# A Mathematical and Simulation-based Analysis of Weighted Linear Regression

Prikken, L.H.A.

*Award date:*
2020

Technische Universiteit
Eindhoven
University of Technology

Department of Mathematics and Computer Science

# A Mathematical and Simulation-based Analysis of Weighted Linear Regression

*Master Thesis*

Levi Prikken BSc

Supervisors:

prof. dr. E.R. van den Heuvel
dr. M. Regis

Version 1.0

Eindhoven, Thursday 30th January, 2020

# Acknowledgements

You are about to read a piece which has been quite an intensive project for me to work on for the past 9 months. I could not have achieved this without the critical, but inspiring attitude of prof. dr. Edwin van den Heuvel. It may have been hard sometimes to keep our noses in the same direction, but I am very happy to have been supervised by him and have learned a lot in the field of applied statistics because of him.

The other person that was essential for me in this project was dr. Marta Regis. She has been of great support to me as my daily supervisor. I am grateful for all the time she put in to help me with my mathematical endeavors and detailed feedback on my drafts, but above all, for the thoughtful mental support when I became overwhelmed by it all sometimes and really needed the encouragement. She kept me grounded and motivated to continue all the way until the end!

During my stay at the CLSA in Canada for three months, I was honoured to have received the pleasant guidance in learning about epidemiology by dr. Lauren Griffith, who has always stood ready to assist me when needed. I thank the many colleagues that welcomed and supported me during my stay, with special mentions to Amparo Casanova, MD PhD, Donna Fitzpatrick-Lewis, MSW and Alexandra Mayhew, PhDC.

In my personal life, I have received the faithful support of loved ones and friends, with special mentions to Zeno Kapitein, Thomas Wiepking and my mother, Freya Vaes. They gave me a boost of confidence when I didn't feel I was doing the best I could, that made it able to push through it.

Thank you all for the lessons taught, the support provided and the distractions given when needed.

# Abstract

Regression analysis is a key component in research. Normally the analysis is performed on a sample from the population, and one wants to draw conclusions about the whole population. It is important that samples are representative of the population, or wrong conclusions will be drawn. Many population studies have a complex sample design, which might create an unrepresentative sample. To account for complex sample designs in the analyses, survey weights have been created. These weights adjust for possible biases in the results, but might cause a larger variability in the estimates. Next to this, weighted regression can have multiple purposes (sample design or heterogeneity correction), which causes confusion which software is intended for which analysis. Therefore, weights have been a discussion topic over the past decades and no general agreement has been found, nor guidelines for proper usage are available.

This thesis thoroughly investigates this topic, showing advantages and drawbacks of the inclusion of weights in linear regression. The main goal is supported by a mathematical analysis of a simple regression case, which shows promising insights in the weighted structure when compared to the unweighted one. We also study the effect of including weight in the analysis in a simulation that reproduces the complexities of a longitudinal population study in a controlled setting. Our results seem to indicate that unweighted analysis should be preferred, due to it's simplicity and lower variance, while keeping a good coverage probability. Therefore we include recommendations on proper model selection and the choice of software consequently. Our results are still of limited validity, but they open the way to a more in-depth understanding of the role of weights and their proper use.

# Contents

# Chapter 1

# Introduction

In fields like epidemiology, regression analyses are often performed on samples taken from a population, and the results are inferred to the whole population. However, if a sample is unrepresentative of its population, by choice or by chance, the estimates will be biased towards the oversampled group(s) [1]. Such a problem may be overcome in two ways: improving the sampling method to increase the likelihood of getting a representative sample from the overall population, and including survey sampling weights in the analysis of the association between variables. Often the two methods are also combined to take into account the sampling choices in the analysis. A wide literature on sampling methods exists (see Elfil and Negida for reference), but the focus of the present work is on weighted regression [3].

When using sampling weights, each relevant category or stratum in the sample is given a certain weight, or importance, that reflects the actual representation of this category in the population. These weights are meant to correct the estimates of association for imbalance, taking into account how the categories have been sampled. If strata means or strata characteristics are systematically different, a case can be made to include weights [3].

However, Gelman stated in the first sentence of his paper; "survey weighting is a mess". Various names are used for the same thing and certain terms are used with multiple applications. For instance, one of the major uses of weights in regression is to adjust for unequal variances in the dependent variable across observations [5]. The more precise outcomes have higher weights, and the less precise outcomes have lower weights. The weights are used to make the outcome homoscedastic. These weights are not comparable to survey weights, that try to correct non-representative samples. When people talk about a weighted regression analysis it is unclear if they want to correct for these unequal variances or for a sample design. This confusing terminology makes it unclear what weighted regression does and how to compute accurate estimates and standard errors. The confusion in terminology is also reflected in software. Various software programs (SAS and R amongst others) use different formulae for estimation with weights, and including weights in a regression analysis can be tricky if the goal behind the software's tools is not fully understood. For instance, some procedures assume that the user applies weights to correct for heteroscedasticity, and therefore uses different formulae for the calculation of the variance matrix and the standard error on the regression coefficients.

Focusing on sampling weights, researchers have investigated and discussed the use of weights for years [3][6][7][8][9]. According to Kish and Frankel weights were mainly developed for use in descriptive statistics, like means, proportions and percentages. Here the role of weights is clear, since estimates for the population values will be biased when weights are not used. Researchers have also argued to use weights for estimation of relations and associations [11]. Some advantages are that the estimate will be design-consistent [12] and design-unbiased [13], while the unweighted estimator will have a bias proportional to the correlation between the characteristic of interest and the population proportions [14].

On the other hand, one of the larger setbacks of incorporating weights is the complexity of using them. One needs to carefully investigate their case, before doing any analysis on their data. This complexity causes a lot of confusion and misconceptions like "weights are needed for means, but not for regression, because these are model-based" [13]. Next to the complexity, there is also evidence supporting the idea that incorporating weights increases the covariance matrix and the standard error of the regression coefficients [13][15]. Kish and Korn and Graubard add that in general there is a trade-off between the potential larger bias of unweighted estimators and the potential larger variability of weighted estimators.

Please note that this is all based on the assumption that the weights do give a better representation of the population, so they have to be calculated correctly based on accurate population totals.

Researchers have developed tests to determine for a specific case if weights do have significant impact on the result. Bollen et al. have created an overview of the created tests and elaborates on how usable they currently are, but concludes that many tests still need to be properly validated through theory or even simulation. No clear rules are discovered or created for the use of survey weights in regression, so Korn and Graubard called for a consensus on an appropriate method for incorporating weights in regression.

Many population studies incorporate weights in their analyses to account for their complex sampling design. The literature does not give conclusive answer if they should include weights or not, which presents a problem for researchers who are working with these kinds of data sources. It makes it unclear to them which analysis gives valid results and should therefore be used. The Canadian Longitudinal Study on Aging (CLSA) is one of these complex sampling design studies that incorporates weights. Research on this dataset has been conducted by the University of Waterloo amongst others, to investigate the effect of weights on a regression analysis in the CLSA. They encountered no significant difference in the estimate and observed a higher variance for the weighted analysis. The latter was also found in literature. This raises questions about how we should interpret these results.

All of this previous research presents the opportunity for a fresh mind to take a look at some of the thoughts on weighting. The main goal of this thesis is to research what the main advantages and drawbacks are of the usage of weights in regression together with investigating when weights should be incorporated in regression. This goal will be supported by performing a mathematical analysis in a basic controlled setting to identify the main differences between an unweighted and a weighted analysis given in Chapter 3. A set of simple simulation studies, based on the mathematics, will be included as well to illustrate the benefits and drawbacks of using weights in a simple controlled setting. However, this mathematical analysis has its limitations when it comes to more complex datasets. Therefore, a large simulation study is performed attempting to reproduce the complexity of a longitudinal population study in a controlled setting, and its results will be discussed in Chapter 4. The CLSA will serve as the longitudinal population study for the simulation study, so more information about the CLSA can be found in Chapter 2. The goal of the simulation is to conclude whether it's appropriate and better to use weights in authentic datasets with highly complex structures. This will entail checking if the estimation for the association is more representative of the population with or without the use weights. And if weights are recommended, what can be said if there is deviation in the weights compared to what the actual weights, according to your population, should be. To summarize our findings about the mathematics and the simulation study, Chapter 5 will present our recommendations and advice for researchers in population studies and future research.

# Chapter 2

# Case study

## 2.1 CLSA

The Canadian Longitudinal Study on Aging (CLSA)[17], is a large, long-term study of more than 50.000 Canadian individuals who are aged between 45 and 85 at the time of recruitment, and followed for at least 20 years or until death. The study started in 2010 with recruiting participants and collecting their data with the aim to find ways to help people to live long and well, and understand why some people age healthy, while others do not.

### 2.1.1 Sampling

The CLSA exists of two components: CLSA Tracking, a set of 20.000 people from across 10 Canadian provinces; and CLSA Comprehensive, a set of 30.000 people living within 25-50 km of one of the 11 Data Collection Sites (DCS) across 7 Canadian provinces. The main difference is that CLSA Tracking entails a 60 minute interview, while for CLSA Comprehensive there is an in-home interview and DCS visits for additional tests and questions. The CLSA contains pre-specified strata that account for the stratified sample design on province (Alberta, British Columbia, Manitoba, Newfoundland and Labrador, Nova Scotia, Ontario and Quebec) and on low education or non-low education in that area. So CLSA dataset contains a total of 14 strata. This is combined into a variable called 'WGHTS_GEOSTRAT_COM'. In software it is recommended to incorporate a strata statement for the sampling design. For both components, individuals were selected with a single-stage sampling method on the strata through random digit dialling (RDD) and mailouts from Provincial Health Registries (HR) [18]. For CLSA Tracking participants were also recruited through the Canadian Community Health Survey (CCHS) and for CLSA Comprehensive also participants were recruited through a previous longitudinal study in Quebec called NuAge [18].

In this thesis we only focus on CLSA Comprehensive, due to the advantage of a large sample size and usage in other research to make results comparable. If we now speak of the CLSA data, we mean the CLSA Comprehensive dataset with this.

For more details about the CLSA data, please read Canadian Longitudinal Study on Aging.

### 2.1.2 CLSA survey weights

Usage of sampling weights is common in survey analysis due to oversampling of certain groups and non-response. This is no exception for CLSA, which incorporates weights to ensure that the sample is a good reflection of the population.

In general the weights are created by first creating design weights. These weights account for the sample selection. Therefore, these weights are most of the time defined as the inverse inclusion probability. This means the sum of the design weights of the sample will equal the size of the population. Design weights are constructed to get a design-unbiased estimator [9]. In survey samples, non-response is also a common issue. For instance, McCabe and West showed for a population survey in the United States that higher age, being male, being Asian or Hispanic, and having lower education all substantially increase the probability of non-response. Therefore, one needs to increase the weights of CLSA participants

from these kinds of non-responsive groups. This can be done by calibrating the weights, where they are adjusted to force internal estimates to be consistent with external measures through auxiliary information [20]. In the CLSA the weights are re-calibrated to the sum of the targeted Canadian population for sex, age group (45-54, 55-64, 65,74, 75+), province and higher probability of having low or non-low educated people in that area using the information supplied by Statistics Canada. The effect that this can have is that some weights become extremely large. In these small number of cases, the weights with the highest values are trimmed, or set equal to the second highest values within their provinces. The reason for this trimming is to reduce variance of the estimator and prevent too much influence from specific observations due to the imbalance [13]. These calibrated weights are referred to as the inflation weights. The sum of the inflation weights will be equal to the total size of the population. The main purpose of using the inflation weights is to calculate descriptive statistics from CLSA, like means and percentages, for the Canadian population.

Another set of weights in CLSA is the analytical weights. These weights are similar to the inflation weights. However, the sum of the analytical weights will not sum up to the entire population size, but will have a mean weight of 1 within each province. This means the analytical weights are proportional to the inflation weights within each province. The analytical weights sum up to the sample size. The reason for this rescaling is two-fold. On one hand a computational advantage is gained, the closer the weight matrix is to the identity matrix. On the other hand, it makes sure computation of the variance of the estimate is not increased or decreased proportionally compared to the unweighted case. To illustrate this consider a set of weights $\mathbf{W}_s$ for which $\frac{1}{n}\sum_{j=1}^{n} w_{s,j} = s > 1$ and the weights $\mathbf{W}_1$ which are defined as $w_{1,j} = \frac{w_{s,j}}{s}$. These weights are proportional to $\mathbf{W}_s$ and $\frac{1}{n}\sum_{j=1}^{n} w_{1,j} = \frac{1}{s} \cdot \frac{1}{n}\sum_{j=1}^{n} w_{s,j} = 1$. Then the weighted residual sum of squares

$$WSS_{\epsilon,s} = \sum w_{s,j}\,\epsilon_j^2 = \sum s \cdot w_{1,j}\,\epsilon_j^2 = s \cdot \sum w_{1,j}\,\epsilon_j^2 = s \cdot WSS_{\epsilon,1} > WSS_{\epsilon,1} \tag{2.1}$$

So the weighted residual squares of $\mathbf{W}_s$ is larger then $\mathbf{W}_1$ simply due to the larger weights. Using no weights is equivalent to having all weights equal to 1, so also having a mean of 1. When using no weights or the analytical weights, both will have a mean of 1, making the residual sum of squares more comparable in size. Therefore, it's better to use the analytical weights in regression analysis.

For more information on the calculation of the weights, please read Canadian Longitudinal Study on Aging.

### 2.1.3 Data Description

The CLSA presents an excellent base for researchers due to a large amount of participants and the many interesting areas the CLSA contains data on. Therefore, it is important for researchers to have insight in the correct usage of weights. In this thesis we take a specific case of association to investigate. This will serve as a representation of other association research. The specific case we will consider is the influence of the CLSA weights on the association between social support availability and two domains of cognitive functions; memory and executive function. The research for this association is relevant to the CLSA and their collaboration with the University of Waterloo, because it investigates links between factors of social engagement, vulnerability, and cognitive function.

The set of variables that will be used for this research are given in Table 2.1.

The weight covariate set was chosen in accordance with the CLSA's recommendation of including sex, age group and province in regression models, as well as with findings of previous research into social support and cognition-related outcomes [21][22]. The University of Waterloo recommended to include these independent covariates as possible confounders of the association. Please note that the independent variable 'education' is not the same as the stratum of the probability of low / non-low educated people in that area. Also it is not possible to incorporate the probability of low / non-low educated people in that area as an independent covariate in the model, since it says something about the area that participant lives in and not about the participant itself. However, it will be included as a stratum, as recommended by the CLSA.

### 2.1.4 Results

The linear regression model takes the dependent variable $\mathbf{y}$, as Memory or Executive Function, and the weight covariates, the independent covariates and the exposure from Table 2.1. The University of

| Variable type | Name | Math notation | Type | Range |
|---|---|---|---|---|
| ID | ID | ID | - | - |
| Weight covariates | Sex | S | binary | 0/1 |
| | Age group | A | ordinal | 1 - 4 |
| | Province | P | categorical | 1 - 7 |
| Independent covariates | Education | Edu | binary | 0/1 |
| | Smoke | Smo | binary | 0/1 |
| | Alcohol | Alc | Ordinal | 1 - 3 |
| | Hypertension | Hyp | binary | 0/1 |
| | Diabetes | Dia | binary | 0/1 |
| | Any help required with ADL* | ADL | binary | 0/1 |
| | Any help required with IADL* | IADL | binary | 0/1 |
| | Depressive symptoms | DS | discrete | 1 - 30 |
| Exposure | Overall Social Support | OSS | continuous | 0 - 5 |
| Dependent variables | Memory | M | continuous | -** |
| | Executive function | EF | continuous | -** |

Table 2.1: Table of used variables

* ADL stands here for Activities of Daily Living, and IADL stands for Instrumental Activities of Daily Living.
** A set of cognitive tests where z-scores (mean = 0 and standard deviation = 1) are obtained for each cognitive test by subtracting the mean test score from the participants' raw score and dividing the difference by the standard deviation of the mean test score. For the calculation of the value for the dependent variables, the domain-specific scores are summed.

Waterloo recommended not to use Memory as a dependent variable, so in our research we will only consider Executive Function. The used model, we will call model 0, is given in Equation (2.2).

$$\begin{aligned} \mathbf{y}_{EF} = \; & \beta_S \cdot \mathbf{x}_S + \cdots + \beta_{P_7} \cdot \mathbf{x}_{P_7} \\ & + \beta_{Edu} \cdot \mathbf{x}_{Edu} + \cdots + \beta_{DS} \cdot \mathbf{x}_{DS} \\ & + \beta_{OSS} \cdot \mathbf{x}_{OSS} + \boldsymbol{\epsilon}, \end{aligned} \tag{2.2}$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma_\epsilon^2)$. The main interest is in the association $\beta_{OSS}$.

This model is implemented in SAS using the procedure PROC SURVEYREG. SURVEYREG is a procedure designed for regression analysis on survey sample data. It can handle complex survey sample designs, including designs with stratified sample design, clustering, and unequal weighting. The procedure computes regression coefficients and their respective variance-covariance matrix. Weights can be assigned in the procedure to correct for misrepresentation of the population. These weights are implemented through a weighted least squares method [23]. The variance is calculated using an Taylor series estimation method.

The outcome of the research the University of Waterloo already performed is shown in Table 2.2. Within brackets is the 95% confidence interval of the estimation.

| | **Unweighted analysis** | **Weighted analysis** |
|---|---|---|
| Memory | 0.1600 (0.1298, 0.1901) | 0.1548 (0.1193, 0.1903) |
| standard error | 0.0154 | 0.0181 |
| Executive function | 0.3607 (0.2956, 0.4258) | 0.3613 (0.2860, 0.4366) |
| standard error | 0.0332 | 0.0384 |

Table 2.2: Regression results of Waterloo

For both responses, the regression coefficient of the unweighted and weighted analysis are quite similar.

Memory has a difference of 0.0052, and Executive function of 0.0006. The standard error appears to be slightly lower for the unweighted estimate, which is conform to what researchers have stated as well.

The main conclusion they drew from these results is, even though the unweighted analysis had a lower standard error, it did not account for the complex survey sampling design in the CLSA. Therefore, failing to account for oversampling of certain types of participants can systematically underestimate the population variance and standard error [24]. They reason that especially in smaller samples, the width of the confidence interval is adjusted properly with a weighted analysis, and would therefore give a more conservative, but correct reflection of the situation.

# Chapter 3

# Mathematical analysis

To get a good understanding of the usage of weights, we examine various cases of a simple linear regression analysis. Each case will differ slightly to get a good impression of the impact of some of these changes. In this chapter general mathematical computations will be displayed of the estimate and the variance of the estimators for an unweighted, as well as a weighed analysis. To make the analysis more tangible, a set of interesting cases are considered that highlight the effects of including weights in the regression.

Consider a finite population $P$ of size $N$. We have taken a sample $S$ of size $n$ from the population, where each observation $i$ consists out of a response value $y_i$ and a column vector $\mathbf{x}_i$. Here all $x_{1i}$ are 1.

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{3.1}$$

In this sample there are $n_m$ males and $n_f = n - n_m$ females present. Sex plays the role of strata, where a simple random sample is taken for males and females separately. We denote $\mathbf{x}_m = (x_1, \ldots, x_{n_m})$ and $\mathbf{x}_f = (x_{n_m+1}, \ldots, x_n)$. A linear model is considered for the relation between the response and the exposure; $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. However, an interesting case would be where males and females have a different error variance $\sigma_\epsilon^2$.

For the weighted analysis, we will consider a simple creation of weights, where we will take the inverse selection probability for each participant. In this selection we only consider sex of importance. In the population $P$ the proportion of males is $p_m$ and of females is $p_f (= 1 - p_m)$. This means the selection probability and corresponding weights are:

$$\pi_m = \frac{\#\text{males in sample}}{n \cdot p_m} = \frac{n_m}{n \cdot p_m} \Rightarrow w_m = \frac{1}{\pi_m} = \frac{n \cdot p_m}{n_m} \tag{3.2}$$

$$\pi_f = \frac{\#\text{females in sample}}{n \cdot p_f} = \frac{n_f}{n \cdot p_f} \Rightarrow w_f = \frac{1}{\pi_f} = \frac{n \cdot p_f}{n_f} \tag{3.3}$$

This gives a matrix $\mathbf{W}$ that is a diagonal matrix with the first $n_m$ observations $w_m$ and the other $n_f$ observations $w_f$.

Some derivations that are used in this chapter are noted below. Please note for simplification the choice was made to not include Bessel's correction for unbiasedness on the variance and covariance.

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j = \frac{n_m}{n}\overline{x}_m + \frac{n_f}{n}\overline{x}_f \tag{3.4}$$

$$\begin{aligned}
Var(\mathbf{x}) &= \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})^2 && (3.5)\\
&= \frac{1}{n}\sum_{j=1}^{n_m}(x_j - \overline{x}_m + \overline{x}_m - \overline{x})^2 + \frac{1}{n}\sum_{j=n_m+1}^{n}(x_j - \overline{x}_f + \overline{x}_f - \overline{x})^2 && (3.6)\\
&= \frac{n_m}{n}Var(\mathbf{x}_m) + \frac{n_m}{n}(\overline{x}_m - \overline{x})^2 + \frac{n_f}{n}Var(\mathbf{x}_f) + \frac{n_f}{n}(\overline{x}_f - \overline{x})^2 && (3.7)\\
&= \frac{n_m}{n}Var(\mathbf{x}_m) + \frac{n_f}{n}Var(\mathbf{x}_f) + \left(\frac{n_m n_f^2}{n^3} + \frac{n_m^2 n_f}{n^3}\right)(\overline{x}_m - \overline{x}_f)^2 && (3.8)\\
&= \frac{n_m}{n}Var(\mathbf{x}_m) + \frac{n_f}{n}Var(\mathbf{x}_f) + \frac{n_m n_f}{n^2}(\overline{x}_m - \overline{x}_f)^2 && (3.9)
\end{aligned}$$

$$\begin{aligned}
Cov(\mathbf{x},\mathbf{y}) &= \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})(y_j - \overline{y}) && (3.10)\\
&= \frac{n_m}{n}Cov(\mathbf{x}_m,\mathbf{y}_m) + \frac{n_m}{n}(\overline{x}_m - \overline{x})(\overline{y}_m - \overline{y}) + \frac{n_f}{n}Cov(\mathbf{x}_f,\mathbf{y}_f) + \frac{n_f}{n}(\overline{x}_f - \overline{x})(\overline{y}_f - \overline{y}) && (3.11)\\
&= \frac{n_m}{n}Cov(\mathbf{x}_m,\mathbf{y}_m) + \frac{n_f}{n}Cov(\mathbf{x}_f,\mathbf{y}_f) + \frac{n_m n_f}{n^2}(\overline{x}_m - \overline{x}_f)(\overline{y}_m - \overline{y}_f) && (3.12)
\end{aligned}$$

## 3.1 Unweighted analysis

### 3.1.1 Estimate

In the unweighted case, there is no difference in approach from a regular least squares approach, so the unweighted estimator is given by;

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{3.13}$$

A more elaborate derivation of the least squares estimate can be found in Appendix C.

This estimator is unbiased, since

$$\begin{aligned}
\text{bias}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}[\hat{\boldsymbol{\beta}}] - \boldsymbol{\beta} && (3.14)\\
&= \mathbb{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] - \boldsymbol{\beta} && (3.15)\\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}[\mathbf{y}] - \boldsymbol{\beta} && (3.16)\\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta} \qquad \text{since we know } E[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta} && (3.17)\\
&= \boldsymbol{\beta} - \boldsymbol{\beta} && (3.18)\\
&= 0 && (3.19)
\end{aligned}$$

This is somewhat a contradiction of what some researchers say. They feel that the bias is not due to the estimation method, but due to the specific way of drawing the sample. This would realise in that the mean $\overline{y}$ will not converge to the expectation of $\mathbf{y}$ and would therefore give a biased estimate.

For the general case described above, the estimator gives;

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \frac{1}{Var(\mathbf{x})}\begin{bmatrix} Var(\mathbf{x})\cdot\overline{y} - \overline{x}\cdot Cov(\mathbf{x},\mathbf{y}) \\ Cov(\mathbf{x},\mathbf{y}) \end{bmatrix} \tag{3.20}$$

The computation of this estimate can be found in Appendix D.

The estimate $\hat{\beta}_1$ can be rewritten to a combination of estimates for males $\hat{\beta}_{1,m} = Cov(\mathbf{x}_m, \mathbf{y}_m)/Var(\mathbf{x}_m)$ and females $\hat{\beta}_{1,f} = Cov(\mathbf{x}_f, \mathbf{y}_f)/Var(\mathbf{x}_f)$, which can be derived from Equation (3.20) applied to the data of respectively only the males and females. Therefore,

$$\hat{\beta}_1 = \frac{Cov(\mathbf{x}, \mathbf{y})}{Var(\mathbf{x})} \tag{3.21}$$

$$\overset{\substack{(3.9)\\(3.12)}}{=} \frac{n_m Cov(\mathbf{x}_m, \mathbf{y}_m) + \frac{n_m n_f}{n}(\overline{y}_f - \overline{y}_m)(\overline{x}_f - \overline{x}_m) + n_f Cov(\mathbf{x}_f, \mathbf{y}_f)}{n_m Var(\mathbf{x}_m) + \frac{n_m n_f}{n}(\overline{x}_f - \overline{x}_m)^2 + n_f Var(\mathbf{x}_f)} \tag{3.22}$$

$$= \frac{n_m Var(\mathbf{x}_m)\hat{\beta}_{1,m} + \frac{n_m n_f}{n}(\overline{y}_f - \overline{y}_m)(\overline{x}_f - \overline{x}_m) + n_f Var(\mathbf{x}_f)\hat{\beta}_{1,f}}{n_m Var(\mathbf{x}_m) + \frac{n_m n_f}{n}(\overline{x}_f - \overline{x}_m)^2 + n_f Var(\mathbf{x}_f)} \tag{3.23}$$

$$= \frac{\frac{n_m}{n} Var(\mathbf{x}_m)\hat{\beta}_{1,m} + \frac{n_m}{n}\frac{n_f}{n}(\overline{y}_f - \overline{y}_m)(\overline{x}_f - \overline{x}_m) + \frac{n_f}{n} Var(\mathbf{x}_f)\hat{\beta}_{1,f}}{\frac{n_m}{n} Var(\mathbf{x}_m) + \frac{n_m}{n}\frac{n_f}{n}(\overline{x}_f - \overline{x}_m)^2 + \frac{n_f}{n} Var(\mathbf{x}_f)} \tag{3.24}$$

where $\frac{n_m}{n}$ and $\frac{n_f}{n}$ represent the proportions of males and females in the sample.

### 3.1.2 Variance (general)

Then the variance of the estimator is:

$$Var(\hat{\boldsymbol{\beta}}) = Var\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y})\right) \tag{3.25}$$

$$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \cdot Var(\mathbf{y}) \cdot \left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)^T \tag{3.26}$$

Most of the time in regression analysis, it is assumed that the residuals are independent and identically distributed. However, for a more general analysis, let us assume the variance of $\mathbf{y}$ and therefore the residuals $\epsilon$ is differently distributed for males and females, meaning $\epsilon_m \sim N(0, \sigma^2_{\epsilon,m})$ and $\epsilon_f \sim N(0, \sigma^2_{\epsilon,f})$. Consider the case where the first $n_m$ observations are from males, and the other $n_f$ observations are from females. Then this gives the covariance matrix of $\mathbf{y}$;

$$Var(\mathbf{y}) = diag(\sigma^2_{\epsilon,m}, \cdots, \sigma^2_{\epsilon,m}, \sigma^2_{\epsilon,f}, \cdots, \sigma^2_{\epsilon,f}) \tag{3.27}$$

Due to the elaborate mathematics, and a particular interest in the variance of the slope, we will focus on entry (2,2) of the covariance matrix in the case considered.

This gives:

$$Var(\hat{\beta}_1) = \frac{\sigma^2_{\epsilon,m} n_m Var(\mathbf{x}_m) + \frac{n_m n_f}{n^2}(\sigma^2_{\epsilon,m} n_f + \sigma^2_{\epsilon,f} n_m)(\overline{x}_m - \overline{x}_f)^2 + \sigma^2_{\epsilon,f} n_f Var(\mathbf{x}_f)}{\left(n_m Var(\mathbf{x}_m) + \frac{n_m n_f}{n}(\overline{x}_f - \overline{x}_m)^2 + n_f Var(\mathbf{x}_f)\right)^2} \tag{3.28}$$

With an elaboration on the computations in Appendix E.

With the standard error on the estimate simply;

$$se(\hat{\beta}_1) = \sqrt{Var(\hat{\beta}_1)} \tag{3.29}$$

Due to the generalisation, the estimation of the error term is more difficult since we cannot simply take the residual sum of squares. One approach could be to estimate the sigma's with;

$$\hat{\sigma}^2_{\epsilon,m} = \frac{1}{n_m - 2}\sum_{i=1}^{n_m}(y_i - \hat{y}_i)^2 \text{ and } \hat{\sigma}^2_{\epsilon,f} = \frac{1}{n_f - 2}\sum_{i=n_m+1}^{n}(y_i - \hat{y}_i)^2$$

However, validity of this idea is not our main focus for now and will be left for further research.

### 3.1.3 Variance ($\sigma^2_{\epsilon,m} = \sigma^2_{\epsilon,f}$)

If the assumption would be made that $\sigma^2_{\epsilon,m} = \sigma^2_{\epsilon,f} = \sigma^2_\epsilon$, then the unweighted variance can be more easily computed. The variance of the estimator simplifies to;

$$
\begin{aligned}
Var(\hat{\boldsymbol{\beta}}) &= Var\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y})\right) && (3.30)\\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \cdot Var(\mathbf{y}) \cdot \left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)^T && (3.31)\\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \cdot \sigma^2_\epsilon \cdot \left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)^T \qquad \text{when } y \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2_\epsilon) && (3.32)\\
&= \sigma^2_\epsilon \cdot (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \qquad \text{since}(\mathbf{X}^T\mathbf{X})^{-1} = \left((\mathbf{X}^T\mathbf{X})^{-1}\right)^T && (3.33)\\
&= \sigma^2_\epsilon \cdot (\mathbf{X}^T\mathbf{X})^{-1} && (3.34)
\end{aligned}
$$

Since we don't know the true value of $\sigma^2_\epsilon$, we take the approximation:

$$\hat{\sigma}^2_\epsilon = \frac{1}{n-2}SS_\epsilon = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.35}$$

So in our case this variance becomes;

$$Var(\hat{\boldsymbol{\beta}}) = \frac{1}{n^2 Var(\mathbf{x})} \cdot \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix} \cdot \sigma^2_\epsilon \tag{3.36}$$

So the variance of the slope is;

$$
\begin{aligned}
Var(\beta_1) &= \frac{n \cdot \sigma^2_\epsilon}{n^2 Var(\mathbf{x})} && (3.37)\\
&= \frac{n \cdot \sigma^2_\epsilon}{n \cdot \sum_{j=1}^n (x_j - \overline{x})^2} && (3.38)\\
&= \frac{\sigma^2_\epsilon}{\sum_{j=1}^n (x_j - \overline{x})^2} && (3.39)
\end{aligned}
$$

To verify the findings from the previous section, if we would continue from Equation (3.28) and take $\sigma_{\epsilon,m} = \sigma_{\epsilon,f} (= \sigma_\epsilon)$, then;

$$
\begin{aligned}
Var(\hat{\beta}_1) &= \frac{\sigma^2_\epsilon n_m Var(\mathbf{x}_m) + \frac{n_m n_f}{n^2}(\sigma^2_\epsilon n_f + \sigma^2_\epsilon n_m)(\overline{x}_m - \overline{x}_f)^2 + \sigma^2_\epsilon n_f Var(\mathbf{x}_f)}{\left(n_m Var(\mathbf{x}_m) + \frac{n_m n_f}{n}(\overline{x}_f - \overline{x}_m)^2 + n_f Var(\mathbf{x}_f)\right)^2} && (3.40)\\
&= \sigma^2_\epsilon \cdot \frac{n_m Var(\mathbf{x}_m) + \frac{n_m n_f}{n^2}(n_f + n_m)(\overline{x}_m - \overline{x}_f)^2 + n_f Var(\mathbf{x}_f)}{\left(n_m Var(\mathbf{x}_m) + \frac{n_m n_f}{n}(\overline{x}_f - \overline{x}_m)^2 + n_f Var(\mathbf{x}_f)\right)^2} && (3.41)\\
&= \sigma^2_\epsilon \cdot \frac{n_m Var(\mathbf{x}_m) + \frac{n_m n_f}{n}(\overline{x}_m - \overline{x}_f)^2 + n_f Var(\mathbf{x}_f)}{\left(n_m Var(\mathbf{x}_m) + \frac{n_m n_f}{n}(\overline{x}_f - \overline{x}_m)^2 + n_f Var(\mathbf{x}_f)\right)^2} && (3.42)\\
&= \sigma^2_\epsilon \cdot \frac{1}{n_m Var(\mathbf{x}_m) + \frac{n_m n_f}{n}(\overline{x}_f - \overline{x}_m)^2 + n_f Var(\mathbf{x}_f)} && (3.43)\\
&\overset{(3.9)}{=} \frac{\sigma^2_\epsilon}{\sum_{j=1}^n (x_j - \overline{x})^2} && (3.44)
\end{aligned}
$$

where $\sigma^2_\epsilon$ is approximated with;

$$\hat{\sigma}^2_\epsilon = \frac{1}{n-2}SS_\epsilon = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.45}$$

The variance in Equation (3.39) and (3.44) are identical as well as compared with basic linear regression books.

Then lastly, this gives as the standard error on this estimate;

$$se(\hat{\beta}_1) = \sqrt{Var(\hat{\beta}_1)} \tag{3.46}$$

## 3.2 Weighted analysis

### 3.2.1 Estimate

For the weighted analysis, the original least squares method will be expanded to a weighted least squares method, which is common in use [3]. This gives the estimate;

$$\hat{\boldsymbol{\beta}}_w = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y} \tag{3.47}$$

A more elaborate derivation of the weighted least squares estimate can be found in Appendix F.

This estimate is unbiased, since

$$\begin{aligned}
\text{bias}(\hat{\boldsymbol{\beta}}_w) &= \mathbb{E}[\hat{\boldsymbol{\beta}}_w] - \boldsymbol{\beta} && (3.48)\\
&= \mathbb{E}[(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}] - \boldsymbol{\beta} && (3.49)\\
&= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbb{E}[\mathbf{y}] - \boldsymbol{\beta} && (3.50)\\
&= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta} \qquad \text{since we know } E[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta} && (3.51)\\
&= \boldsymbol{\beta} - \boldsymbol{\beta} && (3.52)\\
&= 0 && (3.53)
\end{aligned}$$

This unbiasedness is consistent with literature.

For the case considered, the estimate gives;

$$\begin{aligned}
&(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y} = \\
&\quad \frac{1}{C_w} \cdot \Bigg( w_m^2 n_m^2 Var(\mathbf{x}_m) \cdot \begin{bmatrix} \hat{\beta}_{0,m} \\ \hat{\beta}_{1,m} \end{bmatrix} \\
&\qquad + w_m w_f n_m n_f \cdot \begin{bmatrix} Var(\mathbf{x}_f)\cdot\overline{y}_m + Var(\mathbf{x}_m)\cdot\overline{y}_f + \overline{x}_m(\overline{x}_m\cdot\overline{y}_f - \overline{x}\overline{y}_f) + \overline{x}_f(\overline{x}_f\cdot\overline{y}_m - \overline{x}\overline{y}_m) \\ Var(\mathbf{x}_m)\hat{\beta}_{1,m} + Var(\mathbf{x}_f)\hat{\beta}_{1,f} + (\overline{y}_f - \overline{y}_m)(\overline{x}_f - \overline{x}_m) \end{bmatrix} \\
&\qquad + w_f^2 n_f^2 Var(\mathbf{x}_f) \cdot \begin{bmatrix} \hat{\beta}_{0,f} \\ \hat{\beta}_{1,f} \end{bmatrix} \Bigg) \qquad (3.54)
\end{aligned}$$

where

$$C_w = w_m n_m (w_m n_m + w_f n_f) Var(\mathbf{x}_m) + w_m w_f n_m n_f (\overline{x}_m - \overline{x}_f)^2 + w_f n_f (w_m n_m + w_f n_f) Var(\mathbf{x}_f) \tag{3.55}$$

and $\hat{\boldsymbol{\beta}}_m$ and $\hat{\boldsymbol{\beta}}_f$ the regression coefficients if a least squares estimate was produced separately for males and females. So when we express the slope in a combination of the male and female regression slope, this gives;

$$\hat{\beta}_{1,w} = \frac{w_m n_m (w_m n_m + w_f n_f) Cov(\mathbf{x}_m, \mathbf{y}_m) + w_m w_f n_m n_f (\overline{y}_f - \overline{y}_m)(\overline{x}_f - \overline{x}_m) + w_f n_f (w_m n_m + w_f n_f) Cov(\mathbf{x}_f, \mathbf{y}_f)}{w_m n_m (w_m n_m + w_f n_f) Var(\mathbf{x}_m) + w_m w_f n_m n_f (\overline{x}_m - \overline{x}_f)^2 + w_f n_f (w_m n_m + w_f n_f) Var(\mathbf{x}_f)} \tag{3.56}$$

$$= \frac{w_m n_m (w_m n_m + w_f n_f) Var(\mathbf{x}_m) \hat{\beta}_{1,m} + w_m w_f n_m n_f (\overline{y}_f - \overline{y}_m)(\overline{x}_f - \overline{x}_m) + w_f n_f (w_m n_m + w_f n_f) Var(\mathbf{x}_f) \hat{\beta}_{1,f}}{w_m n_m (w_m n_m + w_f n_f) Var(\mathbf{x}_m) + w_m w_f n_m n_f (\overline{x}_m - \overline{x}_f)^2 + w_f n_f (w_m n_m + w_f n_f) Var(\mathbf{x}_f)} \tag{3.57}$$

For an elaboration on this computation, please see Appendix G.

What is noticeable from this expression is that the weights for a subgroup and the amount of participants from that subgroup are inversely proportionate. They cancel each other out, making the estimate some sort of average between the male estimation and female estimation. This way of thinking feels organic, since weights try to adjust for the imbalance in the dataset, and the weights make sure $n_m$ or $n_f$ does not influence the outcome too much if they become larger.

To further illustrate this way of thinking, if the definitions of $w_m$ and $w_f$ would be inserted, we get;

$$(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y} =$$

$$\frac{\begin{array}{c} p_m^2 Var(\mathbf{x}_m)\cdot\begin{bmatrix}\hat{\beta}_{0,m}\\ \hat{\beta}_{1,m}\end{bmatrix} \\ +p_m p_f\cdot\begin{bmatrix} Var(\mathbf{x}_f)\cdot\overline{y}_m + Var(\mathbf{x}_m)\cdot\overline{y}_f + \overline{x}_m(\overline{x}_m\cdot\overline{y}_f - \overline{xy}_f) + \overline{x}_f(\overline{x}_f\cdot\overline{y}_m - \overline{xy}_m) \\ Var(\mathbf{x}_m)\hat{\beta}_{1,m} + Var(\mathbf{x}_f)\hat{\beta}_{1,f} + (\overline{y}_f - \overline{y}_m)(\overline{x}_f - \overline{x}_m)\end{bmatrix} \\ + p_f^2 Var(\mathbf{x}_f)\cdot\begin{bmatrix}\hat{\beta}_{0,f}\\ \hat{\beta}_{1,f}\end{bmatrix}\end{array}}{p_m Var(\mathbf{x}_m) + p_m p_f(\overline{x}_m - \overline{x}_f)^2 + p_f Var(\mathbf{x}_f)} \tag{3.58}$$

For the slope, this can be rewritten to

$$\hat{\beta}_{1,w} = \frac{p_m Var(\mathbf{x}_m)\hat{\beta}_{1,m} + p_m p_f(\overline{y}_f - \overline{y}_m)(\overline{x}_f - \overline{x}_m) + p_f Var(\mathbf{x}_f)\hat{\beta}_{1,f}}{p_m Var(\mathbf{x}_m) + p_m p_f(\overline{x}_m - \overline{x}_f)^2 + p_f Var(\mathbf{x}_f)} \tag{3.59}$$

For an elaboration see Appendix G.1.

When using these inverse inclusion weights, the weighted estimate seems to be independent of the sample size $n$ or the amount of males $n_m$ and females $n_f$, but is determined by the different proportions between males and females in the population and adjusts the estimate accordingly. In the unweighted estimate in Equation (3.23) the formula is similar but does not utilise the proportions in the population, but the proportions in your sample. So when the proportions of your sample $n_m/n$ and $n_f/n$ are very different from the population proportions $p_m$ and $p_f$ the estimates could be quite different. Also, as a sanity check, if we would only want to consider the male regression estimate, $p_f$ would be 0, and we would indeed get the estimate if we would only include males.

### 3.2.2 Variance (general)

The variance of the estimator is:

$$Var(\hat{\boldsymbol{\beta}}_w) = Var\left((\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}\right) \tag{3.60}$$

$$= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\cdot Var(\mathbf{y})\cdot\left((\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\right)^T \tag{3.61}$$

$$= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\cdot Var(\mathbf{y})\cdot\mathbf{W}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1} \tag{3.62}$$

However, again consider a more general analysis. Let us assume the error is different for males and females, meaning $\epsilon_m \sim N(0,\sigma^2_{\epsilon,m})$ and $\epsilon_f \sim N(0,\sigma^2_{\epsilon,f})$. Consider the case where the first $n_m$ observations are from males, and the other $n_f$ observations are from females. Then this gives the covariance matrix of $\mathbf{y}$;

$$Var(\mathbf{y}) = diag(\sigma^2_{\epsilon,m},\cdots,\sigma^2_{\epsilon,m},\sigma^2_{\epsilon,f},\cdots,\sigma^2_{\epsilon,f}) \tag{3.63}$$

Due to the elaborate mathematics, and a particular interest in the variance of the slope, we will focus on entry (2,2) of the covariance matrix in the case considered.

This gives;

$$Var(\hat{\beta}_{1,w}) = \frac{\begin{array}{c}\sigma^2_{\epsilon,m}w_m^2 n_m(w_m n_m + w_f n_f)^2 Var(\mathbf{x}_m) \\ + w_m^2 w_f^2 n_m n_f(\sigma^2_{\epsilon,m}n_f + \sigma^2_{\epsilon,f}n_m)(\overline{x}_m - \overline{x}_f)^2 + \sigma^2_{\epsilon,f}w_f^2 n_f(w_m n_m + w_f n_f)^2 Var(\mathbf{x}_f)\end{array}}{\left(w_m n_m(w_m n_m + w_f n_f)Var(\mathbf{x}_m) + w_m w_f n_m n_f(\overline{x}_m - \overline{x}_f)^2 + w_f n_f(w_m n_m + w_f n_f)Var(\mathbf{x}_f)\right)^2} \tag{3.64}$$

For an elaboration on the computations see Appendix H.

What can be noticed here is that the influence of the weights lies in the difference in proportions between the $w_m$ and $w_f$. Above as well as below in the fraction, there is always a fourth power of weights. As an example, let us say $w_m = 10$ and $w_f = 2$. If the weights would be $w_m = 20$ and $w_f = 4$, this multiplicative factor of 2 can be extracted out of each $w$ above and below the fraction line since the adjustment is a constant applied to all weights and can be divided out of the equation, giving the same

estimate as the other weights. Newsom et al. supports this statement where normalisation of the weights does not influence the estimate or the variance when using inverse selection probabilities for the weights.

So when again inserting the definitions of $w_m$ and $w_f$, we get;

$$Var(\hat{\beta}_{1,w}) = \frac{\begin{array}{l}\sigma^2_{\epsilon,m}(\frac{n\cdot p_m}{n_m})^2 n_m(\frac{n\cdot p_m}{n_m}n_m + \frac{n\cdot p_f}{n_f}n_f)^2 Var(\mathbf{x}_m) \\ \quad + (\frac{n\cdot p_m}{n_m})^2(\frac{n\cdot p_f}{n_f})^2 n_m n_f(\sigma^2_{\epsilon,m}n_f + \sigma^2_{\epsilon,f}n_m)(\overline{x}_m - \overline{x}_f)^2 + \sigma^2_{\epsilon,f}(\frac{n\cdot p_f}{n_f})^2 n_f(\frac{n\cdot p_m}{n_m}n_m + \frac{n\cdot p_f}{n_f}n_f)^2 Var(\mathbf{x}_f)\end{array}}{\left(\frac{n\cdot p_m}{n_m}n_m(\frac{n\cdot p_m}{n_m}n_m + \frac{n\cdot p_f}{n_f}n_f)Var(\mathbf{x}_m) + \frac{n\cdot p_m}{n_m}\frac{n\cdot p_f}{n_f}n_m n_f(\overline{x}_m - \overline{x}_f)^2 + \frac{n\cdot p_f}{n_f}n_f(\frac{n\cdot p_m}{n_m}n_m + \frac{n\cdot p_f}{n_f}n_f)Var(\mathbf{x}_f)\right)^2} \tag{3.65}$$

$$= \frac{n^4}{n^4}\frac{\begin{array}{l}\sigma^2_{\epsilon,m}\frac{p_m^2}{n_m}(p_m + p_f)^2 Var(\mathbf{x}_m) \\ \quad + \frac{p_m^2}{n_m}\frac{p_f^2}{n_f}(\sigma^2_{\epsilon,m}n_f + \sigma^2_{\epsilon,f}n_m)(\overline{x}_m - \overline{x}_f)^2 + \sigma^2_{\epsilon,f}\frac{p_f^2}{n_f}(p_m + p_f)^2 Var(\mathbf{x}_f)\end{array}}{\left(p_m(p_m + p_f)Var(\mathbf{x}_m) + p_m p_f(\overline{x}_m - \overline{x}_f)^2 + p_f(p_m + p_f)Var(\mathbf{x}_f)\right)^2} \tag{3.66}$$

$$= \frac{\sigma^2_{\epsilon,m}\frac{p_m^2}{n_m}Var(\mathbf{x}_m) + \frac{p_m^2}{n_m}\frac{p_f^2}{n_f}(\sigma^2_{\epsilon,m}n_f + \sigma^2_{\epsilon,f}n_m)(\overline{x}_m - \overline{x}_f)^2 + \sigma^2_{\epsilon,f}\frac{p_f^2}{n_f}Var(\mathbf{x}_f)}{\left(p_m Var(\mathbf{x}_m) + p_m p_f(\overline{x}_m - \overline{x}_f)^2 + p_f Var(\mathbf{x}_f)\right)^2} \tag{3.67}$$

With standard error;

$$se(\hat{\beta}_{1,w}) = \sqrt{Var(\hat{\beta}_{1,w})} \tag{3.68}$$

In the estimate, all $n_m$'s and $n_f$'s were canceled out with the weights, but now there are still some left. Therefore, when increasing the sample sizes, meaning increasing $n_f$ and $n_m$, they ensure that the variance overall is decreasing. This finding is also supported by Newsom et al..

### 3.2.3 Variance ($\boldsymbol{\sigma^2_{\epsilon,m} = \sigma^2_{\epsilon,f}}$)

If again the assumption would be made that $\sigma^2_{\epsilon,m} = \sigma^2_{\epsilon,f} = \sigma^2_\epsilon$, then the weighted variance becomes the following;

The variance of the estimator is:

$$Var(\hat{\boldsymbol{\beta}}_w) = Var\left((\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}\right) \tag{3.69}$$

$$= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\cdot Var(\mathbf{y})\cdot\left((\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\right)^T \tag{3.70}$$

$$= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\cdot\sigma^2_\epsilon\cdot\left((\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\right)^T \quad \text{since we know } y \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2_\epsilon) \tag{3.71}$$

$$= \sigma^2_\epsilon\cdot(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{W}^T\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1} \tag{3.72}$$

$$= \sigma^2_\epsilon\cdot(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^2\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1} \quad \text{since } \mathbf{W} \text{ is diagonal, we know } \mathbf{W} = \mathbf{W}^T \tag{3.73}$$

As can be seen easily, is that not a lot can be simplified as was the case in the unweighted analysis. This is due to this extra $\mathbf{W}$ matrix.

When $\sigma^2_{\epsilon,m} = \sigma^2_{\epsilon,f} = \sigma^2_\epsilon$, then;

$$Var(\hat{\beta}_1) = \frac{\sigma^2_\epsilon\frac{p_m^2}{n_m}Var(\mathbf{x}_m) + \frac{p_m^2}{n_m}\frac{p_f^2}{n_f}(\sigma^2_\epsilon n_f + \sigma^2_\epsilon n_m)(\overline{x}_m - \overline{x}_f)^2 + \sigma^2_\epsilon\frac{p_f^2}{n_f}Var(\mathbf{x}_f)}{\left(p_m Var(\mathbf{x}_m) + p_m p_f(\overline{x}_m - \overline{x}_f)^2 + p_f Var(\mathbf{x}_f)\right)^2} \tag{3.74}$$

$$= \sigma^2_\epsilon\cdot\frac{\frac{p_m^2}{n_m}Var(\mathbf{x}_m) + \frac{p_m^2}{n_m}\frac{p_f^2}{n_f}(n_f + n_m)(\overline{x}_m - \overline{x}_f)^2 + \frac{p_f^2}{n_f}Var(\mathbf{x}_f)}{\left(p_m Var(\mathbf{x}_m) + p_m p_f(\overline{x}_m - \overline{x}_f)^2 + p_f Var(\mathbf{x}_f)\right)^2} \tag{3.75}$$

$$= \sigma^2_\epsilon\cdot\frac{\frac{p_m^2}{n_m}Var(\mathbf{x}_m) + \frac{p_m^2}{n_m}\frac{p_f^2}{n_f}n(\overline{x}_m - \overline{x}_f)^2 + \frac{p_f^2}{n_f}Var(\mathbf{x}_f)}{\left(p_m Var(\mathbf{x}_m) + p_m p_f(\overline{x}_m - \overline{x}_f)^2 + p_f Var(\mathbf{x}_f)\right)^2} \tag{3.76}$$

where again $\sigma_\epsilon^2$ will can be approximated with;

$$\hat{\sigma_\epsilon}^2 = \frac{1}{n-2} WSS_\epsilon = \frac{1}{n-2} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$$

Then lastly, this gives as the standard error on the estimate;

$$se(\hat{\beta}_{1,w}) = \sqrt{Var(\hat{\beta}_{1,w})} \tag{3.77}$$

The conclusions as in the previous section are still valid, but taking this assumption does not add new information, except for the extraction of the $\sigma_\epsilon^2$.

### 3.2.4 Variance (heteroscedasticity)

As explained in the introduction, weights have also been created with the purpose to counteract the heteroscedasticity of the data. This is a different goal then what we aim to use weights for in this thesis, namely sampling design correction. It is also not possible to use two sets of weights simultaneously, one for correction of the heteroscedasticity and one for correction of the sample design. Therefore, we will not consider this kind of weights in the next section where we'll illustrate some example cases. However, many software tools for regression analysis support these other kinds of weights as well or even assume the user will only incorporate weights for heteroscedastic purposes. To make sure users recognize the different usages of weights in software documentation, we will elaborate on this kind of weight briefly. A common structure used is to choose the weights the inverse observation variance.

$$w_i = \frac{1}{\sigma_i^2} \tag{3.78}$$

In the case of assuming different epsilon's for males and females, this would mean;

$$w_m = \frac{1}{\sigma_{\epsilon,m}^2} \text{ and } w_f = \frac{1}{\sigma_{\epsilon,f}^2} \tag{3.79}$$

That means in the variance computation

$$Var(\mathbf{y}) \cdot \mathbf{W}^T = diag(\sigma_{\epsilon,m}^2, \cdots, \sigma_{\epsilon,m}^2, \sigma_{\epsilon,f}^2, \cdots, \sigma_{\epsilon,f}^2) \cdot diag\left(\frac{1}{\sigma_{\epsilon,m}^2}, \cdots, \frac{1}{\sigma_{\epsilon,m}^2}, \frac{1}{\sigma_{\epsilon,f}^2}, \cdots, \frac{1}{\sigma_{\epsilon,f}^2}\right) \tag{3.80}$$

$$= 1 \tag{3.81}$$

So,

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}_w) &= Var\left((\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}\right) && (3.82)\\ &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W} \cdot Var(\mathbf{y}) \cdot \mathbf{W}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1} && (3.83)\\ &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W} \cdot 1 \cdot \mathbf{X}\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1} && (3.84)\\ &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{X}\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1} && (3.85)\\ &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1} && (3.86) \end{aligned}$$

Then;

$$Var(\hat{\boldsymbol{\beta}}_w) = \frac{1}{C_w} \begin{bmatrix} w_m \sum_{i=1}^{n_m} x_i^2 + w_f \sum_{i=n_m+1}^{n} x_i^2 & -\left(w_m \sum_{i=1}^{n_m} x_i + w_f \sum_{i=n_m+1}^{n} x_i\right) \\ -\left(w_m \sum_{i=1}^{n_m} x_i + w_f \sum_{i=n_m+1}^{n} x_i\right) & w_m n_m + w_f n_f \end{bmatrix} \tag{3.87}$$

## 3.3 Numerical cases

To illustrate the differences between weighted and unweighted survey analysis, some examples are created. Again consider the general case described in the beginning of this chapter. We consider a sample of $n = 500$ people where there are $n_m$ males and $n_f$ females. The variable of interest for the males $x_m$ are sampled from a Normal distribution $N(\mu_m, \sigma_m^2)$ and for females $x_f$ from a Normal distribution $N(\mu_f, \sigma_f^2)$. The linear model considered for the relation between the response and the exposure for males is; $y_i = \beta_{0,m} + \beta_1 \cdot x_{m,i} + \epsilon_{m,i}$, where $\epsilon_{m,i} \sim N(0, \sigma_{\epsilon,m}^2)$ and $\beta_1 = 10$. and equivalently for females is; $y_i = \beta_{0,f} + \beta_1 \cdot x_{f,i} + \epsilon_{f,i}$, where $\epsilon_{f,i} \sim N(0, \sigma_{\epsilon,f}^2)$. This relation will be approximated with one regression line $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ where the assumption is made that the $\sigma_{\epsilon,m}^2 = \sigma_{\epsilon,f}^2$.

Some differences that would interesting to include are:

- **The amount of males and females in your sample**
  Often, a sample is not a perfect representation of the population based on the occurrence of some attributes in the data. The regression could be influenced if there are, for instance, more males then females in the sample, and especially if males have different properties then females. We would like to consider some variations in the amount of males and females in the sample.

  1. $n_m = 50$ and $n_f = 450$
  2. $n_m = 249$ and $n_f = 251$
  3. $n_m = 450$ and $n_f = 50$

- **The assumptions about the males and female properties**
  In research it is common that the properties for males and females differ. For instance males are most of the time taller then females. Then the assumption that the distribution of the $x_m$'s and $x_f$'s is equal would be invalid. Also the variance for males and females could differ. Some combinations can be made that could be a potential realisation of the sample;

  a. same distribution; $x_m \sim N(10, 1)$, $x_f \sim N(10, 1)$ and $\sigma_{\epsilon,m}^2 = \sigma_{\epsilon,f}^2 = 1$
  b. different mean; $x_m \sim N(100, 1)$, $x_f \sim N(10, 1)$ and $\sigma_{\epsilon,m}^2 = \sigma_{\epsilon,f}^2 = 1$
  c. different mean and sd; $x_m \sim N(100, 10)$, $x_f \sim N(10, 1)$ and $\sigma_{\epsilon,m}^2 = \sigma_{\epsilon,f}^2 = 1$
  d. different error; $x_m \sim N(10, 1)$, $x_f \sim N(10, 1)$, $\sigma_{\epsilon,m}^2 = 4$ and $\sigma_{\epsilon,f}^2 = 1$
  e. different mean, sd and error; $x_m \sim N(100, 10)$, $x_f \sim N(10, 1)$, $\sigma_{\epsilon,m}^2 = 4$ and $\sigma_{\epsilon,f}^2 = 1$

- **Intercept** The value for $\beta_0$ can also vary. One sex could have a higher intercept value, but still the same association. Therefore we consider two cases;

  i. same intercept; $\beta_{0,m} = \beta_{0,f} = 20$
  ii. different intercept; $\beta_{0,m} = 30$ and $\beta_{0,f} = 20$

- **Inclusion of weights**

  uw. an unweighted analysis, so all weights = 1.
  w. a weighted analysis, where the weights are the inverse inclusion probability as defined in Equation (3.2) and (3.3).

One might wonder why the slope is not varied in these cases. If males and females have a different slope, this simple model is misspecified, since an interaction term between the exposure and sex would be necessary. This would contradict the situation and mathematics computed above, so will not be considered in this case analysis.

A simple program is written in R, that calculates the estimate and variance as described above. We create 10.000 datasets that produce an equal amount of estimates $\hat{\boldsymbol{\beta}}_j$ and corresponding variances. The mean value will be reported, to counteract possible outlier results. Also the coverage is calculated with formula;

$$coverage(\hat{\boldsymbol{\beta}}) = \frac{1}{10000} \sum_{j=1}^{10000} \mathbb{1}\left(\boldsymbol{\beta} \in \left[\hat{\boldsymbol{\beta}}_j - t_{\alpha/2} \cdot se(\hat{\boldsymbol{\beta}}_j), \hat{\boldsymbol{\beta}}_j + t_{1-\alpha/2} \cdot se(\hat{\boldsymbol{\beta}}_j)\right]\right) \tag{3.88}$$

with $\alpha = 0.05$ for a 95% confidence interval.

When creating the results it was found out that situation *ii* did not give feasible results, since the estimation had a problem fitting only one line for a population that should have two different lines (at least for a different intercept). Especially when the mean of the distribution of the $x_m$ and $x_f$ were far apart, the coverage even became 0. To correct for this, sex needs to be included in the model as well, just like mentioned before for the different slopes, but that would not present more information on the case we have considered in the beginning of this chapter. Therefore, the results will not consider situation *ii*.

### 3.3.1 Results

This presents the results for the intercept in Table 3.1 and for the slope in Table 3.2.

| | | i. equal intercept | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | a. same distribution | | b. different mean | | c. different mean, sd | | d. different error | | e. diff. error, mean, sd | |
| #people | Measures | uw. | w. | uw. | w. | uw. | w. | uw. | w. | uw. | w. |
| 1. | Estimate | 20,0056 | 20,0113 | 20,0002 | 20,0002 | 20,0002 | 20,0004 | 20,0078 | 20,0204 | 20,0003 | 20,0007 |
| 50m - 450f | Standard error | 0,4504 | 0,7432 | 0,0547 | 0,0544 | 0,0545 | 0,0563 | 0,5133 | 1,1697 | 0,0621 | 0,0886 |
| | Coverage | 0,9496 | 0,9463 | 0,9504 | 0,9466 | 0,9488 | 0,9432 | 0,9492 | 0,8870 | 0,9508 | 0,9843 |
| | RSE | 0,9998 | 0,9953 | 0,9998 | 0,9953 | 0,9998 | 0,9952 | 1,1396 | 1,5664 | 1,1375 | 1,5663 |
| 2. | Estimate | 19,9987 | 19,9987 | 20,0006 | 20,0006 | 20,0006 | 20,0006 | 19,9990 | 19,9990 | 20,0007 | 20,0007 |
| 249m - 251f | Standard error | 0,4501 | 0,4501 | 0,0704 | 0,0704 | 0,0699 | 0,0699 | 0,7106 | 0,7115 | 0,1104 | 0,1105 |
| | Coverage | 0,9528 | 0,9529 | 0,9523 | 0,9523 | 0,9522 | 0,9523 | 0,9513 | 0,9510 | 0,9968 | 0,9968 |
| | RSE | 0,9994 | 0,9994 | 0,9994 | 0,9994 | 0,9994 | 0,9994 | 1,5780 | 1,5799 | 1,5780 | 1,5798 |
| 3. | Estimate | 20,0044 | 20,0090 | 20,0013 | 20,0012 | 20,0014 | 20,0013 | 20,0072 | 20,0104 | 20,0017 | 20,0013 |
| 450m - 50f | Standard error | 0,4500 | 0,7422 | 0,1570 | 0,1563 | 0,1490 | 0,1544 | 0,8657 | 1,1782 | 0,2868 | 0,2450 |
| | Coverage | 0,9487 | 0,9490 | 0,9474 | 0,9460 | 0,9480 | 0,9453 | 0,9486 | 0,9934 | 0,9990 | 0,9981 |
| | RSE | 0,9994 | 0,9944 | 0,9994 | 0,9943 | 0,9994 | 0,9944 | 1,9225 | 1,5784 | 1,9236 | 1,5785 |

Table 3.1: Case analysis results of the intercept

| | | i. equal intercept | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | a. same distribution | | b. different mean | | c. different mean, sd | | d. different error | | e. diff. error, mean, sd | |
| #people | Measures | uw. | w. | uw. | w. | uw. | w. | uw. | w. | uw. | w. |
| 1. | Estimate | 9,9995 | 9,9989 | 10,0000 | 10,0000 | 10,0000 | 10,0000 | 9,9992 | 9,9980 | 10,0000 | 10,0000 |
| 50m - 450f | Standard error | 0,0448 | 0,0739 | 0,0017 | 0,0016 | 0,0016 | 0,0016 | 0,0511 | 0,1164 | 0,0019 | 0,0026 |
| | Coverage | 0,9489 | 0,9473 | 0,9499 | 0,9483 | 0,9495 | 0,9477 | 0,9495 | 0,8858 | 0,7547 | 0,8894 |
| | RSE | 0,9998 | 0,9953 | 0,9998 | 0,9953 | 0,9998 | 0,9952 | 1,1396 | 1,5664 | 1,1375 | 1,5663 |
| 2. | Estimate | 10,0001 | 10,0001 | 10,0000 | 10,0000 | 10,0000 | 10,0000 | 10,0001 | 10,0001 | 10,0000 | 10,0000 |
| 249m - 251f | Standard error | 0,0448 | 0,0448 | 0,0010 | 0,0010 | 0,0010 | 0,0010 | 0,0707 | 0,0708 | 0,0015 | 0,0016 |
| | Coverage | 0,9508 | 0,9506 | 0,9499 | 0,9500 | 0,9496 | 0,9496 | 0,9511 | 0,9508 | 0,9472 | 0,9474 |
| | RSE | 0,9994 | 0,9994 | 0,9994 | 0,9994 | 0,9994 | 0,9994 | 1,5780 | 1,5799 | 1,5780 | 1,5798 |
| 3. | Estimate | 9,9996 | 9,9991 | 10,0000 | 10,0000 | 10,0000 | 10,0000 | 9,9993 | 9,9990 | 10,0000 | 10,0000 |
| 450m - 50f | Standard error | 0,0448 | 0,0738 | 0,0017 | 0,0016 | 0,0016 | 0,0016 | 0,0861 | 0,1172 | 0,0030 | 0,0026 |
| | Coverage | 0,9505 | 0,9500 | 0,9488 | 0,9453 | 0,9477 | 0,9451 | 0,9509 | 0,9939 | 0,9971 | 0,9932 |
| | RSE | 0,9994 | 0,9944 | 0,9994 | 0,9943 | 0,9994 | 0,9944 | 1,9225 | 1,5784 | 1,9236 | 1,5785 |

Table 3.2: Case analysis results of the slope

Looking at the results, it's difficult to state that one of the two analyses gives consistently better results then the other.

When for instance considering cases *1.i.a*, *1.i.d*, *3.i.a* and *3.i.d* (marked in purple), the average standard error of the slope for the weighted analysis is reasonably higher then the unweighted analysis. However, the coverage is still approximately 95% for both analysis, which would indicate that neither underestimate the standard error. This is true except for *1.i.d*, where the weighted estimate still has a higher standard error, but also has lower coverage. So from these results the recommendation would be to use the unweighted analysis in cases similar to these, since the standard error is lower and does not underestimate it.

Also when the data is already quite nicely distributed, assuming the population has a 50-50 proportions males and females, like in case 2, the effect of adding weights is negligible. For simplicity, an unweighted analysis would then be recommended.

However, when the difference between the properties for males and females starts to increase, the unweighted analysis starts underestimating the standard error, and the weighted becomes better. This can be seen in cases *1.i.e* (marked in yellow), where the standard error of the slope of the unweighted analysis is reasonable lower then in the weighted analysis, but the coverage has also dropped 20%, which shows the standard error is underestimated. Then the weighted standard error performs better.

In case *3.i.e* the weighted analysis even had a lower standard error then the unweighted analysis, while still maintaining a good coverage.

The current hypothesis is that when differences between subgroups are increased, the weighted estimator starts to show some of its advantages.

## 3.4 Conclusion

The overall conclusion that this chapter has shown is that the unweighted analysis performed better in most of the tested situations. However, the weighted regression showed some promise as well, so it is setting-dependent when a weighted analysis would be better then an unweighted analysis, and vice versa. In the performed case analysis, only a few cases are considered on a very basic level, so it is difficult to generalize these findings to larger, more complex studies. To be able to advise researchers in weight usage in larger studies, the mathematical analysis will not be sufficient. A simulation study specific to similar data, model and sample design is needed to give some more insight into the process.

# Chapter 4

# Simulation study

## 4.1 Simulation Design

After having explored the mathematical properties of weighted and unweighted regression and illustrated their performance on a simple controlled study with varying proportions, we perform a larger simulation study. The aim is to extend our formal check to more complex datasets (including various correlations that normally occur in practice) and provide recommendations based on our conclusions. We consider the CLSA dataset and create replicates by introducing some variability, while keeping the inherent structure of the dataset. Then we investigate the performance of weighted and unweighted regression in the study of the association between the Overall Social Support (OSS) and the Executive Function (EF), introduced in Section 2.1.3. Since the CLSA dataset is complex, choices and assumptions need to be made when creating datasets of similar complexity to preserve the structure and internal relationships between variables.

In this chapter we make such choices explicit and explain the set-up of the simulation study. We present the general set-up visually and then go through each step in detail in the following sections. The procedure is also illustrated via a pseudocode of the algorithm. At the end of this chapter we present and illustrate the results of our simulation study.

### 4.1.1 General set-up

In this section we introduce the general set-up of the simulation study illustrated in Figure 4.1.

We consider the CLSA dataset, which has been elaborately introduced in Section 2.1.3, and can be seen as box A in Figure 4.1. We define the vector of 'true' regression coefficients, $\boldsymbol{\beta}_{TRUE} = (\beta_{A_1,TRUE}, \dots, \beta_{OSS,TRUE})$, as the vector of parameter values that one would get when fitting a unweighted regression model on the complete data from the Canadian population (box B), see red arrow (2).

To make this analysis possible, the data on the entire population has to be available, but we only have a sample, CLSA (box A). One possible way is to upscale the CLSA dataset using a set of weights $\mathbf{W}_T$ to approximate the Canadian population (illustrated with red arrow (1)). $\mathbf{W}_T$ are the weights that would upscale the CLSA to the Canadian population and then get $\beta_{\mathbf{W}_T,TRUE}$ via unweighted regression (red arrow (2)). However, this is computationally intensive and requires large memory. An alternative is to include weights $\mathbf{W}_T$ in a weighted regression analysis. With this method we would overcome the whole generation process and still get the estimated 'true' coefficients $\boldsymbol{\beta}_{\mathbf{W}_T,TRUE}$ (blue arrow (3)). In fact, using these weights $\mathbf{W}_T$ in a weighted regression model on the sample (box A) will produce the same $\beta_{\mathbf{W}_T,TRUE}$ as performing an unweighted regression on the estimated complete population (box B). In our simulation study we take the blue route to get the 'true' coefficients. A detailed elaboration on the choices made for $\mathbf{W}_T$ is presented in Section 4.1.2.

Then we generate $m$ replicates of the CLSA dataset (boxes $C_1$ to $C_m$) by altering the CLSA dataset, as indicated with blue line (4). A detailed description of the procedure of this generation and the assumptions made are provided in Section 4.1.3. We fit weighted and unweighted linear regression models on each of the $m$ datasets, with different weight options. Details on this part are provided in Section 4.1.4. Finally, we compare the estimated regression coefficients $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_w$ from the unweighted and weighted

analysis to the 'true' estimates $\boldsymbol{\beta}_{\mathbf{W}_T,TRUE}$ in terms of bias, variance and the coverage probability (see Section 4.1.5). A pseudo algorithm of the simulation presented in this chapter is available in Section 4.1.7.

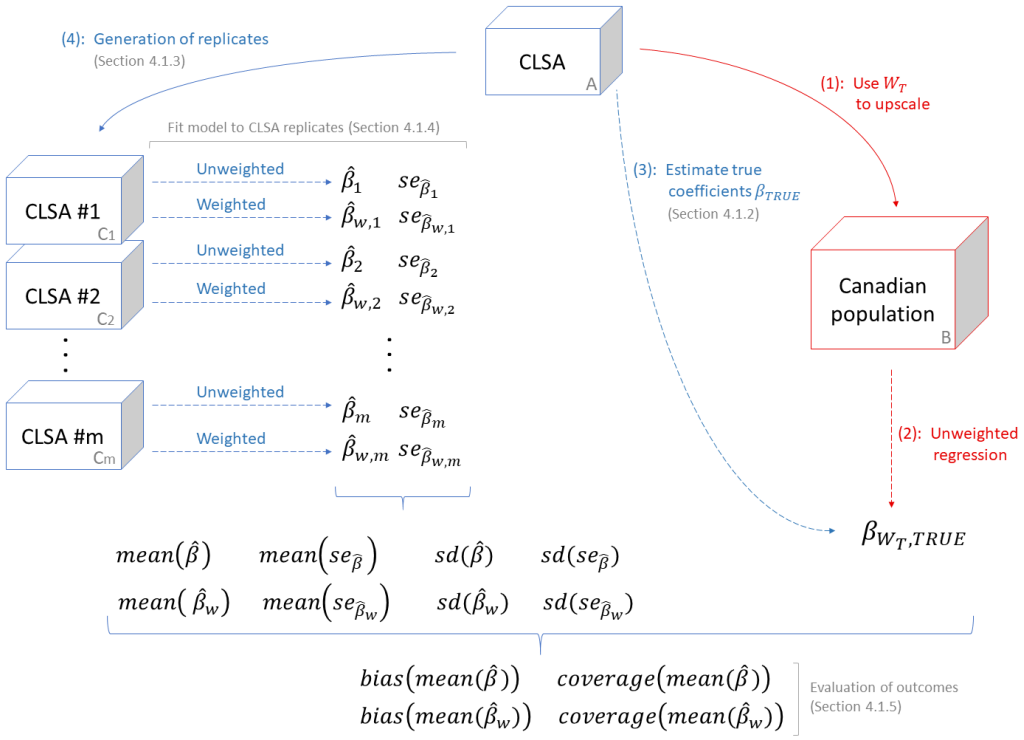


Figure 4.1: Set-up of simulation
The boxes denote datasets; continuous arrows denote generation of datasets; dashed arrows denote estimation of regression coefficients; the blue paths are taken in this simulation, the red are not; sections refer to parts of this chapter relative to that specific step.

### 4.1.2 Estimate True Coefficients $\boldsymbol{\beta}_{TRUE}$

We would like to use a set of weights $\mathbf{W}_T$ that best represent the population, but every set of calculated weights will always be an estimation of the actual ones. The CLSA weights are no exception, so we take the CLSA analytical weights as a basis and add some white noise to them that would represent the deviation with the actual weights. That is, we assume the estimated weights are centered, meaning half of the new weights will be lower then originally and half will be higher, but the amount of variability will differ, and we investigate the effect corresponding to different noise levels. Note that the procedure for calculating weights is very elaborate, so we cannot implement calculation of new weights.

We chose to perturbate the weights with multiplicative error that is random and non-negative. To keep the weights centered and to be able to better compare the estimates of the standard error, the analytical weights with noise still have to be centered, so have a median of 1, and cannot become negative. Therefore, the weights will be multiplied with a log-normal distribution, where the corresponding normal distribution has a mean of 0 and some variance $\sigma_n^2$. The log-normal distribution will be appropriate, since it can never go negative due to the exponential function. Note that the expected median of the

new weights with noise $\mathbf{W}_{noise(\sigma_n)}$ is still the same, however, the expectation changes:

$$\begin{aligned}
\mathbb{E}[\mathbf{W}_{noise(\sigma_n)}] &= \mathbb{E}[\mathbf{W}_{analytical} \cdot L], \text{ where } L \sim \text{Lognormal}(\mu = 0, \sigma_n^2) && (4.1)\\
&= \mathbb{E}[\mathbf{W}_{analytical}] \cdot \mathbb{E}[L] && (4.2)\\
&= \bar{\mathbf{W}}_{analytical} \cdot \exp(0 + \sigma_n^2/2) && (4.3)\\
&= 1 \cdot \exp(\sigma_n^2/2) && (4.4)\\
&\neq 1, \text{ unless } \sigma_n = 0, \text{ when no noise is added.} && (4.5)
\end{aligned}$$

The last thing to consider is the size of the $\sigma_n$. It is beneficial to choose various sigma's in order to see the effect of the uncertainty in estimating weights. This is especially interesting when the weighted analysis would perform better then the unweighted analysis. Then it can be examined how much the weights may differ from the actual weights in order to still perform better then the unweighted analysis. The selected values for $\sigma_n$ are provided in Table 4.1. The largest mean change in weights is $1.13 = exp(1/8)$.

| $\sigma_n$ | normal interval | log-normal interval |
|---|---|---|
| 0.05 | $[-0.098; 0.098]$ | [0.907; 1.103] |
| 0.125 | $[-0.245; 0.245]$ | [0.783; 1.278] |
| 0.25 | $[-0.409; 0.490]$ | [0.613; 1.632] |
| 0.5 | $[-0.980; 0.980]$ | [0.375; 2.664] |

Table 4.1: choices for $\sigma_n$

So to summarize, the various weights $\mathbf{W}_T$ that will be used to estimate the $\boldsymbol{\beta}_{\mathbf{W}_T,TRUE}$ are; analytical weights with perturbation $\sigma_n = \{0, 0.05, 0.125, 0.25, 0.5\}$, inflation weights and unweighted.

### 4.1.3 Generation of replicates

In this section we elaborate more on the choices we make to get the replicates of the CLSA datasets. We are adding variability while keeping the inherent structure and prevalence of the data attributes.

As explained in Section 2.1.2, the CLSA weights assigned to each participant are based on sex, age group, province and high probability of low/non-low educated people in that province, together also known as the weight variables. If the value of some of these weight variables change, or the proportions between the groups of these variables in the sample change, the weights should change accordingly. For example consider the inflation weights. When only taking a subsample of the CLSA dataset, the weights should be recalculated. The CLSA weights have been trimmed and calibrated with several iterations, thus the recalculation of weights to new values according to the different proportions among variables would require a study on itself. Therefore, we chose to fix the weights, and also the weight variables. This way we ensure that no recalculation of the weights is needed and the current values of the weights remain appropriate (see **Generation of replicates** algorithm - step 2.a in Section 4.1.7).

On the other hand, we want to perturbate the original dataset through the independent and exposure variables. For these, we generate new values while accounting for existing correlations. To illustrate the importance of preserving correlations, we make a small example. Consider having smoking and lung cancer as the independent and exposure variable respectively. When generating these values independently and at random, there is a chance that in the new dataset relatively more people suddenly don't smoke, but still have lung cancer. This affects the inherent structure and correlations between these variables and thus conclusions drawn in further analysis, since the influence of smoking on lung cancer will have decreased in the new dataset. One way to cope with this problem is to generate new values while taking the correlations into account. This can be achieved, for example, by using a Gaussian copula (see **Data Generation** algorithm - step 1 in Section 4.1.7). We draw random samples from the uniform distribution $\boldsymbol{u} = (u_1, ..., u_d)$ and then use the joint distribution $\Phi_R$ to generate new values for the independent and exposure variables, that comply with the original correlation.

The Gaussian copula's cumulative function is defined as;

$$C_R^{\text{Gauss}}(u) = \Phi_R\left(\Phi^{-1}(u_1), \ldots, , \Phi^{-1}(u_d)\right), \tag{4.6}$$

with $\Phi_R$ the joint cumulative distribution function of a multivariate normal distribution with mean vector zero and covariance matrix equal to the correlation matrix $R$ and $\Phi$ the cumulative distribution function of a standard normal. We construct a 1000 CLSA replicates. Each replicate will keep the CLSA's ID values, the weights variables and the CLSA weights (see **Data Generation** algorithm - step 2.a in Section 4.1.7). For each row of each new dataset, uniform values for all the independent and exposure variables are sampled according to the copula. These values are transformed back using the empirical cumulative functions (see **Data Generation** algorithm - step 2.b.i in Section 4.1.7). The response value is then calculated for each new combination of independent and exposure variables, given the weight variables, the previously defined $\boldsymbol{\beta}_{TRUE}$, and model 0 as defined in Section 2.1.4 (see **Data Generation** algorithm - step 2.b.iii in Section 4.1.7).

To experiment with the different levels of explained variance, we consider two values for $\sigma_\epsilon$ corresponding to $R^2 \approx 0.7$ and to $R^2 \approx 0.3$. This corresponds to the two values $\sigma_\epsilon = 0.75$ and $\sigma_\epsilon = 2$ respectively. The structure of each generated replicate of the CLSA dataset is provided schematically in Figure 4.2.

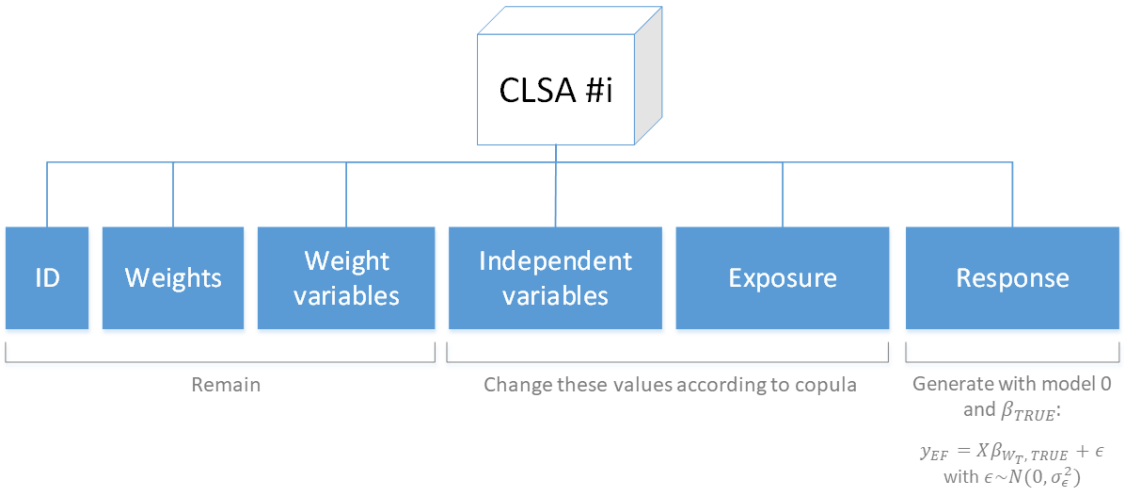


Figure 4.2: Elaboration on the variables for generation of CLSA $\#i$

### 4.1.4 Fit model to CLSA replicates

When the alternative datasets are created, we can produce estimates for the regression coefficients of the population using regression models. We again consider model 0. To evaluate the impact of leaving out some of the variables, we also define another model, model 1, which is equivalent to model 0, but the variable *province* is not included. This presents the following two considered models: (see **Fit model to CLSA replicates** algorithm - step 1 in Section 4.1.7)

**Model 0:**

$$\begin{aligned}\mathbf{y}_{EF} = \; & \beta_S \cdot \mathbf{x}_S + \cdots + \beta_{A_4} \cdot \mathbf{x}_{A_4} \\ & + \beta_{P_1} \cdot \mathbf{x}_{P_1} + \cdots + \beta_{P_7} \cdot \mathbf{x}_{P_7} \\ & + \beta_{Edu} \cdot \mathbf{x}_{Edu} + \cdots + \beta_{DS} \cdot \mathbf{x}_{DS} \\ & + \beta_{OSS} \cdot \mathbf{x}_{OSS} + \boldsymbol{\epsilon}\end{aligned} \tag{4.7}$$

with $\boldsymbol{\epsilon} \sim N(0, \sigma_\epsilon^2)$.

**Model 1:**

$$\begin{aligned}\mathbf{y}_{EF} = \; & \beta_S \cdot \mathbf{x}_S + \cdots + \beta_{A_4} \cdot \mathbf{x}_{A_4} \\ & + \beta_{Edu} \cdot \mathbf{x}_{Edu} + \cdots + \beta_{DS} \cdot \mathbf{x}_{DS} \\ & + \beta_{OSS} \cdot \mathbf{x}_{OSS} + \boldsymbol{\epsilon}\end{aligned} \tag{4.8}$$

with $\boldsymbol{\epsilon} \sim N(0, \sigma_\epsilon^2)$.

A set of analysis weights $\mathbf{W}_A$ are used in a weighted regression analysis on each CLSA replicate, that produces a vector of estimates and their standard error for each replicate. The analysis weights will be the CLSA weights (inflation and analytical) and unweighted (where the weights equal 1).

The estimation of the regression coefficients is done with a (weighted) least squares method, as in Chapter 3. Giving for each dataset CLSA $\#i$

$$\hat{\boldsymbol{\beta}}_{W_A,i} = (\mathbf{X}^T\mathbf{W}_A\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_A\,\mathbf{y} \tag{4.9}$$

and the standard error $se(\hat{\boldsymbol{\beta}}_{W_A,i})$. (see **Fit model to CLSA replicates** algorithm - step 2.a in Section 4.1.7) More on the computation of the standard error in software can be found in Section 4.1.6.

The estimates and their standard error are then combined to receive measures for comparison between different weights and fits. These measures are the mean estimate, the standard deviation of the estimates, the mean standard error, the bias and the coverage.

### 4.1.5 Evaluation of outcomes

In this section we provide the measures to evaluate the different fits in the simulation study. Using the estimates and the standard errors, the bias of the estimator and the coverage can be derived, where

$$b(\hat{\beta}_{\mathbf{W}_A,OSS}) = \left| \left( \frac{1}{\#runs} \sum_{i=1}^{\#runs} \hat{\beta}_{\mathbf{W}_A,OSS,i} \right) - \beta_{\mathbf{W}_T,OSS,TRUE} \right| \tag{4.10}$$

and

$$coverage(\hat{\beta}_{\mathbf{W}_A,OSS}) = \frac{1}{\#runs} \sum_{i=1}^{\#runs} \mathbb{1}\left( \beta_{\mathbf{W}_T,OSS,TRUE} \in \left[ \hat{\beta}_{\mathbf{W}_A,OSS,i} - t_{\alpha/2} \cdot se(\hat{\beta}_{\mathbf{W}_A,OSS,i}) \;,\; \hat{\beta}_{\mathbf{W}_A,OSS,i} + t_{1-\alpha/2} \cdot se(\hat{\beta}_{\mathbf{W}_A,OSS,i}) \right] \right), \tag{4.11}$$

with $\alpha = 0.05$ for a 95% confidence interval, $t_{\alpha/2}$ quantile from t distribution and with the number of runs equal to 1000.

In the introduction the goals of the simulation where briefly highlighted, but now these can be given in more detail.

1. The main goal is to see for which type of weight, if any, the bias $b(mean(\hat{\beta}_{\mathbf{W}_A,OSS}))$ in estimating the association between the response and the exposure is minimal.
2. See for which type of weight, if any, the mean standard error $mean(se(\hat{\beta}_{\mathbf{W}_A,OSS}))$ of the estimates is the lowest, while taking into account the coverage $coverage(\hat{\beta}_{\mathbf{W}_A,OSS})$.
3. If it appears that a weighted analysis does perform better, how robust and consistent is the estimate with its standard error and coverage if the CLSA weights are further removed from the actual weights (CLSA weights with some noise)? Is there a turning point when weighting isn't appropriate anymore?

### 4.1.6 Program usage

There are various linear regression procedures in SAS, like PROC GLM and PROC GENMOD, which also have the possibility to include weights. However, not every procedure uses weights in the same way. Some procedures aim to use weights for the correction of heteroscedasticity, like described in Section 3.2.4, where the variance can be simplified to $\sigma_\epsilon^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$. However, for survey analysis this estimation is incorrect and specific procedures for survey analysis have to be used to ensure this. Most survey regression procedures however are very complex and let clustering and stratification be taken

into account. Due to these extra effects the theoretical formula described in Section 3.2.2 is not usable anymore and variance estimation methods need to be considered.

In the procedure used in the research of Section 2.1.4, PROC SURVEYREG is used and recommended for complex survey analysis. This procedure has three possible variance estimation methods; Taylor series linearization and replication methods; Balanced Repeated Replication (BRR) and Jackknife. In this simulation the variance will be estimated using the Taylor series method, because this is one of the most straightforward methods for estimation, is less computationally intensive then the alternatives and is commonly recommended in literature.

Linearization entails approximating a nonlinear statistic with a linear function of the observations by using first-order Taylor Series expansions. Then, an easily found variance estimator of the linear approximation is used as an estimator of the variance of the nonlinear statistic [26]. More on the details of linearization can be found in literature, but is beyond the scope of this thesis.

### 4.1.7 Pseudo algorithm

Combining all previously explained parts, the pseudocode of the simulations will be the following.

**Estimate True Coefficients** $\boldsymbol{\beta}_{TRUE}$ (Diagonal matrix of the true weights $W_T$):

1. Consider Model 0:

$$\begin{aligned}\mathbf{y}_{EF} = &\ \beta_S \cdot \mathbf{x}_S + \cdots + \beta_{A_4} \cdot \mathbf{x}_{A_4} && (4.12)\\ &+ \beta_{P_1} \cdot \mathbf{x}_{P_1} + \cdots + \beta_{P_7} \cdot \mathbf{x}_{P_7} && (4.13)\\ &+ \beta_{Edu} \cdot \mathbf{x}_{Edu} + \cdots + \beta_{DS} \cdot \mathbf{x}_{DS} && (4.14)\\ &+ \beta_{OSS} \cdot \mathbf{x}_{OSS} + \boldsymbol{\epsilon} && (4.15)\\ = &\ \boldsymbol{\beta}^T \mathbf{X} + \boldsymbol{\epsilon}, && (4.16)\end{aligned}$$

with $\boldsymbol{\epsilon} \sim N(0, \sigma^2)$.
Determine the estimate of $\boldsymbol{\beta}_{TRUE}$ by applying a weighted regression using weights $\mathbf{W}_T$ on the CLSA data using PROC SURVEYREG:

$$\boldsymbol{\beta}_{\mathbf{W}_T,TRUE} = (\mathbf{X}^T \mathbf{W}_T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}_T \mathbf{y} \tag{4.17}$$

**end**

**Generation of replicates** ($\boldsymbol{\beta}_{\mathbf{W}_T,TRUE}$ , #runs, $\sigma_\epsilon$, correlation matrix $R$):

1. Define a Gaussian copula with correlation matrix $R$;

$$C_R^{\text{Gauss}}(\mathbf{u}) = \Phi_R\left(\Phi^{-1}(u_{Edu}), \ldots, \Phi^{-1}(u_{DS}), \Phi^{-1}(u_{OSS})\right) \tag{4.18}$$

where $\Phi$ is the cumulative distribution function of a standard normal and $\Phi_R$ is the joint cumulative distribution function of a multivariate normal distribution with mean vector zero and covariance matrix equal to the correlation matrix $R$.

2. **For $i = 1$ to *#runs***:

   a. Define CLSA $\#i$ by copying ID, weight variables, CLSA inflation weights and CLSA analytical weights from the CLSA dataset and therefore, keeping them fixed.

   b. **For $k = 1$ to #nrows(CLSA $\#i$)**:

      i. Generate a vector $\mathbf{u}_k = (u_{Edu,k}, \ldots, u_{DS,k}, u_{OSS,k})$ by sampling from $C_R^{\text{Gauss}}(\mathbf{u})$, and translate the uniform values to the original using the empirical cumulative distributions;

$$(x_{Edu,k}, \ldots, x_{DS,k}, x_{OSS,k}) = (\hat{F}^{-1}_{Edu,n}(u_{Edu,k}), \ldots, \hat{F}^{-1}_{DS,n}(u_{DS,k}), \hat{F}^{-1}_{OSS,n}(u_{OSS,k})) \tag{4.19}$$

      ii. Append $(x_{Edu,k}, \ldots, x_{DS,k}, x_{OSS,k})$ to CLSA $\#i$ at row $k$.

iii. Calculate the response using the true regression coefficients $\boldsymbol{\beta}_{\mathbf{W}_T,TRUE}$ of the weight variables, independent variables and the exposure:

$$\begin{aligned} y_{EF,k} = \; & \beta_{\mathbf{W}_T,S,TRUE} \cdot x_{S,k} + \cdots + \beta_{\mathbf{W}_T,P_7,TRUE} \cdot x_{P_7,k} && (4.20)\\ & + \beta_{\mathbf{W}_T,Edu,TRUE} \cdot x_{Edu,k} + \cdots + \beta_{\mathbf{W}_T,DS,TRUE} \cdot x_{DS,k} && (4.21)\\ & + \beta_{\mathbf{W}_T,OSS,TRUE} \cdot x_{OSS,k} + \epsilon_k, && (4.22) \end{aligned}$$

with $\epsilon_k \sim N(0, \sigma^2_\epsilon)$.

iv. Append $y_{EF,k}$ to CLSA $\#i$ at row $k$.

**end**

**end**

**end**

**Fit model to CLSA replicates** ($\boldsymbol{\beta}_{\mathbf{W}_T,TRUE}$ , diagonal matrix of the analysis weights $\mathbf{W}_A$, new CLSA datasets, model choice (0/1)):
The amount of new CLSA datasets equals the #runs.

1. Consider the regression models of the response $EF$ on the weight variables, independent variables and exposure $OSS$:
**If *model choice* = 0:**
Model 0:

$$\begin{aligned} \mathbf{y}_{EF} = \; & \beta_S \cdot \mathbf{x}_S + \cdots + \beta_{A_4} \cdot \mathbf{x}_{A_4} && (4.23)\\ & + \beta_{P_1} \cdot \mathbf{x}_{P_1} + \cdots + \beta_{P_7} \cdot \mathbf{x}_{P_7} && (4.24)\\ & + \beta_{Edu} \cdot \mathbf{x}_{Edu} + \cdots + \beta_{DS} \cdot \mathbf{x}_{DS} && (4.25)\\ & + \beta_{OSS} \cdot \mathbf{x}_{OSS} + \boldsymbol{\epsilon} && (4.26)\\ = \; & \boldsymbol{\beta}^T \mathbf{X} + \boldsymbol{\epsilon}, && (4.27) \end{aligned}$$

with $\boldsymbol{\epsilon} \sim N(0, \sigma^2)$.
**end**
**If *model choice* = 1:**
Model 1:

$$\begin{aligned} \mathbf{y}_{EF} = \; & \beta_S \cdot \mathbf{x}_S + \cdots + \beta_{A_4} \cdot \mathbf{x}_{A_4} && (4.28)\\ & + \beta_{Edu} \cdot \mathbf{x}_{Edu} + \cdots + \beta_{DS} \cdot \mathbf{x}_{DS} && (4.29)\\ & + \beta_{OSS} \cdot \mathbf{x}_{OSS} + \boldsymbol{\epsilon} && (4.30)\\ = \; & \boldsymbol{\beta}^T \mathbf{X} + \boldsymbol{\epsilon}, && (4.31) \end{aligned}$$

with $\boldsymbol{\epsilon} \sim N(0, \sigma^2)$.
**end**

2. **For $i$ = 1 to *#runs*:**
Define $\mathbf{X}$ as the matrix of the observations of the weight variables, independent variables and exposure and $\mathbf{y}$ as the vector of the observations of the response.

   a. Perform a weighted linear regression analysis using weights $\mathbf{W}_A$ through a weighted least squares estimator to find the $\hat{\boldsymbol{\beta}}$ for dataset CLSA $\#i$ ( $\hat{\boldsymbol{\beta}}_{\mathbf{W}_A,i}$ ) using PROC SURVEYREG. Therefore,

$$\hat{\boldsymbol{\beta}}_{\mathbf{W}_A,i} = (\mathbf{X}^T \mathbf{W}_A \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}_A \, \mathbf{y} \qquad (4.32)$$

   and the $se(\hat{\boldsymbol{\beta}}_{\mathbf{W}_A,i})$ found from variance estimation method Taylor linearization.

3. Calculate measures for comparison of the relation between the response and the exposure:

- $mean(\hat{\beta}_{\mathbf{W}_A,OSS}) = \frac{1}{\#runs} \sum_{j=1}^{\#runs} \hat{\beta}_{\mathbf{W}_A,OSS,j}$

- $sd(\hat{\beta}_{\mathbf{W}_A,OSS}) = \sqrt{\frac{1}{\#runs}\sum_{j=1}^{\#runs}\left(\hat{\beta}_{\mathbf{W}_A,OSS,j} - mean(\hat{\beta}_{\mathbf{W}_A,OSS})\right)^2}$
- $mean(se(\hat{\beta}_{\mathbf{W}_A,OSS})) = \sqrt{\frac{1}{\#runs}\sum_{j=1}^{\#runs} se^2(\hat{\beta}_{\mathbf{W}_A,OSS,j})}$
- $bias(\hat{\beta}_{\mathbf{W}_A,OSS}) = \left|mean(\hat{\beta}_{\mathbf{W}_A,OSS}) - \beta_{\mathbf{W}_T,OSS,TRUE}\right|$
- $coverage(\hat{\beta}_{\mathbf{W}_A,OSS}) = \frac{1}{\#runs}\sum_{j=1}^{\#runs}\mathbb{1}\left(\beta_{\mathbf{W}_T,OSS,TRUE} \in \left[\hat{\beta}_{\mathbf{W}_A,OSS,j} - t_{\alpha/2}\cdot se(\hat{\beta}_{\mathbf{W}_A,OSS,j}), \hat{\beta}_{\mathbf{W}_A,OSS,j} + t_{1-\alpha/2}\cdot se(\hat{\beta}_{\mathbf{W}_A,OSS,j})\right]\right)$ with $\alpha = 0.05$ for a 95% confidence interval

**end**

**end**

### 4.1.8 Choice of parameters

1. **#runs**: 1000.
2. **R**: The correlation matrix present in the CLSA dataset.
3. $\mathbf{W}_T$: $\boldsymbol{\beta}_{\mathbf{W}_T,TRUE}$ will be determined by running a weighted regression on the CLSA data using a set of weights $\mathbf{W}_T$. These weights are;
   - Unweighted $\mathbf{I}$, (all weights = 1)
   - Analytical weights with noise ($\sigma_n = \{0, 0.05, 0.125, 0.25, 0.5\}$) $\mathbf{W}_{noise(\sigma_n)}$
   - Inflation weights $\mathbf{W}_{inflation}$
4. $\mathbf{W}_A$: the possible weights used in the analysis will be the known CLSA weights;
   - Unweighted $\mathbf{I}$
   - Analytical weights $\mathbf{W}_{analytical}$
   - Inflation weights $\mathbf{W}_{inflation}$
5. $\boldsymbol{\sigma}_\epsilon$: For each different $\mathbf{W}_T$ variation, a different seed is used for drawing values from $N(0, \sigma_\epsilon^2)$, with $\sigma_\epsilon = \{0.75, 2\}$.
6. **Model choice**: {0, 1}

## 4.2 Results

The result tables of the simulation can be found in the appendix B.

The bias is always reasonably low, especially taking the relative size of the standard error of the estimates into account. It's maximum is namely 0.001152. The standard error does seem to vary with different weights and settings. The coverage of the estimates seems sufficiently high and is close to the nominal (ranging between 0.931 and 0.965). This indicates that in about 95% of the cases the 'true' parameter $\boldsymbol{\beta}_{OSS,TRUE}$ lies within the confidence interval.

### 4.2.1 Different $\mathbf{W}_A$

The three different regression models considered are unweighted, using inflation weights, and using analytical weights. There is not an absolute best in terms of bias. However the mean standard error of the inflation weights is always the largest, followed by the analytical weights and finally the unweighted. This is probably because the spread between the inflation weights is relatively larger than the spread between the analytical weights. This could show why the estimates in regression with the inflation weights have a larger standard error. The estimates obtained from such regression fluctuate more than the others do, due to the larger spread in the weights.

### 4.2.2 Different $\boldsymbol{\sigma_\epsilon}$

As expected, the bias and the mean standard error of the estimates are always lower for the lower $\sigma_\epsilon$, since there is less deviation in the data. The coverage however is comparable.

### 4.2.3 Different analysis model

In general there doesn't seem to be a large difference between the complete regression model and the one missing explanatory variables. The bias is sometimes lower for the regular model and sometimes for the underspecified model. However, the mean standard error is always slightly higher for the underspecified model.

### 4.2.4 Different $\mathbf{W}_T$

Between all the different variations of the $\mathbf{W}_T$, the bias fluctuates slightly, but the mean standard error remains similar across each variation. It only starts differing when the chosen $\mathbf{W}_A$ changes.

There seem to be no correlation between the increase in the bias, when having more variation in the weights.

# Chapter 5

# Discussion

## 5.1 Conclusion

### 5.1.1 Literature

Weighted regression is used to correct for non-representative sampling and heterogeneity in the data. When restricting the research to weighted regression to correct for non-representative sampling, authors seem to agree that it is beneficial in case of complex survey designs. The most important benefit is that weights can correct for unequal sampling probabilities, stratification and clustering, common structures in designs in epidemiology. Weighted regression eliminates the bias due to non-representative sampling. On the other hand, the literature is still divided on the topic: many state that adding weights also increases the variance of the estimates, limiting the advantage of the better estimate with an increased confidence interval.

### 5.1.2 Mathematics

In Chapter 3 we have considered a simple scenario to illustrate the structure of unweighted and weighted least squares estimates, highlighting the role of weighting from a mathematical point of view. From our analysis, it results that both estimators are unbiased. This result is not what expected from the literature (see Section 5.1.1). Researchers state in fact that estimates from unweighted regression are biased. It's important to highlight the fact that the estimator (both for weighted and unweighted) is unbiased. However, the bias mentioned by researchers can be due to having a non-representative sample, where there are large differences between subgroups. The results from our small simulation study in Section 3.3.1 are also unbiased. This could be due to the fact that the settings are very simple.

What the computation also shows is that the estimate is in both cases some sort of average estimate of the group contribution (gender in our case). For the unweighted analysis the group contributes according to the proportion of the group itself in the sample $n_g/n$, while for the weighted analysis the group contributes according to the proportion of the group in the population $p_g$. The results from the weighted analyses are thus valid for the whole population, while the results from the unweighted analysis are valid for the sample at hand only.

A comparison between variances is harder to see. It seems the ratio between the population proportion and the number present in the sample became of importance. Due to time constraints we could not dig further into this structure. It would be interesting to rewrite the expression differently for new insights. For example, rewrite the $n_m$ and $n_f$ in the unweighted variance from Equation (3.42) to the proportions of the sample.

Due to this complexity in the variance, it was difficult to find a general conclusion, so some small case simulation examples were considered to show some possible outcomes. There it was seen that the unweighted method performed better or similarly to the weighted analysis in terms of standard error and coverage, except for some specific settings. In such specific settings the weighted analysis performed better and one explanation for it could be due to the conservative nature of the weighted analysis combined with the increasing differences and complexity of the group (males and females) distributions

considered in that case. There, for both methods the coverage dropped significantly, but due to the higher standard error of the weighted estimate, the coverage was higher than of the unweighted analysis.

However, it was shown and proven for the weighted estimate and variance, that the scale of the weights does not matter for the result. If all weights upscaled by with a certain factor, this will not interfere with the outcome. This argues against common beliefs about the necessity that the sum of the weights equals the population size, or the sample size, or that they sum up to one. The proportional difference between the weights is what is important.

### 5.1.3 Simulation

A more complex case study was then considered to investigate the performance of unweighted and weighted regression in a case study closer to the ones encountered in practical scenarios. The basis of this larger simulation was the CLSA study.

All the results were close to unbiased, and had a good coverage probability. The estimation differed mostly in their standard error.

In the unweighted analysis the standard error came out slightly lower than the standard error when using any set of weights. Unweighted estimates have the smaller variance while preserving a coverage probability close to the nominal one, and therefore are the preferred choice.

A reason why the results of the analyses using different weights did not differ greatly, could the fact that the subgroups in the CLSA don't differ much from each other in the association being investigated. As a consequence, weighting would not have a large effect there. Another possibility is that the sample itself (the CLSA data) isn't that unrepresentative of the population and therefore, the results between the two analyses are not that different. Under this assumption, the CLSA is comparable with case *2* in Section 3.3.1, where there is also not a big difference between the sample and the population in terms of proportion.

In our simulation study we have also investigated what the effect of uncertainty on the weights is. Namely, what is the effect of a larger variance of the weights on the outcome of the analysis. Contrarily to our expectations, the weighted analysis is robust to uncertainty on the sample weights, making the results of weighted regression reliable also when estimating the weights.

### 5.1.4 General

Combining all previous results we can conclude that the literature, the mathematics and the simulation present similar outcomes. From what observed in the analysed cases one could conclude that the unweighted analysis performs better than the weighted one, also in case of complex sample designs. However, this does not necessarily mean that this conclusion holds for each situation. There are other situations that are interesting to address to have a complete view on the weighted/unweighted model choice. For instance, when analysing the mathematics understanding the role of weight covariates that are also included in the model itself, or analysing other associations and population studies in the simulation to see if they present similar results.

Our current recommendation for the specific case of the CLSA would be to not include weights in regression analysis. Our advice would be to perform some analyses a priori for each new association, to see if the results of the unweighted and the weighted analysis are similar, and based on that include weights whenever the results would suggest the inclusion. However, one of the main existing problems about inclusion of weights is that the usage often stays unclear with respect to software. This results in wrong conclusions about survey datasets, and unclear interpretation of the results. For researchers unfamiliar with the software they are using, our recommendation would be to not include them in the research.

Our hypothesis is that weighting becomes necessary when the distributions of the subgroups that are considered are very different from each other. Then the weights start to matter and could give a potential larger standard error, but would account for those differences better. So when not knowing anything about the properties of the subgroups, still a case can be made to use weights in view of a conservative approach (with increased estimate error when compared to the unweighted analysis). They might give less certain results, but it is known for certain that the effects of different subgroups is accounted for.

## 5.2 Recommendations

When trying to say something about a descriptive statistic, research has already shown that the inclusion of weights is recommended. When drawing conclusions about the sample instead of the population, weights should not be included.

If one feels that inclusion of survey weights is necessary, they should also be fully aware of the situation considered, the way the sample is gathered and have the skills to apply the necessary analytic techniques to perform the analysis, and to interpret the results and formulate the conclusions. Check if there is reason to suspect the sample is unrepresentative of the population and if those subgroups differ greatly in properties included in the research. Also, when deciding to include weights, the variables that must be included in the calibration of weights should be clear.

Particular attention should be paid in the choice of the software. The current recommendation for survey analysis is to use survey-specific tools. For survey analysis, even when you don't need stratification or clustering, do not use PROC GLM or PROC REG in SAS or lm() in R for analysis with survey weights, or be careful in interpreting the estimate errors. These tools assume a different purpose for the weights and therefore use a different calculation for the variance, i.e. $Var(\hat{\boldsymbol{\beta}}_w) = \sigma_\epsilon^2 \cdot (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$, which is not the variance of the estimator in the regression model with survey weights.

## 5.3 Limitations

The replication of the CLSA study implied a set of choices. Such choices have shaped our research and may have an influence on the obtained results.

The weight variables are currently not considered in the copula in Section 4.1.3, and therefore it was assumed that there was no correlation between the independent variables and the weight variables. The reason for this assumption is that the weight variable values were fixed. If the weight variable value would change, the weight value itself would also change. When drawing new observations, the weight variable values could not be changed since changing them would result in inaccurate weights. So the new observations had to be conditioned in a way on the already known values of the weight variables.

In the variance estimation we used that linearization was an appropriate and easy method to implement in the simulation for finding the variance of the estimators. This method has not been compared with other options because of mathematical and computational limitations.

We assumed homogeneity across groups, and did not investigate heterogeneity across groups. We expect that heterogeneity could have had an effect on the results as well.

In this thesis no analysis has been performed on the correctness of the CLSA weights, as this was falling outside the scope of the present work.

## 5.4 Further research

For further research, it would be interesting to check other mathematical cases to get to the essence when unweighted or possibly weighted performs better. In fact, our results already show clearly the contribution of weights in the least squares estimators. A good starting point would be to check the current computations made in this thesis for possible additional explanations. The inclusion of sex as a covariate in the mathematical model would be an interesting next step.

The current conclusions could be supported by considering other associations in the CLSA study to verify if similar conclusions still hold, and thus the generality of our results.

This research has focused on the usage of weights in linear regression, but the results could be different for other forms of regression, like logistic regression. Further research will be needed to investigate the effect of the inclusion of weights in other types of models.

On a more practical basis, there should be more research put into tests that will be able to properly compare a weighted and an unweighted analysis, to suggest if weighting is needed in that specific case or not.

# Bibliography


[1] D. Holt, T. M. F. Smith, and P. D. Winter. Regression Analysis of Data from Complex Surveys. *Journal of the Royal Statistical Society*, 143(4):474–487, 1980. 1

[2] Mohamed Elfil and Ahmed Negida. Sampling methods in Clinical Research; an Educational Review. *Emergency (Tehran, Iran)*, 5(1):e52, 2017. ISSN 2345-4563. doi: 10.22037/emergency.v5i1.15215. URL http://www.ncbi.nlm.nih.gov/pubmed/28286859{%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5325924. 1

[3] Greg J Duncan and H Dumouchel. Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, 78(383):535–543, 1983. 1, 11

[4] Andrew Gelman. Struggles with survey weighting and regression modeling. *Statistical Science*, 22 (2):153–164, 2007. ISSN 08834237. doi: 10.1214/088342306000000691. 1

[5] James Randolph Jr Knaub. Heteroscedasticity and Homoscedasticity. *Encyclopedia of Measurement and Statistics*, pages 431–432, 2007. doi: 10.4135/9781412952644.n201. 1

[6] J. S. Williams. The Variance of Weighted Regression Estimators. *Journal of the American Statistical Association*, 62(320):1290–1301, 1967. 1

[7] K Brewer and R Mellor. The Effect of Sample Structure on Analytical Surveys. *Australian & New Zealand Journal of Statistics*, 15:145–152, 1973. 1

[8] Danny Pfeffermann. The Role of Sampling Weights When Modeling Survey Data. *International Statistical Review / Revue Internationale de Statistique*, 61(2):317, 1993. ISSN 03067734. doi: 10.2307/1403631. 1

[9] JNK Rao, M Hidiroglou, W. Yung, and M. Kovacevic. Role of Weights in Descriptive and Analytical Inferences from Survey Data: An Overview. *Journal of the Indian Society of Agricultural Statistics*, 64(2):129–135, 2010. 1, 3

[10] Leslie Kish and Martin Richard Frankel. Inference from Complex Samples. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(1):1–22, 1974. doi: 10.1111/j.2517-6161.1974.tb00981.x. 1

[11] Pierre Lavallee and Jean-Francois Beaumont. Why We Should Put Some Weight on Weights. *Survey Methods: Insights from the Field*, pages 1–18, 2015. ISSN 2296-4754. doi: 10.13094/SMIF-2015-00001. 1

[12] CE Sarndal, B Swensson, and JH Wretman. *Model Assisted Survey Sampling*. Springer, 1992. 1

[13] Leslie Kish. Weighting for Unequal Pi. *Journal of Official Statistics*, 8(2):183–200, 1992. 1, 4

[14] J. N. K. Rao. Alternative Estimators in PPS Sampling for Multiple Characteristics. *The Indian Journal of Statistics*, 28(1):47–60, 1966. 1

[15] E. L. Korn and B. I. Graubard. Examples of Differing Weighted and Unweighted Estimates from a Sample Survey. *American Statistical Association*, 49(3):291–295, 1995. 1, 2

[16] Kenneth Bollen, Paul Biemer, Alan Karr, Stephen Tueller, and Marcus E. Berzofsky. Are Survey Weights Needed? A Review of Diagnostic Tests in Regression Analysis. *Ssrn*, 2016. doi: 10.1146/annurev-statistics-011516-012958. 2

[17] Parminder S. Raina, Christina Wolfson, Susan A. Kirkland, Lauren E. Griffith, Mark Oremus, Christopher Patterson, Holly Tuokko, Margaret Penning, Cynthia M. Balion, David Hogan, Andrew Wister, Hélène Payette, Harry Shannon, and Kevin Brazil. The canadian longitudinal study on aging (CLSA). *Canadian Journal on Aging*, 28(3):221–229, 2009. ISSN 07149808. doi: 10.1017/S0714980809990055. 3

[18] Canadian Longitudinal Study on Aging. Sampling and Computation of Response Rates and Sample Weights for the Tracking (Telephone Interview) Participants and Comprehensive Participants, 2017. URL https://www.clsa-elcv.ca/doc/1041. 3, 4

[19] Sean Esteban McCabe and Brady T. West. Selective nonresponse bias in population-based survey estimates of drug use behaviors in the United States. *Social Psychiatry and Psychiatric Epidemiology*, 51(1):141–153, 2016. ISSN 09337954. doi: 10.1007/s00127-015-1122-2. 3

[20] Phillip S. Kott. An Introduction to Calibration Weighting for Establishment Surveys. *Proceedings of the Fourth International Conference of Establishment Surveys*, pages 8–10, 2012. URL http://www.amstat.org/meetings/ices/2012/papers/302286.pdf. 4

[21] Dmitry Kats, Mehul D. Patel, Priya Palta, Michelle L. Meyer, Alden L. Gross, Eric A. Whitsel, David Knopman, Alvaro Alonso, Thomas H. Mosley, and Gerardo Heiss. Social support and cognition in a community-based cohort: The Atherosclerosis Risk in Communities (ARIC) study. *Age and Ageing*, 45(4):475–480, 2016. ISSN 14682834. doi: 10.1093/ageing/afw060. 4

[22] Tjalling Jan Holwerda, Dorly J.H. Deeg, Aartjan T.F. Beekman, Theo G. Van Tilburg, Max L. Stek, Cees Jonker, and Robert A. Schoevers. Feelings of loneliness, but not social isolation, predict dementia onset: Results from the Amsterdam Study of the Elderly (AMSTEL). *Journal of Neurology, Neurosurgery and Psychiatry*, 85(2):135–142, 2014. ISSN 1468330X. doi: 10.1136/jnnp-2012-302755. 4

[23] SAS Institute Inc. SAS/STAT® User's Guide, Version 8, 1999. 5

[24] Bethany A. Bell, Anthony J. Onwuegbuzie, John M. Ferron, Qun G. Jiao, Susan T. Hibbard, and Jeffrey D. Kromrey. Use of design effects and sample weights in complex health survey data: A review of published articles using data from 3 commonly used adolescent health surveys. *American Journal of Public Health*, 102(7):1399–1405, 2012. ISSN 00900036. doi: 10.2105/AJPH.2011.300398. 6

[25] J Newsom, R N Jones, and S M Hofer. *Longitudinal Data Analysis: A Practical Guide for Researchers in Aging, Health, and Social Sciences*. Multivariate Applications Series. Taylor & Francis, 2013. ISBN 9781136705472. URL https://books.google.nl/books?id=uve3Q4gI8EgC. 13

[26] Keith Rust. Variance Estimation for Complex Estimators in Sample Surveys. *Journal of Official Statistics*, 1(4):381–397, 1985. 23

[27] Roger B Nelson. *An introduction to copulas*. Springer Science & Business Media, 2007. ISBN 978-0-387-98623-4. 32

# Appendix A

# Elaboration on the copula

A copula describes the dependence of random variables and takes the original distribution of each independent variable into account.

$C : [0,1]^d \to [0,1]$ is a joint cumulative distribution function of a $d$-dimensional random vector with uniform marginals [27]. In our case, the marginal distributions are not uniform. Using a transformation this can still be achieved. Let $x \sim F_1$, then $F_1(x)$ is uniformly distributed. This can be easily shown by;

$$\mathbb{P}(F_1(x) \leq u) = P(x \leq F_1^{-1}(u)) = F_1(F_1^{-1}(u)) = u \tag{A.1}$$

So each marginal distribution of the independent variables can be transformed to a uniform distribution for usage in the copula, and can also be transformed back. In the simulation no general distribution will be assumed, but the empirical cumulative distribution will be estimated and used for all independent variables.

For the choice which form of the copula to take, the aim is to take the correlation between the independent variables into account, which can be done in a Gaussian copula with a correlation matrix $R$. For given uniform independent variables $u_i \in [0,1]$, together with the corresponding correlation matrix $R \in [0,1]^{d \times d}$, where $d$ equals the amount of independent variables, the Gaussian copula's cumulative function is defined as;

$$C_R^{\text{Gauss}}(u) = \Phi_R\left(\Phi^{-1}(u_1), \ldots, , \Phi^{-1}(u_d)\right), \tag{A.2}$$

with $\Phi_R$ the joint cumulative distribution function of a multivariate normal distribution with mean vector zero and covariance matrix equal to the correlation matrix $R$ and $\Phi^{-1}$ the inverse cumulative distribution function of a standard normal.

# Appendix B

# Simulation results

## B.1 Model 0 results

| runs | Model | $\sigma\_\epsilon$ | W_T | W_A | $\beta\_(W\_T,TRUE)$ | MEAN est | STD est | bias | MEAN stdErr | Coverage |
|---|---|---|---|---|---|---|---|---|---|---|
| 1000 | 0 | 0,75 | analytical | analytical | 0,3585 | 0,3586 | 0,0075 | 0,000024 | 0,007568 | 0,948 |
| 1000 | 0 | 0,75 | analytical | inflation | 0,3585 | 0,3586 | 0,0083 | 0,000121 | 0,008411 | 0,964 |
| 1000 | 0 | 0,75 | analytical | unweighted | 0,3585 | 0,3585 | 0,0067 | 0,000033 | 0,006637 | 0,944 |
| 1000 | 0 | 0,75 | inflation | analytical | 0,3669 | 0,3670 | 0,0074 | 0,000089 | 0,007571 | 0,952 |
| 1000 | 0 | 0,75 | inflation | inflation | 0,3669 | 0,3670 | 0,0084 | 0,000084 | 0,008413 | 0,944 |
| 1000 | 0 | 0,75 | inflation | unweighted | 0,3669 | 0,3670 | 0,0065 | 0,000083 | 0,006642 | 0,96 |
| 1000 | 0 | 0,75 | noise(0.05) | analytical | 0,3568 | 0,3565 | 0,0077 | 0,000230 | 0,007574 | 0,949 |
| 1000 | 0 | 0,75 | noise(0.05) | inflation | 0,3568 | 0,3565 | 0,0084 | 0,000295 | 0,008410 | 0,946 |
| 1000 | 0 | 0,75 | noise(0.05) | unweighted | 0,3568 | 0,3565 | 0,0068 | 0,000322 | 0,006643 | 0,942 |
| 1000 | 0 | 0,75 | noise(0.125) | analytical | 0,3577 | 0,3578 | 0,0074 | 0,000131 | 0,007571 | 0,959 |
| 1000 | 0 | 0,75 | noise(0.125) | inflation | 0,3577 | 0,3579 | 0,0084 | 0,000186 | 0,008411 | 0,95 |
| 1000 | 0 | 0,75 | noise(0.125) | unweighted | 0,3577 | 0,3578 | 0,0065 | 0,000149 | 0,006641 | 0,955 |
| 1000 | 0 | 0,75 | noise(0.25) | analytical | 0,3435 | 0,3435 | 0,0078 | 0,000007 | 0,007570 | 0,945 |
| 1000 | 0 | 0,75 | noise(0.25) | inflation | 0,3435 | 0,3434 | 0,0088 | 0,000052 | 0,008414 | 0,946 |
| 1000 | 0 | 0,75 | noise(0.25) | unweighted | 0,3435 | 0,3434 | 0,0069 | 0,000080 | 0,006638 | 0,941 |
| 1000 | 0 | 0,75 | noise(0.5) | analytical | 0,3554 | 0,3549 | 0,0077 | 0,000427 | 0,007566 | 0,947 |
| 1000 | 0 | 0,75 | noise(0.5) | inflation | 0,3554 | 0,3551 | 0,0086 | 0,000277 | 0,008406 | 0,943 |
| 1000 | 0 | 0,75 | noise(0.5) | unweighted | 0,3554 | 0,3551 | 0,0066 | 0,000241 | 0,006637 | 0,953 |
| 1000 | 0 | 0,75 | unweighted | analytical | 0,3576 | 0,3577 | 0,0078 | 0,000094 | 0,007569 | 0,949 |
| 1000 | 0 | 0,75 | unweighted | inflation | 0,3576 | 0,3576 | 0,0086 | 0,000035 | 0,008411 | 0,942 |
| 1000 | 0 | 0,75 | unweighted | unweighted | 0,3576 | 0,3576 | 0,0068 | 0,000061 | 0,006641 | 0,942 |
| 1000 | 0 | 2 | analytical | analytical | 0,3585 | 0,3586 | 0,0201 | 0,000064 | 0,020181 | 0,948 |
| 1000 | 0 | 2 | analytical | inflation | 0,3585 | 0,3589 | 0,0222 | 0,000323 | 0,022429 | 0,964 |
| 1000 | 0 | 2 | analytical | unweighted | 0,3585 | 0,3584 | 0,0179 | 0,000088 | 0,017698 | 0,944 |
| 1000 | 0 | 2 | inflation | analytical | 0,3669 | 0,3672 | 0,0198 | 0,000238 | 0,020190 | 0,952 |
| 1000 | 0 | 2 | inflation | inflation | 0,3669 | 0,3672 | 0,0224 | 0,000225 | 0,022434 | 0,944 |
| 1000 | 0 | 2 | inflation | unweighted | 0,3669 | 0,3672 | 0,0172 | 0,000223 | 0,017711 | 0,960 |
| 1000 | 0 | 2 | noise(0.05) | analytical | 0,3568 | 0,3562 | 0,0205 | 0,000613 | 0,020196 | 0,949 |
| 1000 | 0 | 2 | noise(0.05) | inflation | 0,3568 | 0,3560 | 0,0224 | 0,000786 | 0,022425 | 0,946 |
| 1000 | 0 | 2 | noise(0.05) | unweighted | 0,3568 | 0,3559 | 0,0181 | 0,000859 | 0,017714 | 0,942 |
| 1000 | 0 | 2 | noise(0.125) | analytical | 0,3577 | 0,3580 | 0,0199 | 0,000349 | 0,020190 | 0,959 |
| 1000 | 0 | 2 | noise(0.125) | inflation | 0,3577 | 0,3582 | 0,0224 | 0,000495 | 0,022428 | 0,950 |
| 1000 | 0 | 2 | noise(0.125) | unweighted | 0,3577 | 0,3581 | 0,0174 | 0,000398 | 0,017709 | 0,955 |
| 1000 | 0 | 2 | noise(0.25) | analytical | 0,3435 | 0,3435 | 0,0209 | 0,000017 | 0,020185 | 0,945 |
| 1000 | 0 | 2 | noise(0.25) | inflation | 0,3435 | 0,3433 | 0,0233 | 0,000138 | 0,022436 | 0,946 |
| 1000 | 0 | 2 | noise(0.25) | unweighted | 0,3435 | 0,3432 | 0,0185 | 0,000213 | 0,017700 | 0,941 |
| 1000 | 0 | 2 | noise(0.5) | analytical | 0,3554 | 0,3542 | 0,0206 | 0,001140 | 0,020177 | 0,947 |
| 1000 | 0 | 2 | noise(0.5) | inflation | 0,3554 | 0,3546 | 0,0229 | 0,000739 | 0,022417 | 0,943 |
| 1000 | 0 | 2 | noise(0.5) | unweighted | 0,3554 | 0,3547 | 0,0177 | 0,000642 | 0,017699 | 0,953 |
| 1000 | 0 | 2 | unweighted | analytical | 0,3576 | 0,3578 | 0,0207 | 0,000252 | 0,020185 | 0,949 |
| 1000 | 0 | 2 | unweighted | inflation | 0,3576 | 0,3577 | 0,0230 | 0,000092 | 0,022428 | 0,942 |
| 1000 | 0 | 2 | unweighted | unweighted | 0,3576 | 0,3577 | 0,0182 | 0,000162 | 0,017710 | 0,942 |

## B.2 Model 1 results

| runs | Model | $\sigma\_\epsilon$ | W_T | W_A | $\beta\_(W\_T,TRUE)$ | MEAN est | STD est | bias | MEAN stdErr | Coverage |
|---|---|---|---|---|---|---|---|---|---|---|
| 1000 | 1 | 0,75 | analytical | analytical | 0,3585 | 0,3585 | 0,0086 | 0,000005 | 0,008515 | 0,942 |
| 1000 | 1 | 0,75 | analytical | inflation | 0,3585 | 0,3587 | 0,0091 | 0,000187 | 0,009123 | 0,953 |
| 1000 | 1 | 0,75 | analytical | unweighted | 0,3585 | 0,3585 | 0,0078 | 0,000027 | 0,007482 | 0,944 |
| 1000 | 1 | 0,75 | inflation | analytical | 0,3669 | 0,3670 | 0,0086 | 0,000074 | 0,008521 | 0,953 |
| 1000 | 1 | 0,75 | inflation | inflation | 0,3669 | 0,3671 | 0,0092 | 0,000155 | 0,009129 | 0,948 |
| 1000 | 1 | 0,75 | inflation | unweighted | 0,3669 | 0,3670 | 0,0075 | 0,000090 | 0,007487 | 0,951 |
| 1000 | 1 | 0,75 | noise(0.05) | analytical | 0,3568 | 0,3565 | 0,0086 | 0,000242 | 0,008520 | 0,94 |
| 1000 | 1 | 0,75 | noise(0.05) | inflation | 0,3568 | 0,3566 | 0,0092 | 0,000222 | 0,009128 | 0,947 |
| 1000 | 1 | 0,75 | noise(0.05) | unweighted | 0,3568 | 0,3565 | 0,0077 | 0,000314 | 0,007488 | 0,944 |
| 1000 | 1 | 0,75 | noise(0.125) | analytical | 0,3577 | 0,3578 | 0,0085 | 0,000122 | 0,008534 | 0,952 |
| 1000 | 1 | 0,75 | noise(0.125) | inflation | 0,3577 | 0,3579 | 0,0091 | 0,000261 | 0,009121 | 0,948 |
| 1000 | 1 | 0,75 | noise(0.125) | unweighted | 0,3577 | 0,3578 | 0,0075 | 0,000160 | 0,007500 | 0,956 |
| 1000 | 1 | 0,75 | noise(0.25) | analytical | 0,3435 | 0,3435 | 0,0087 | 0,000004 | 0,008505 | 0,952 |
| 1000 | 1 | 0,75 | noise(0.25) | inflation | 0,3435 | 0,3435 | 0,0094 | 0,000020 | 0,009124 | 0,945 |
| 1000 | 1 | 0,75 | noise(0.25) | unweighted | 0,3435 | 0,3434 | 0,0078 | 0,000069 | 0,007473 | 0,945 |
| 1000 | 1 | 0,75 | noise(0.5) | analytical | 0,3554 | 0,3549 | 0,0086 | 0,000439 | 0,008385 | 0,947 |
| 1000 | 1 | 0,75 | noise(0.5) | inflation | 0,3554 | 0,3552 | 0,0092 | 0,000212 | 0,009007 | 0,94 |
| 1000 | 1 | 0,75 | noise(0.5) | unweighted | 0,3554 | 0,3551 | 0,0075 | 0,000233 | 0,007367 | 0,949 |
| 1000 | 1 | 0,75 | unweighted | analytical | 0,3576 | 0,3577 | 0,0087 | 0,000075 | 0,008431 | 0,931 |
| 1000 | 1 | 0,75 | unweighted | inflation | 0,3576 | 0,3577 | 0,0092 | 0,000100 | 0,009033 | 0,944 |
| 1000 | 1 | 0,75 | unweighted | unweighted | 0,3576 | 0,3576 | 0,0077 | 0,000056 | 0,007412 | 0,946 |
| 1000 | 1 | 2 | analytical | analytical | 0,3585 | 0,3586 | 0,0206 | 0,000033 | 0,020555 | 0,950 |
| 1000 | 1 | 2 | analytical | inflation | 0,3585 | 0,3589 | 0,0225 | 0,000379 | 0,022705 | 0,957 |
| 1000 | 1 | 2 | analytical | unweighted | 0,3585 | 0,3584 | 0,0184 | 0,000089 | 0,018031 | 0,948 |
| 1000 | 1 | 2 | inflation | analytical | 0,3669 | 0,3672 | 0,0204 | 0,000219 | 0,020566 | 0,950 |
| 1000 | 1 | 2 | inflation | inflation | 0,3669 | 0,3672 | 0,0227 | 0,000294 | 0,022713 | 0,948 |
| 1000 | 1 | 2 | inflation | unweighted | 0,3669 | 0,3672 | 0,0178 | 0,000224 | 0,018044 | 0,957 |
| 1000 | 1 | 2 | noise(0.05) | analytical | 0,3568 | 0,3562 | 0,0209 | 0,000627 | 0,020570 | 0,946 |
| 1000 | 1 | 2 | noise(0.05) | inflation | 0,3568 | 0,3561 | 0,0228 | 0,000712 | 0,022704 | 0,946 |
| 1000 | 1 | 2 | noise(0.05) | unweighted | 0,3568 | 0,3559 | 0,0184 | 0,000854 | 0,018046 | 0,941 |
| 1000 | 1 | 2 | noise(0.125) | analytical | 0,3577 | 0,3580 | 0,0204 | 0,000343 | 0,020572 | 0,951 |
| 1000 | 1 | 2 | noise(0.125) | inflation | 0,3577 | 0,3582 | 0,0227 | 0,000573 | 0,022703 | 0,948 |
| 1000 | 1 | 2 | noise(0.125) | unweighted | 0,3577 | 0,3581 | 0,0178 | 0,000409 | 0,018047 | 0,958 |
| 1000 | 1 | 2 | noise(0.25) | analytical | 0,3435 | 0,3435 | 0,0211 | 0,000008 | 0,020554 | 0,950 |
| 1000 | 1 | 2 | noise(0.25) | inflation | 0,3435 | 0,3434 | 0,0235 | 0,000068 | 0,022710 | 0,949 |
| 1000 | 1 | 2 | noise(0.25) | unweighted | 0,3435 | 0,3433 | 0,0188 | 0,000200 | 0,018028 | 0,943 |
| 1000 | 1 | 2 | noise(0.5) | analytical | 0,3554 | 0,3542 | 0,0210 | 0,001152 | 0,020497 | 0,947 |
| 1000 | 1 | 2 | noise(0.5) | inflation | 0,3554 | 0,3547 | 0,0231 | 0,000673 | 0,022649 | 0,949 |
| 1000 | 1 | 2 | noise(0.5) | unweighted | 0,3554 | 0,3547 | 0,0180 | 0,000633 | 0,017982 | 0,947 |
| 1000 | 1 | 2 | unweighted | analytical | 0,3576 | 0,3578 | 0,0211 | 0,000232 | 0,020522 | 0,939 |
| 1000 | 1 | 2 | unweighted | inflation | 0,3576 | 0,3577 | 0,0232 | 0,000153 | 0,022665 | 0,943 |
| 1000 | 1 | 2 | unweighted | unweighted | 0,3576 | 0,3577 | 0,0186 | 0,000157 | 0,018011 | 0,940 |

# Appendix C

# Least squares estimate

The optimal solution $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)$ is when the residual $\epsilon_i$ is minimized, meaning the residual sum of squares is minimal. Therefore take the estimate $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} SS_\epsilon$, where $SS_\epsilon$ is the error function, which equals

$$\begin{aligned} SS_\epsilon = \sum_{i=1}^{n} \left|\epsilon_i\right|^2 = \sum_{i=1}^{n} \left|y_i - \hat{y}_i\right|^2 = \sum_{i=1}^{n} \left|y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right|^2 &= \left\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right\|^2 \\ &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\hat{\boldsymbol{\beta}} - (\mathbf{X}\hat{\boldsymbol{\beta}})^T\mathbf{y} - (\mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \mathbf{y}^T\mathbf{y} - 2(\mathbf{X}^T\mathbf{y})^T\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} \end{aligned}$$

Now minimize $SS_\epsilon$ by taking the gradient and setting it to zero.

$$\begin{aligned} \nabla S = &-2\mathbf{X}^T\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) = 0 \\ \Rightarrow &-2\mathbf{X}^T\mathbf{y} - 2\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = 0 \\ \Rightarrow &\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y} \\ \Rightarrow &\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \end{aligned}$$

# Appendix D

# General case unweighted estimator

For the general case described at the beginning of Chapter 3, the estimate

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{D.1}$$

gives

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \tag{D.2}$$

$$\Rightarrow (\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{C} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix} \tag{D.3}$$

where $C = n \cdot \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2 = n^2 \cdot \overline{x^{(2)}} - n^2 \cdot \overline{x}^2 = n^2 Var(\mathbf{x})$. Then

$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix} \tag{D.4}$$

Combining these gives the estimate

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \frac{1}{n^2 Var(\mathbf{X})} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix} \tag{D.5}$$

$$= \frac{1}{n^2 Var(\mathbf{x})} \begin{bmatrix} \sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i \\ n \cdot \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i \end{bmatrix} \tag{D.6}$$

$$= \frac{n^2}{n^2 Var(\mathbf{x})} \begin{bmatrix} \overline{x^{(2)}} \cdot \overline{y} - \overline{x} \cdot \overline{xy} \\ \overline{xy} - \overline{x} \cdot \overline{y} \end{bmatrix} \tag{D.7}$$

$$= \frac{1}{Var(\mathbf{x})} \begin{bmatrix} Var(\mathbf{x}) \cdot \overline{y} + \overline{x}^2 \cdot \overline{y} - \overline{x} \cdot \overline{xy} \\ Cov(\mathbf{x}, \mathbf{y}) \end{bmatrix} \tag{D.8}$$

$$= \frac{1}{Var(\mathbf{x})} \begin{bmatrix} Var(\mathbf{x}) \cdot \overline{y} - \overline{x} \cdot Cov(\mathbf{x}, \mathbf{y}) \\ Cov(\mathbf{x}, \mathbf{y}) \end{bmatrix} \tag{D.9}$$

# Appendix E

# General unweighted variance computation

The variance of the estimators is

$$Var(\boldsymbol{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \cdot Var(\mathbf{y}) \cdot \left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)^T = \tag{E.1}$$

So,

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \frac{1}{C}\begin{bmatrix} \sum_{i=1}^n x_i^2 - x_1\sum_{i=1}^n x_i & \dots & \sum_{i=1}^n x_i^2 - x_n\sum_{i=1}^n x_j \\ nx_1 - \sum_{i=1}^n x_i & \dots & nx_n - \sum_{i=1}^n x_i \end{bmatrix} \tag{E.2}$$

So,

$$\begin{aligned}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \cdot Var(\mathbf{y}) \cdot \left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)^T = \\ \frac{1}{(n^2Var(\mathbf{x}))^2}\begin{bmatrix} \sigma^2_{\epsilon,m}\cdot\left(\sum_{i=1}^n x_i^2 - x_1\sum_{i=1}^n x_i\right) & \dots & \sigma^2_{\epsilon,f}\cdot\left(\sum_{i=1}^n x_i^2 - x_n\sum_{i=1}^n x_j\right) \\ \sigma^2_{\epsilon,m}\cdot\left(nx_1 - \sum_{i=1}^n x_i\right) & \dots & \sigma^2_{\epsilon,f}\cdot\left(nx_n - \sum_{i=1}^n x_i\right) \end{bmatrix} \\ \cdot \begin{bmatrix} \sum_{i=1}^n x_i^2 - x_1\sum_{i=1}^n x_i & nx_1 - \sum_{i=1}^n x_i \\ \vdots & \vdots \\ \sum_{i=1}^n x_i^2 - x_n\sum_{i=1}^n x_j & nx_n - \sum_{i=1}^n x_i \end{bmatrix}\end{aligned} \tag{E.3}$$

This gives on entry (2,2);

$$Var(\hat{\beta}_1) = \frac{1}{(n^2Var(\mathbf{x}))^2}\left(\sigma^2_{\epsilon,m}\left(nx_1 - \sum_{i=1}^n x_i\right)^2 + \dots + \sigma^2_{\epsilon,f}\left(nx_n - \sum_{i=1}^n x_i\right)^2\right) \tag{E.4}$$

$$= \frac{1}{(n^2Var(\mathbf{x}))^2}\left(\sigma^2_{\epsilon,m}\sum_{j=1}^{n_m}\left(nx_j - \sum_{i=1}^n x_i\right)^2 + \sigma^2_{\epsilon,f}\sum_{j=n_m+1}^{n}\left(nx_j - \sum_{i=1}^n x_i\right)^2\right) \tag{E.5}$$

$$= \frac{1}{(n^2Var(\mathbf{x}))^2}\left(\sigma^2_{\epsilon,m}\sum_{j=1}^{n_m} n^2\left(x_j - \overline{x}\right)^2 + \sigma^2_{\epsilon,f}\sum_{j=n_m+1}^{n} n^2\left(x_j - \overline{x}\right)^2\right) \tag{E.6}$$

$$= \frac{1}{(nVar(\mathbf{x}))^2}\left(\sigma^2_{\epsilon,m}\sum_{j=1}^{n_m}\left(x_j - \overline{x}\right)^2 + \sigma^2_{\epsilon,f}\sum_{j=n_m+1}^{n}\left(x_j - \overline{x}\right)^2\right) \tag{E.7}$$

$$\overset{(3.7)}{=} \frac{1}{(nVar(\mathbf{x}))^2}\left(\sigma^2_{\epsilon,m}n_m(Var(\mathbf{x}_m) + (\overline{x}_m - \overline{x})^2) + \sigma^2_{\epsilon,f}n_f(Var(\mathbf{x}_f) + (\overline{x}_f - \overline{x})^2)\right) \tag{E.8}$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2_{\epsilon,m} n_m \left(Var(\mathbf{x}_m) + (\overline{x}_m - \overline{x})^2\right) + \sigma^2_{\epsilon,f} n_f \left(Var(\mathbf{x}_f) + (\overline{x}_f - \overline{x})^2\right)}{\left(n_m Var(\mathbf{x}_m) + \frac{n_m n_f}{n}(\overline{x}_f - \overline{x}_m)^2 + n_f Var(\mathbf{x}_f)\right)^2} \tag{E.9}$$

$$= \frac{\sigma^2_{\epsilon,m} n_m Var(\mathbf{x}_m) + \sigma^2_{\epsilon,m} n_m (\overline{x}_m - \overline{x})^2 + \sigma^2_{\epsilon,f} n_f Var(\mathbf{x}_f) + \sigma^2_{\epsilon,f} n_f (\overline{x}_f - \overline{x})^2}{\left(n_m Var(\mathbf{x}_m) + \frac{n_m n_f}{n}(\overline{x}_f - \overline{x}_m)^2 + n_f Var(\mathbf{x}_f)\right)^2} \tag{E.10}$$

$$\overset{(3.4)}{=} \frac{\sigma^2_{\epsilon,m} n_m Var(\mathbf{x}_m) + \sigma^2_{\epsilon,m} n_m (\overline{x}_m - \frac{n_m}{n}\overline{x}_m - \frac{n_f}{n}\overline{x}_f)^2 + \sigma^2_{\epsilon,f} n_f Var(\mathbf{x}_f) + \sigma^2_{\epsilon,f} n_f (\overline{x}_f - \frac{n_m}{n}\overline{x}_m - \frac{n_f}{n}\overline{x}_f))^2}{\left(n_m Var(\mathbf{x}_m) + \frac{n_m n_f}{n}(\overline{x}_f - \overline{x}_m)^2 + n_f Var(\mathbf{x}_f)\right)^2} \tag{E.11}$$

$$= \frac{\sigma^2_{\epsilon,m} n_m Var(\mathbf{x}_m) + \sigma^2_{\epsilon,m} n_m (\frac{n_f}{n})^2 (\overline{x}_m - \overline{x}_f)^2 + \sigma^2_{\epsilon,f} n_f Var(\mathbf{x}_f) + \sigma^2_{\epsilon,f} n_f (\frac{n_m}{n})^2 (\overline{x}_m - \overline{x}_f)^2}{\left(n_m Var(\mathbf{x}_m) + \frac{n_m n_f}{n}(\overline{x}_f - \overline{x}_m)^2 + n_f Var(\mathbf{x}_f)\right)^2} \tag{E.12}$$

$$= \frac{\sigma^2_{\epsilon,m} n_m Var(\mathbf{x}_m) + \frac{n_m n_f}{n^2}(\sigma^2_{\epsilon,m} n_f + \sigma^2_{\epsilon,f} n_m)(\overline{x}_m - \overline{x}_f)^2 + \sigma^2_{\epsilon,f} n_f Var(\mathbf{x}_f)}{\left(n_m Var(\mathbf{x}_m) + \frac{n_m n_f}{n}(\overline{x}_f - \overline{x}_m)^2 + n_f Var(\mathbf{x}_f)\right)^2} \tag{E.13}$$

# Appendix F

# Weighted least squares

The weighted least squares method entails

$\hat{\boldsymbol{\beta}}_w = \arg\min_{\boldsymbol{\beta}} WSS_\epsilon$, where $WSS_\epsilon$ is the weighted error function, which equals

$$\begin{aligned} WSS_\epsilon = \sum_{i=1}^{n} w_i \cdot \left|\epsilon_i\right|^2 = \sum_{i=1}^{n} w_i \cdot \left|y_i - \hat{y}_i\right|^2 = \sum_{i=1}^{n} w_i \cdot \left|y_i - (\hat{\beta}_{w,0} + \hat{\beta}_{w,1} x_i)\right|^2 &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_w)^T \mathbf{W} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_w) \\ &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\hat{\boldsymbol{\beta}}_w - (\mathbf{X}\hat{\boldsymbol{\beta}}_w)^T\mathbf{y} - (\mathbf{X}\hat{\boldsymbol{\beta}}_w)^T(\mathbf{X}\hat{\boldsymbol{\beta}}_w) \\ &= \mathbf{y}^T\mathbf{y} - 2(\mathbf{X}^T\mathbf{y})^T\hat{\boldsymbol{\beta}}_w - \hat{\boldsymbol{\beta}}_w^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}_w \end{aligned}$$

Now minimize $WSS_\epsilon$ by taking the gradient and setting it to zero.

$$\begin{aligned} \nabla S = &-2\mathbf{X}^T\mathbf{W}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_w\right) = 0 \\ \Rightarrow &-2\mathbf{X}^T\mathbf{W}\mathbf{y} - 2\mathbf{X}^T\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}}_w = 0 \\ \Rightarrow &\mathbf{X}^T\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}}_w = \mathbf{X}^T\mathbf{W}\mathbf{y} \\ \Rightarrow &\hat{\boldsymbol{\beta}}_w = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y} \end{aligned}$$

# Appendix G

# General case weighted estimator

For the case considered, the estimate gives;

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y} = \tag{G.1}$$

So;

$$\mathbf{X}^T\mathbf{W}\mathbf{X} = \begin{bmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{bmatrix} \cdot diag(w_m, \dots, w_m, w_f, \dots, w_f) \cdot \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \tag{G.2}$$

$$= \begin{bmatrix} w_m n_m + w_f n_f & w_m \sum_{i=1}^{n_m} x_i + w_f \sum_{i=n_m+1}^{n} x_i \\ w_m \sum_{i=1}^{n_m} x_i + w_f \sum_{i=n_m+1}^{n} x_i & w_m \sum_{i=1}^{n_m} x_i^2 + w_f \sum_{i=n_m+1}^{n} x_i^2 \end{bmatrix} \tag{G.3}$$

$$\Rightarrow (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1} = \frac{1}{C_w} \begin{bmatrix} w_m \sum_{i=1}^{n_m} x_i^2 + w_f \sum_{i=n_m+1}^{n} x_i^2 & -\left(w_m \sum_{i=1}^{n_m} x_i + w_f \sum_{i=n_m+1}^{n} x_i\right) \\ -\left(w_m \sum_{i=1}^{n_m} x_i + w_f \sum_{i=n_m+1}^{n} x_i\right) & w_m n_m + w_f n_f \end{bmatrix} \tag{G.4}$$

where

$$C_w = (w_m n_m + w_f n_f) \cdot \left( w_m \sum_{i=1}^{n_m} x_i^2 + w_f \sum_{i=n_m+1}^{n} x_i^2 \right) - \left( w_m \sum_{i=1}^{n_m} x_i + w_f \sum_{i=n_m+1}^{n} x_i \right)^2 \tag{G.5}$$

$$= (w_m n_m + w_f n_f) \cdot \left( w_m n_m \cdot \overline{x_m^{(2)}} + w_f n_f \cdot \overline{x_f^{(2)}} \right) - (w_m n_m \cdot \overline{x}_m + w_f n_f \cdot \overline{x}_f)^2 \tag{G.6}$$

$$= w_m^2 n_m^2 \left( \overline{x_m^{(2)}} - \overline{x}_m^2 \right) + w_m w_f n_m n_f (\overline{x_m^{(2)}} + \overline{x_f^{(2)}} - 2 \cdot \overline{x}_m \cdot \overline{x}_f) + w_f^2 n_f^2 \left( \overline{x_f^{(2)}} - \overline{x}_f^2 \right) \tag{G.7}$$

$$= w_m^2 n_m^2 Var(\mathbf{x}_m) + w_m w_f n_m n_f \left( Var(\mathbf{x}_m) + Var(\mathbf{x}_f) + (\overline{x}_m - \overline{x}_f)^2 \right) + w_f^2 n_f^2 Var(\mathbf{x}_f) \tag{G.8}$$

$$= w_m n_m (w_m n_m + w_f n_f) Var(\mathbf{x}_m) + w_m w_f n_m n_f (\overline{x}_m - \overline{x}_f)^2 + w_f n_f (w_m n_m + w_f n_f) Var(\mathbf{x}_f) \tag{G.9}$$

Please note that we assume here that the definition of the variance is with division by $n$, and not the unbiased form $n-1$. Then

$$\mathbf{X}^T\mathbf{W}\mathbf{y} = \begin{bmatrix} w_m & \dots & w_f \\ w_m x_1 & \dots & w_f x_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} w_m \sum_{i=1}^{n_m} y_i + w_f \sum_{i=n_m+1}^{n} y_i \\ w_m \sum_{i=1}^{n_m} x_i y_i + w_f \sum_{i=n_m+1}^{n} x_i y_i \end{bmatrix} \tag{G.10}$$

Combining these gives the estimate

$$\hat{\boldsymbol{\beta}}_w = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y} \tag{G.11}$$

$$= \frac{1}{C_w}\begin{bmatrix} w_m\sum_{i=1}^{n_m} x_i^2 + w_f\sum_{i=n_m+1}^{n} x_i^2 & -\left(w_m\sum_{i=1}^{n_m} x_i + w_f\sum_{i=n_m+1}^{n} x_i\right) \\ -\left(w_m\sum_{i=1}^{n_m} x_i + w_f\sum_{i=n_m+1}^{n} x_i\right) & w_m n_m + w_f n_f \end{bmatrix} \tag{G.12}$$

$$\cdot \begin{bmatrix} w_m\sum_{i=1}^{n_m} y_i + w_f\sum_{i=n_m+1}^{n} y_i \\ w_m\sum_{i=1}^{n_m} x_i y_i + w_f\sum_{i=n_f+1}^{n} x_i y_i \end{bmatrix} \tag{G.13}$$

$$= \frac{1}{C_w}\begin{bmatrix} w_m n_m \cdot \overline{x_m^{(2)}} + w_f n_f \cdot \overline{x_f^{(2)}} & -(w_m n_m \cdot \overline{x}_m + w_f n_f \cdot \overline{x}_f) \\ -(w_m n_m \cdot \overline{x}_m + w_f n_f \cdot \overline{x}_f) & w_m n_m + w_f n_f \end{bmatrix} \begin{bmatrix} w_m n_m \cdot \overline{y}_m + w_f n_f \cdot \overline{y}_f \\ w_m n_m \cdot \overline{xy}_m + w_f n_f \cdot \overline{xy}_f \end{bmatrix} \tag{G.14}$$

$$= \frac{1}{C_w}\cdot\Bigg(w_m^2 n_m^2 \cdot \begin{bmatrix} \overline{x_m^{(2)}}\cdot\overline{y}_m - \overline{x}_m\cdot\overline{xy}_m \\ \overline{xy}_m - \overline{x}_m\cdot\overline{y}_m \end{bmatrix} + w_m w_f n_m n_f \cdot \begin{bmatrix} \overline{x_f^{(2)}}\cdot\overline{y}_m + \overline{x_m^{(2)}}\cdot\overline{y}_f - \overline{x}_m\cdot\overline{xy}_f - \overline{x}_f\cdot\overline{xy}_m \\ \overline{xy}_m + \overline{xy}_f - \overline{x}_m\cdot\overline{y}_f - \overline{x}_f\cdot\overline{y}_m \end{bmatrix} + w_f^2 n_f^2 \cdot \begin{bmatrix} \overline{x_f^{(2)}}\cdot\overline{y}_f - \overline{x}_f\cdot\overline{xy}_f \\ \overline{xy}_f - \overline{x}_f\cdot\overline{y}_f \end{bmatrix}\Bigg)$$

$$= \frac{1}{C_w}\cdot\Bigg(w_m^2 n_m^2 \cdot \begin{bmatrix} (Var(\mathbf{x}_m) + \overline{x}_m^2)\cdot\overline{y}_m - \overline{x}_m\cdot\overline{xy}_m \\ Cov(\mathbf{x}_m,\mathbf{y}_m) \end{bmatrix} + w_m w_f n_m n_f \cdot \begin{bmatrix} (Var(\mathbf{x}_f) + \overline{x}_f^2)\cdot\overline{y}_m + (Var(\mathbf{x}_m) + \overline{x}_m^2)\cdot\overline{y}_f - \overline{x}_m\cdot\overline{xy}_f - \overline{x}_f\cdot\overline{xy}_m \\ Cov(\mathbf{x}_m,\mathbf{y}_m) + \overline{x}_m\cdot\overline{y}_m + Cov(\mathbf{x}_f,\mathbf{y}_f) + \overline{x}_f\cdot\overline{y}_f - \overline{x}_m\cdot\overline{y}_f - \overline{x}_f\cdot\overline{y}_m \end{bmatrix} + w_f^2 n_f^2 \cdot \begin{bmatrix} (Var(\mathbf{x}_f) + \overline{x}_f^2)\cdot\overline{y}_f - \overline{x}_f\cdot\overline{xy}_f \\ Cov(\mathbf{x}_f,\mathbf{y}_f) \end{bmatrix}\Bigg)$$

$$= \frac{1}{C_w}\cdot\Bigg(w_m^2 n_m^2 \cdot \begin{bmatrix} Var(\mathbf{x}_m)\cdot\overline{y}_m - \overline{x}_m\cdot Cov(\mathbf{x}_m,\mathbf{y}_m) \\ Cov(\mathbf{x}_m,\mathbf{y}_m) \end{bmatrix} + w_m w_f n_m n_f \cdot \begin{bmatrix} Var(\mathbf{x}_f)\cdot\overline{y}_m + Var(\mathbf{x}_m)\cdot\overline{y}_f + \overline{x}_m(\overline{x}_m\cdot\overline{y}_f - \overline{xy}_f) + \overline{x}_f(\overline{x}_f\cdot\overline{y}_m - \overline{xy}_m) \\ Cov(\mathbf{x}_m,\mathbf{y}_m) + Cov(\mathbf{x}_f,\mathbf{y}_f) + (\overline{y}_f - \overline{y}_m)(\overline{x}_f - \overline{x}_m) \end{bmatrix} + w_f^2 n_f^2 \cdot \begin{bmatrix} Var(\mathbf{x}_f)\cdot\overline{y}_f - \overline{x}_f\cdot Cov(\mathbf{x}_f,\mathbf{y}_f) \\ Cov(\mathbf{x}_f,\mathbf{y}_f) \end{bmatrix}\Bigg)$$

$$= \frac{1}{C_w}\cdot\Bigg(w_m^2 n_m^2 Var(\mathbf{x}_m)\cdot\begin{bmatrix} \hat{\beta}_{0,m} \\ \hat{\beta}_{1,m} \end{bmatrix} + w_m w_f n_m n_f \cdot \begin{bmatrix} Var(\mathbf{x}_f)\cdot\overline{y}_m + Var(\mathbf{x}_m)\cdot\overline{y}_f + \overline{x}_m(\overline{x}_m\cdot\overline{y}_f - \overline{xy}_f) + \overline{x}_f(\overline{x}_f\cdot\overline{y}_m - \overline{xy}_m) \\ Var(\mathbf{x}_m)\hat{\beta}_{1,m} + Var(\mathbf{x}_f)\hat{\beta}_{1,f} + (\overline{y}_f - \overline{y}_m)(\overline{x}_f - \overline{x}_m) \end{bmatrix} + w_f^2 n_f^2 Var(\mathbf{x}_f)\cdot\begin{bmatrix} \hat{\beta}_{0,f} \\ \hat{\beta}_{1,f} \end{bmatrix}\Bigg) \tag{F.15}$$

So for the slope this gives:

$$\hat{\beta}_{1,w} = \frac{w_m^2 n_m^2 Cov(\mathbf{x}_m,\mathbf{y}_m) + w_m w_f n_m n_f (Cov(\mathbf{x}_m,\mathbf{y}_m) + Cov(\mathbf{x}_f,\mathbf{y}_f) + (\overline{y}_f - \overline{y}_m)(\overline{x}_f - \overline{x}_m)) + w_f^2 n_f^2 Cov(\mathbf{x}_f,\mathbf{y}_f)}{w_m n_m (w_m n_m + w_f n_f) Var(\mathbf{x}_m) + w_m w_f n_m n_f (\overline{x}_m - \overline{x}_f)^2 + w_f n_f (w_m n_m + w_f n_f) Var(\mathbf{x}_f)} \tag{G.15}$$

$$= \frac{w_m n_m (w_m n_m + w_f n_f) Cov(\mathbf{x}_m,\mathbf{y}_m) + w_m w_f n_m n_f (\overline{y}_f - \overline{y}_m)(\overline{x}_f - \overline{x}_m) + w_f n_f (w_m n_m + w_f n_f) Cov(\mathbf{x}_f,\mathbf{y}_f)}{w_m n_m (w_m n_m + w_f n_f) Var(\mathbf{x}_m) + w_m w_f n_m n_f (\overline{x}_m - \overline{x}_f)^2 + w_f n_f (w_m n_m + w_f n_f) Var(\mathbf{x}_f)} \tag{G.16}$$

## G.1 Substitution of $w_m$ and $w_f$

Taking equation (F.15) as a base;

$$(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y} = \tag{G.17}$$

$$\frac{\begin{array}{r} w_m^2 n_m^2 Var(\mathbf{x}_m)\cdot\begin{bmatrix}\hat{\beta}_{0,m}\\ \hat{\beta}_{1,m}\end{bmatrix} \\ +w_m w_f n_m n_f \cdot \begin{bmatrix} Var(\mathbf{x}_f)\cdot\overline{y}_m + Var(\mathbf{x}_m)\cdot\overline{y}_f + \overline{x}_m(\overline{x}_m\cdot\overline{y}_f - \overline{x}\overline{y}_f) + \overline{x}_f(\overline{x}_f\cdot\overline{y}_m - \overline{x}\overline{y}_m) \\ Var(\mathbf{x}_m)\hat{\beta}_{1,m} + Var(\mathbf{x}_f)\hat{\beta}_{1,f} + (\overline{y}_f - \overline{y}_m)(\overline{x}_f - \overline{x}_m)\end{bmatrix} \\ + w_f^2 n_f^2 Var(\mathbf{x}_f)\cdot\begin{bmatrix}\hat{\beta}_{0,f}\\ \hat{\beta}_{1,f}\end{bmatrix}\end{array}}{w_m n_m(w_m n_m + w_f n_f)Var(\mathbf{x}_m) + w_m w_f n_m n_f(\overline{x}_m - \overline{x}_f)^2 + w_f n_f(w_m n_m + w_f n_f)Var(\mathbf{x}_f)} = \tag{G.18}$$

$$\frac{\begin{array}{r} (\frac{n\cdot p_m}{n_m})^2 n_m^2 Var(\mathbf{x}_m)\cdot\begin{bmatrix}\hat{\beta}_{0,m}\\ \hat{\beta}_{1,m}\end{bmatrix} \\ +\frac{n\cdot p_m}{n_m}\frac{n\cdot p_f}{n_f} n_m n_f \cdot \begin{bmatrix} Var(\mathbf{x}_f)\cdot\overline{y}_m + Var(\mathbf{x}_m)\cdot\overline{y}_f + \overline{x}_m(\overline{x}_m\cdot\overline{y}_f - \overline{x}\overline{y}_f) + \overline{x}_f(\overline{x}_f\cdot\overline{y}_m - \overline{x}\overline{y}_m) \\ Var(\mathbf{x}_m)\hat{\beta}_{1,m} + Var(\mathbf{x}_f)\hat{\beta}_{1,f} + (\overline{y}_f - \overline{y}_m)(\overline{x}_f - \overline{x}_m)\end{bmatrix} \\ + (\frac{n\cdot p_f}{n_f})^2 n_f^2 Var(\mathbf{x}_f)\cdot\begin{bmatrix}\hat{\beta}_{0,f}\\ \hat{\beta}_{1,f}\end{bmatrix}\end{array}}{\frac{n\cdot p_m}{n_m} n_m(\frac{n\cdot p_m}{n_m} n_m + \frac{n\cdot p_f}{n_f} n_f)Var(\mathbf{x}_m) + \frac{n\cdot p_m}{n_m}\frac{n\cdot p_f}{n_f} n_m n_f(\overline{x}_m - \overline{x}_f)^2 + \frac{n\cdot p_f}{n_f} n_f(\frac{n\cdot p_m}{n_m} n_m + \frac{n\cdot p_f}{n_f} n_f)Var(\mathbf{x}_f)} = \tag{G.19}$$

$$\frac{n^2}{n^2}\frac{\begin{array}{r} (p_m^2 Var(\mathbf{x}_m)\cdot\begin{bmatrix}\hat{\beta}_{0,m}\\ \hat{\beta}_{1,m}\end{bmatrix} \\ +p_m p_f \cdot \begin{bmatrix} Var(\mathbf{x}_f)\cdot\overline{y}_m + Var(\mathbf{x}_m)\cdot\overline{y}_f + \overline{x}_m(\overline{x}_m\cdot\overline{y}_f - \overline{x}\overline{y}_f) + \overline{x}_f(\overline{x}_f\cdot\overline{y}_m - \overline{x}\overline{y}_m) \\ Var(\mathbf{x}_m)\hat{\beta}_{1,m} + Var(\mathbf{x}_f)\hat{\beta}_{1,f} + (\overline{y}_f - \overline{y}_m)(\overline{x}_f - \overline{x}_m)\end{bmatrix} \\ + p_f^2 Var(\mathbf{x}_f)\cdot\begin{bmatrix}\hat{\beta}_{0,f}\\ \hat{\beta}_{1,f}\end{bmatrix}\end{array}}{p_m(p_m + p_f)Var(\mathbf{x}_m) + p_m p_f(\overline{x}_m - \overline{x}_f)^2 + p_f(p_m + p_f)Var(\mathbf{x}_f)} = \tag{G.20}$$

$$\frac{\begin{array}{r} p_m^2 Var(\mathbf{x}_m)\cdot\begin{bmatrix}\hat{\beta}_{0,m}\\ \hat{\beta}_{1,m}\end{bmatrix} \\ +p_m p_f \cdot \begin{bmatrix} Var(\mathbf{x}_f)\cdot\overline{y}_m + Var(\mathbf{x}_m)\cdot\overline{y}_f + \overline{x}_m(\overline{x}_m\cdot\overline{y}_f - \overline{x}\overline{y}_f) + \overline{x}_f(\overline{x}_f\cdot\overline{y}_m - \overline{x}\overline{y}_m) \\ Var(\mathbf{x}_m)\hat{\beta}_{1,m} + Var(\mathbf{x}_f)\hat{\beta}_{1,f} + (\overline{y}_f - \overline{y}_m)(\overline{x}_f - \overline{x}_m)\end{bmatrix} \\ + p_f^2 Var(\mathbf{x}_f)\cdot\begin{bmatrix}\hat{\beta}_{0,f}\\ \hat{\beta}_{1,f}\end{bmatrix}\end{array}}{p_m Var(\mathbf{x}_m) + p_m p_f(\overline{x}_m - \overline{x}_f)^2 + p_f Var(\mathbf{x}_f)} \tag{G.21}$$

When considering only entry (2,2), this can be rewritten as follows;

$$\hat{\beta}_{1,w} = \frac{p_m^2 Var(\mathbf{x}_m)\hat{\beta}_{1,m} + p_m p_f\left(Var(\mathbf{x}_m)\hat{\beta}_{1,m} + Var(\mathbf{x}_f)\hat{\beta}_{1,f} + (\overline{y}_f - \overline{y}_m)(\overline{x}_f - \overline{x}_m)\right) + p_f^2 Var(\mathbf{x}_f)\hat{\beta}_{1,f}}{p_m Var(\mathbf{x}_m) + p_m p_f(\overline{x}_m - \overline{x}_f)^2 + p_f Var(\mathbf{x}_f)} \tag{G.22}$$

$$= \frac{p_m(p_m + p_f)Var(\mathbf{x}_m)\hat{\beta}_{1,m} + p_m p_f(\overline{y}_f - \overline{y}_m)(\overline{x}_f - \overline{x}_m) + p_f(p_m + p_f)Var(\mathbf{x}_f)\hat{\beta}_{1,f}}{p_m Var(\mathbf{x}_m) + p_m p_f(\overline{x}_m - \overline{x}_f)^2 + p_f Var(\mathbf{x}_f)} \tag{G.23}$$

$$= \frac{p_m Var(\mathbf{x}_m)\hat{\beta}_{1,m} + p_m p_f(\overline{y}_f - \overline{y}_m)(\overline{x}_f - \overline{x}_m) + p_f Var(\mathbf{x}_f)\hat{\beta}_{1,f}}{p_m Var(\mathbf{x}_m) + p_m p_f(\overline{x}_m - \overline{x}_f)^2 + p_f Var(\mathbf{x}_f)} \tag{G.24}$$

# Appendix H

# General weighted variance computation

The variance of the weighted estimators is

$$Var(\hat{\boldsymbol{\beta}}_w) = Var\left((\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}\right) \tag{H.1}$$

$$= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W} \cdot Var(\mathbf{y}) \cdot \left((\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\right)^T \tag{H.2}$$

So,

$$(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W} = \tag{H.3}$$

$$\frac{1}{C_w}\begin{bmatrix} w_m\sum_{i=1}^{n_m}x_i^2 + w_f\sum_{i=n_m+1}^{n}x_i^2 & -w_m\sum_{i=1}^{n_m}x_i - w_f\sum_{i=n_m+1}^{n}x_i \\ -w_m\sum_{i=1}^{n_m}x_i - w_f\sum_{i=n_m+1}^{n}x_i & w_m n_m + w_f n_f \end{bmatrix}\cdot\begin{bmatrix} w_m & \dots & w_m & w_f & \dots & w_f \\ w_m x_1 & \dots & w_m x_{n_m} & w_f x_{n_m+1} & \dots & w_f x_n \end{bmatrix} = \tag{H.4}$$

$$\frac{1}{C_w}\begin{bmatrix} w_m(w_m\sum_{i=1}^{n_m}x_i^2 + w_f\sum_{i=n_m+1}^{n}x_i^2) + w_m x_1(-w_m\sum_{i=1}^{n_m}x_i - w_f\sum_{i=n_m+1}^{n}x_i) & \dots & w_f(w_m\sum_{i=1}^{n_m}x_i^2 + w_f\sum_{i=n_m+1}^{n}x_i^2) + w_f x_n(-w_m\sum_{i=1}^{n_m}x_i - w_f\sum_{i=n_m+1}^{n}x_i) \\ w_m(-w_m\sum_{i=1}^{n_m}x_i - w_f\sum_{i=n_m+1}^{n}x_i) + w_m x_1(w_m n_m + w_f n_m) & \dots & w_f(-w_m\sum_{i=1}^{n_m}x_i - w_f\sum_{i=n_m+1}^{n}x_i) + w_f x_n(w_m n_m + w_f n_m) \end{bmatrix} = \tag{H.5}$$

$$\frac{1}{C_w}\begin{bmatrix} w_m(w_m n_m\cdot\overline{x_m^{(2)}} + w_f n_f\cdot\overline{x_f^{(2)}}) + w_m x_1(-w_m n_m\overline{x}_m - w_f n_f\overline{x}_f) & \dots & w_f(w_m n_m\cdot\overline{x_m^{(2)}} + w_f n_f\cdot\overline{x_f^{(2)}}) + w_f x_n(-w_m n_m\overline{x}_m - w_f n_f\overline{x}_f) \\ w_m(-w_m n_m\overline{x}_m - w_f n_f\overline{x}_f) + w_m x_1(w_m n_m + w_f n_f) & \dots & w_f(-w_m\overline{x}_m - w_f n_f\overline{x}_f) + w_f x_n(w_m n_m + w_f n_f) \end{bmatrix} = \tag{H.6}$$

$$\frac{1}{C_w}\begin{bmatrix} w_m\left(w_m n_m(\overline{x_m^{(2)}} - x_1\overline{x}_m) + w_f n_f(\overline{x_f^{(2)}})x_1\overline{x}_f)\right) & \dots & w_f\left(w_m n_m(\overline{x_m^{(2)}} - x_n\overline{x}_m) + w_f n_f(\overline{x_f^{(2)}})x_n\overline{x}_f)\right) \\ w_m\left(w_m n_m(x_1-\overline{x}_m) + w_f n_f(x_1-\overline{x}_f)\right) & \dots & w_f\left(w_m n_m(x_n-\overline{x}_m) + w_f n_f(x_n-\overline{x}_f)\right) \end{bmatrix} \tag{H.7}$$

So,

$$Var(\hat{\boldsymbol{\beta}}_w) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\cdot Var(\mathbf{y})\cdot\left((\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\right)^T \tag{H.8}$$

$$= \frac{1}{C_w}\begin{bmatrix} w_m\sigma^2_{\epsilon,m}\left(w_m n_m(\overline{x_m^{(2)}} - x_1\overline{x}_m) + w_f n_f(\overline{x_f^{(2)}})x_1\overline{x}_f)\right) & \dots & w_f\sigma^2_{\epsilon,f}\left(w_m n_m(\overline{x_m^{(2)}} - x_n\overline{x}_m) + w_f n_f(\overline{x_f^{(2)}})x_n\overline{x}_f)\right) \\ w_m\sigma^2_{\epsilon,m}\left(w_m n_m(x_1-\overline{x}_m) + w_f n_f(x_1-\overline{x}_f)\right) & \dots & w_f\sigma^2_{\epsilon,f}\left(w_m n_m(x_n-\overline{x}_m) + w_f n_f(x_n-\overline{x}_f)\right) \end{bmatrix} \tag{H.9}$$

$$\cdot\frac{1}{C_w}\begin{bmatrix} w_m\left(w_m n_m(\overline{x_m^{(2)}} - x_1\overline{x}_m) + w_f n_f(\overline{x_f^{(2)}})x_1\overline{x}_f)\right) & w_m\left(w_m n_m(x_1-\overline{x}_m) + w_f n_f(x_1-\overline{x}_f)\right) \\ \vdots & \vdots \\ w_f\left(w_m n_m(\overline{x_m^{(2)}} - x_n\overline{x}_m) + w_f n_f(\overline{x_f^{(2)}})x_n\overline{x}_f)\right) & w_f\left(w_m n_m(x_n-\overline{x}_m) + w_f n_f(x_n-\overline{x}_f)\right) \end{bmatrix} \tag{H.10}$$

This gives on entry (2,2);

$$Var(\hat{\beta}_{1,w}) = \frac{1}{C_w^2}\left[\sigma_{\epsilon,m}^2 w_m^2 \left(w_m n_m(x_1-\overline{x}_m)+w_f n_f(x_1-\overline{x}_f)\right)^2 + \cdots + \sigma_{\epsilon,f}^2 w_f^2 \left(w_m n_m(x_n-\overline{x}_m)+w_f n_m(x_n-\overline{x}_f)\right)^2\right] \tag{H.11}$$

$$= \frac{1}{C_w^2}\left[\sigma_{\epsilon,m}^2 w_m^2 \sum_{j=1}^{n_m}\left(w_m n_m(x_j-\overline{x}_m)+w_f n_f(x_j-\overline{x}_f)\right)^2 + \sigma_{\epsilon,f}^2 w_f^2 \sum_{j=n_m+1}^{n}\left(w_m n_m(x_j-\overline{x}_m)+w_f n_f(x_j-\overline{x}_f)\right)^2\right] \tag{H.12}$$

$$= \frac{1}{C_w^2}\left[\sigma_{\epsilon,m}^2 w_m^2 \left(w_m^2 n_m^2 \sum_{j=1}^{n_m}(x_j-\overline{x}_m)^2 + 2w_m w_f n_m n_f \sum_{j=1}^{n_m}(x_j-\overline{x}_m)(x_j-\overline{x}_f) + w_f^2 n_f^2 \sum_{j=1}^{n_m}(x_j-\overline{x}_f)^2\right)\right. \tag{H.13}$$

$$\left. + \sigma_{\epsilon,f}^2 w_f^2 \left(w_m^2 n_m^2 \sum_{j=n_m+1}^{n}(x_j-\overline{x}_m)^2 + 2w_m w_f n_m n_f \sum_{j=n_m+1}^{n}(x_j-\overline{x}_m)(x_j-\overline{x}_f) + w_f^2 n_f^2 \sum_{j=n_m+1}^{n}(x_j-\overline{x}_f)^2\right)\right] \tag{H.14}$$

$$= \frac{\begin{array}{l}\sigma_{\epsilon,m}^2 w_m^2 \left(w_m^2 n_m^2 \sum_{j=1}^{n_m}(x_j-\overline{x}_m)^2 + 2w_m w_f n_m n_m \sum_{j=1}^{n_m}(x_j-\overline{x}_m)(x_j-\overline{x}_m+\overline{x}_m-\overline{x}_f) + w_f^2 n_f^2 \sum_{j=1}^{n_m}(x_j-\overline{x}_m+\overline{x}_m-\overline{x}_f)^2\right) \\ + \sigma_{\epsilon,f}^2 w_f^2 \left(w_m^2 n_m^2 \sum_{j=n_m+1}^{n}(x_j-\overline{x}_f+\overline{x}_f-\overline{x}_m)^2 + 2w_m w_f n_m n_f \sum_{j=n_m+1}^{n}(x_j-\overline{x}_f+\overline{x}_f-\overline{x}_m)(x_j-\overline{x}_f) + w_f^2 n_f^2 \sum_{j=n_m+1}^{n}(x_j-\overline{x}_f)^2\right)\end{array}}{\left(w_m^2 n_m^2 Var(\mathbf{x}_m) + w_m w_f n_m n_f\left(Var(\mathbf{x}_m)+Var(\mathbf{x}_f)+(\overline{x}_m-\overline{x}_f)^2\right) + w_f^2 n_f^2 Var(\mathbf{x}_f)\right)^2} \tag{H.15}$$

(H.16)

(H.17)

$$= \frac{\begin{array}{l}\sigma_{\epsilon,m}^2 w_m^4 n_m^2 \sum_{j=1}^{n_m}(x_j-\overline{x}_m)^2 + 2\sigma_{\epsilon,m}^2 w_m^3 w_f n_m n_f \sum_{j=1}^{n_m}(x_j-\overline{x}_m)^2 + w_m^2 w_f^2 \sigma_{\epsilon,m}^2 n_f^2 \left(\sum_{j=1}^{n_m}(x_j-\overline{x}_m)^2 + n_m(\overline{x}_m-\overline{x}_f)^2\right) \\ + w_m^2 w_f^2 \sigma_{\epsilon,f}^2 n_m^2 \left(\sum_{j=n_m+1}^{n}(x_j-\overline{x}_f)^2 + n_f(\overline{x}_m-\overline{x}_f)^2\right) + 2\sigma_{\epsilon,f}^2 w_m w_f^3 n_m n_f \sum_{j=n_m+1}^{n}(x_j-\overline{x}_f)^2 + \sigma_{\epsilon,f}^2 w_f^4 n_f^2 \sum_{j=n_m+1}^{n}(x_j-\overline{x}_f)^2\end{array}}{\left(w_m^2 n_m^2 Var(\mathbf{x}_m) + w_m w_f n_m n_f\left(Var(\mathbf{x}_m)+Var(\mathbf{x}_f)+(\overline{x}_m-\overline{x}_f)^2\right) + w_f^2 n_f^2 Var(\mathbf{x}_f)\right)^2} \tag{H.18}$$

(H.19)

$$= \frac{\begin{array}{l}\sigma_{\epsilon,m}^2 w_m^4 n_m^3 Var(\mathbf{x}_m) + 2\sigma_{\epsilon,m}^2 w_m^3 w_f n_m^2 n_f Var(\mathbf{x}_m) + w_m^2 w_f^2 \sigma_{\epsilon,m}^2 n_f^2 n_m Var(\mathbf{x}_m) + w_m^2 w_f^2 \sigma_{\epsilon,m}^2 n_m n_f^2 (\overline{x}_m-\overline{x}_f)^2 \\ + w_m^2 w_f^2 \sigma_{\epsilon,f}^2 n_m^2 n_f Var(\mathbf{x}_f) + w_m^2 w_f^2 \sigma_{\epsilon,f}^2 n_m^2 n_f (\overline{x}_m-\overline{x}_f)^2 + 2\sigma_{\epsilon,f}^2 w_m w_f^3 n_m n_f^2 Var(\mathbf{x}_f) + \sigma_{\epsilon,f}^2 w_f^4 n_f^3 Var(\mathbf{x}_f)\end{array}}{\left(w_m^2 n_m^2 Var(\mathbf{x}_m) + w_m w_f n_m n_f\left(Var(\mathbf{x}_m)+Var(\mathbf{x}_f)+(\overline{x}_m-\overline{x}_f)^2\right) + w_f^2 n_f^2 Var(\mathbf{x}_f)\right)^2} \tag{H.20}$$

$$Var(\hat{\beta}_1) = \frac{\begin{array}{l}\sigma^2_{\epsilon,m} w_m^2 n_m (w_m^2 n_m^2 + 2 w_m w_f n_m n_f + w_f^2 n_f^2) Var(\mathbf{x}_m) \\ \quad + w_m^2 w_f^2 n_m n_f (\sigma^2_{\epsilon,m} n_f + \sigma^2_{\epsilon,f} n_m)(\overline{x}_m - \overline{x}_f)^2 + \sigma^2_{\epsilon,f} w_f^2 n_f (w_m^2 n_m^2 + 2 w_m w_f n_m n_f + w_f^2 n_f^2) Var(\mathbf{x}_f)\end{array}}{\left(w_m n_m (w_m n_m + w_f n_f) Var(\mathbf{x}_m) + w_m w_f n_m n_f (\overline{x}_m - \overline{x}_f)^2 + w_f n_f (w_m n_m + w_f n_f) Var(\mathbf{x}_f)\right)^2} \tag{H.21}$$

$$\tag{H.22}$$

$$= \frac{\begin{array}{l}\sigma^2_{\epsilon,m} w_m^2 n_m (w_m n_m + w_f n_f)^2 Var(\mathbf{x}_m) \\ \quad + w_m^2 w_f^2 n_m n_f (\sigma^2_{\epsilon,m} n_f + \sigma^2_{\epsilon,f} n_m)(\overline{x}_m - \overline{x}_f)^2 + \sigma^2_{\epsilon,f} w_f^2 n_f (w_m n_m + w_f n_f)^2 Var(\mathbf{x}_f)\end{array}}{\left(w_m n_m (w_m n_m + w_f n_f) Var(\mathbf{x}_m) + w_m w_f n_m n_f (\overline{x}_m - \overline{x}_f)^2 + w_f n_f (w_m n_m + w_f n_f) Var(\mathbf{x}_f)\right)^2} \tag{H.23}$$