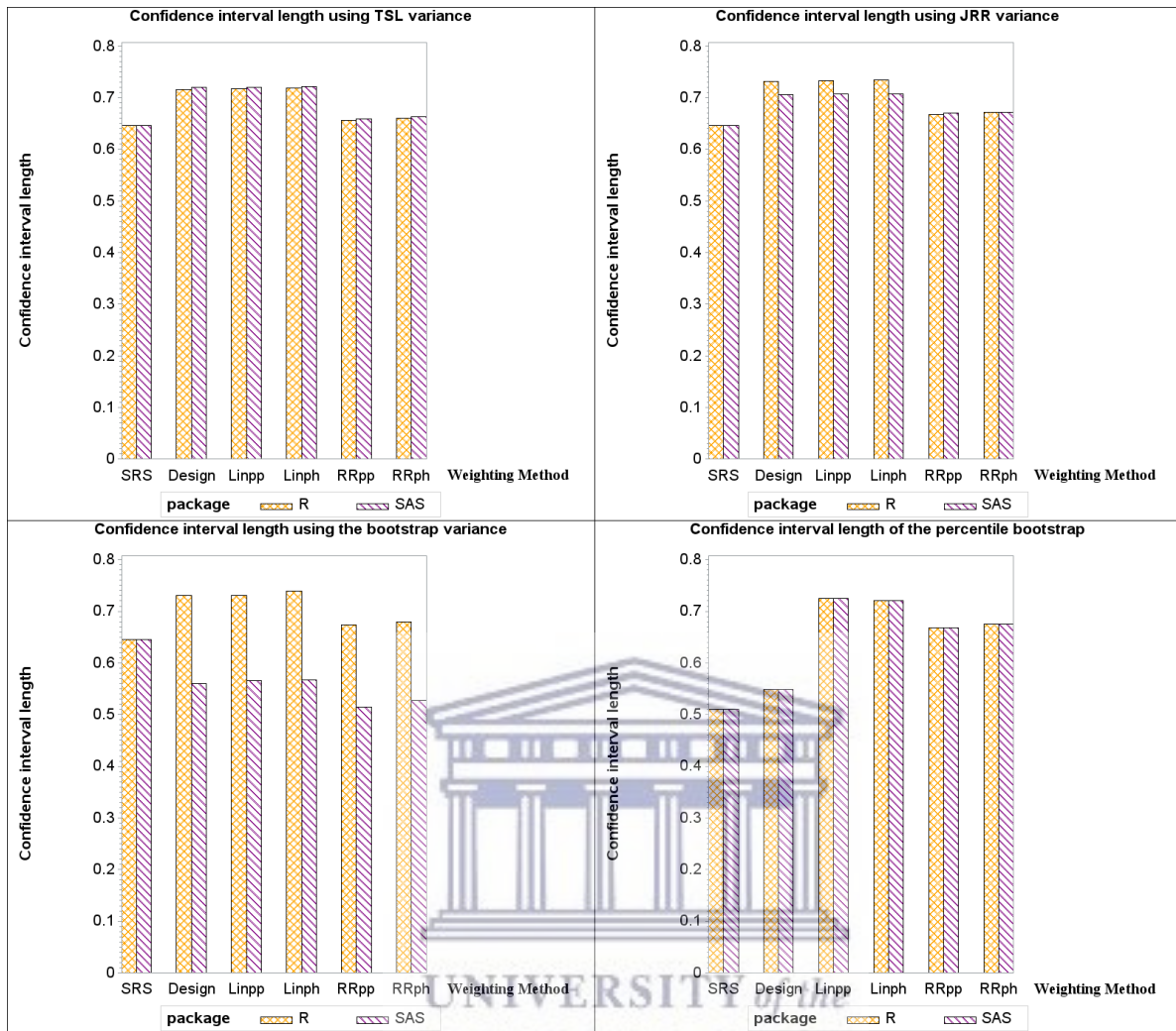


Figure D. 6: The confidence interval lengths for β_8 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

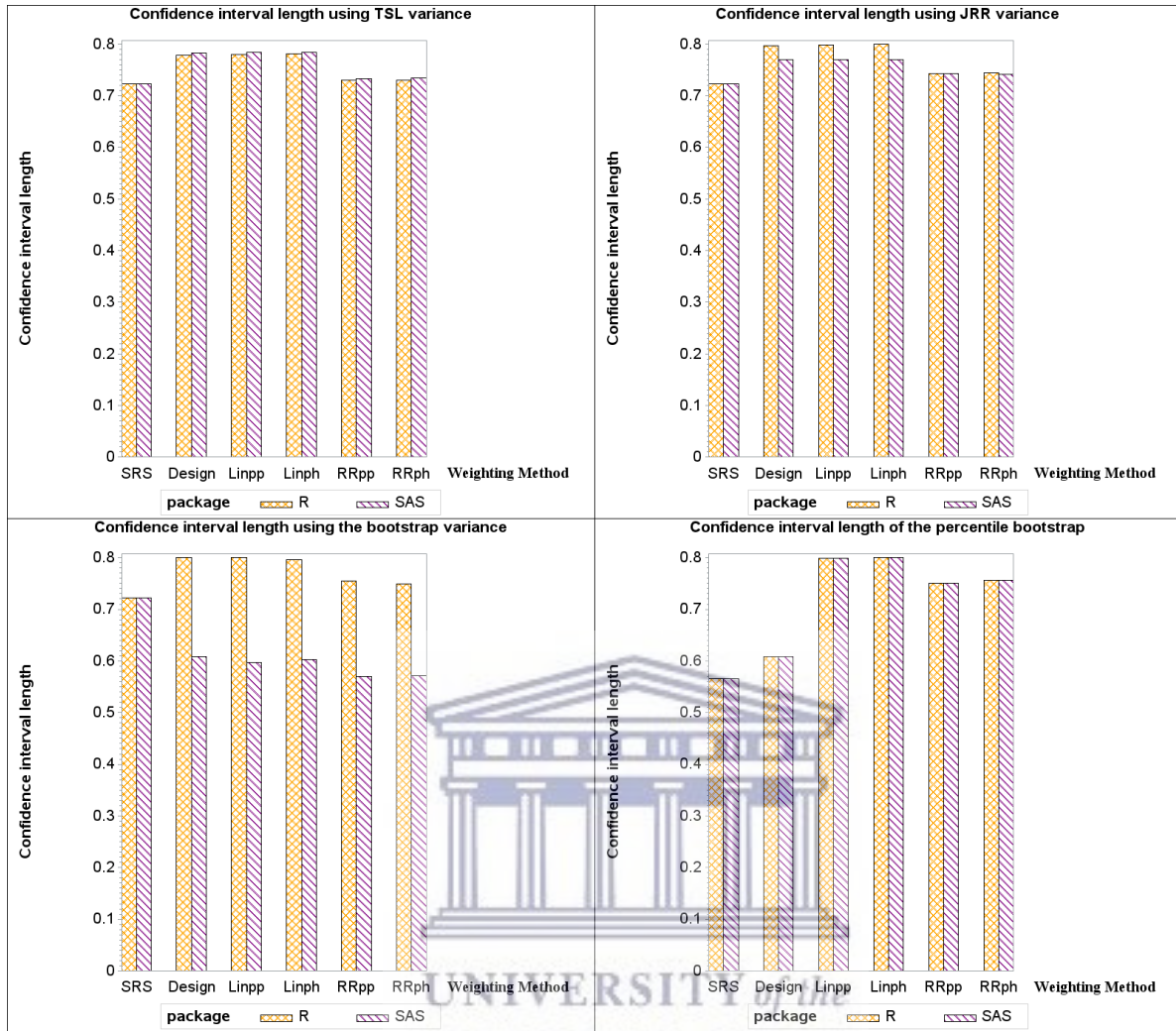


Figure D. 7: The confidence interval lengths for β_9 under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

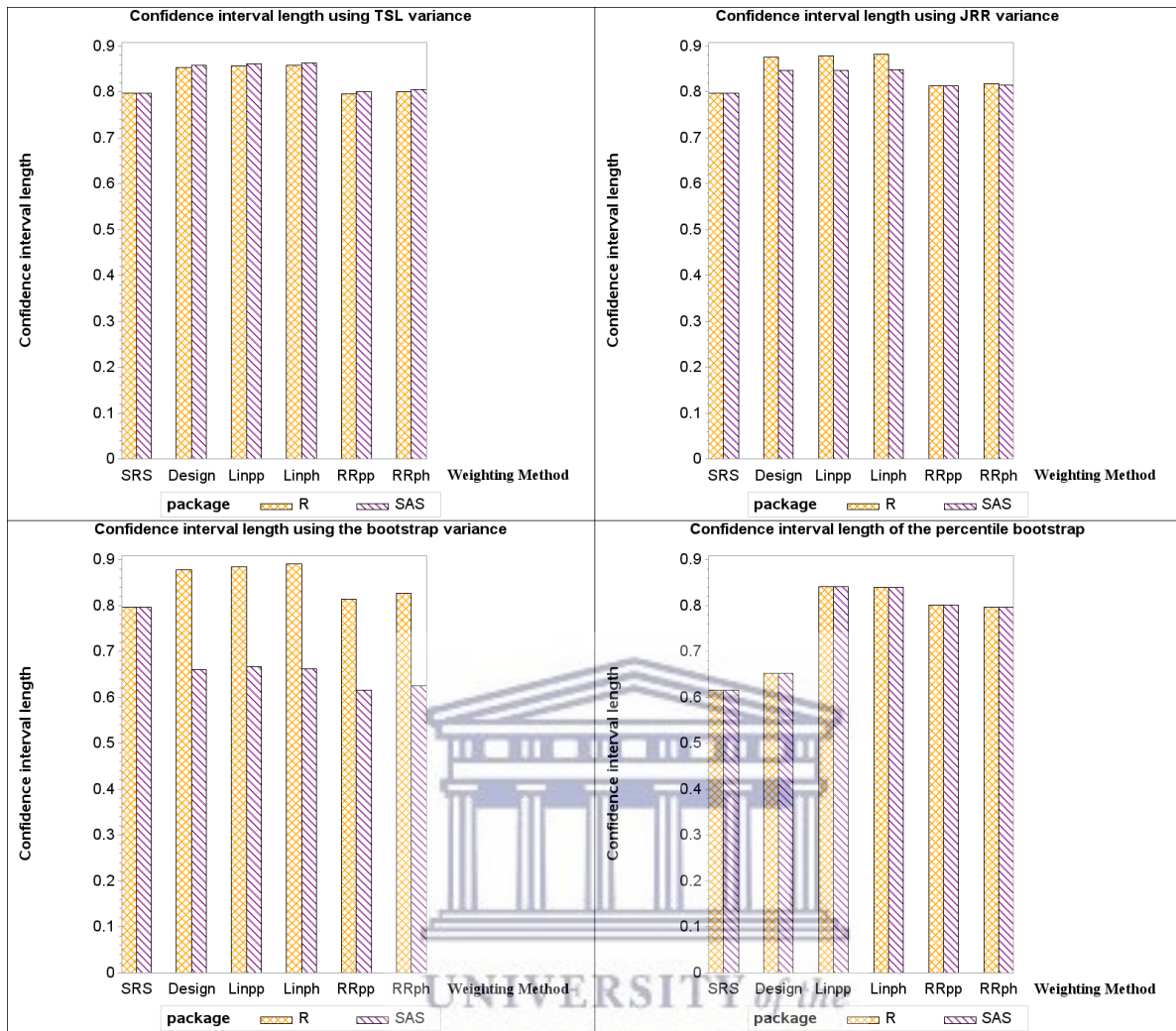


Figure D. 8: The confidence interval lengths for β_{10} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

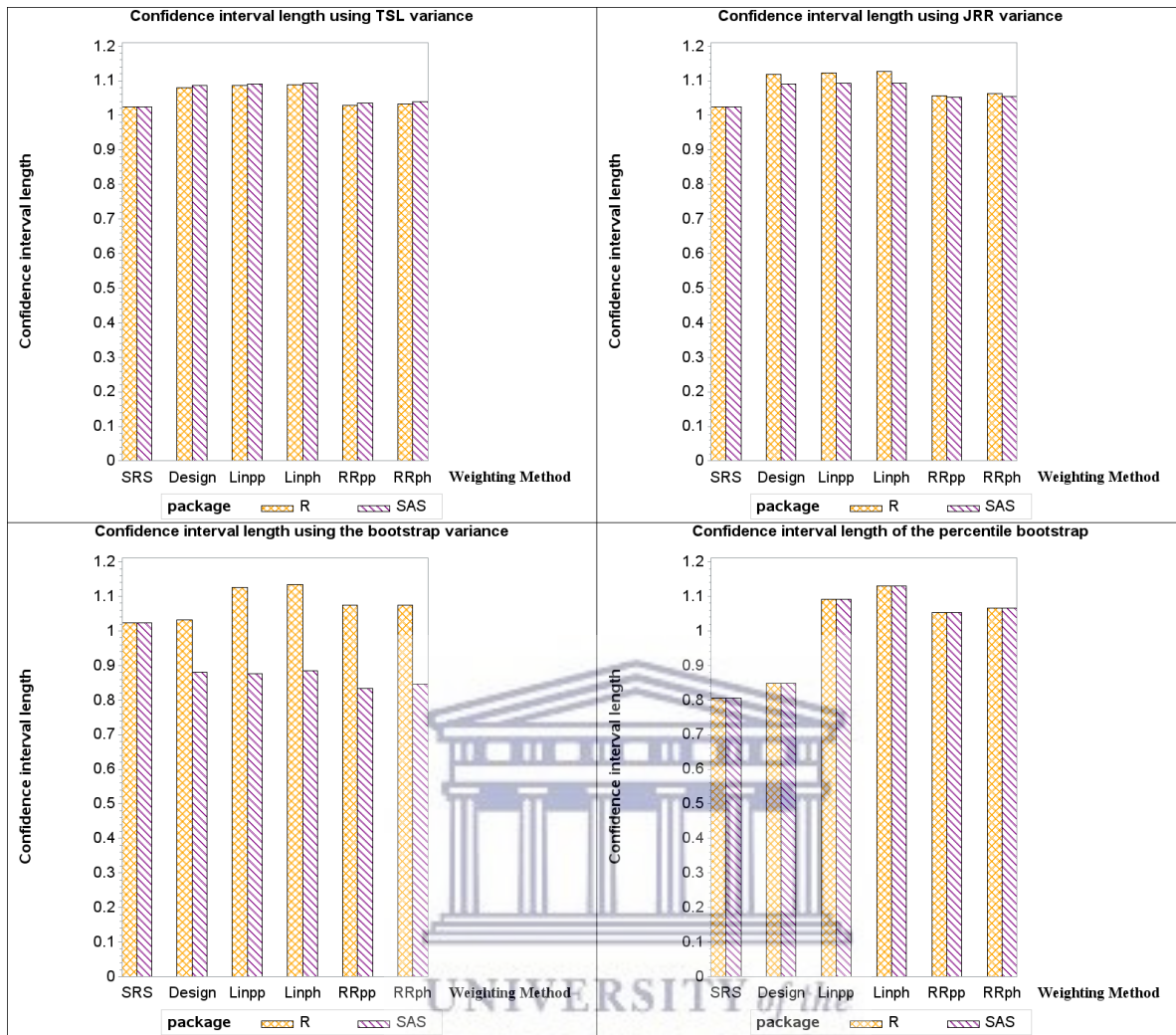


Figure D. 9: The confidence interval lengths for β_{11} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

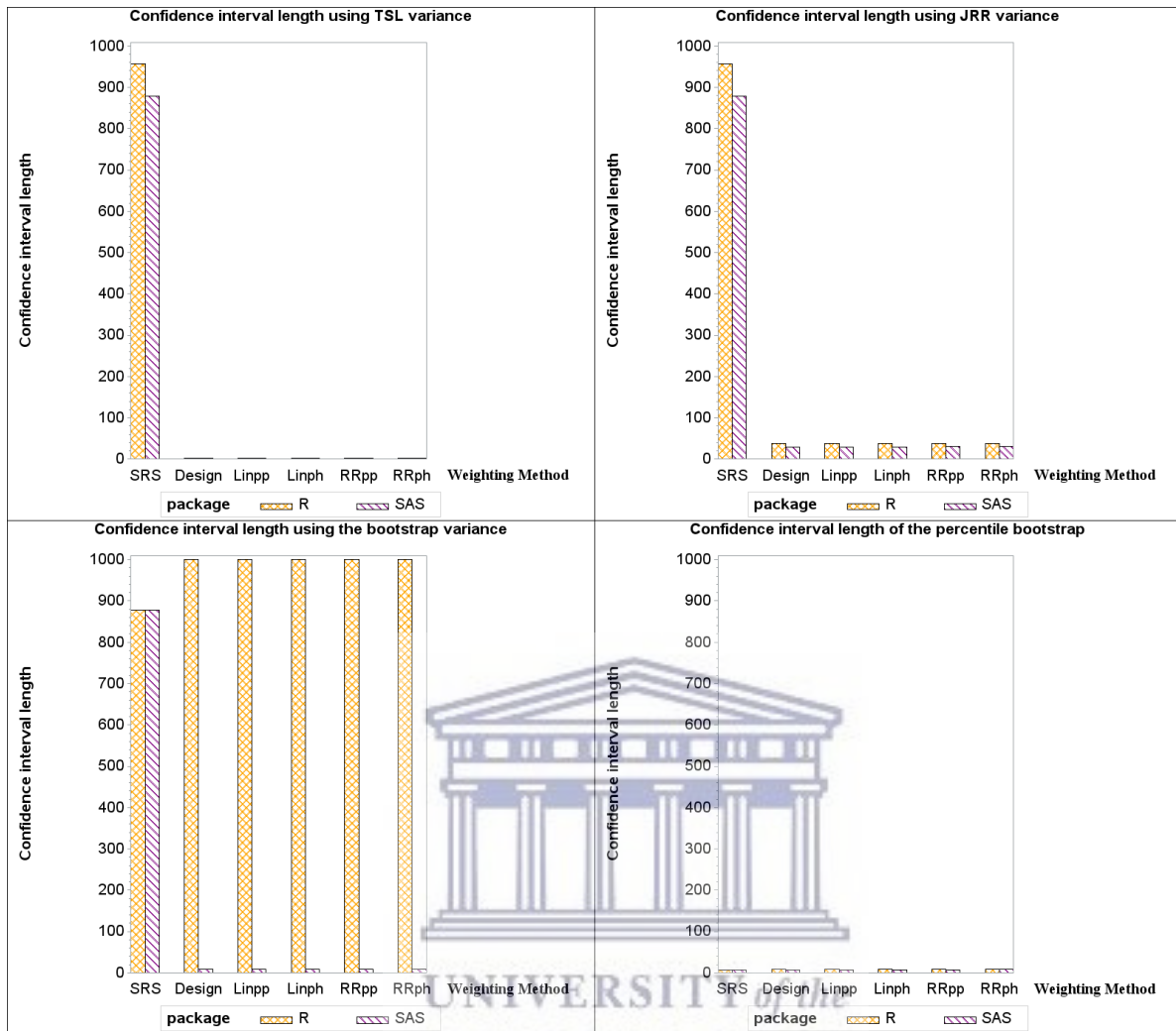


Figure D. 10: The confidence interval lengths for β_{12} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

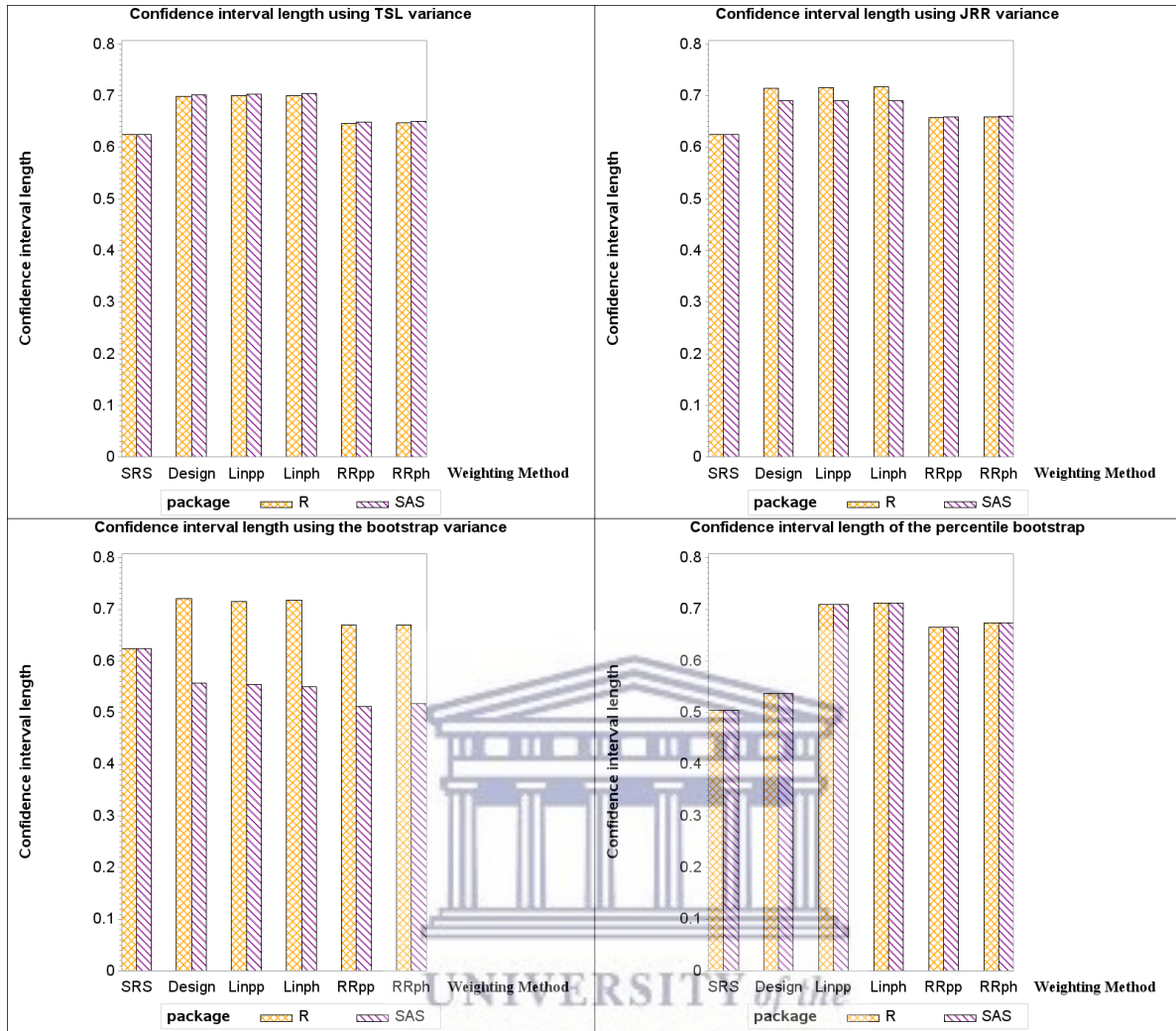


Figure D. 11: The confidence interval lengths for β_{13} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

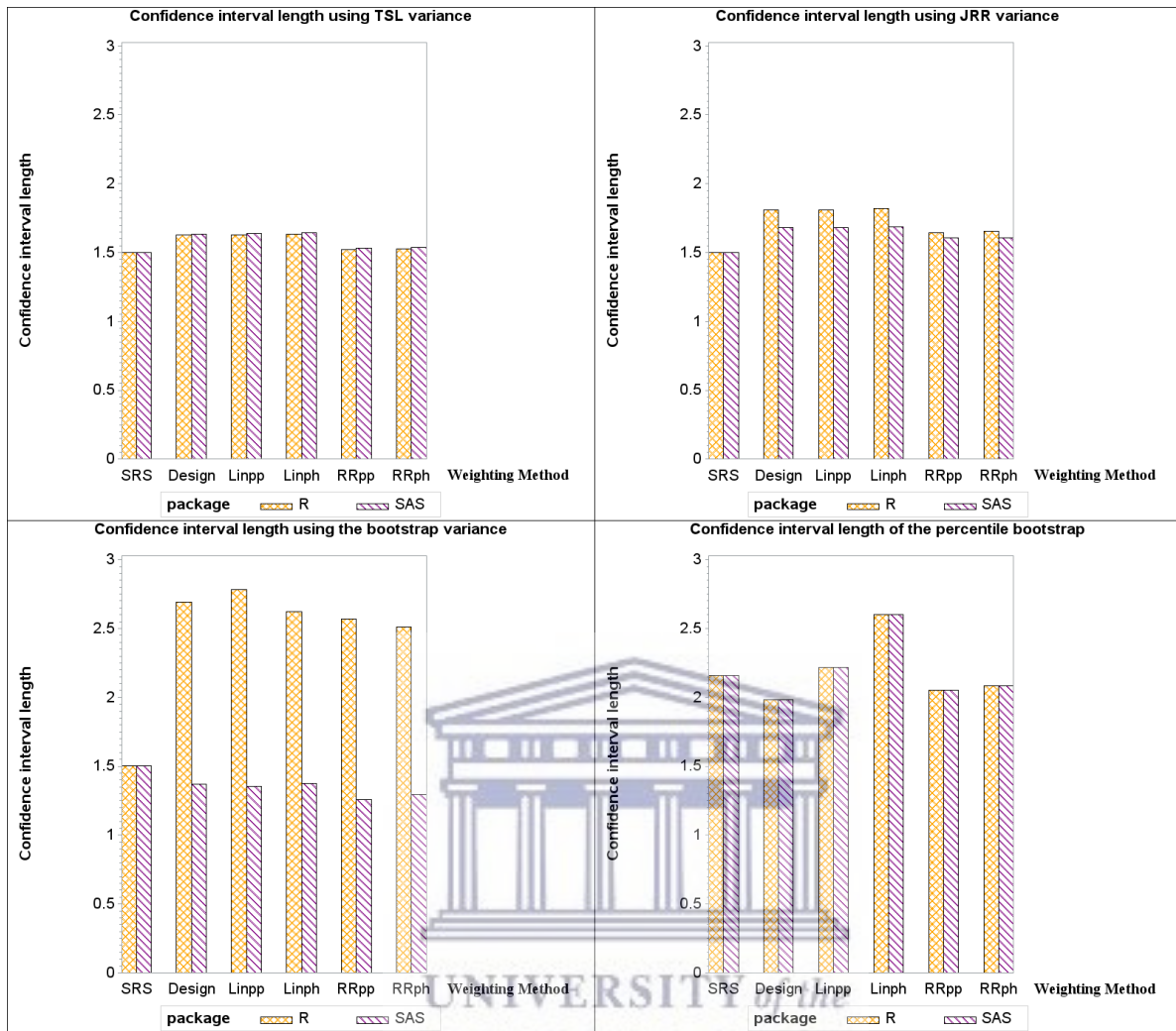


Figure D. 12: The confidence interval lengths for β_{14} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

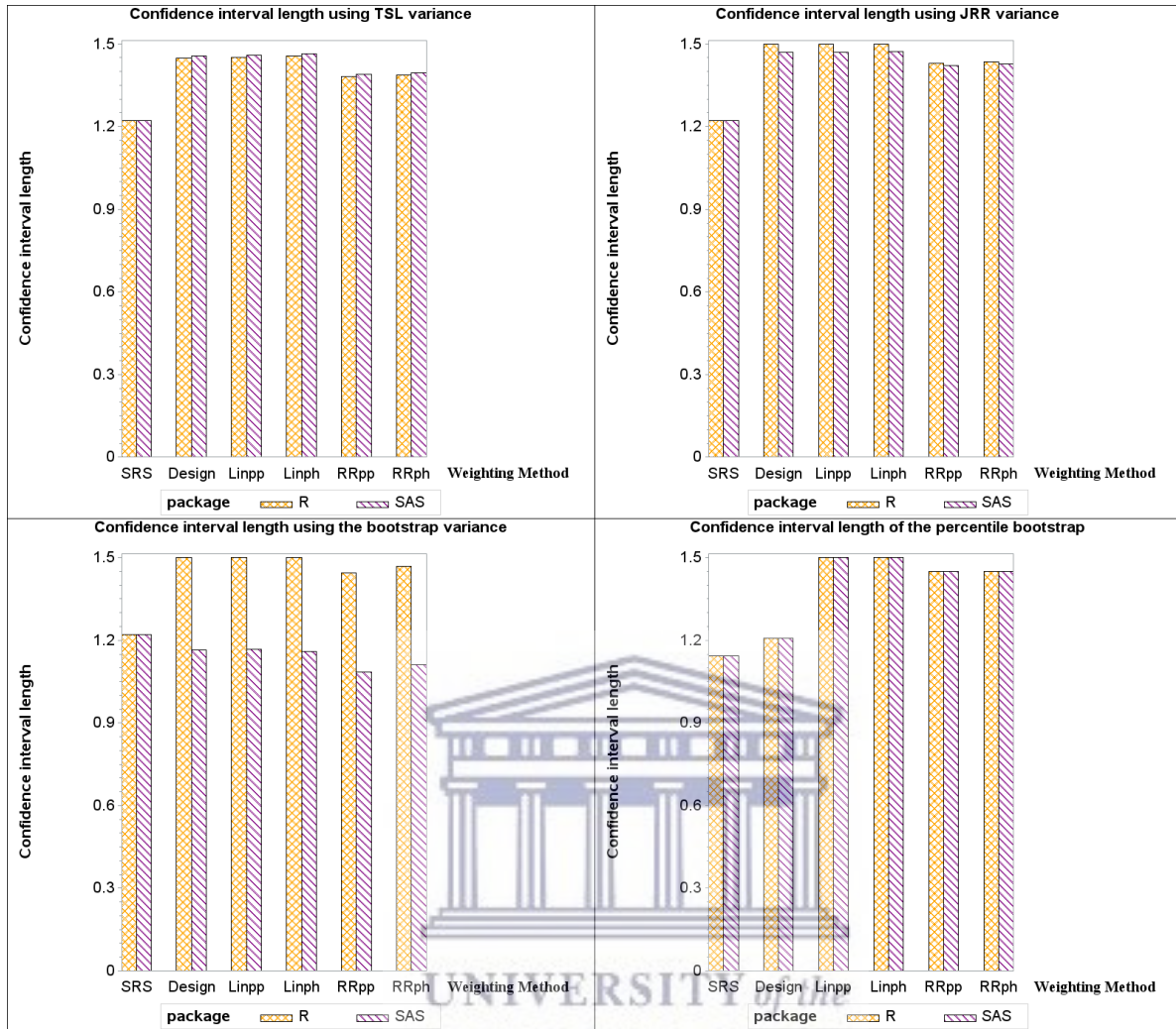


Figure D. 13: The confidence interval lengths for β_{15} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

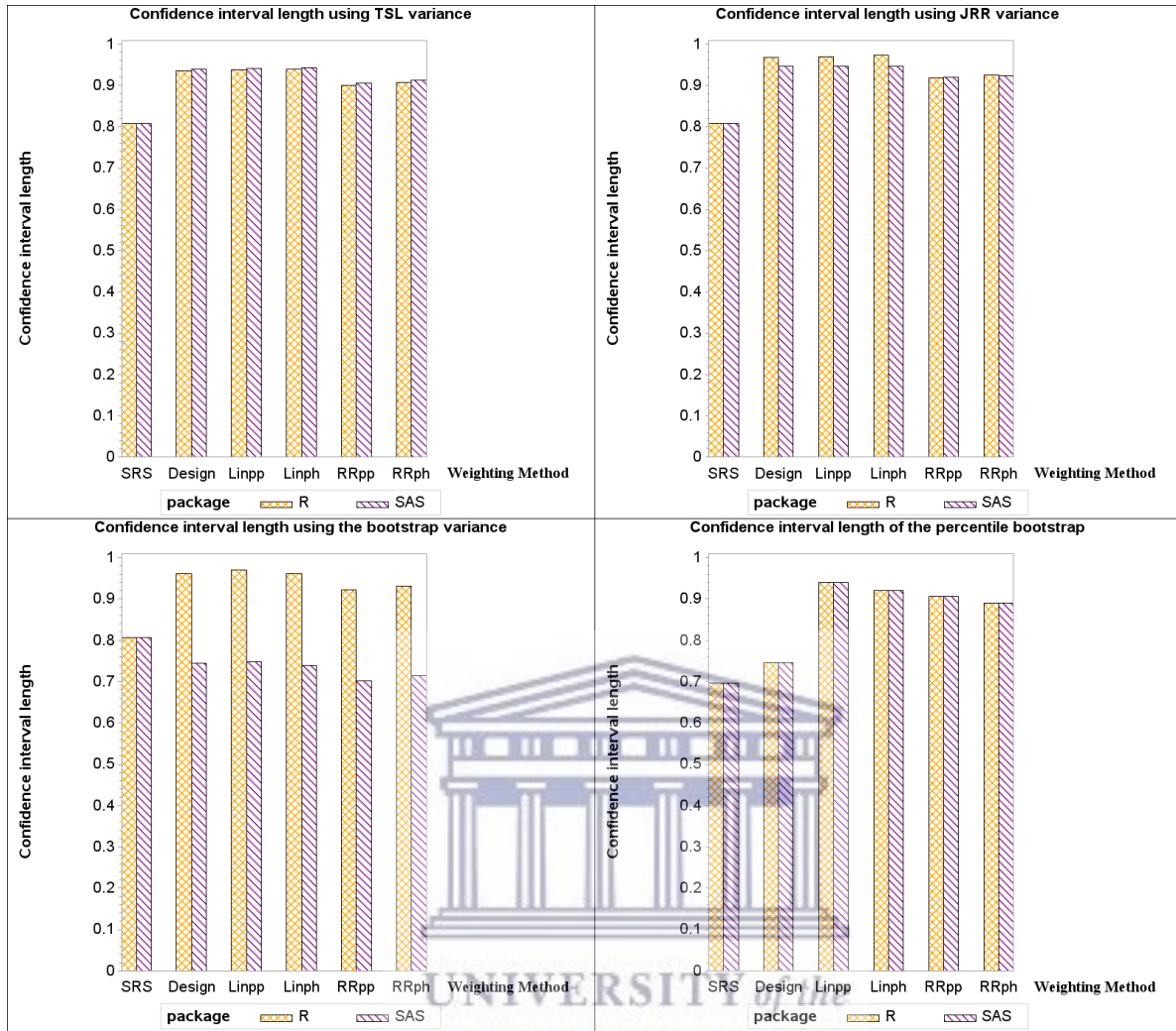


Figure D. 14: The confidence interval lengths for β_{16} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

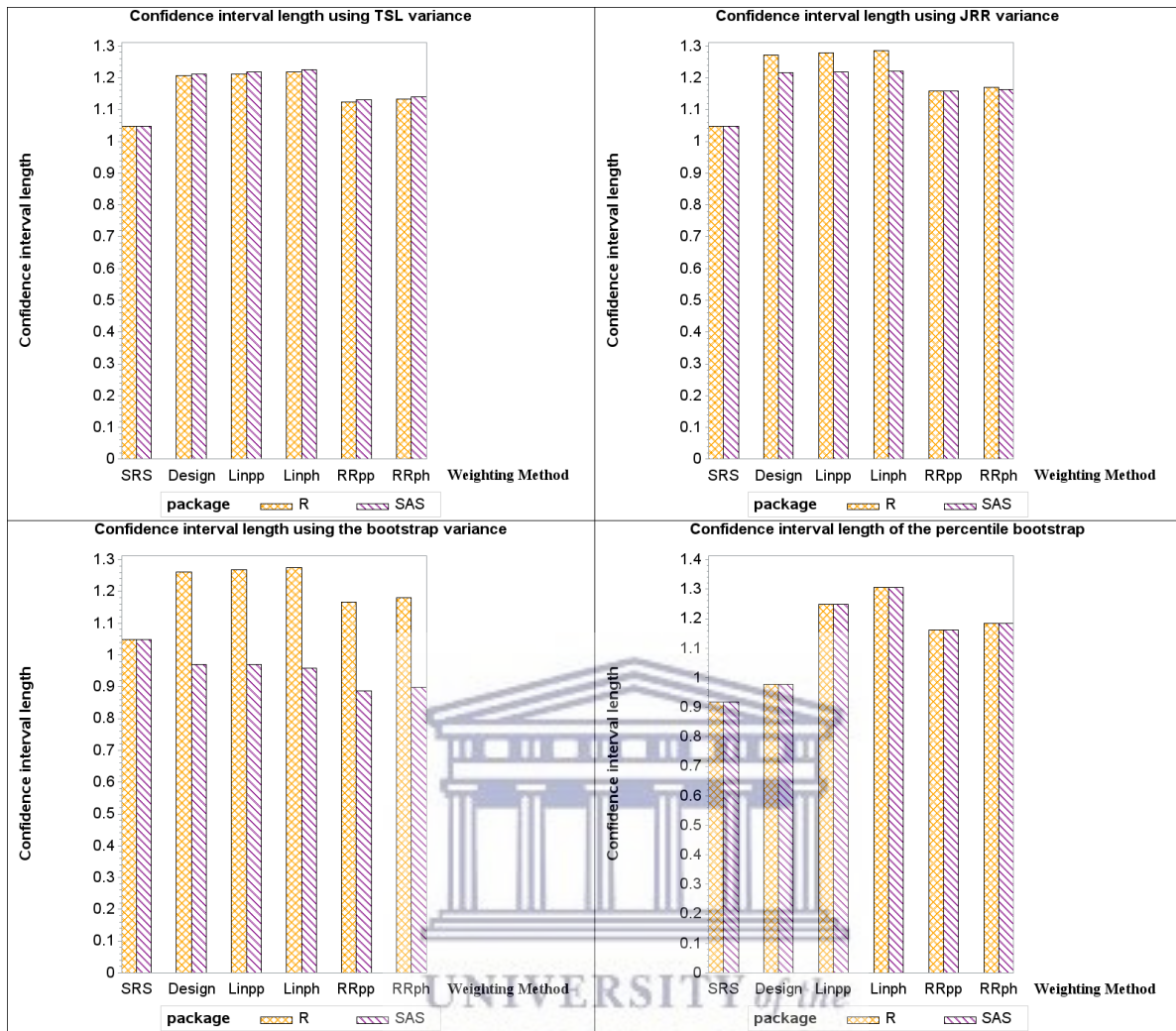


Figure D. 15: The confidence interval lengths for β_{17} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

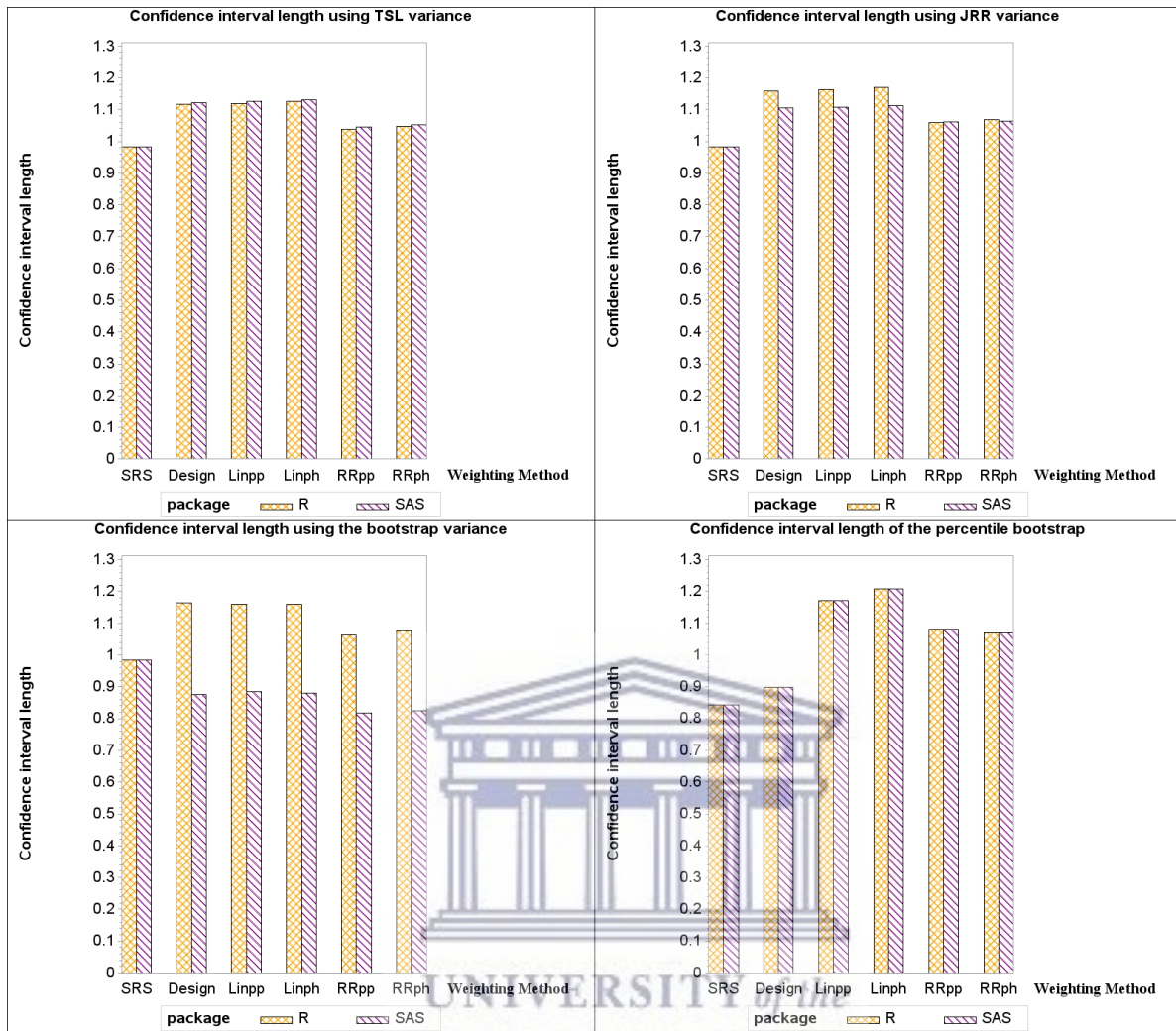


Figure D. 16: The confidence interval lengths for β_{18} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

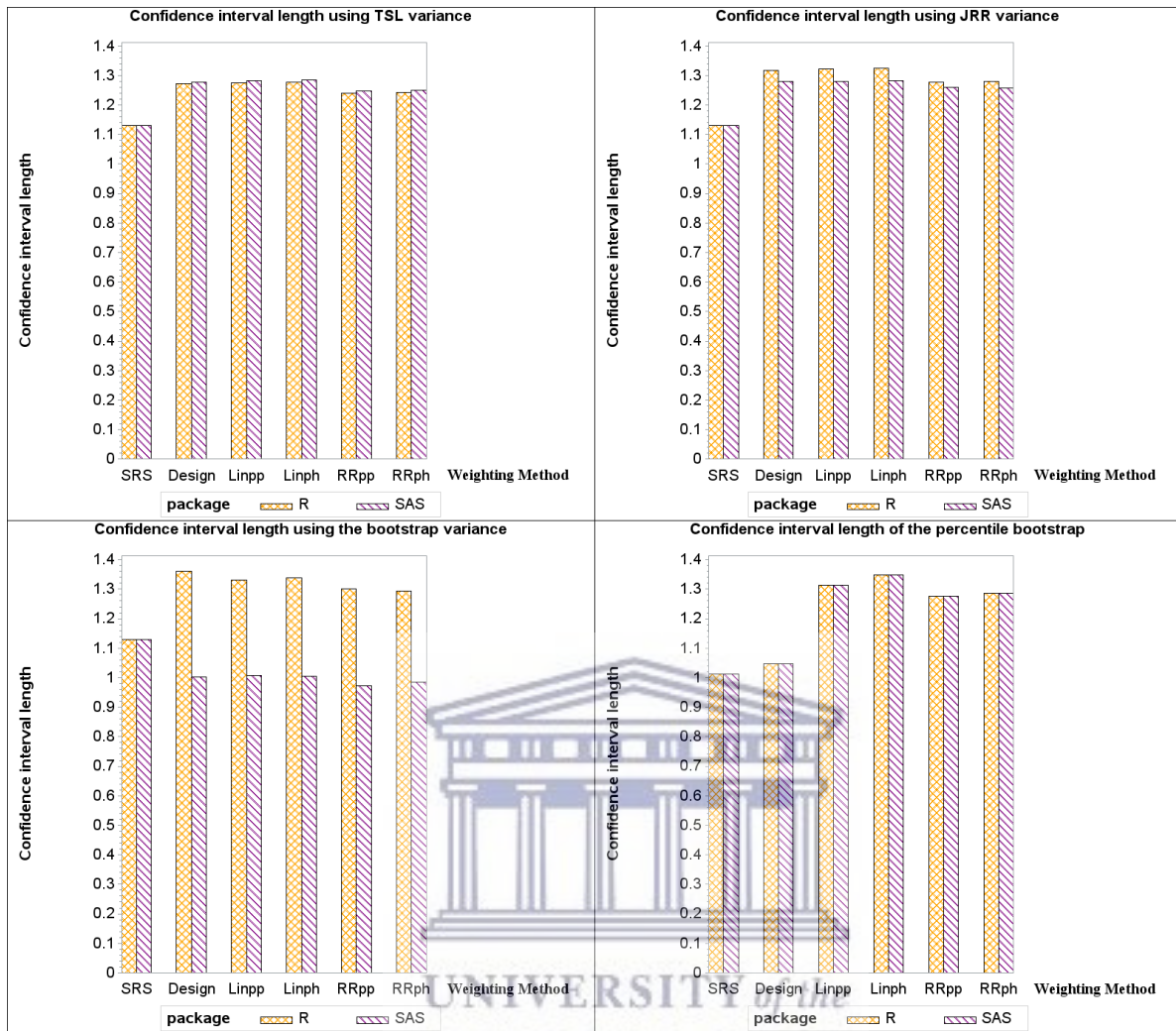


Figure D. 17: The confidence interval lengths for β_{19} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

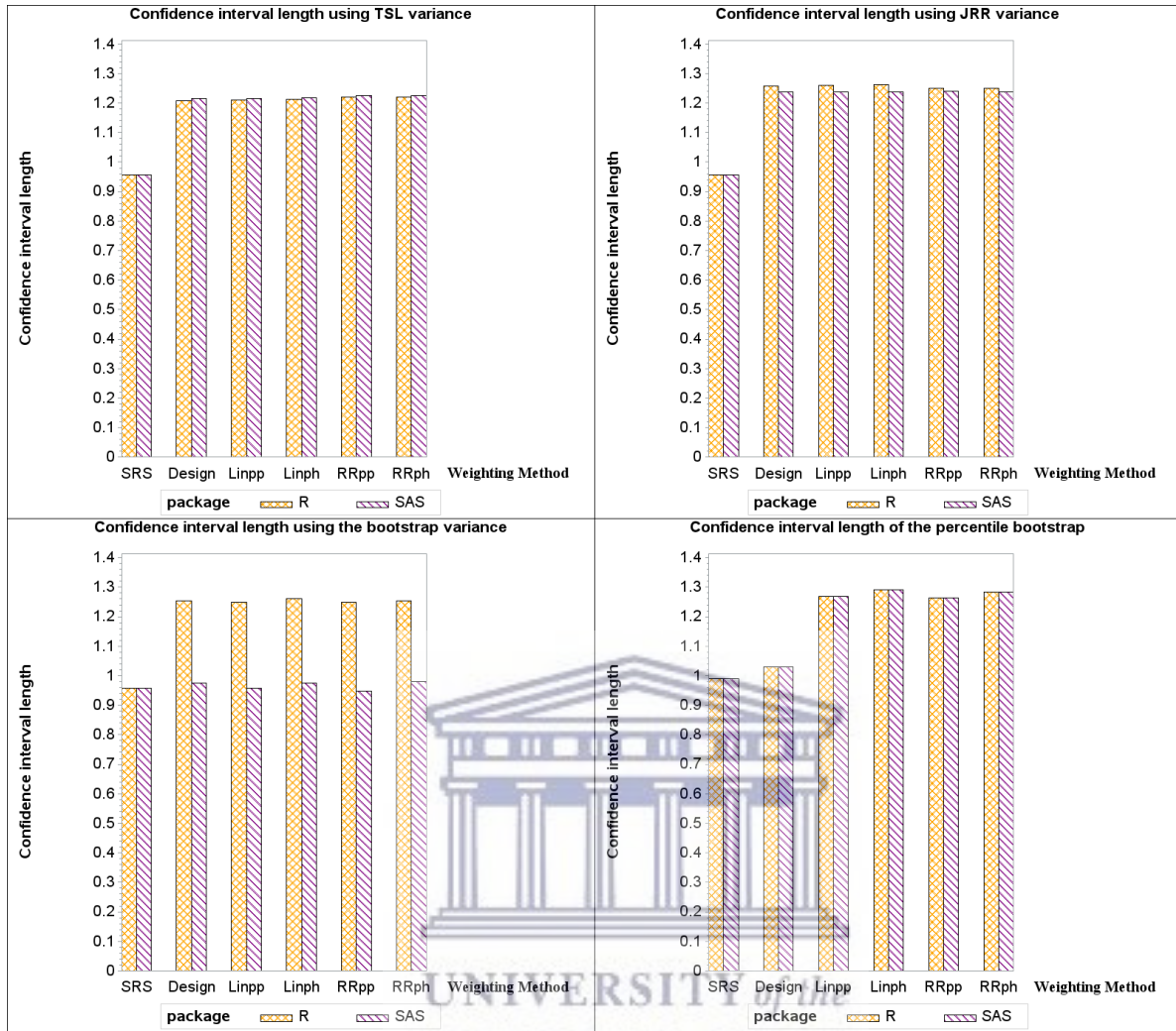


Figure D. 18: The confidence interval lengths for β_{20} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.

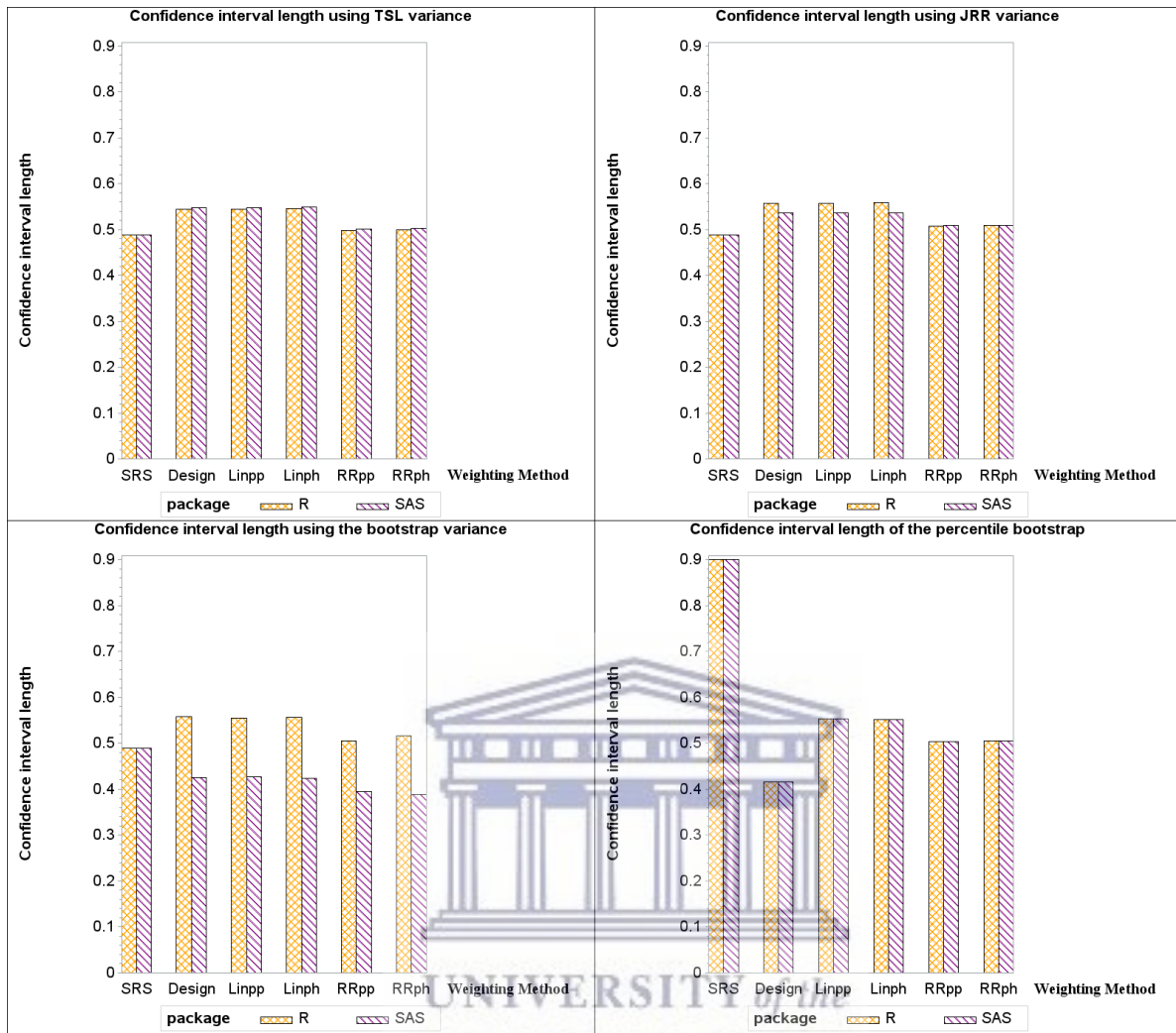


Figure D. 19: The confidence interval lengths for β_{21} under SRS and other weighting methods using TSL, JRR, the bootstrap estimated variances and for the bootstrap percentile interval are shown for SAS and R.