Understanding the relationship between intellectual disability, poverty and health variables in the UK and Brazil

Laura Jane Barlow, M.Sci (Hons)

Department of Mathematics and Statistics

Submitted for the degree of Doctor of Applied Social Statistics at Lancaster university.

January 2021

Lancaster University

Abstract

The overall aim of this project was to determine whether or not there is a relationship between intellectual disability and poverty and health variables in Brazil and the UK. Through determining this relationship, the aim was to then try to profile a child who is at greater risk of intellectual disability in order to result in a quicker diagnosis and earlier access to the support and resources available for children with intellectual disabilities.

In order to investigate this relationship in Brazil, The Pesquisa Nacional de Saúde (PNS) was used. For the UK, data from the Millennium Cohort Study (MCS) was used.

In order to account for the complex survey design, both model-based and design-based approaches to analysis were investigated. Simulations were run to compare the various methods and recommendations about which methods to use in various scenarios were made.

Due to the large number of variables available in the two data sets, methods of variable selection were examined. Both stepwise selection based on Akaike Information Criterion (AIC) and the lasso were compared through simulations. It was found that although these methods resulted in different models being selected, the inference made based on the selected models did not vary much between the two methods.

To conduct the analysis of the PNS and the MCS a design-based approach was taken. Stepwise selection using AIC was used for variable selection and sampling weights were used when calculating the coefficient estimates and standard errors.

After the analysis of the PNS data, a potential profile of a child likely to have an intellectual disability in Brazil was found to be: a child who is unable to read and write with poor general health and multiple visits to doctor within a 12 month period. In the UK, it was found to be: a child in a family who requires extra support in the form of benefits along with a poorer general health which limits daily activities.

Acknowledgements

Firstly, I would like to thank my supervisor Juhyun Park for her guidance and support throughout this project. I would also like to thank Rodrigo de Moraes and Dirley dos Santos for their help in understanding the PNS and for being amazing hosts during my month in Brazil. Thank you also to Gillian Lancaster for introducing me to this project. Acknowledgements to the Economic and Social Research Council for funding this work.

One of the best things to come out of the last four years is the people I have met along the way. Thank you to all of the people in B18, both past and present, who have created such a friendly working environment. Thank you for listening to my constant moaning and for the advice and support you have all given to me over the years. I definitely wouldn't have made it to the end without you all.

Thank you to Helen for constantly making me laugh until cry and keeping me sane(ish) over the last many many years. Thank you also to Rach who believed in me even when I didn't believe in myself and whose never ending support has made completing this project possible. Thank you to both of you for listening to my many rants and always being there with a board game.

Finally, none of this would have been possible without the love and support from my parents. Thank you for always being there for me when I need you and for the pride you have in me for even my smallest achievements. I love you both so much.

Contents

List of Figures

List of Tables

List of Acronyms

AAIDD	American Association on Intellectual and Developmental Disabilities
AIC	Akaike Information Criteria
ANOVA	Analysis of Variance
BIC	Bayes Information Criteria
CAIDS-Q	Child and Adolescent Intellectual Disability Screening Questionnaire
CANTAB	Cambridge Neuropsychological Test Automated Battery
ESRC	Economic and Social Research Council
FACS	Family and Child Study
ICD	International Statistical Classification of Diseases
IQ	Intelligence Quotient
LASSO	Least Absolute Shrinkage and Selection Operator
LRT	Likelihood Ratio Test
MCP	Minimax Concave Penalty
MCS	Millennium Cohort Study
MLE	Maximum Likelihood Estimator
MSE	Mean Squared Error
PNS	Pesquisa Nacional de Saúde
PPH	Permanent Private Households
PPS	Probability Proportional to Size
PSU	Primary Sampling Unit
SCAD	Smoothly Clipped Absolute Deviation
SEN	Special Educational Needs
SENCO	Special Educational Needs Co-ordinator
SEP	Socio-economic position
SIPD	Integrated Household Surveys System
SRS	Simple Random Sampling
WHO	World Health Organisation

Chapter 1

Introduction

In recent years the population of both the UK and Brazil has increased rapidly and as a result challenges have arisen in many areas such as health, welfare, education and housing.

Intellectually disabled children in Brazil face barriers of a family, social and educational nature. Historically, in Brazil, the education available to disabled children was inadequate and in many cases non-existent. Between 1889 and 1920 there were only 7 state schools in Brazil for children with an intellectual disability (Jannuzzi, 2005). More recently however, after much persistence from parents, children with intellectual disability are now integrated into regular education or special services.

A similar trend has been seen in the UK. Up to the 1970's many children with an intellectual disability did not live with their families and instead were institutionalised. In more recent years, however, the majority of children with an intellectual disability live with their families and also attend schools which are inclusive of their needs (Scior and Werner, 2015).

Researching the equity of provision of education will provide a valuable insight into the relationship between educational attainment and the demand for education for children who are intellectually disabled. Also, a comparison between Brazil and the UK will provide a contrast to determine whether educational policies for disabled children can be improved upon in either of the two countries.

1.1 Definition and diagnosis of intellectual disability

The World Health Organisation (WHO) defines intellectual disability as "a significantly reduced ability to understand new or complex information and to learn and apply new skills (impaired intelligence). This results in a reduced ability to cope independently (impaired social functioning), and begins before adulthood, with a lasting effect on development".

An intellectual disability is characterised by the impairment of skills which contribute to the overall level of intelligence (Ke and Liu, 2015). Examples of such skills include cognitive, language, motor and social skills.

There are numerous terms used for intellectual disability which vary from country to country. These terms include, but are not restricted to, learning disability, special educational needs and mental retardation.

Diagnosis of an intellectual disability can happen at a variety of stages of a child's life: as a baby, at school age or at the transition from childhood to adulthood. If a baby is born with a syndrome which commonly results in an intellectual disability, then a diagnosis may be received within the first year of the child's life. When a child enters school, if their progress does not align with expectations or the progress of their peers, then a diagnosis may be made. Finally, a diagnosis may be made if a child struggles with the independence associated with transitioning from childhood to adulthood (McKenzie, 2013).

1.1.1 UK

In the UK, the term special educational needs or learning disability is most commonly used. This term should not be confused with the term learning difficulty. A learning disability is a condition that affects all areas of life, whereas a learning difficulty (such as dyslexia) is a condition which causes an obstacle to a specific form of learning (such as reading and spelling) but does not affect IQ (FPLD, 2017).

1.1.1.1 Definition

In "A Working Definition of Learning Disabilities", a paper written by Emerson and Heslop in 2010, it is stated that a child in the UK is classified as having a learning

disability if they meet any of the following criteria:

1. *"They have been identified within education services as having a Special Educational Need (SEN) associated with 'moderate learning difficulty' or 'profound multiple learning difficulty'. Children aged 7 or older should be at the School Action Plus stage of assessment or have a statement of SEN. Younger children should also be included if they are at the School Action stage of assessment of SEN.*
2. *They score lower than two standard deviations below the mean on a validated test of general cognitive functioning (equivalent to an IQ score of less than 70) or general development. Care should, however, be taken when considering the results of tests, especially tests carried out in English on children below the age of 7 living in bi-lingual households or households where English is not spoken.*
3. *They have been identified as having learning disabilities on locally held disability registers (including registers held by GP practices or Primary Care Trusts)."* (Emerson and Heslop, 2010)

'School Action' is the support that a child in the UK receives when it is felt that a child is not progressing adequately despite differentiated teaching (Dauncey, 2015). This support includes the involvement of a Special Educational Needs Co-ordinator (SENCO) to help to aid the child's learning. If a child still doesn't progress adequately, then additional support may be given. This additional support is 'School Action Plus'.

1.1.1.2 Diagnosis

For a person to be diagnosed as having an intellectual disability in the UK they must meet three criteria. First, they must have an IQ of 69 or lower. Next, they must have great difficulties in areas such as self-care or safety. Finally, the onset of these problems must have been during childhood. In order to evaluate these criteria, assessment should be carried out by a qualified psychologist using a standardised test (British Psychological Society, 2000).

One such test that is used in the UK to assess whether or not a child between the ages of eight and sixteen has an intellectual disability is the Child and Adolescent Intellectual Disability Screening Questionnaire (CAIDS-Q). This questionnaire is made up of seven items and hence is relatively quick to complete. During an evaluation of the properties

of the CAIDS-Q, the questionnaire was found to be highly accurate in identifying a child with intellectual disability with sensitivity of 100% and specificity ranging from 83% to 94% (McKenzie et al., 2018).

The CAIDS-Q includes items which are likely to be associated with a child having an intellectual disability such as the ability to tell the time and literacy skills. Responses to each of the items are recorded as yes or no with one point being given to a response of yes (with the exception of two items which are given one point if the response given is no). The score is then converted into a percentage and the higher the score, the less likely a child is to have an intellectual disability (McKenzie et al., 2012).

1.1.2 Brazil

In Brazil, the terms intellectual disability, intellectual developmental disorder or learning disability are most commonly used. Previously the term mental retardation has also been used.

1.1.2.1 Definition

In Brazil, the American Association on Intellectual and Developmental Disabilities (AAIDD) definition of intellectual disability is used (Carvalho and Forrester-Jones, 2016). That is, *"significant limitations in intellectual functioning and in adaptive behaviour as expressed in practical, social and conceptual skills originating before the age of 18"*.

A Portuguese paper written by Ke and Liu in 2015 states that according to the Diagnostic and Statistical Manual of Mental Disorders (DSM) and the International Statistical Classification of Diseases and Related Health Problems (ICD) there are three basic criteria which must be met for an individual to be diagnosed with an intellectual disability. The conditions are as follows:

- They have intellectual functioning significantly below average. This can be determined by an IQ score of 70 or lower.
- The individual has deficits or impairments in functioning in at least two of the following areas: communication, self-care, home living, social/interpersonal skills, use of resources in the community, self-direction, academic skills, work, leisure, health and safety.

- The onset is before the age of 18 (Ke and Liu, 2015).

1.1.2.2 Diagnosis

Diagnosis of intellectual disability in Brazil is conducted using ICD-10. The ICD-10 guidelines for diagnosing intellectual disability state that an individual should present with a reduced level of intellectual functioning and therefore have a reduced ability to adapt to the demands of daily life.

According to these guidelines, an individual with an IQ of between 50 and 69 has a mild intellectual disability, an individual with an IQ of between 35 and 49 has a moderate intellectual disability, an individual with an IQ of between 20 to 34 has a severe intellectual disability and an individual with an IQ of below 20 has a profound intellectual disability. To determine the IQ of an individual a standardized intelligence test which is appropriate for the individual's level of functioning should be used (WHO, 1992). However, there is no diagnosis measure that has been normed on the population of Brazil and the methodology for diagnosis suggested by ICD-10 is generally not followed correctly (Oakland, 2004).

Since most services for people with intellectual disabilities in Brazil are mainly privately funded, with only a few state funded, the assessment process of diagnosing intellectual disability is very inconsistent. This means that many people who have an intellectual disability in Brazil never get a diagnosis (Carvalho and Forrester-Jones, 2016).

1.2 Prevalence of Intellectual Disability

A meta-analysis conducted in 2011 by Maulik et al. found the global prevalence of intellectual disability to be 1.04% (Maulik et al., 2011). The rate of intellectual disability globally varies across an assortment of factors: gender (prevalence is higher in males), income (the highest prevalence occurs in low and middle income countries) and environment (prevalence is higher in urban areas than in rural areas) (Ke and Liu, 2015).

1.2.1 UK

Although there is no record of the exact number of people living with an intellectual disability in the UK, in 2011 it was estimated that the number was approximately 1.191 million people (roughly 1.9% of the population). Of these, 286,000 were estimated to be children (Emerson et al., 2012).

1.2.2 Brazil

Data collected in the 2010 census showed that around 45.6 million people in Brazil have a disability of some sort. It was also found that 1.4% of these people have a learning or mental disability (Carvalho and Forrester-Jones, 2016). It is thought however, that this figure may be inaccurate as the assessment instrument used was said to be difficult to understand and socio-culturally insensitive. Also, within the 2010 census people with mental illness were counted as having an intellectual disability and it failed to include individuals with an undeclared disability (Carvalho and Forrester-Jones, 2016).

1.3 Education Structure

Education is commonly thought of as a basic right which should be available for all children, with the majority of countries worldwide now having a range of ages for which education is compulsory. The funding for education in many countries comes from public resources and national governments. After the availability of basic education becoming a global priority in the mid 20th Century the rate of illiteracy worldwide declined greatly (Roser and Ortiz-Ospina, 2016).

1.3.1 UK

The structure of education in the UK is comprised of four main levels: primary education, secondary education, further education and higher education. Primary education is for children aged between 4 and 11 and secondary education is for children aged between 11 and 16. Further education may be entered once a student finishes their secondary education and higher education, which is university level, may be entered after further education has been completed.

In 1944 education was made compulsory until the age of 15. In 1972 the age was raised and now education is compulsory until the age of 16 in the whole of the UK (Norris, 2007). In 2013, the law was amended slightly in England adding a requirement that from age 16 to 18 an individual must either stay in full time education, start an apprenticeship or spend 20 hours per week working or volunteering whilst in part time education.

In the UK, all children between the ages of 5 and 16 are eligible for a free place at a state school. The funding for state schools comes from local authorities or from the government. There are four types of state schools in the UK. Community schools are funded by the state and hiring of staff for the school is the responsibility of the local education authority. Foundation schools are also state funded however the hiring of the employees of the school is the responsibility of the governors (Wood, 2006). Grammar schools can be funded by local authorities, a foundation body or a trust and admit children dependant on their academic ability. Finally, academies are run by trusts and are independent to local authorities.

A national curriculum was created in 1988 to provide a framework for education between the ages of 5 and 18 in England and Wales. The national curriculum includes a set of subjects for schools to follow in order to ensure that children across the country are all learning the same things. The equivalent in Scotland is called the Curriculum for Excellence programme and in Northern Ireland is known as the common curriculum (British Council, 2013). Community schools, foundation schools and grammar schools tend to follow the national curriculum whereas academies have the freedom to follow a different curriculum.

In addition to the free state schools, there are also private schools in the UK. Private schools are not funded by the government or local authorities and therefore do not have to follow the national curriculum.

1.3.2 Brazil

The structure of education in Brazil was introduced in 1971 and is comprised of three levels: elementary school (ensino fundamental), high school (ensino médio) and higher education (ensino superior). Elementary school is for children aged 6 to 14 and is compulsory for all children between the ages of 7 and 14 years old. High school is

for children between the ages of 15 and 17 and higher education is university level and takes place after all other schooling is complete (Meyer, 2010).

Before this structure was put into place, education was only compulsory up to fourth grade (approximately 11 years old) and education beyond this point was available only to children from higher income families. The rate of illiteracy in Brazil before 1971 was estimated to be 33% (Mantoan and Valente, 1998).

Education at municipality, state and federal level in Brazil is overseen by a system of ministries and government offices. Early childhood education is provided and regulated by municipalities whereas primary and secondary level education is the responsibility of states and federal districts. Nationally, The Ministry of Education is responsible for establishing policies and regulating public and private schools and provides technical and financial support for education systems within municipalities and states (Stanek, 2013).

Classes are provided to cover a range of areas: communication and expression (Portuguese), social studies (geography, history, political science), science (mathematics, physical-biological science) and educational practices (physical education, art, health education, civic and moral education) (Mantoan and Valente, 1998).

1.4 Intellectual Disability and Poverty

Inadequate prenatal care, inadequate medical care during delivery, malnutrition, accidents, physical abuse, childhood diseases and inherited syndromes are all factors that can lead to a child having an intellectual disability. Many of these factors are linked to poverty and thus there is evidence that poverty is a cause of intellectual disability and may be preventable through public health measures (Block, 2007).

1.4.1 UK

There have been a various studies into the relationship between poverty and intellectual disability in the UK. Studies have been conducted using data from the Family and Child Study (FACS) and the Millennium Cohort Study (MCS).

One paper written in 2010 for the Journal of Intellectual and Developmental Disability entitled “Poverty transitions among families supporting a child with intellectual

disability" outlines three potential pathways to the relationship between intellectual disability and poverty. Firstly, if a family has a child with an intellectual disability it may be more likely to enter poverty and less likely to escape poverty. Secondly, if a child grows up in poverty, they are more exposed to a range of hazards which can increase the risk of developing health conditions or impairments which are related to disability. Finally, there may be "third factors" which lead to an increase in risk for both poverty and intellectual disability. Examples of such "third factors" are poor parental health or parental intellectual disability (Emerson et al., 2010).

This study used data from the FACS and identified a child has having an intellectual disability if they responded "yes" to either *"Does ⟨ name of child ⟩ have any long-standing illness or disability?"* or *"Has ⟨ name of child ⟩ been identified at school as having SEN?"* . In addition to this, an answer of "yes" also had to be recorded to either *"Do/Does/Will this problem/ any of these problems affect ⟨ name of child ⟩'s ability to attend school or college regularly?"* or *"Do/Does/Will this problem/ any of these problems cause you to spend more time caring for ⟨ name of child ⟩ compared with a fully-fit child of a similar age?"*.

In order to measure poverty, two different methods were used. The first was income poverty which was based on equivalised household incomes. Equivalised income accounts for the different sizes and compositions of households by dividing the household's total income by it's equivalent size. The second was hardship which was based on a family's access to assets and resources.

The paper identifies three potential events which are believed to be associated with a family including a child with an intellectual disability entering income poverty: an increase in the number of dependent children in the family; the main informant of the questionnaire developing a disability of some kind and the number of adults working 16 or more hours per week increasing. After analysis it was found that the first two of these events were found to be statistically significant however the third was not.

In addition to income poverty, the paper also highlighted three potential trigger events potentially related to a family entering hardship. These events are: separation, an increase in the number of dependent children in the family and if the occupational status of the family decreases. Analysis found that none of these events were statistically significant in relation to the risk of a family entering hardship.

The probability of a family exiting income poverty was also investigated in this paper. Three possible events which could be associated with a family exiting poverty were identified: if the occupational status of the family increased; if the health of the informant of the questionnaire improved and if the number of adults working 16 or more hours per week increased. Only the first of these three potential trigger events was found to be significant during the analysis.

The paper summarises that families who are supporting a child with an intellectual disability or any other type of disability are more likely to be living in income poverty and hardship when compared to a family who is not supporting a child with an intellectual disability.

Also, when taking into account the initial poverty status of a family, it was found that a family including a child with an intellectual disability is both more likely to enter hardship and less likely to transition out of hardship in a 12 month period, when compared to a family which does not support a child with an intellectual disability.

These findings are consistent with the initial analysis in a later article from The Journal of Social Policy with the title "Child Disability and the Dynamics of Family Poverty, Hardship and Financial Strain: Evidence from the UK". In addition to the two indicators used in the previously mentioned study to measure poverty (income poverty and hardship) this study used a third indicator: financial strain - based on a self-reported (by the adult informant) evaluation of the level of financial strain experienced. This study also used data from the FACS and so a child was identified as having a disability if they met the same conditions as previously mentioned.

The article hypothesised that when compared to a family who is not supporting a child with a disability, a family supporting a disabled child will: be in poverty for a greater proportion of yearly intervals; have an increased chance of entering poverty and have a reduced chance exiting poverty (Shahtahmasebi et al., 2011).

In addition to the initial analysis, a further analysis comparing the association between poverty and child disability, whilst taking into account possible confounding variables, was also conducted. The proposed confounding variables included: composition of the household, age and sex of informant, general health status of informant, presence of long-standing illness or disability, smoking status of informant, occupational status of the household, academic attainment of household and neighbourhood deprivation.

When an analysis was completed including these potential confounding variables it was found that any associations found previously were reduced, eliminated or reversed meaning that a family supporting a child with a disability is no more likely to exit or enter poverty than a family who is not supporting a child with a disability when they have similar levels of resources.

A further paper which was written for Public Health England in 2015 used data from the MCS to "summarise current knowledge about the determinants of health inequalities experienced by children with learning disabilities in the UK" (Emerson, 2015). In this study it was discovered that the majority of children in the UK who do not have a learning disability had not experienced income poverty at three or more of the initial five waves of the MCS. The paper states that there is already known evidence of an inequality between children with an learning disability and children with no learning disability. Children with a learning disability are known to be 'significantly more likely than their non-disabled peers to be living in households characterised by low socio economic position (SEP) and poverty" (Emerson, 2015).

It finds that although this association is not exclusive to learning disabilities, the relationship is especially strong between child disability and low SEP for children with learning disabilities. This is particularly true for children with less severe learning disabilities. Disabled children (including those with an intellectual disability) are, at any point in time, at a greater risk of poverty. Over time, they are also more likely to enter poverty and remain in poverty and less likely to escape from poverty.

As a likely result of the apparent association between intellectual disability and poverty, children with a learning disability are more likely to experience a variety of hazards that can be detrimental to their health. These hazards include, but are not exclusive to: inadequate nutrition, poor housing conditions and family, peer and community violence (Emerson, 2015).

1.4.2 Brazil

There have been very few studies conducted in Brazil to evaluate the relationship between intellectual disability and poverty. In Brazil only 7% of articles in the mental health field between 1999 and 2003 discuss mental health in children and this percentage is further decreased when looking at intellectual disability (Razzouk et al., 2006).

A paper entitled "Perspectives of intellectual disability in Latin American countries: epidemiology, policy and services for children and adults" discusses the lack of research into intellectual disability in Brazil among other Latin American countries. It states that the majority of studies published regarding intellectual disabilities in Brazil tend to be specific to small regions and produce only descriptive statistics (Mercadante et al., 2009). For example a study of 500 children, aged between 6 and 12, was conducted and found that 4% of students had an IQ of below 70 (Assis, 2009). A study linking education and intellectual disability looked at two groups of 44 students and concluded that intellectual disability was the greatest predictor of a child dropping out of school (Tramontina et al., 2002).

A further paper entitled "Poverty, disabilities and violence" examines social inequalities in Latin America, particularly Brazil, regarding chronic poverty in the disabled community and its connection with violence. The paper states that "the links between poverty and disability - that poverty causes disability and disability causes poverty has not yet been addressed" (Marinho, 2009). The major barrier in studying this relationship currently is defining poverty. When looking at poverty it is important to not only consider the lack of material assets but to take a broader view and also consider the lack of resources available. Despite the fact people with disabilities may not be the poorest, they may suffer from lack of access to healthcare, education and employment.

This study found that based on data from the 2000 Census, there were approximately 9 million disabled people in Brazil (not specifically intellectually disabled) who had a monthly salary of between 100 and 200 US dollars.

Marinho states that there is still insufficient interest in the relationship between intellectual disability and poverty in Brazil and emphasises the need to bring this topic to social and political sectors in order to reduce poverty in the disabled community.

1.5 Motivation

Despite various studies being conducted into the relationship between poverty and intellectual disability in the UK, there has been very little research conducted for the population of Brazil. Some studies in the UK confirm a relationship between intellectual disability and poverty, however others conclude that when accounting for other variables,

this relationship no longer exists. Therefore in order to fully understand the relationship between intellectual disability and poverty in both a developed and a developing country, a new analysis will be conducted.

In 2011 it was estimated that only 6.6% of adults with an intellectual disability in the UK were in some form of paid employment with the majority of employment being part-time (Emerson et al., 2012). This suggests that despite recent advances in the policies for disabled individuals, there are still improvements which can be made. If it is possible to profile a child that may be likely to have an intellectual disability, early intervention is more likely to be an option.

Early intervention is when a child receives support for their disability in the early years of their lives, however according to Mencap, many children with intellectual disabilities currently do not receive the support that they need. Without early intervention, an intellectually disabled individual may be more likely to have poor outcomes in life resulting in higher financial costs not only to the family but also to society (Cooper et al., 2014).

There is a belief in Brazil that there is a relationship between poverty and intellectual disability however there is a lack of research into this subject. In particular there have been no large scale studies in Brazil which examine the relationship between intellectual disability poverty. A person living in poverty faces limitations with their access to services and resources such as education. Therefore, without studies regarding intellectual disabilities the progression of education for disabled children will be halted (Mantoan and Valente, 1998).

International comparisons are important since they allow the countries involved to gain inspiration from each other in order to improve. Both the UK and Brazil have adopted similar policies in recent years regarding the education of children with intellectual disabilities and therefore, comparing the two countries will provide insight into how both systems can be improved.

1.6 Data Sources

In order to examine the relationship between poverty variables and intellectual disability in Brazil, data from the Brazilian National Health Survey, The Pesquisa Nacional de

Saúde (PNS) will be used. The PNS was created in partnership with the Brazilian Institute of Geography (IBGE) by the Ministry of Health with the objectives to "produce national level data about the health status and lifestyles of the Brazilian population and also data about health care regarding access, use of health services, preventative actions, continuity of care and health care funding" (Souza-Júnior et al., 2015).

The PNS was first conducted in 2013 and used a multi-stage, probability sampling design in order to provide estimates of various characteristics of the population of Brazil.

In regards to the UK, data from the The Millennium Cohort Study (MCS) will be used. The MCS is a longitudinal study which is collecting data from approximately 19,000 children who were born in the UK between September 2000 and January 2002. The study is funded by the Economic and Social Research council (ESRC) and certain Government departments and so far there have been seven surveys conducted. These have occurred when the children were aged nine months, three years, five years, seven years, eleven years old, fourteen and seventeen years old.

Data concerning the child, the child's siblings and the child's parents has been collected. Particular topics covered by the data include: parenting; school choice; child behaviour and cognitive development; employment of parents; education of parents and income and poverty.

The study was developed to see the impact that family context (the home setting and family characteristics e.g. parental stress and parenting practices) in the early years of a child's life has on the child's development and outcomes throughout their childhood, their adolescent years and further into adulthood.

All analyses and simulations in this project will be conducted using R.

1.7 Sampling weights

If a survey has a complex design then there are many factors which may lead to sampling weights being included within the data. In order for any analysis conducted on the sample to fully describe the population, these sampling weights may need to be included.

The main purposes of sample weights when considering data with a complex sampling structure are: to compensate when individuals or households do not have an equal probability of inclusion in the survey, to compensate for non-responses and to adjust for

certain characteristics in order to ensure the sample conforms to the entire population (Yansaneh, 2003).

Both the Millennium Cohort Study (MCS) and the Pesquisa Nacional de Saúde (PNS) have complex survey designs and both include sampling weights. This raises the question of whether or not these weights should be incorporated into the analysis of these data sets and, if so, how it is best to incorporate them.

1.8 Variable Selection

When a survey such as the PNS or MCS is conducted, generally information about a wide variety of topics is collected. This in turn can lead to an extensive data set with a large number of possible predictors. When this is the case, some sort of variable selection will be necessary. Variable selection aims to remove any unnecessary predictors from the model.

There are a number of motivations for conducting variable selection. Occam's Razor says that if there are multiple explanations for something then the simplest explanation is the best. This principle can be applied to regression models and therefore we want to find the smallest possible model which fits the data (Wears and Lewis, 1999).

If more predictors than necessary are included in the regression model, it will add noise to the estimation of the parameters of interest. Also, the risk of collinearity is increased if there are too many covariates included.

There are a number of ways in which variable selection can be conducted. These include: background knowledge, using information criteria and introducing a penalty to the likelihood. In particular, the recent developments in the penalty methods suggest that it is possible to incorporate variable selection in the estimation so that an automatic variable selection is possible. Arguably, the Lasso (Tibshirani, 1996a) is one of the first methods developed and is still the most popular. This raises the question of which method of variable selection should be used to analyse the survey data from the PNS and the MCS.

1.9 Research questions and objectives

The research questions to be answered in this thesis are:

1. How do poverty and health variables interrelate with intellectual disability in Brazil and the UK?

2. When is it appropriate to include sampling weights and how should they be used when necessary?

3. Which methods of variable selection are commonly used when analysing survey data? Specifically, how has the lasso been adapted beyond linear regression with continuous covariates for use with more complex data?

4. Is it possible to profile different types of children that need lower and higher levels of support to aid in identifying subgroups for selective interventions to alleviate inequalities in education?

The specific objectives are to:

1. Review the current practice used to analyse data from a complex survey design, specifically how it is advised to incorporate the sampling weights.

2. Conduct simulations to show the effect that using sampling weights has on coefficient estimates under varying sampling schemes.

3. Review the current practices used to conduct variable selection and see how they can be extended for use on data from a complex survey.

4. Conduct simulations to compare the resulting models when using different methods of variable selection.

5. Review the current practices for conducting variable selection for a complex survey sample.

6. Compute a relevant analysis to determine the relationship between intellectual disability and the poverty and health variables in both Brazil and the UK.

1.10 Outline of thesis

Chapter 2 gives information about the Brazilian National Health Survey (Pesquisa Nacional de Saúde (PNS)). The data from this survey will be used to answer the research

questions regarding Brazil. The aims and content of the survey are discussed. Information is given regarding how the sample of the PNS was selected and furthermore how the sampling weights provided have been calculated. Exploratory analysis has been conducted. Within this, the variable of interest has been selected and potential variables that could be indicators of poverty have been established.

Chapter 3 looks at the Millennium Cohort Study (MCS). The data from the fifth sweep of the MCS conducted in 2013 will be used to answer the research questions regarding the UK. Similarly to the previous chapter, the aims and content of the survey are discussed as well as the sampling scheme and how the sampling weights have arisen. Exploratory analysis has been conducted in which the response variable has been selected, similar variables to those found in the PNS have been established and further variables which may be indicators of poverty have been identified.

Chapter 4 looks at when and how sampling weights are used when analysing survey data. The different methods for calculating sampling weights are described. Literature regarding when it is appropriate to use sampling weights for descriptive statistics and regression modelling is discussed along with the proposed methods to conduct such analyses. Both model-based and design-based methods are examined. Potential tests to determine whether or not the use of sampling weights is necessary when using linear regression modelling are described. Model-based and design-based methods are compared through simulation studies for a variety of scenarios. A test for determining the appropriateness of using weights in a linear regression model is adapted for use with a logistic regression model.

Chapter 5 examines various methods of variable selection. Variable selection using background knowledge, information criteria and penalised likelihoods are discussed. The least absolute shrinkage and selection operator (lasso) is defined and the ways in which it has been adapted to account for binary response variables and variables with a grouping structure are discussed. Simulations comparing the lasso with step-wise selection have been computed. The need for centering and standardising the design matrix is also examined.

Chapter 6 combines methods from the two previous chapters the way in which the lasso can be used for survey data from a finite population is discussed. Various design-based models are fit to the data from the PNS and the MCS and the results of each are

examined to determine the similarities and differences in inference between the resulting models.

Chapter 7 uses stepwise selection based on AIC to determine models to describe the relationship between intellectual disability and poverty and health variables in the two countries. Sample weights are used when calculating the estimates for the coefficients. The resulting models are then interpreted and the results for both countries are compared and contrasted.

Chapter 8 concludes the findings of the previous chapters. The profile of a child with a potential intellectual disability is determined and potential recommendations to policies are made. Finally, future work is discussed.

Chapter 2

Data - Brazil (The Pesquisa Nacional de Saúde (PNS))

This chapter will discuss the dataset that will be used for the Brazilian analysis. The sampling scheme of the PNS will be discussed along with how the sampling weights have been calculated. The content of the survey will be also be described. Finally, exploratory analysis will be conducted in which the response variable will be identified, proxy variables for poverty will be examined and the remaining variables of the PNS will also be explored.

2.1 Obtaining the sample for the PNS

The target of geographical coverage of the PNS was the entire country. The sample was made up of people living in permanent private households (PPH) within one of the census tracts of the 2010 Geographic operating base (Souza-Júnior et al., 2015).

Certain areas of the country were excluded from the sample such as indigenous villages, military bases, camp sites, jails, nursing homes and hospitals.

2.1.1 The Master Sample

The sample selected for the PNS was a subsample of the Integrated Household Surveys System's (SIPD) Master Sample (Damacena et al., 2013). The Master Sample is made up of primary sampling units (PSUs) which are used for various studies. A detailed description of how the Master Sample was created is given later in the chapter.

2.1.2 Sample Size

After taking many factors into account, including the effect of the sampling plan, it was calculated that the size of the sample for each of the geographical areas should be at least 900 households. Therefore the estimated sample size, based on a 20% non-response rate, was approximately 80,000 households.

During the study a total of 81,167 households were visited. Out of these households, 69,994 were occupied. In total 64,348 household interviews and 60,202 individual interviews with a randomly selected resident were conducted.

Since the main interest of this project is intellectual disability in children the data will be reduced to include only responses regarding children aged between 5 and 18. Therefore the sample size used throughout this project is 45,517.

2.2 Survey design and content

When conducting the PNS, the questionnaire was divided into three sections each with a different respondent. Within each section, the questionnaire was further divided into modules.

2.2.1 Part 1 - Household section

The first section of the questionnaire concerned the whole household and questions were answered by the head of the household. The modules in this section of the questionnaire asked questions regarding the following:

- **Module A:** Address information and household demographics
- **Module B:** Home visits from the Family Health Team and Endemic Disease Agent.

Module A includes questions regarding the location of the home, the type of housing, the materials used in the walls/ floor/ roof, access to water and access to other appliances such as TV/ fridge/ computer.

Module B asks whether the family has a family health plan and whether they have recently received visits from the Family Health Team and Endemic Disease Agents.

2.2.2 Part 2 - Household residents

The second section of the PNS was answered by all of the residents of the household. If a member of the household was absent during the interview, the questions could be answered on their behalf by the head of the household. For children, answers were supplied by a parent or guardian.

The modules of the second section of the questionnaore asked questions relating to the following subjects:

- **Module C:** General characteristics of residents
- **Module D:** Education characteristics of people 5 years or older
- **Module E:** Work of household members (aged 14 years or older)
- **Module F:** Household income
- **Module G:** People with disabilities
- **Module I:** Health plan coverage
- **Module J:** Health service utilisation
- **Module K:** Health of individuals 60 years or older and mammography coverage for women over 50
- **Module L:** Children under 2.

More specifically, module C concerns the general characteristics of the individual such as gender, age and race.

Module D was asked to only those aged over five years old and concerned the educational characteristics of the individual. Questions were asked about whether or not school is attended, whether the individual can read and write and the highest level of education an individual has received.

Due to issues in data collection, answers to modules E and F are not available.

Module G is concerned with disabilities and asks whether or not an individual has an intellectual, physical, hearing or visual disability along with asking about how the disability has an effect on every day tasks if relevant. Since the main focus of this project is intellectual disability, this module will be of high interest and the question which asks

"Do you have an intellectual disability?" will serve as a binary response when conducting logistic regression.

Module I asks about whether the individual has a health plan and if they do what the plan covers, how long the health plan has been held and how much the health plan costs.

Module J asks about how much the individual uses health services in particular, the general health status of an individual, when the last time a doctor/ dentist was consulted and whether or not an individual has been hospitalised in the last 12 months.

Module K was asked to residents over the age of 50 and so is not of interest in this project. Similarly, module L concerned only children under the age of two and so is also of no interest in this project.

2.2.3 Part 3 - Individual

Part 3 of the questionnaire was issued to only one member of each household. This resident was selected at random from all the residents of the household aged 18 or over. Questions in this section related to the following topics:

- **Module M:** Other work characteristics and social support
- **Module N:** Perception of health status
- **Module O:** Accidents and violence
- **Module P:** Lifestyle
- **Module Q:** Chronic diseases
- **Module R:** Women's health (aged 18 or over)
- **Module S:** Prenatal care (women who gave birth between 28/07/11 and 27/07/13)
- **Module U:** Oral health
- **Module X:** Health care

Since this section was asked to residents over the age of 18, the data collected in this section of the questionnaire is of no interest in this project.

2.3 Sampling weights in The PNS

Due to the complex survey structure of the PNS, sampling weights are provided in the dataset. The following section describes the sampling plan in more detail and how the sampling weights have been calculated.

2.3.1 Stratification of The Master Sample

As mentioned previously, the Master Sample is a collection of census tracts or groups of census tracts that have been selected for use as primary sampling units (PSUs) in various studies.

PSUs were selected by stratification across four different criteria:

1. **Administrative** - stratifies by state and also within the state based on whether the area is a capital city, metropolitan region or integrated economic development region.

2. **Geographic** - divides the state capital cities and any larger cities into further strata.

3. **Area situation** - separates geographic strata into rural and urban areas.

4. **Statistical** - classifies the rural and urban areas into similar strata by household income and number of permanent private households (PPHs) (Souza-Júnior et al., 2015).

2.3.2 Probability of a PSU being selected in the Master Sample

PSUs from each strata were selected using probability proportional to size (PPS) sampling where the number of PPHs were used to determine the size of each PSU.

The probability of a certain cluster being selected when using PPS sampling is

$$\frac{a \times b}{c}$$

where:

- a is the cluster population,

- b is the number of clusters sampled and
- c is the total population (sum of the population of all clusters).

Therefore, for the Master Sample, the probability of PSU i being selected in strata h is:

$$\frac{N_{hi} \times m_h}{N_h}$$

where:

- N_{hi} is the population of PSU i,
- m_h is the number of PSUs to be sampled in strata h and
- N_h is the total population of all PSUs in strata h.

2.3.3 Weight of the primary sampling units in the Master Sample

The sampling weights of each of the PSUs is the inverse of their sampling probability. Therefore PSU i in strata h has a sampling weight of

$$\frac{N_h}{N_{hi} \times m_h},$$

where N_h, N_{hi} and m_h are as previously defined.

2.3.4 Obtaining the PNS sample from the Master Sample

The sample for the PNS was obtained using simple random sampling in three stages:

1. **Stage 1:** Simple random sampling (SRS) was used to select the PSUs from the Master sample. This maintained the stratification of PSUs used in the Master Sample which is as previously discussed.
2. **Stage 2:** Simple random sampling was used to select a fixed number of PPHs from each of the PSUs.
3. **Stage 3:** Simple random sampling was used again to select an individual in each household (aged 18 or over) from a list of residents to complete the final part of the questionnaire (Souza-Júnior et al., 2015).

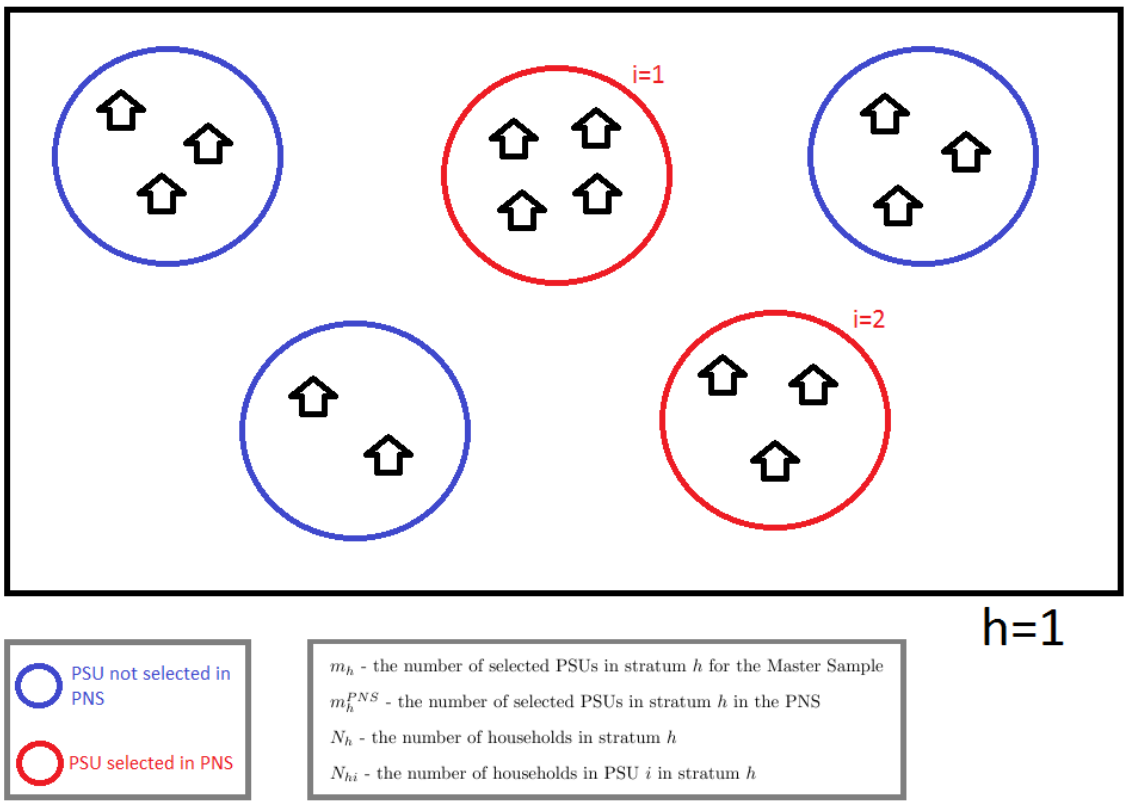

Figure 2.1: A strata of the Master Sample showing the primary sampling units (PSUs) and the individual households within these PSUs.

2.3.5 Probability of a PSU being selected in the PNS sample

Figure 2.1 shows an example of a strata from the Master Sample. The circles represent PSUs with the red indicating that the PSU is selected in the sample for the PNS.

The probability of PSU i being selected in the sample for the PNS is given by the probability of the PSU being selected in the Master sample multiplied by the probability that it is chosen from this sample by SRS.

$$\frac{m_h \times N_{hi}}{N_h} \times \frac{m_h^{PNS}}{m_h}$$

where m_h, N_{hi} and N_h are as previously defined and m_h^{PNS} is the number of PSUs in stratum h that are selected for the PNS.

2.3.6 Weight of PSU i in the PNS sample

The weights of the PSUs in the PNS sample are calculated as the inverse of the sampling probability of selection and are defined as

$$w_{hi} = \frac{N_h}{m_h \times N_{hi}} \times \frac{m_h}{m_h^{PNS}}$$

for PSU i in stratum h (Paulo and Freitas, 2014a).

2.3.7 Basic weights of households in the PNS sample

The households in each of the PSUs were selected by simple random sampling therefore the probability of a household in PSU i and stratum h being selected is given that this PSU and stratum has been selected:

$$\frac{n_{hi}}{N_{hi}^*}$$

where n_{hi} is the number of households selected in PSU i, stratum h and N_{hi}^* is the total number of PPHs in PSU i, stratum h.

Therefore the weight of a household within a PSU is given by:

$$w_{j|hi} = \frac{N_{hi}^*}{n_{hi}}$$

Combining the weight of a PSU and a household within the PSU gives a basic household weight of:

$$w_{hij} = w_{hi} \times w_{j|hi} = \frac{N_h}{m_h \times N_{hi}} \times \frac{m_h}{m_h^{PNS}} \times \frac{N_{hi}^*}{n_{hi}}$$

for household j in PSU i in stratum h (Paulo and Freitas, 2014a).

2.3.8 Weights correcting for non-responses and to population calibration

The final weight of a household in the PNS sample accounts for the weight of the corresponding PSU, adjusts for households with non-responses and also calibrates estimates with population totals with the use of other sources (Paulo and Freitas, 2014a).

2.3.8.1 Correcting for non-responses

Non-responses may have occurred for a variety of reasons including refusal by the informant or an inability to contact a resident of a chosen household. An adjustment for this

is given by the following weight:

$$w^{*}_{hij} = w_{hij} \times \frac{n^{*}_{hi}}{n^{**}_{hi}}$$

where n^{*}_{hi} is the number of selected households with residents in PSU i, stratum h and n^{**}_{hi} is the number of households in which an interview was held in PSU i, stratum h (Paulo and Freitas, 2014a).

2.3.8.2 Population calibration

An additional adjustment was made to account for the results of research from other sources. This adjustment means that when estimating the total population at certain geographic levels, the estimates concur with the population estimates produced by the Population Coordination and Social Indicators (COPIS) of the Research Board (Paulo and Freitas, 2014a).

Based on information from 27th July 2013, the calibration can be seen in the following expression:

$$w^{**}_{hij} = w^{*}_{hij} \times \frac{P^{tri}_{a}}{\hat{P}^{tri}_{a}}$$

where P^{tri}_{a} is the population estimates produced by COPIS at geographic level a and $\hat{P}^{tri}_{a}$ is the population estimated obtained with the survey data at geographic level a.

2.3.9 Final Household Weight

The final weight of a household which accounts for the probability of selection of the corresponding PSU, compensates for non-responses and calibrates for known population totals is given by (Paulo and Freitas, 2014b):

$$w^{**}_{hij} = \frac{1}{m_h} \times \frac{N_h}{N_{hi}} \times \frac{m_h}{m^{PNS}_{h}} \times \frac{N^{*}_{hi}}{n_{hi}} \times \frac{P^{tri}_{a}}{\hat{P}^{tri}_{a}}.$$

2.4 Exploratory analysis

As mentioned previously, the variable of interest in this project is intellectual disability. Since the variable of interest in this project is binary, logistic regression will be used

to examine the relationship between intellectual disability and the other socio-economic variables in the data. The methods for analysis of data from a survey with a complex sampling structure will be discussed in more detail in later chapters.

In order, to understand what is available in the PNS data, exploratory analysis will be carried out. This will involve looking at the frequency of the response variable and looking at the other socio-economic variables in more detail.

Since interest lies in children with intellectual disabilities, the data set was cut down to include only people younger than 18 years old. Since Module D of the PNS was only filled in for people over the age of the 5, the lower bound of the age range was adjusted to account for this. Therefore, the following analysis was conducted on data from the PNS for children aged between 5 and 18 years old.

2.4.1 The response variable - Intellectual disability

Module G of the PNS questionnaire is interested in disabilities. Information was collected regarding intellectual disabilities, physical disabilities, hearing impairments and visual impairments.

The following question was asked to/about each member of the household:

Does ⟨ name ⟩ have an intellectual disability?

The possible answers to this question were "yes" or "no" providing a binary variable.

Initial exploratory analysis of this variable showed that out of 45,517 children, 404 of them have been recognised as having an intellectual disability.

Since less than 1% of the sample size has an intellectual disability, it may be relevant to also look at the other disabilities (physical, hearing, visual) which were asked about in the questionnaire. This can be done with the aim to see whether there is any correlation between intellectual disability and any of the other three disabilities and as a result, whether a further group of people could be used to help to model any variables significant to intellectual disability.

Table 2.1 shows whether or not the children who have been classified as having an intellectual disability have also been classified as having a further disability. It can been seen, however, that out of the 404 children with an intellectual disability 327 of them have no further disability. This means that it is unlikely that combining the disability groups will give any further insight to the factors that influence intellectual disability.

Table 2.1: Disabilities (physical, hearing or visual) the children in the PNS sample with an intellectual disability have.

	Hearing disability				
	Yes		No		
	Visual disability				
Physical disability	Yes	No	Yes	No	**Total**
Yes	1	2	5	36	44
No	7	10	16	327	360
Total	8	12	21	363	404

2.4.2 Using Module A to measure poverty

In order to examine the relationship between intellectual disability and poverty, it needs to be determined how poverty will be measured using the available data. Generally, to measure poverty, a cut off called the 'poverty line' is used. A poverty line of 60% of the countries median income is widely used and it is recommended that this value is used for cross-national comparisons (Eurostat, 1998).

Module E of the PNS contains the question:

What is the gross monthly income usually received for this work?

Based on the recommendation from Eurostat, this answer to this question by the members of the child's household could be used to determine whether or not the child is classified as living in poverty. However, information from Module E (among others) is missing from the dataset meaning that this method is no longer an option.

In a paper written for the Joseph Rowntree Foundation in 2014, the following broad definition of poverty is used:

"When a person's resources (mainly their material resources) are not sufficient to meet their minimum needs (including social participation)" (Goulden and D'arcy, 2014).

Based on this definition of poverty, a person is said to be in poverty if the resources that they have are not sufficient to meet their basic needs. Module A of the PNS is concerned with household demographics and includes questions regarding the material of the walls and roof and the availability of running water to the household. Some of the items covered in this module of the questionnaire may be useful as a proxy to determine the socio-economic position of the household that the children live in.

2.4.2.1 Combining of levels in Module A

Since in the PNS there is no information available regarding income and as such it is difficult to use the standard measure of poverty, other variables must be looked at. There are certain variables within Module A of the PNS that can be recoded to give a better insight into whether a person in Brazil is perceived to be living in poverty or not.

If the household in which a person is living is constructed with unsuitable materials then perhaps this could be seen as a sign of poverty. From a book published by IBGE, it is said that certain household characteristics can be classed as "adequate" or "inadequate". In particular, the material of the walls, roof and floor along with water supply, outlet of bathroom, waste collection and origin of electricity can all be defined in this way (Fundação Instituto Brasileiro de Geografia and Estatística. Departamento de População and Indicadores Sociais, 1998). Using the recoded variables with combined levels instead of the original categorical variables will also aim to provide a model that is more interpretable without losing too much information. Table 2.2 shows how the original answers to these questions can be adapted into either "adequate" or "inadequate".

Furthermore, it is asked whether the home contains a variety of items. These items include: stove, TV, refrigerator, video/DVD player, washing machine, telephone, microwave, car, computer and internet.

Based on recommendations from the same book published by IBGE, these goods can be grouped into basic goods and status goods. It is recommended that basic goods include stove, TV and refrigerator and that status goods include video/DVD player, washing machine, telephone, microwave and car.

The two questions about whether or not a home has a computer and whether or not the home has internet has also been grouped into one variable with responses: has computer and internet, has computer with no internet, no computer has internet and no computer or internet.

Table 2.2: The renaming of certain household variables in Module A of the PNS

Material of Walls	Coated masonry	Adequate
	Uncoated masonary	Adequate
	Suitable wood for construction	Adequate
	Uncoated taipa	Inadequate
	Seized wood	Inadequate
	Straw	Inadequate
	Other	Inadequate
Material of Roof	Roof tile	Adequate
	Concrete slab	Adequate
	Suitable wood for construction	Adequate
	Metal sheet	Adequate
	Seized wood	Adequate
	Straw	Inadequate
	Other	Inadequate
Material of Floor	Carpet	Adequate
	Ceramic tile or stone	Adequate
	Suitable wood for construction	Adequate
	Cement	Adequate
	Seized wood	Inadequate
	Dirt	Inadequate
	Other	Inadequate
Water Supply	General distribution	Adequate
	Well or spring on the property	Inadequate
	Well or spring off the property	Inadequate
	Car tanker	Inadequate
	Water stored in rain tanks	Inadequate
	Rain water stored otherwise	Inadequate
	Rivers, lakes or streams	Inadequate
	Other	Inadequate
Outlet of Bathroom	General sewage	Adequate
	Septic tank	Adequate
	Fossa rudimentary	Inadequate
	Ditch	Inadequate
	Straight to river, lake or stream	Inadequate
	Other	Inadequate
Waste Collection	Collected directly by housekeeping	Adequate
	Collected in housekeeping bucket	Adequate
	Burned on the property	Inadequate
	Buried on the property	Inadequate
	Thrown onto wasteland	Inadequate
	Thrown into river, lake or sea	Inadequate
	Other	Inadequate
Electricity	General network	Adequate
	Other source (generator etc)	Inadequate
	No electricity	Inadequate

2.4.2.2 Module A and Intellectual Disability

After combining the levels of some of the responses to the questions in Module A, some exploratory analysis can be conducted in order to get a basic idea of the relationship between these and the response variable. Table 2.3 shows the summary of the results of the exploratory analysis conducted. Sampling weights were not used when conducting the exploratory analysis.

It can be seen that when fitted alone with intellectual disability very few of the

variables from Module A are found to be significant at the 5% level.

The type of home that a child lives in is found to be significant at the 10% level with the odds of a child having an intellectual disability reducing if they live in an apartment or lodging compared to a house.

If the material of the walls, roof and floor are inadequate, then the odds of a child having an intellectual disability are increased compared to if the materials are adequate. However, only the material of the walls is found to be significant at the 5% level with the material of the roof found to be significant at the 10% level and the floor at the 15% level.

If a home does not have a kitchen, the odds of a child having an intellectual disability are found to increase compared to when a home does have a kitchen. The remaining variables are found not to be significant at the 5% or 10% level.

From this initial analysis it appears that there may be evidence of a relationship between intellectual disability and some poverty variables. How adequately a house is constructed appears to have the strongest relationship with the odds that a child has an intellectual disability. There appears to be little to no relationship between a families access to both basic goods and status goods with whether or not a child has an intellectual disability.

2.4.3 Exploratory Analysis of the remaining modules

The remaining modules of the PNS focus on topics such as health care, education and other disabilities.

Table 2.4 summarises the results of the exploratory analysis conducted on these modules. The percentage of the sample belonging to each level of a variable has been calculated and then this percentage has been split across whether a child has an intellectual disability or not. Then each of the variables has been included as the only explanatory variable in a logistic regression model with intellectual disability as the response.

It can be seen that many of the variables are found to be significant at the 5% level. If a child is female the odds of them having an intellectual disability are lower than if a child is male. This is a common finding among studies into intellectual disabilities (Lai et al., 2012). The race of a child was not found to be significant.

If a child is unable to read and write, the odds of that child having an intellectual

Table 2.3: The percentage of the population in each of the newly grouped levels in Module A, the percentage split across whether the child has an intellectual disability or not and the results of a chi-squared test and a logistic regression model including each variable separately with intellectual disability as the response variable.

			Int. Dis		Chi-Sq		Log. Reg			
		%	No	Yes	p-value		Est.	Odds Ratio	p-value	
Type of Home	House	92.61	91.72	0.89	0.01608	*		1		
	Apartment	7.01	6.99	0.03			-0.81	0.45	0.06	.
	Lodging	0.38	0.38	0			-1.64	0.19	0.10	.
Walls	Adequate	97.03	96.42	0.87	0.0008916	**		1		
	Inadequate	2.70	2.65	0.05			0.67	1.95	0.03	*
Roof	Adequate	98.39	97.50	0.89	0.79			1		
	Inadequate	1.61	1.58	0.03			0.62	1.86	0.10	.
Floor	Adequate	96.80	95.92	0.88	0.04	*		1		
	Inadequate	3.20	3.16	0.04			0.42	1.52	0.15	
Water Supply	Primary supply adequate	79.96	79.21	0.75	0.25			1		
	Primary supply inadequate - adequate secondary supply	0.76	0.75	0.01			0.44	1.56	0.49	
	Primary supply inadequate - inadequate secondary supply	2.99	2.99	0.01			-1.31	0.27	0.01	**
	Primary supply inadequate - no secondary supply	16.29	16.13	0.16			0.03	1.03	0.90	
Running water	Yes	90.30	89.50	0.80	0.96			1		
	No	9.70	9.58	0.12			0.30	1.35	0.22	
Drinking water	Filtered	45.40	44.99	0.41	0.41			1		
	Boiled	1.56	1.53	0.02			0.45	1.57	0.39	
	Treated differently at home	2.19	2.17	0.03			0.36	1.43	0.46	
	Industry mineralised	16.76	16.60	0.16			0.05	1.05	0.82	
	Without treatment at home	34.09	33.79	0.30			-0.03	0.97	0.87	
Kitchen	Yes	97.77	96.88	0.89	0	***		1		
	No	2.23	2.20	0.03			0.55	1.74	0.04	*
Sewage	Adequate	69.68	69.04	0.64	0.84			1		
	Inadequate	30.32	30.04	0.28			0	1	0.98	
Waste	Adequate	85.13	84.33	0.80	0.45			1		
	Inadequate	14.87	14.75	0.12			-0.17	0.84	0.46	
Electricity	Adequate	98.55	97.63	0.91	0.03	*		1		
	Inadequate	1.45	1.45	0.01			-0.87	0.42	0.16	
Basic Goods	One or less	1.17	1.16	0.01	0.38			1		
	Two	3.66	3.63	0.03			0.17	1.19	0.81	
	Three	95.17	94.29	0.88			0.23	1.25	0.68	
Status Goods	None	10.94	10.83	0.11	0.09	.		1		
	One	24.45	24.16	0.30			0.20	1.22	0.36	
	Two	17.60	17.48	0.12			-0.39	0.68	0.11	
	Three	15.81	15.72	0.09			-0.51	0.60	0.04	*
	Four	15.85	15.73	0.12			-0.23	0.79	0.48	
	Five	15.35	15.17	0.18			0.16	1.18	0.61	
Computer and Internet	Computer with internet	42.44	42.14	0.30	0.01	*		1		
	Computer without internet	7.67	7.62	0.05			0	1.005	0.99	
	Internet with no computer	2.83	2.81	0.02			-0.02	0.98	0.95	
	No computer, no internet	47.05	46.51	0.54			0.49	1.64	0.02	*
Domestic Employee	Yes	5.27	5.19	0.07	0.36			1		
	No	94.73	93.89	0.85			-0.42	0.66	0.31	
Pet	Yes	64.52	63.93	0.58				1		
	No	35.48	35.15	0.34	0.93		0.04	1.04	0.81	

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 " 1

disability was found to be significantly higher than for a child who is able to read and write. If a child is educated, the odds of intellectual disability are reduced compared to an uneducated child.

When looking at the other forms of disabilities, it was found that if a child has a physical, hearing or visual disability the odds of them having an intellectual disability are increased. It was previously noted in Table 2.1 that there are very few children who have another disability in addition to an intellectual disability.

The health status of a child was also found to significantly affect the odds of a child having an intellectual disability. If a child was reported as having average health the odds of them having an intellectual disability are increased compared to a child with above average health. The odds are increased further if a child is reported to have below average health.

If a child's health was found to limit daily activities the odds of intellectual disability are increased. If a child has been diagnosed with a chronic illness the odds are also increased.

When the length of time that a child last saw a doctor increases, the odds of that child having an intellectual disability is decreased. Also if no health care was sought for the child in the previous two weeks to when the survey was conducted, the odds of intellectual disability are reduced compared to a child who did seek health care.

If a child was hospitalised or received emergency care at home in the last 12 months the odds of the child having an intellectual disability are greater than for a child who was not.

From this analysis, it appears that there are relationships between many of the health variables and intellectual disability. There is evidence that there is a relationship between sex and intellectual disability however, there is no evidence of a relationship between race and intellectual disability. There is also evidence of a relationship between level of education and intellectual disability.

2.5 Issues to consider during analysis

When it comes to analysing the data from the PNS with regards to intellectual disability there are a number of issues which need to be considered.

Table 2.4: Exploratory analysis conducted on the variables from the remaining modules of the PNS

			Int. Dis		Chi-Sq		Log. Reg.			
		%	No	Yes	p-value		Est.	Odds Ratio	p-value	
Visits from family	Monthly	28.06	27.79	0.27	0.67			1		
health team	Every two months	7.32	7.28	0.04			-0.49	0.61	0.13	
	From 2 to 4 times	8.22	8.17	0.05			-0.49	0.61	0.14	
	Once	6.92	6.84	0.08			0.21	1.23	0.48	
	Never received	10.04	9.97	0.07			-0.41	0.66	0.19	
	Not registered, don't know	39.44	39.04	0.40			0.05	1.05	0.84	
Visits from	Monthly	20.29	20.04	0.24	0	***		1		
endemic agent	Every two months	13.84	13.7	0.14			-0.15	0.86	0.61	
	From 2 to 4 times	18.34	18.13	0.20			-0.08	0.92	0.78	
	Once	18.31	18.15	0.16			-0.34	0.71	0.19	
	Never received	29.2	29.05	0.17			-0.70	0.50	0.01	*
Sex	Male	50.76	50.20	0.57	0	***		1		
	Female	49.24	48.88	0.35			-0.44	0.64	0.01	*
Race	White	40.58	40.22	0.36	0.52			1		
	Non-white	59.42	58.86	0.56			0.08	1.09	0.65	
Can read and	Yes	88.17	87.86	0.31	0	***		1		
write	No	11.83	11.22	0.61			2.72	15.2	0	***
Level of	Uneducated	27.40	26.88	0.52	0	***		1		
education	Completed elementary school	54.43	54.05	0.38			-1.02	0.36	0	***
	Completed high school	17.06	17.04	0.02			-2.61	0.07	0	***
	Graduated/ Masters	1.11	1.11	0			-2.38	0.09	0	***
Physical	Yes	0.53	0.40	0.13	0	***		1		
disability	No	99.47	98.68	0.79			-3.71	0.02	0	***
Hearing	Yes	0.45	0.40	0.05	0	***		1		
disability	No	99.55	98.68	0.87			-2.63	0.07	0	***
Visual	Yes	1.47	1.41	0.06	0	***		1		
disability	No	98.53	97.67	0.86			-1.54	0.21	0	***
Health status	Above average	88.78	88.30	0.49	0	***		1		
	Average	10.22	9.89	0.33			1.79	6.01	0	***
	Below average	1	0.89	0.11			3.10	22.12	0	***
Health limited	Yes	5.22	5.11	0.11	0	***		1		
activities	No	94.78	93.97	0.81			-0.89	0.41	0	***
Chronic Illness	Yes	4.87	4.36	0.51	0	***		1		
	No	95.13	94.72	0.41			-3.29	0.04	0	***
Usually sees the	Yes	79.66	78.85	0.81	0.01	**		1		
same doctor	No	20.34	20.23	0.11			-0.65	0.52	0	***
Last consulted	In the last 12 months	64.28	63.52	0.76	0	***		1		
doctor	Between 1 and 2 years ago	20.50	20.39	0.11			-0.77	0.46	0.03	*
	More than 2 years ago	13.58	13.54	0.04			-1.39	0.25	0	***
	Never	1.63	1.63	0			-2.88	0.06	0.11	
Last consulted	In the last 12 months	52.26	51.82	0.44	0.05	*		1		
dentist	Between 1 and 2 years ago	18.27	18.12	0.15			-0.03	0.97	0.91	
	More than 2 years ago	13.86	13.7	0.16			0.31	1.36	0.2	
	Never	15.61	15.44	0.17			0.28	1.32	0.21	
Sought health care	Yes	9.78	9.57	0.21	0	***		1		
in last 2 weeks	No	90.22	89.51	0.71			-1.02	0.36	0	***
Hospitalised in the	Yes	3.22	3.10	0.12	0	***		1		
last 12 months	No	96.78	95.98	0.80			-1.52	0.22	0	***
Emergency care at	Yes	0.68	0.65	0.03	0	***		1		
home	No	99.32	98.43	0.89			-1.67	0.19	0	***
Tried alternative	Yes	2.76	2.7	0.06	0.20			1		
treatment	No	97.24	96.38	0.86			-0.85	0.43	0.05	*

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 " 1

Firstly, when making a comparison between Brazil and the UK, similar variables in each data set will need to be used. When it comes to the variables in module A of the PNS it may be difficult to find any corresponding variables in the UK dataset since subjects such as the material of the walls aren't commonly asked in UK surveys.

As described above, the PNS has a complex sampling structure and therefore each observation in the data set has a sampling weight. When analysing the relationship between intellectual disability and the variables in the data, it will need to be determined whether or not these sampling weights are required and, if so, how they should be used.

Finally, despite some of the variables being combined to create new ones, there is still a relatively large number of variables to consider when looking at the relationship that they have with intellectual disability. Therefore, some sort of variable selection will need to be considered.

Chapter 3

Data - UK (The Millennium Cohort Study (MCS))

This chapter will discuss the dataset that will be used for the UK based analysis. First, the aims of the MCS will be discussed. Next the sampling scheme and calculation of the sampling weights will be described. Finally, exploratory analysis will be conducted to examine the response variable and any variables that correspond to the variables in the Brazilian dataset. Issues to be considered in further analysis is also discussed.

3.1 Aims of the Survey

When designing the surveys to be used in the MCS, the following five principles were strongly considered:

- The study should provide data regarding children living in all four countries of the UK (England, Wales, Scotland and Northern Ireland).
- The study should provide adequate data for specific sub-groups of children. These subgroups include: children living in disadvantaged circumstances, children of ethnic minorities and children living in the three smaller countries of the UK.
- The study should provide data not only on the child but also on the child's family circumstances and the environment in which they grow up, including socio-economic factors.
- The children in the study should be born within a single 12 month period.

- Any child born within the selected time period of the study should have a known, non-zero probability of being selected to be in the sample (Joshi et al., 2002).

The ESRC previously funded two other cohort studies: The National Child Development Study in 1958 and the British Cohort Study in 1970. When a proposal was made to the ESRC for the MSC it was stated that it should differ from the previous studies in the following ways:

- Instead of a one week time period, the sample should include a sample of children born in a 12 month time period, to include births in all seasons.
- The birth dates in the sample should include children born in the year 2000.
- The sample should include children from the whole of the UK.
- Data collected should focus strongly on the families social and economic circumstances.
- 15,000 should be the target sample size.
- The length of the interview will be controlled by a budget of £1.7 million.
- A sample design which over-samples ethnic minorities should be considered.
- The first of the surveys should be carried out when the children are all around 6 months old (Hansen, 2012).

The aims of each of the surveys differ slightly from each other since the information to be captured in each wave of the study changed to be more relevant to the age of the child at the time of the survey.

3.1.1 First Survey - 9 months old

Originally it was planned that the first survey would be conducted when the child was aged 6 months. However, this target was unattainable and instead it was chosen to postpone the survey until the child was 9 months old. In total 18, 553 families took part in the first survey of the MCS.

In the first survey, information was collected regarding the circumstances of both the mother's pregnancy and the birth of the child. Information was also collected on the social and economic background of the families.

The objectives of the first survey were proposed in March 2000 by the Centre for Longitudinal Studies and included:

- To obtain an insight to the initial health, social and economic advantages and disadvantages that children born in the new millennium will face.
- To obtain a baseline for comparison of data collected in further studies.
- To collect information on topics which have not been covered by previous studies such as the father's participation in the care and development of the child.
- To focus strongly on the mother, father and siblings of the child, recording how they have adapted to the newcomer in the family.

3.1.2 Second Survey - 3 years old

The second survey was conducted when the children were 3 years old. The data collected in this survey allowed any changes in circumstances since the first survey to become apparent.

In addition to the 18,553 families who took part in the first survey, a further 1,389 families were contacted to participate in the second survey.

Objectives of the second survey included:

- To measure the physical, cognitive, social and emotional development of the child.
- To look at changes in the family since the child was 9 months old.
- To collect information on any older siblings that the child has.
- To compile data which is comparable to previous cohort studies in the UK and other studies from outside of the UK.
- To collect data from families who had moved into survey areas after the original sample had been decided.

3.1.3 Third Survey - 5 years old

When the children were 5 years old, the third survey was conducted. The sample for the third survey was made up of any family who had responded to at least one of the previous surveys. This survey had the following objectives:

- To continue to measure the physical, cognitive and behavioural development of the child.
- To assess the child's experiences of starting primary school.
- To continue to collect data on the siblings of the child.
- To contact any families who have responded to a previous survey (regardless of whether they have responded to all earlier sweeps).

3.1.4 Fourth Survey - 7 years old

The fourth survey was conducted when the children in the sample were 7 years old and had similar objectives to the third survey with the addition of:

- To directly ask the children about their thoughts and experiences.

In total, the number of families eligible to participate in the fourth survey of the MCS was 19,244. However due to exclusion due to ineligibility, refusal, untraceability or sensitive family circumstances, the survey was issued to 17,031 families.

3.1.5 Fifth Survey - 11 years old

The fifth survey took place when the children were aged 11. There were 19,244 potential families eligible to take part in this survey however due to death, emigration, refusal or untraceability, the total number of families who responded to this sweep of the MCS was 13,287. Therefore, information was collected for 13,469 cohort members in total.

The respondents for this survey were the cohort members, their parents and their teachers. The aims were mostly similar to the surveys conducted in previous years with the addition of:

- Collect information regarding puberty
- Collect information about the child's attitude towards smoking, drinking and anti-social behaviours
- Collect information about the child's final year of primary school.

Since this survey was conducted between January 2012 and February 2013, when making comparisons with analysis of the PNS, data from the fifth survey of the MCS will be used.

3.1.6 Sixth Survey - Aged 14

The sixth survey of the MCS was conducted in 2015 when the children were 14 years old. In total 11,726 families took part in the survey and data was collected for 11,872 children.

This survey was the first conducted whilst the cohort members were teenagers and hence some more age-appropriate questions such as those regarding alcohol, drugs, puberty and romantic relationships were added to the cohort member self-completion questionnaire.

A further addition to this survey was activity monitoring. All cohort members in Scotland, Wales and Northern Ireland along with 81% of English cohort members wore a wrist activity monitor for two days along with self-completing an activity diary.

Saliva samples for DNA extraction and genotyping were collected from all cohort members as well as their natural parents.

3.1.7 Seventh Survey - Aged 17

The seventh survey of the MCS was conducted in 2019 when the cohort members were 17 years old. In total 10,625 families took part in the survey and data was collected for 10,757 cohort members.

In previous surveys a major focus of many of the questionnaires was schooling. Since at age 17 many of the cohort members will have made major decisions in relation to education and employment, a principle aim of this survey was to collect data regarding this transition.

This survey aimed to build a picture of the daily life of the cohort members with regards to: relationships with parents; family and peers; risky behaviours; social media engagement and effort on activities such as education/school.

3.2 Sampling scheme for the MCS

The population of interest for the MCS is any child born between September 2000 and January 2002 who at the age of nine months was living in the UK and eligible to receive Child Benefit. This population includes children living in non-household circumstances such as hostels at nine months as well as children who were born outside of the UK but

were UK residents at the age of 9 months.

Children in the study were found using the Child Benefit records. In England and Wales children born between 1 September 2000 and 31 August 2001 were sampled. In Scotland and Northern Ireland children born between 23 November 2000 and 11 January 2002 were sampled. The initial time period of 12 months was extended in Scotland and Northern Ireland due to a shortage in numbers which was noticed during fieldwork.

Any child born within the correct time period and living in one of the roughly 400 electoral wards of the UK at 9 months old were eligible for the study. Children who died before the age of 9 months, who emigrated out of the UK before the age of 9 months or who were not residents of the UK at the age of 9 months were excluded from the sample (Cullis, 2007).

Although the sampling technique aimed to give an accurate representation of the whole population, particular sub-groups were over-sampled intentionally. This was done to ensure that there was accurate representation of: children living in the smaller countries of the UK (Scotland, Wales and Northern Ireland); children living in disadvantaged areas and areas in England with a large ethnic minority population (30% or more of the population in 1991 was Black or Asian). The Child Poverty Index, defined as the percentage of children under 16 in an electoral ward living in families receiving at least one type of means tested benefit, was used during the stratification. The poorest 25% of wards based on the Child Poverty Index were over-sampled (Hansen, 2012).

3.2.1 Stratification of the population

The population was stratified by country (England, Wales, Scotland and Northern Ireland) then, within these strata, further stratification was conducted.

In England the population was stratified into three strata. The first of these strata was "ethnic minority" which included wards in which at least 30% of the total population was either black or Asian. The next stratum was "disadvantaged" which included wards which fell into the poorest 25% of wards based on the Child Poverty Index. The final stratum was "advantaged" which included wards which fell into the wealthiest 25% of wards based on the Child Poverty Index. If a ward was included in the "ethnic minority" stratum it was excluded from the remaining two strata.

In Wales, Scotland and Northern Ireland, the population was stratified into two

strata: "disadvantaged" and "advantaged". The criteria for inclusion in these strata were equivalent to the strata in England.

Since the wards in the UK vary greatly in size, particularly in England, in some instances multiple wards were combined to create "superwards" which had a minimum of 24 expected births in a year. In order to be combined, wards had to be bordering each other in the same district and be within the same stratum. If this was not possible then non-bordering within the same district were combined. Furthermore if this was not possible then non-bordering wards in different districts could be combined. Under no circumstances were wards in different strata combined.

3.2.2 Sample size

The target sample size of the MCS was 15,000. Sampling based on expected number of births would have resulted in the sample sizes in Wales, Scotland and Northern Ireland being too small for meaningful analysis. Instead, 1,500 children were chosen to be sampled from these three countries meaning that 10,500 children were to be sampled from England.

Initially, in England, half of the children were to be sampled from advantaged wards, a quarter from disadvantaged wards and the final quarter from ethnic wards. In Wales, Scotland and Northern Ireland half of the children were to be sampled from advantaged wards and the other half from disadvantaged wards.

After additional resources were allocated to the study, further children could be sampled from each country. In England, 35 additional disadvantaged wards were selected. In Wales the target sample size was doubled with the additional 1500 children to be sampled from disadvantaged wards. In Scotland, the target sample size was increased to 2,500 with 500 of the additional children to be sampled from advantaged wards and the remaining 500 to be sampled from disadvantaged wards. Finally in Northern Ireland the target sample size was increased to 2,000 with the additional children to be selected from disadvantaged wards. The final target sample size was 20,646 children (Plewis et al., 2007).

3.2.3 Obtaining the sample

The population in England was ordered by the standard regions (South East, London, North West, East of England, West Midlands, South West, Yorkshire and the Humber, East Midlands and North East) and then within these regions ordered by ward size (largest to smallest). In Scotland, four regions (South, Central, North East and North West) were used and then similarly to England, the wards were listed in descending order. Wales and Northern Ireland were not divided into regions and instead were listed only by ward size.

Then, systematic sampling was used within each stratum and country to select the wards to be sampled. The sampling interval was established based on the ratio of the number of wards in the population to the required number of wards in the sample.

Once the wards were selected, a list of all eligible children living in the wards was collected. The list was created based on the Child Benefit register. A letter was sent out to all families with an eligible child in the selected wards inviting them to participate in the survey. In total, 18,553 families agreed to take part in the study.

3.2.4 Sampling weights

Since the probability of selection varies depending on whether the child comes from an advantaged or disadvantaged ward, sampling weights are given in order to adjust for this.

Since there was no sub-sampling within wards, the weights for each child in the same ward are equal. There are different weights provided depending on whether analysis is based on an individual country or based on the whole of the UK.

The basic weight for a ward was calculated as the inverse of the sampling fraction applied to each stratum. This weight was then scaled so that the mean sampling weight was equal to one.

The sampling weights for each stratum are given in Table 3.1. Both the weights for when analysis of a single country is to be conducted and for when analysis for the whole of the UK is to be conducted are given.

The weights were then further adjusted to account for any bias arising from non-responses. There are a variety of ways in which these biases can occur.

Table 3.1: The sampling weights for each stratum of the MCS

Country	Stratum	Weight for country analysis	Weight for UK analysis
England	Advantaged	1.32	2.00
	Disadvanaged	0.71	1.09
	Ethnic	0.24	0.37
Wales	Advantaged	1.77	0.62
	Disadvantaged	0.65	0.23
Scotland	Advantaged	1.23	0.93
	Disadvantaged	0.75	0.57
Northern Ireland	Advantaged	1.41	0.47
	Disadvantaged	0.76	0.25

The first of these ways is that there is an under-representation of families who have recently moved. Since it is thought that families who move around more have different characteristics to whose who don't move very often if at all it is important to account for this bias.

Next, there were some losses in the Child Benefit sample due to some families being excluded for numerous reasons which could lead them to be classed as a sensitive case. These reasons include: if there had been an infant death in the family in the past five years; the child had been taken into the care system and the family had already been selected for another survey. If the Child Benefit was to be paid directly into a bank account, then records may not always show a change of address and so non-response may also occur due to this.

Non-response also occurred when the survey was conducted in the field. It was found that families in ethnic wards in England and advantaged wards in Northern Ireland were less likely to respond. It was also found that if the claimant of Child Benefits had the title "miss" they were less likely to respond. If the Child Benefit was paid into a bank account and the age of the mother was over 33, then it was found that the family was more likely to respond to the survey.

Finally, there was an over-representation of children born in winter in Scotland and Northern Ireland due to the time period of eligible births being increased in these countries. A weight was given to the families with a child born between 24 November and 11 January in 2000 or 2001 to account for this.

A separate weight was created to adjust for each of these biases and similarly to the weights for the strata, they were standardised to have mean one in each case. The

four weights can then be multiplied together to give a total non-response weight for each observation. This weight was then combined with the stratum to give an overall sampling weight (Plewis et al., 2007).

3.3 Contents of the Survey

As mentioned previously, in order to make comparisons with the analysis of the PNS, data from the fifth survey of the MCS (MCS5) will be used. This part of the study was made up of a variety of questionnaires. The household demographic module was asked in the form of an interview and the respondent could be either the mother, father or guardian of the child.

The "main" parent survey was asked in the form of an interview and in most cases was answered by the mother or mother figure of the child. The contents of this interview were as follows:

- **Module FC:** Family content
- **Module ES:** Early education, schooling and childcare
- **Module AB:** Child and family activities and child behaviour
- **Module PA:** Parenting activities
- **Module CH:** Child Health
- **Module PH:** Parental health
- **Module EI:** Employment, education and income
- **Module HA:** Housing and local area
- **Module OM:** Other matters
- **Module OS:** Older siblings
- **Module Z:** Consents and contact information.

The "partner" survey was also asked in the form of an interview and in most cases was answered by the father or father figure of the child. The following modules were contained with this interview:

- **Module FC:** Family context
- **Module ES:** Early education, schooling and childcare
- **Module PA:** Parenting activities
- **Module PH:** Parental health
- **Module EI:** Employment, education and income
- **Module OM:** Other matters
- **Module Z:** Consents and contact information.

Both the mother and father of the child were asked to fill in a self-completed questionnaire. The contents of this survey included:

- **Module SC:** Self completion
  - Strengths and difficulties questionnaire (main respondent only)
  - Discipline (main respondent only)
  - Relationship with cohort member
  - Cohort member's pubertal development (main respondent only)
  - Attitudes, racial harassment and discrimination, anti-social behaviour and consumerism
  - Mental health
  - AUDIT (alcohol use disorders identification test)
  - Relationship with partner
  - Life satisfaction.

The child's height, weight and body fat were all measured and the children all completed some assessments to evaluate their cognitive development. These assessments were made up of the British Ability Scale (testing verbal similarities), the Cambridge Neuropsychological Test Automated Battery (CANTAB) spatial working memory task and the CANTAB gambling task (testing decision making).

In addition to the assessments, the children were asked to respond to a self-completed questionnaire to obtain an insight into their thoughts about a variety of topics. This questionnaire contained questions about the following topics:

- Activities outside school
- Internet and social networking
- Life satisfaction, happiness and self-esteem
- Friends and unsupervised time
- Pocket money, family financial position and materialism
- Anti-social behaviours
- Secondary school
- Attitudes
- Other children (including bullying)
- Risky behaviours (including smoking and alcohol)
- Mental health
- Future ambitions.

In England and Wales, the teacher of each of the children was also asked to self-complete a questionnaire. This questionnaire covered the following topics:

- Child's abilities and behaviour
- Suspension and truancy
- Cohort member's profile (including English as an additional language (EAL), Special educational needs (SEN), help and support, peers, bullying)
- Move to secondary school
- Future education
- Parents
- Class groupings and setting
- Child's class
- Teacher profile (Mostafa et al., 2014).

3.4 Exploratory Analysis

Similarly to in the previous chapter, some initial exploratory analysis of the MCS5 will be conducted. Unlike the PNS, the MCS only contains data regarding children and therefore the sample does not need to be cut down to accommodate for this.

In this section, the response variable will be looked into as well as the other available variables in the dataset. It will be determined whether or not there are any variables in the data which correspond to the variables in the PNS in order to make comparisons between the two countries.

3.4.1 The response variable

Once again, the response variable of interest is whether or not the child has an intellectual disability. In the MCS the term used for intellectual disability is special educational needs (SEN). There are a few questions which ask about whether or not a child has SEN within the survey.

In some previous studies into children with intellectual disability in the UK, a child's score in the cognitive tests have been used to determine whether or not the child has an intellectual disability. For example, using principle component analysis, the first component, based on all age standardised test scores from the cognitive tests, accounting for 63% of the score variance was extracted. A child was then classified as having an intellectual disability if they scored lower than two standard deviations below the mean of this first component (Emerson et al., 2016).

However, since the purpose of analysis in this project is to make an international comparison between Brazil and the UK, a different response variable will be used. In the PNS, no cognitive tests were carried out and hence the response variable was chosen to be the answer to a question in which the parent or guardian of the child was asked whether or not the child had an intellectual disability. Therefore, a combination of the questions asked to the main respondents of the MCS5 about whether or not a child has a SEN will be used in order to make the response variables as comparable as possible between the two countries.

In the teacher survey the question "Does the child have Special Educational Needs?" is asked. However, since the teacher survey was only conducted in England and Wales

this question does not provide information for all of the children in the study.

The interview asked to the main respondent contained the question "has the child's school told you that your child has special needs?". The frequency of responses to this question can be found in Table 3.2.

Table 3.2: The frequency of responses to the question "has the child's school told you that your child has special needs?" from the "main" interview.

	Frequency
Yes	1429
No	12013

In addition to this question, the main respondent was also asked "Does the child have a statement of SEN?". The frequency of responses to this question can be found in Table 3.3. In this case, the "not applicable" responses correspond to the children who answered no, not applicable or don't know to the previous question.

Table 3.3: The frequency of responses to the question "does the child have a statement of SEN?" from the "main" interview.

	Frequency
Not applicable	12013
Yes	629
No	747
Child is currently being assessed	53

Since the definition of intellectual disability in the UK is that a child has been identified as having a SEN within education services, the response to this question will be used as the response variable of interest. The second question was only asked to parents who responded yes to the first question and so the responses of "not applicable" can be recoded as "no". A child who is currently being assessed will be also classed as not having SEN since at the time of the survey they had not yet been given a statement of SEN.

After recoding some of the responses, the response variable is now binary with 12,813 children being recorded as not having SEN and 629 children being reported as having SEN.

3.4.2 Variables corresponding to the PNS

When trying to make an international comparison between Brazil and the UK, similar variables should be used in the analysis of each country.

Module A of the PNS concerned the household, with many of the questions asking about the type of materials parts of the house were constructed from. These variables have been used in order to determine whether the house a child is living in is adequate or inadequate and hence whether the child may be living in poverty. There are no variables of this nature in the MCS and so a different way to asses the living conditions of a child will be looked into. Variables in MCS5 directly relating to variables from Module A of the PNS are: number of rooms in the home, number of cars in the home, number of people in the home, whether the home has a computer and whether the home has internet. Similarly to the PNS, whether a house has a computer and internet was asked as two separate questions. For analysis they will be combined to allow direct comparison between the two countries.

Data is available regarding the general characteristics of each child which correspond to the variables in the PNS. Sex, age and race are all available. Since all of the children in the MCS were all born within a specified period of time, the ages of the children in the study only range between 10 and 12. This differs to the PNS since the range of ages of children in that study was chosen to be 5-18. In MCS5 race was available as a factor with six levels. However, to make it comparable to the PNS, the number of levels will be reduced to two.

In regards to education, a question was asked regarding the school year that a child was in. Since the range of ages in the MCS is small, unlike the PNS, there is only very little variation in the level of education that the children are currently at.

Similarly to the PNS, there are questions in the MCS regarding disabilities. There are variables available showing whether or not a child has a visual disability, a hearing disability or a physical disability. Also, there are questions asking about the general health status of the child, whether their health limits their daily activities and whether they have been diagnosed with a long term illness. Data is also available regarding whether a child has been hospitalised or consulted a dentist in the last twelve months.

3.4.3 Additional variables

When looking at the available variables in the PNS data, the majority of the variables found that could be related to whether or not a family is living in poverty concern the adequacy of the household. There are no variables of this nature in the MCS. Therefore

different ways to establish whether or not a family is living in poverty in the UK need to be investigated.

In the MCS5 data there is a variable regarding damp in the home which may show in part the living conditions of a family. The parent was asked to rate how much of a problem they have with damp in their home from no damp to great problem. If the response to this question was "great problem" it may suggest that the family live in inadequate conditions.

In MCS5, the parent was asked whether the family is receiving any payments from the following: jobseekers allowance, income support, sickness or disability, child benefit, tax credits, any type of family related benefit, housing benefit or any other state benefit. Since the ideas behind means tested benefits are to help to redistribute resources from richer people to poorer people, if a family is receiving a benefit of some sort it may suggest that the family is not as well off financially as others (Finn and Goodship, 2014). Since the sample of the MCS primarily came from the Child Benefit records and Child Benefits is not means tested, this type of benefit will not be included when analysing. A number of different types of benefits are considered. According to the UK government website, the benefits considered are received by individuals in the following circumstances:

- job seeker's allowance - when out of work and actively looking for a job,
- income support - when on a low income or have no income and have a low amount of money in savings,
- sickness support - when extra costs of living are incurred as a result of a long-term health problem or disability,
- tax credit - when on a low income,
- family benefits - when on maternity leave, when seeking child care or if in full time education with a child under 15,
- housing benefits - if unemployed or on low income.

The parent interview asks the question "How well would you say your family is managing financially?" with the options to reply: living comfortably, doing alright, just getting by, finding it quite difficult and finding it very difficult. Although this question

does not show quantitatively the financial situation of a family it does show how the family feel that they are coping. The broad definition of poverty is "When a person's resources (mainly their material resources) are not sufficient to meet their minimum need" (Goulden and D'arcy, 2014). Therefore, a negative response to this question may suggest that a family is living in poverty.

The more quantitative definition of poverty compares a families income with a poverty line which is usually a cut-off of 60% of the countries median income. If a families income falls below this cut-off then they are described as living in poverty (Eurostat, 1998). The MCS5 contains an indicator variable which identifies whether a family falls below this line or not.

3.4.4 Exploratory Analysis

Similar exploratory analysis was carried out for the variable described above as for the PNS. The percentage of the sample belonging to each level of a variable has been calculated and then this percentage has been split across whether a child has a SEN or not. Then, each of the variables has been included as the only explanatory variable in a logistic regression mode with SEN as the binary response variable. The results of this exploratory analysis can be found in Table 3.4. Sampling weights were not used when conducting the exploratory analysis.

It can be seen that many of the variables are found to be significant at the 5% level. When looking at the different types of benefits that a family might receive, all of them are found to be significant with the exception of job seekers allowance. If a family receives any of the significant benefits it can be seen that the odds of the child having a SEN is increased. The greatest increase can be seen with other state benefits and the lowest can be seen with tax credit.

If a parent feels like they are finding it very difficult to manage financially it can be seen that the odds of their child having a SEN is increased compared to if they feel like they are coping comfortably. The greater they feel like they are struggling, the greater the odds of their child having a SEN.

Whether or not the house that a child is living in has a problem with damp was not found to be significant at the 5% level.

If a families income falls below the poverty line, the odds of the child having an

intellectual disability are increased. This, along with the other variables which could be related with a child living in poverty shows that there may be a relationship between poverty and SEN in the UK.

Similarly to the results of the exploratory analysis of the PNS, it was found that the odds of a child having a SEN is reduced if the child is female compared to if the child is male. Also similarly to the Brazilian data, race was found not to be significant.

Similarly to the PNS, other disabilities are found to be significant. If a child has another disability then the odds of them having a SEN are increased. Also if a child has a long-term illness the odds that that they have a SEN are increased.

The general health status of a child was found to significantly affect the odds of a child having a SEN. The worse the general health status of the child is reported to be, the greater the odds of the child having a SEN. If it was reported that the childs health limits their daily activities then the odds of SEN are increased compared to if their health does not limit their daily activities. This along with the other health variables show that there is strong evidence of a possible relationship between a child's health and the odds that they have SEN.

3.5 Further Analysis

Once again, the MCS has a complex sampling structure and each of the observations has its own sampling weight. Therefore when analysing the data to determine whether or not there is a relationship between a child having a SEN and other variables in the data, it must first be decided whether or not these sampling weights need to be used and if so, how they should be used.

Also, despite being less than in the PNS, there is a relatively large number of variables to consider. A method of variable selection will need to be chosen in order to further analyse the data and produce a model showing the relationship between SEN and the variables from the MCS.

Table 3.4: Exploratory analysis conducted on the variables from the MCS.

Variable		SEN Percentage	No	Yes	Chi-Sq. p-value		Log. Reg Estimate	Odds ratio	p-value	
Job seekers	No	96.67	92.66	4.01	0.41			1.00		
	Yes	3.33	3.21	0.12			-0.15	0.86	0.61	
Income support	No	93.46	90.14	3.32	0.01	*		1.00		
	Yes	6.54	5.74	0.80			1.33	3.78	0.00	***
Sickness support	No	93.31	90.11	3.20	0.00	***		1.00		
	Yes	6.69	5.77	0.93			1.51	4.53	0.00	***
Tax credit	No	53.82	52.19	1.63	0.00	***		1.00		
	Yes	46.18	43.69	2.49			0.60	1.82	0.00	***
Family benefits	No	98.33	94.42	3.91	0.06	.		1.00		
	Yes	1.67	1.46	0.21			1.24	3.46	0.00	***
Housing benefits	No	84.38	81.42	2.96	0.00	***		1.00		
	Yes	15.62	14.46	1.16			0.79	2.20	0.00	***
Other benefits	No	98.77	94.96	3.91	0.01	*		1.00		
	Yes	1.23	0.92	0.31			2.13	8.41	0.00	***
Coping financially	Comfortably	21.17	20.61	0.55	0.04	*		1.00		
	Alright	34.34	33.05	1.29			0.37	1.45	0.03	*
	Getting by	30.08	28.67	1.40			0.60	1.82	0.00	***
	Some difficulty	10.29	9.73	0.56			0.76	2.14	0.00	***
	Very difficult	4.13	3.81	0.32			1.13	3.10	0.00	***
Damp	No damp	83.81	80.48	3.33	0.64			1.00		
	Small problem	6.93	6.63	0.31			0.11	1.12	0.58	
	Some problem	7.01	6.65	0.36			0.27	1.31	0.13	
	Large problem	2.24	2.12	0.12			0.33	1.39	0.23	
Sex	Male	50.25	47.35	2.90	0.00	***		1.00		
	Female	49.75	48.52	1.22			-0.89	0.41	0.00	***
Age	10	34.34	32.89	1.44	0.46			1.00		
	11	65.27	62.64	2.64			0.11	1.12	0.70	
	12	0.39	0.35	0.04			1.05	2.86	0.03	*
School year	Year 6	95.28	91.42	3.86	0.81			1.00		
	Different year	4.72	4.46	0.26			0.32	1.38	0.08	.

Table 3.5: Continuation of Table 3.5.

Variable		SEN Percentage	No	Yes	Chi-Sq. p-value		Log. Reg Estimate	Odds ratio	p-value	
Computer and Internet	Computer and Internet	95.48	91.66	3.82	0.99			1.00		
	Computer, no internet	1.51	1.42	0.09			0.37	1.45	0.27	
	No computer, internet	1.43	1.31	0.12			0.75	2.12	0.01	*
	No computer, no internet	1.59	1.49	0.10			0.48	1.62	0.12	
Health status	Excellent	61.07	59.28	1.79	0.01	**		1.00		
	Very good	27.13	25.97	1.16			0.39	1.48	0.00	***
	Good	8.78	8.05	0.72			1.08	2.94	0.00	***
	Fair	2.57	2.26	0.30			1.49	4.44	0.00	***
	Poor	0.45	0.31	0.14			2.72	15.18	0.00	***
Long term illness	No	85.90	84.07	1.83	0.00	***		1.00		
	Yes	14.10	11.81	2.29			2.19	0.11	0.00	***
Vision problems	No	98.97	95.13	3.83	0.00	***		1.00		
	Yes	1.03	0.74	0.29			2.27	9.68	0.00	***
Hearing problems	No	99.15	95.29	3.86	0.03	*		1.00		
	Yes	0.85	0.59	0.26			2.41	11.13	0.00	***
Mobility problems	No	98.33	94.82	3.52	0.00	***		1.00		
	Yes	1.67	1.06	0.61			2.73	15.33	0.00	***
Health limits activities	No	92.37	90.34	2.04	0.00	***		1.00		
	Yes	7.63	5.54	2.09			2.82	16.78	0.00	***
Hospitalised in last year	Not at all	33.40	32.24	1.16	0.25			1.00		
	Less than once a month	20.66	20.06	0.60			-0.17	0.84	0.27	
	Once or twice a month	25.09	24.09	1.00			0.14	1.15	0.30	
	Once or twice a week	8.54	8.12	0.42			0.37	1.45	0.05	*
	Several times a week	5.70	5.42	0.28			0.35	1.42	0.09	.
	Every/almost every day	6.61	5.94	0.67			1.14	3.13	0.00	***
Seen dentist in last year	No	92.81	89.15	3.66	0.30			1.00		
	Yes	7.19	6.72	0.46			0.52	1.68	0.00	***
Race	White	88.25	84.57	3.68	0.65			1.00		
	Non-white	11.75	11.31	0.44			-0.11	0.90	0.45	
Below poverty line	No	85.98	82.79	3.19				1.00		
	Yes	14.02	13.09	0.93	0.01	**	0.62	1.86	0.00	***

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 " 1

Chapter 4

Methodology and Analysis - Sampling Weights

This chapter will aim to answer the research question: When is it appropriate to include sampling weights and how should they be used when necessary?

It will begin by discussing how sampling weights are calculated for complex survey designs. Next a review of the current practices used to analyse data from a complex survey design, specifically how it is advised to incorporate the sampling weights will be conducted. Then, simulations will be conducted with the aim to show the effect that using sampling weights has on coefficient estimates under various sampling schemes. This chapter will also investigate whether the F-test, used by DuMouchel and Duncan (2008) to determine the appropriateness of using sampling weights for linear regression, can be extended using the likelihood ratio test for a logistic regression model.

4.1 How are sampling weights calculated?

Once survey data has been collected it must be appropriately weighted before analysis. This process involves adding a new variable to the data for each respondent. The weight given to an individual shows the projection of this individual to the population. For example, if a person in the sample has a sample weight of 100, this person will represent 100 people in the population. The sample weights given in a data set will generally adjust for unequal sampling probabilities, non-responses, coverage errors and various other sampling biases (Bollen et al., 2016).

4.1.1 Base weights

Sampling weights are usually provided when the probability of selection is unequal across individuals in the population. In the simplest case, the sampling weight will be the inverse of the probability of selection. For example, if an individual, i, has probability π_i of being selected in the sample then the base weight given to this individual w_i can be calculated as $w_i = 1/\pi_i$.

In a simple random sample the weight given to every individual will be equal and can be calculated as N/n where N is the size of the population and n is the sample size. For a stratified sample, the weights across different strata will vary but the weight of each individual within the same strata will be constant. For example, for an individual in strata k the base weight can be calculated as N_k/n_k where N_k is the population size of the strata and n_k is the sample size for the strata.

If a survey has a multi-stage design then the base weight has to reflect the probability of selection across each stage of the design. For example, in a three-stage design in which firstly a primary sampling unit (PSU) is selected, secondly a household is selected and thirdly an individual within the household is selected then the overall probability of selection is calculated as the product of the probability of selection at each stage. The weight for the individual is then calculated as the inverse of the overall probability.

4.1.2 Adjusting for non-response

In a survey it is unlikely that a response will be obtained for every question from every sampled unit. If a household or individual fails to respond to any of the questions of the survey, this is known as unit or total non-response. If a household or individual fails to respond to some of the questions but does provide some data then this is known as item non-response (Yansaneh, 2003). It is important that the non-response rate is clearly reported in surveys along with the method in which this rate was calculated.

If non-response is by chance then it will not cause too many problems. It may result in larger confidence intervals as a result of the smaller sample size but it will not result in bias. However it is important to account for non-response whilst calculating sampling weights since there is likely to be systematic differences between those who respond and those who do not and if this is ignored any estimates produced from an analysis will

be biased. In order for any bias caused by non-response to be reduced, either the non-response rate needs to be small or that there needs to be minimal differences between those who respond and those who do not (De Leeuw et al., 2012).

The primary method in which item non-response bias is reduced is imputation. Imputation is a process which fills in the data missing due to non-response with plausible values in order to create a complete data set. This ensures that the original sample size is maintained (Durrant et al., 2005).

To reduce bias caused by unit or total non-response there are three standard procedures. The first of these procedures is to adjust the weights to compensate for non-responses. Secondly, when establishing the required sample size, a larger sample size than is actually required can be used with the extra units making up a "reserve" sample from which replacements for non-responses can be selected. Finally, substitution can be used. If no response is obtained from a household then it is replaced in the sample by another household which is similar with respect to the characteristic of interest. In general it is recommended that adjusting the weights is the method used to compensate for unit non-response (Yansaneh, 2003).

There are generally four stages to adjusting weights for non-response. Initially the base weights are calculated to account for unequal sampling probabilities. Next, the sample is partitioned into subgroups and the weighted response rates are computed for each subgroup. Then, the non-response adjustment is calculated as the inverse of the subgroup response rate. Finally the non-response adjusted weight is calculated as the product of the base weight and the non-response adjustment weight.

4.1.3 Adjusting for under-coverage

When a survey is designed there is a specific population of interest which is known as the target population. A sample frame is a list of members of the target population. A coverage error occurs when the sample frame and the target population do not match. Under-coverage occurs when not all members of the target population are included in the sample frame. Over-coverage occurs when one or more units from the target population appear more than once in the sample frame (De Leeuw et al., 2012).

Since the design of the majority of household surveys is multi-stage, coverage errors may occur at three levels. The first of these levels is the primary sampling unit (PSU)

level, which is generally a geographical area such as a state or a county. The second is at household level in which households from within the selected PSUs are sampled from a list of all available households. Lastly, is at the individual level in which a sample of one or more people is selected from a list of available residents.

Under-coverage at PSU level is usually minimal since in general the geographical areas which make up the PSUs cover the entire geographic extent of the target population. However in some instances during the design stage of a survey, some PSUs will be excluded from the sample frame due to inaccessibility as a result of civil unrest or natural disasters for example. An additional reason for a PSU to be excluded is cost. If a PSU is a remote area containing very few households, then it may be decided that it is too costly to cover the area considering it will only represent a very small amount of the target population. During the design stage it should be explicitly stated if any PSUs have been excluded from the sample frame.

At the household level the amount of under-coverage is generally larger and therefore a more serious concern. One reason is that in some circumstances it is difficult to define what is meant by a household or dwelling unit. For example, how is a multi-unit structure such as a block of apartments handled during sampling? If a house is unoccupied during the design stage but occupied at the time of data collection this may cause it to be missed from the sample frame. The way in which institutions, such as hospitals and prisons, are handled may also add confusion and increase under-coverage (Groves et al., 2011).

Generally, under-coverage at the household level is more common in surveys conducted in developing countries. This is due to the fact that most census data does not provide full details of sampling units at either the household or person level and commonly out of date listings are used to construct the sample frame (Yansaneh, 2003).

Under-coverage at the individual level may arise since generally sampling frames identify housing units but not people within each household. Residency rules should be established before a survey is conducted in order to create a correct residency list for each household.

A common residency rule is the de jure rule in which the list of residents is constructed from the people who "usually" reside within a household. There are a number of problems with this rule however as it may not work in all circumstances: some people

may have no usual address and some may have more than one; some households are complex and the list of "usual" residents may not be easily defined; there may be disagreements within a household about who resides there and who does not (Martin and De La Puente, 1993).

There are two ways in which under-coverage can be corrected. The first of these is reduction, improving the listings procedures used to create the sample frame in order to reduce under-coverage. The second is to adjust the sample weights in order to compensate for under-coverage.

One method of reduction can be used when it is known in advance that there are some ineligible units listed on the sampling frame for example unoccupied households. If there is an approximation to the prevalence of ineligible units within the sample frame, then the sample size can be increased in order to account for this and decrease under-coverage. A further method of reduction is to use multiple sampling frames. For example an older list of housing units can be enhanced by a list of newly constructed housing units (Groves et al., 2011).

If there are reliable controls available for the entire population as well as specified subgroups of the population an attempt can be made to adjust the weights of each sampling unit so that the sum of the weights equal the sum of the weights of the controls within the specified subgroups. This technique is known as post-stratification and adjusts for both non-response and under-coverage simultaneously (Yansaneh, 2003).

4.2 When should sampling weights be used?

Before analysing survey data it should be understood when and how sampling weights should be used. In some circumstances it is essential to include sampling weights during analysis, for example when the objective is to estimate a population mean. However, in other cases such as regression modelling, whether or not weights should be used is a much more complex question.

4.2.1 Descriptive statistics

When the aim is to summarise features of a data set descriptive statistics such as means may be used. When estimating descriptive statistics, it is generally accepted that sam-

pling weights should always be used (De Leeuw et al., 2012).

Including sampling weights allows population totals to be estimated and leads to means being representative of the target populations. Methods for computing descriptive statistics using sampling weights will be discussed later in the chapter.

4.2.2 Regression modelling

When the objective of data analysis is to determine whether or not there is a relationship between two or more variables, regression modelling may be used. As previously mentioned, whether or not sampling weights should be used when fitting a regression modelling is not as clear cut as it is when estimating descriptive statistics.

There are three situations in which weighting may be necessary during regression: calculating estimates which correct for heteroskedasticity (when the variability of the error terms is not constant); correcting for endogenous sampling and identifying partial effects in the presence of unmodeled heterogeneity effects (Solon et al., 2013).

When correcting for heteroskedastic error terms, the use of weights will aim to ensure a greater precision of coefficients in both linear and non-linear regression models. It is proposed that weighting is done based on group or strata population in order to correct for population-size-related heteroskedasticity in the group or strata error terms. However, it has been found that in many cases the use of weights in this way, makes estimates much less precise (Solon et al., 2013). This finding agrees with the discussion in a paper written by William T. Dickens. Dickens stated that individual error terms are likely to be correlated due to group specific error components so weighting in this way will not be appropriate (Dickens, 1990).

Solon et al. recommend that, in addition to reporting heteroskedasticity-robust standard error estimates, that it is good practice to report both weighted and unweighted estimates.

Endogenous sampling occurs when the selection probability is related to the response variable. If this is the case then an analysis which ignores the selection probabilities may lead to biases in the estimated regression parameters (Lohr, 2009). Weighting by the inverse probability of selection can reduce the bias present without the use of weights. Weighting is not necessary if the sampling probability varies across strata and the strata are included within the model since the error term should no longer be related to the

error term (Solon et al., 2013).

When identifying average partial effects in the presence of unmodeled heterogeneity effects, it is suggested that the heterogeneity should be investigated to determine how to best account for it rather than trying to average it out through the use of weights. One way in which to do this is to, if possible, use a fully saturated model (Solon et al., 2013).

Chromy and Abeyasekera express that when analysing data from a household survey with a complex sampling design weights should be used in order to adjust for unequal sampling probabilities and non-responses in addition to properly estimate the precision of any estimates. However, if weights are to be ignored, assumptions are required. One such assumption is that the design of the sample generated an equal probability sample. Assumptions such as this are most reasonable when the analysis conducted is the application of a regression model to study the relationship between a dependent variable and one or more independent explanatory variables (Chromy and Abeyasekera, 2005).

When considering a binary response variable, standard statistical methods are often inappropriate due to clustering and stratification in the sampling design. In particular, the chi-squared and likelihood ratio tests largely increase the type I error rate when there is strong intra-cluster correlation present. Therefore some adjustments to standard methods are required so that inferences made from survey data are valid (Roberts et al., 1987).

In many cases is it suggested that when it comes to fitting a regression model to survey data that two models are fitted, one with weights and one without.

Thomas Lumley states that it often makes little difference when drawing conclusions whether or not sampling weights are used when fitting regression models. If there is a large difference between the estimates when weights are used it is generally an indication that some particularly influential observations have large sampling weights. This in turn could mean that neither the weighted or unweighted model is reliable. Sensitivity analysis can be performed by removing the most influential observation and refitting the model. Lumley also states that if both the weighted and unweighted estimates are valid (the sampling weights are ignorable) then the weighted estimates will in fact be less precise (Lumley, 2010).

This is consistent with a suggestion made by Winship and Mare. They advise that when conducting regression analysis with data which has sampling weights two models should be estimated - one with unweighted data and the other with weighted data. If the parameter estimates from the two models are similar then the unweighted model is preferred since the estimated standard errors will be correct. If the results of the two models are different then it may indicate that the model is missing non-linear or interaction terms (Winship and Radbill, 1994).

Lohr also agrees, advising that both unweighted and weighted regression models should be fitted. So long as the model fitted is a good one, then the estimates obtained from fitting a regression without weights should be similar to those obtained with weights. If this is not the case then it may suggest that the proposed model does not fit well for some of the population. It is proposed that if the model fitted is a good one then the only difference between a weighted and unweighted model would be apparent in the intercept term (Lohr, 2009).

Chromy and Abeyasekera advise that after a non-survey statistical package has been used to fit a regression model and variable selection has been performed, a survey statistical package, which acknowledges the survey design, should also be used to fit the chosen model. This type of software can be used to fit logistic regression models based on survey data and obtain estimates of parameters and also the standard errors of these parameters. The estimates of the parameters based on the sample data will be estimates of what would be obtained from fitting the model to the entire finite population (Chromy and Abeyasekera, 2005).

4.2.3 Testing whether to use sampling weights

Since advice of whether or not to use sampling weights in regression analysis is largely vague, a test of the appropriateness of weighting would reduce confusion and make the decision more objective. Tests to determine whether or not weighting is required do exist but due to lack of awareness amongst researchers, certain traditions in different fields (either always weighting or never weighting) and the fact that some of these tests are not available in current R packages, these tests are rarely applied. Most of these tests fall into one of two categories: difference in coefficient and weight association tests (Bollen et al., 2016).

Difference in coefficient tests fit both weighted and unweighted regression models and compare the coefficients to determine whether or not the difference between them is significantly different from zero or not. Weight association tests fit one regression model containing both the unweighted covariates along with some transformed form of the covariates and examine the coefficients of the transformed covariates. Difference in coefficient tests give a direct comparison between weighted and unweighted estimates whereas weight association tests do not (Bollen et al., 2016).

4.2.3.1 Difference in coefficient tests

One way to test the differences between the coefficients of a weighted and an unweighted model was proposed by Hahs-Vaughn and Lomax. They suggest that both models are fitted and the confidence intervals of the coefficients are examined. If the confidence intervals for the two models overlap then they state that weighting makes no difference and analysis should continue ignoring the weights (Hahs-Vaughn and Lomax, 2006).

Pfefferman proposed that a test developed by Hausman to identify model misspecifications could be used to compare the coefficients of a weighted and an unweighted model (Pfeffermann, 1993). The underlying idea of the Hausman test is that if a model is specified correctly then two different consistent estimators of a parameter will, as the sample size increases, converge to the same value. If the model has been misspecified then the estimators will diverge.

Assumptions made by Hausman for this test are: both parameter estimates are consistent estimators of the true parameter; both parameter estimators have asymptotic normal distributions and also that the second of the two parameter estimates is asymptotically efficient (Bollen et al., 2016).

The null hypothesis in the Hausman test is that the model has been specified correctly. The Hausman test creates a chi-squared test statistic based on the differences between the coefficients of two models. The inverse of the covariance matrix of these differences is pre-multiplied and post-multiplied by the differences to form this statistic. This test statistic is then compared to a Chi-squared distribution with degrees of freedom equal to the number of coefficients in the model (Hausman, 1978).

Pfeffermann proposed that this test could be used to assess the need for the use of sampling weights in regression models. The coefficients of the weighted and unweighted

models are compared. The covariance matrix used in this test is equal to $[\hat{V}[\hat{\beta}_w] - \hat{V}[\hat{\beta}_u]]$ where $\hat{\beta}_w$ and $\hat{\beta}_u$ are the coefficient matrices of the weighted and unweighted models respectively. This matrix is pre and post multiplied by the vector of differences between the coefficients of the two models and this value is compared to a Chi-squared distribution.

Asparouhav and Muthen also use a Hausman test to compare the differences between weighted and unweighted models however they use a slightly modified covariance matrix to that as used by Pfeffermann. In this case, the matrix to be pre and post multiplied by the vector of differences is given by $[\hat{V}[\hat{\beta}_w] - \hat{V}[\hat{\beta}_u] - 2C]$ where C is defined to be the covariance matrix of the two estimators (Asparouhav and Muthen, 2007).

4.2.3.2 Weight Association Tests

Similarly to some of the difference in coefficient tests discussed above, many weight association tests have been derived from a Hausman test for misspecification. Hausman tests the significance of $\boldsymbol{\beta}_\mathbf{M}$ after fitting the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{X}_\mathbf{M}\boldsymbol{\beta}_\mathbf{M} + \boldsymbol{\epsilon}$, where $\mathbf{Y}$ is the vector of the response variable, $\mathbf{X}$ is the design matrix of the explanatory variables and $\mathbf{X}_\mathbf{M}$ is the design matrix $\mathbf{X}$ which has been transformed in some way. An F-test with the null hypothesis $H_0 : \boldsymbol{\beta}_{\boldsymbol{M}} = 0$ is then used to test whether or not the model has been correctly specified (Hausman, 1978).

Whereas Hausman used this method for testing for misspecification, DuMouchel and Duncan applied this test to determine whether or not weights should be used. They use ordinary least squares to fit $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_\mathbf{u} + \mathbf{X}_\mathbf{w}\boldsymbol{\beta}_\mathbf{w} + \boldsymbol{\epsilon}$ where $\mathbf{Y}$ and $\mathbf{X}$ are as defined above and $\mathbf{X}$ is transformed to $\mathbf{X}_\mathbf{w}$ by applying the sampling weights. Similarly to the Hausman test, an F-test with the null hypothesis $H_0 : \boldsymbol{\beta}_{\boldsymbol{w}} = 0$ is used to determine whether the coefficients of all of the weighted covariates are statistically significantly different to zero. If the null hypothesis is rejected then weights are required. If there is a failure to reject the null hypothesis then it is suggested that a weighted analysis is not necessary (DuMouchel and Duncan, 2008).

Fuller also adapted the Hausman test to determine the appropriateness of using sample weights during analysis. Originally Fuller proposed transforming the design matrix in the same way as DuMouchel and Duncan, however, later he proposed a further adaptation to the test. The model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_\mathbf{u} + \mathbf{W}\boldsymbol{\beta}_\mathbf{w} + \boldsymbol{\epsilon}$ where $\mathbf{W}$ is the vector of the

sample weights, is fitted and the null hypothesis $H_0 : \boldsymbol{\beta_w} = 0$ is once again tested with an F-test. Fuller also suggested that if only certain covariates are of interest, then a subset of $\mathbf{WX}$ can be used in place of $\mathbf{X_M}\boldsymbol{\beta_M}$ in the test originally proposed by Hausman (Fuller, 2009).

A weight association test which differs from the Hausman test was proposed by Pfeffermann and Sverchkov. Residuals from an Ordinary Least Squares regression of the dependent variable on the original covariates (ignoring weights) are calculated and then correlated with the sampling weights. The square and the cube of the residuals are also correlated with the weights. A Fisher's F-test or a bootstrap estimation can then be used to determine whether or not these correlations are zero and therefore whether or not weights are required (Pfeffermann and Sverchkov, 1999).

A further test proposed by Pfeffermann and Sverchkov again calculates residuals from an ordinary least squares regression of the dependent variables on the original covariates and then uses ordinary least squares regression of these residuals on the weight variable. A t-test is then performed to determine whether the coefficient of the weights is statistically significant to zero and hence whether or not the weights should be used (Pfeffermann and Sverchkov, 1999).

These tests have all been developed for assessing the use of weights in linear regression models. If the response is binary and a logistic regression model is used, an F-test can no longer be used. It will be useful to examine whether or not these tests can be further developed to account for binary responses. This will be done in the simulations section later in this chapter.

4.3 How should sampling weights be used?

Once it has been determined whether or not it is appropriate to use sampling weights within an analysis, the question of how best to incorporate the weights arises. When analysing survey data there are three different approaches which can be used: model-based, design-based and model assisted (Lehtonen and Pahkinen, 2004).

4.3.1 Model-based analysis - ignoring weights

In areas of statistics which are not concerned with survey data, generally analysis takes a model-based approach. This means that inference is based on a model which is assumed to describe the relationship between the explanatory variables and the response variable (Lohr, 2009).

This approach does not account for a complex survey design and instead assumes that the data is a result of simple random sampling from an infinite population and is independent and identically distributed.

For a single explanatory variable the linear model used, in the model-based setting, is of the form

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{4.1}$$

where Y_i denotes the response variable and x_i an explanatory variable. The error terms ϵ_i are assumed to satisfy the following conditions:

1. $E(\epsilon_i) = 0$ for all i.
2. $V(\epsilon_i) = \sigma^2$ for all i.
3. $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$.

These assumptions are essential in order to make inference about the true parameters β_0 and β_1 as well as predicted values of the response variable (Lohr, 2009).

If the observations truly follow the chosen model then the sample design should have no effect on the estimates so long as the only relationship between the inclusion probabilities and the response variable is through the explanatory variables (Lohr, 2009).

When the response of interest is continuous, linear regression is the most common model-based approach. In this case, the aim is to estimate $\boldsymbol{\beta}$ in the model $y = \boldsymbol{X}^T \boldsymbol{\beta}$ which is calculated by the ordinary least squares estimator:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}. \tag{4.2}$$

When the response variable is binary, logistic regression is used. The standard logistic regression model is of the form :

$$\log\left(\frac{p}{1-p}\right) = \boldsymbol{X}^T\boldsymbol{\beta} \tag{4.3}$$

where p denotes the probability that the response variable takes a value of 1.

The regression coefficients can be estimated using maximum likelihood estimation. Since it is not possible to explicitly define the values of the coefficients which maximise the likelihood function, an iterative process such as iteratively reweighted least squares (IRWLS) must be used.

4.3.2 Model-based analysis - including weights

If interest lies in the relationship between the mean of one (normally distributed) variable (the response variable Y) and one or more further variables (explanatory variables X) then a linear regression model may be used. The general form of such a model is

$$E[Y] = \beta_0 + \boldsymbol{X}\boldsymbol{\beta} \quad \text{where} \quad Var[Y] = \sigma^2. \tag{4.4}$$

This implies that the variance of the response variable is constant and despite having no influence on the interpretation of β_0 and $\boldsymbol{\beta}$ it does affect the precision of the estimates.

When using a regression model, generally the aim is to identify and estimate an underlying model which could have generated the data. One problem with using a design-based approach is that this is not done. However, results of design-based regression can be justified through the use of an extension of a standard linear model (Kott, 1991).

To obtain estimates for the parameters β_0 and $\boldsymbol{\beta}$ which account for a complex sampling structure, sampling-weighted least squares can be used. This means finding the values of the parameters which minimise the estimate of the population sum of squared residuals:

$$\widehat{RSS} = \sum_{i=1}^{n} \frac{1}{\pi_i}(Y_i - \beta_0 - X_i\beta)^2. \tag{4.5}$$

Solving this, the parameter estimates can be calculated using:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y} \tag{4.6}$$

where $\boldsymbol{X}$ is the $n \times p$ design matrix, $\boldsymbol{W}$ is the $n \times n$ matrix with $\frac{1}{\pi_i}$ on the diagonal

and $\boldsymbol{Y}$ is the vector of responses.

The variance of the parameter estimates is given by:

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}^{-1}). \tag{4.7}$$

4.3.3 Design-based analysis

When analysing complex survey samples, generally a design-based approach is used. This means that a population, whose data values are unknown but treated as fixed, is specified. Since the the sample design of the survey is under the control of the researcher, any sampling probabilities are known. Design-based methods of analysis are used to make estimates concerning the fixed, finite population and are not able to make generalisations for other populations (Lumley, 2010). A design-based approach to analysis acknowledges that the data is a sample from a finite population and gives a sampling weight to each observation. This differs from a model-based approach to analysis which treats the data as a simple random sample from an infinite population (Wheeler et al., 2008).

In order to use design-based methods, there are certain conditions which need to be met:

1. Each individual in the population must have a non-zero probability (π_i) of being selected as part of the sample.

2. This probability should be known for each individual in the sample.

3. Each pair of individuals in the population should have a non-zero probability ($\pi_{i,j}$) of both being included in the sample.

4. This probability should be known for each pair of individuals in the sample (Lumley, 2010).

4.3.3.1 Design-based methods of obtaining population estimates

If interest lies in the population total of a certain measure X, for example income, the Horvitz-Thompson estimator can be used. When the data is from a sample of size n and a sampling weight of $1/\pi_i$ indicates that individual i was selected with probability π_i, the Horvitz-Thompson estimator is given by:

$$\hat{T}_X = \sum_{i=1}^{n} \frac{1}{\pi_i} X_i = \sum_{i=1}^{n} \check{X}_i, \tag{4.8}$$

where $\check{X}_i = \frac{1}{\pi_i} X_i$.

From this, the mean can be estimated using

$$\hat{\mu}_X = \frac{1}{N} \sum_{i=1}^{n} \check{X}_i \tag{4.9}$$

where N is the population size.

For a simple random sample of size n all sampling weights are equal to N/n. This means that the estimate of the mean of X is given by the sample mean:

$$\hat{\mu}_X = \frac{1}{N} \sum_{i=1}^{n} \check{X}_i = \frac{1}{N} \sum_{i=1}^{n} \frac{N}{n} X_i = \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{4.10}$$

4.3.3.2 Design-based methods for investigating causal effects - linear regression

As before, if we are interested in identifying a linear relationship between a response variable and p explanatory variables, multiple linear regression can be used. Again, we wish to estimate $\boldsymbol{\beta}$ in the model $y = \mathbf{x}^T \boldsymbol{\beta}$.

Using a design-based approach this estimate can be found using the following formula:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i \in S} w_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \sum_{i \in S} w_i \mathbf{x}_i y_i \tag{4.11}$$

where S is the sample, w_i is the sampling weight for observation i $\mathbf{x}$ is the $n \times p$ design matrix and y is the vector of responses.

An estimator for the variance of $\hat{\boldsymbol{\beta}}$ is then given by

$$\hat{V}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i \in S} w_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \hat{V} \left(\sum_{i \in S} w_i \mathbf{q}_i \right) \left(\sum_{i \in S} w_i \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \tag{4.12}$$

with

$$\mathbf{q}_i = \mathbf{x}_i (y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}). \tag{4.13}$$

and where $\hat{V}\left(\sum_{i \in S} w_i \mathbf{q}_i \right)$ is an estimate for the variance of a total based on obser-

vations $\{w_i \mathbf{q}_i\}$.

For a simple random sample:

$$\hat{V}\left(\sum_{i\in S} w_i \mathbf{q}_i\right) = (1 - \frac{n}{N})N^2 \frac{s_q^2}{n} \tag{4.14}$$

where

$$s_q^2 = \frac{1}{n-1}\sum_{j\in S}(q_j - \bar{q})^2. \tag{4.15}$$

For a stratified sample:

$$\hat{V}\left(\sum_{i\in S} w_i \mathbf{q}_i\right) = \sum_{h=1}^{H}(1 - \frac{n_h}{N_h})N_h^2 \frac{S_h^2}{n_h} \tag{4.16}$$

where H is the number of strata and

$$s_h^2 = \frac{1}{n-1}\sum_{j\in S_j}(q_{hj} - \bar{q}_h)^2. \tag{4.17}$$

An alternative way of expressing these equations is given by Sarndel et al. In the context of a finite population, $\boldsymbol{\beta}$ can be estimated using the following:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} X_i X_i^T\right)^{-1} \sum_{i=1}^{n} X_i y_i \tag{4.18}$$

where $\mathbf{X}$ is a matrix of covariates and y is the response vector.

In a model-assisted setting, initially $\boldsymbol{\beta}$ is expressed as a vector of population totals and then estimated by substituting the appropriate weighted estimators for these totals.

More specifically, firstly for $j = 1, \ldots, p$, and $j' = 1, \ldots, p$, where p denotes the number of covariates, the population totals for $k \in S$, where S is the sample, can be expressed as

$$t_{jj'} = \sum_{j=1}^{n} X_{jk} X_{j'k} \quad \text{and} \quad t_{j0} = \sum_{j=1}^{n} X_{jk} y_k. \tag{4.19}$$

If we let $\mathbf{T} = \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T$ and $\mathbf{t} = \sum_{i=1}^{n} \mathbf{X}_i y_i$ Equation 4.18 can then be expressed as

$$\hat{\boldsymbol{\beta}} = \mathbf{T}^{-1}\mathbf{t} \tag{4.20}$$

which is a function of totals $t_{jj'}$ and t_{j0}. These totals can be approximated by their weighted estimators

$$\hat{t}_{jj'} = \sum_{j=1}^{n} \frac{X_{ik}X_{j'k}}{\pi_k} \quad \text{and} \quad \hat{t}_{j0} = \sum_{j=1}^{n} \frac{X_{jk}y_k}{\pi_k} \tag{4.21}$$

where π_k is the sampling weight for observation k.

Therefore $\boldsymbol{\beta}$ can now be estimated using:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{j=1}^{n} \frac{X_{ik}X_{j'k}}{\pi_k} \right)^{-1} \left(\sum_{j=1}^{n} \frac{X_{jk}y_k}{\pi_k} \right). \tag{4.22}$$

Once again, this estimate is not unbiased and, similarly to design-based analysis, the variance formula is an approximation (Särndal et al., 2003). The variance estimator is defined as:

$$\hat{V}(\hat{\boldsymbol{\beta}}) = \left(\sum_{k=1}^{n} \frac{X_k X_k^T}{\pi_k} \right)^{-1} \hat{\mathbf{V}} \left(\sum_{k=1}^{n} \frac{X_k X_k^T}{\pi_k} \right)^{-1} \tag{4.23}$$

where $\hat{\mathbf{V}}$ is a symmetric $p \times p$ matrix with elements

$$\hat{v}_{jj'} = \sum_{k=1}^{n} \sum_{l=1}^{n} \check{\Delta}_{kl} \left(\frac{X_{jk}e_k}{\pi_k} \right) \left(\frac{X_{j'l}e_l}{\pi_l} \right) \tag{4.24}$$

where $e_k = y_k - \hat{\boldsymbol{\beta}}X_k$ and

$$\check{\Delta} = 1 - \frac{\pi_k \pi_l}{\pi_{kl}}. \tag{4.25}$$

For a stratified sample, with population size $U\{1, \ldots, N\}$, let U_h be the subset of the index for the elements in the hth stratum, $h = 1, \ldots, H$. Then $U = U_1 \cup \ldots \cup U_H$ and $N_1 + \ldots + N_H = N$.

Let n be the sample size and n_h, $h = 1, \ldots, H$ be the size of hth stratum so that $n_1 + \ldots + n_H = n$.

The probability of observation k and l being selected for the sample is given by:

$$\pi_{k\ell} = \begin{cases} \frac{n_h(n_h-1)}{N_h(N_h-1)}, & k,\ell \in U_h \\ \frac{n_h}{N_h}\frac{n_i}{N_i}, & k \in U_h, \ell \in U_i, h \neq i. \end{cases} \tag{4.26}$$

4.3.3.3 Design-based methods for investigating causal effects - logistic regression

When the response variable Y is binary, logistic regression is used. Similarly to linear regression, a complex sampling design will affect the standard errors of the logistic regression coefficients (Lohr, 2009).

When weights are ignored, maximum likelihood estimation is generally used to obtain estimates for the parameters. This is not the case when sampling weights are used. Instead, quasi-likelihood methods are used.

Starting with the logistic regression model given by Equation 4.3 with mean $E(y_i) = \mu(\beta)$, a quasi-maximum likelihood estimator (MLE) $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ can be found by solving the following equation

$$\hat{T}(\beta) = \sum_{i=1}^{n} w_i u_i(\beta) = 0 \tag{4.27}$$

where w_i is the sampling weight for the i-th observation and

$$u_i(\beta) = [y_i - \mu_i(\beta)]x_i. \tag{4.28}$$

An estimate of the covariance matrix of $\hat{\beta}$ is then given by

$$\widehat{Var}(\hat{\beta}) = [I_O(\hat{\beta})]^{-1} V(\hat{T}) [I_O(\hat{\beta})]^{-1} \tag{4.29}$$

where $I_O(\hat{\beta})$ is the observed information matrix and $V(\hat{T})$ is the estimated covariance matrix of the estimated total $\hat{T}(\beta)$ (Chambers and Skinner, 2003).

4.3.4 Model-assisted methods

A model-assisted approach to survey sample analysis is a combination of a both a model-based and and a design-based approach. A model is used to establish any parameters of interest and then inference is based upon the design of the survey. More specifically, a particular model is fitted because it is believed to be a possible candidate for generating

the population, however the sampling weights are used to estimate the parameters and the sampling design is used to estimate the variance of the estimate (Lohr, 2009).

4.4 Simulations

In order to understand the role of sampling weights when analysing survey data more thoroughly, analysis of data in a variety of different settings will be conducted. Initially different sampling schemes will be looked at then, the proportion of the population sampled will be varied and finally a binary response variable will be considered. For each circumstance, a model-based analysis, both with and without weights, as well as a design-based analysis will be conducted.

In addition to fitting regression models using both model-based and design-based approaches, for each scenario the weight association test proposed by DuMouchel and Duncan will be conducted in order to see whether or not it would be recommended that the sampling weights should be used during regression analysis. Since these tests have been developed for continuous responses, during this section it will be determined whether or not a likelihood ratio test can be used in place of an F-test when considering a binary response.

For the continuous responses, initially a linear model, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$, will be fit using ordinary least squares. Next, an extended model including the original covariates plus the covariates multiplied by the sampling weights , $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}_\mathbf{u} + \mathbf{X}_\mathbf{w}\boldsymbol{\beta}_\mathbf{w}$ will be fit. Then an F-test can be used to test the null hypothesis $H_0 : \boldsymbol{\beta}_{\boldsymbol{w}} = 0$ to determine whether the coefficients of all of the weighted covariates are statistically significantly different to zero. If the null hypothesis is rejected then this suggests that weights are required during regression analysis. If there is a failure to reject the null hypothesis then it is suggested that a weighted during regression analysis is not necessary.

The population data consists of 1000 observations of six continuous variables (one response and five explanatory). Each observation also belongs to one of three strata. The true values of $\boldsymbol{\beta}$ are $\beta_0 = 1.5, \beta_1 = 1, \beta_2 = -0.5, \beta_3 = -1, \beta_4 = -0.5, \beta_5 = 2$.

4.4.1 Different sampling schemes

A variety of different sampling schemes will be considered. In this section the sample size is selected to be 100. Initially a simple random sample will be taken, then stratified samples will be selected using multiple different methods to determine the sample size of each strata.

4.4.1.1 Simple random sampling

For the simple random sample, each of the observations has the same sampling weight. The probability of selection for each observation is 0.1 meaning that the weight of each observation is 10. Because of this it is expected that there will be little to no difference between the three analysis approaches.

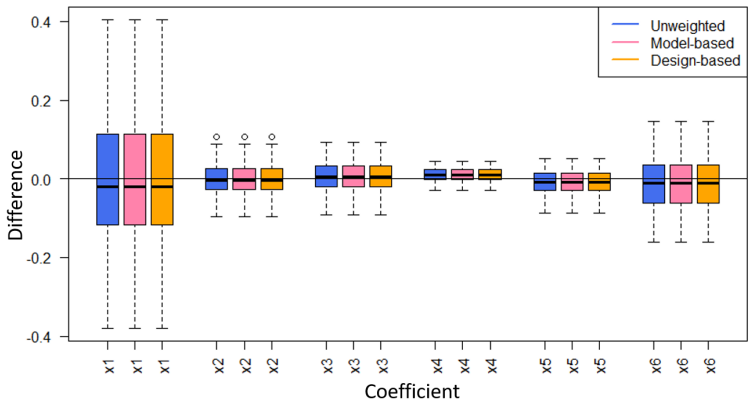

Figure 4.1: The difference between the coefficient estimates and the true values for the simple random sample not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange).

The results from the three different modelling approaches can be seen in Figures 4.1 and 4.2. Figure 4.1 shows the difference between the coefficient estimates and the true values. Since all weights are equal, the coefficient estimates obtained from each of the methods are also equal. It can be seen the range of this difference between simulations is largest for the intercept and smallest for x_4.

Figure 4.2 shows the standard errors for each of the coefficient estimates. It can be seen that, across all three methods, the range in standard errors for the coefficient of the intercept term is the largest and is smallest for the coefficient of x_4. This was to be expected as it supports what was seen in the previous figure. Dissimilarly from the

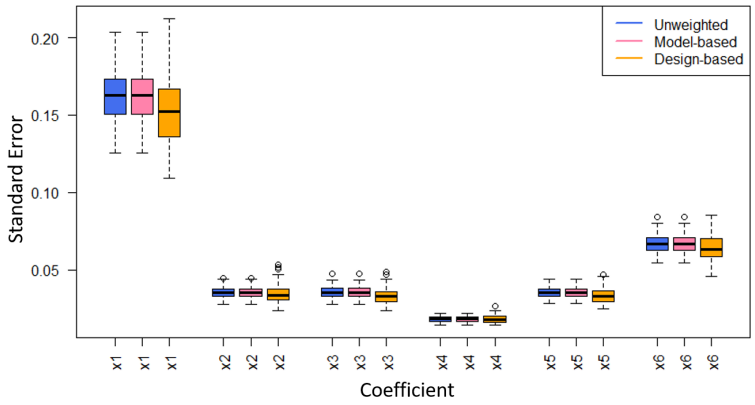

Figure 4.2: The standard errors of the coefficient estimates for the simple random sample not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange).

coefficient estimates, there is a difference in the range of the standard errors between methods. For all of the coefficients the median standard error calculated using the design-based method is smaller. The range of the standard errors however, is larger for the design-based models than it is for the two model-based approaches.

F-test

Since under simple random sampling, all sampling weights are equal, it is not possible to fit the extended model due to singularity. For a simple random sample it is generally not recommended that sampling weights are necessary however, the three methods discussed above have been conducted as a reference.

4.4.1.2 Stratified sampling

There are a variety of different ways in which the sample size selected from each strata in a stratified random sample can be selected. Selecting a fixed number, proportional allocation and optimal allocation will be considered. An underlying relationship between the response variable and strata will also be considered to determine how sampling weights effect the results of analysis both when strata is ignored and when strata is included.

4.4.1.2.1 Fixed number in each strata

The first stratified sample selected will select a fixed number of observations from each

of the strata. In the sample, the frequency of observations in strata 1, 2 and 3 is 204, 358 and 438 respectively. It is decided that the sample will consist of 30 observations from strata 1, 40 observations from strata 2 and 30 observations from strata 3.

Since a simple random sample of fixed size is selected from each strata, each observation within the same strata will have the same sampling weight. The probability of selection for each strata is $\pi_{i1} = 0.14706$, $\pi_{i2} = 0.11173$ and $\pi_{i3} = 0.06849$. Therefore the corresponding weights are $w_{i1} = 6.8$, $w_{i2} = 8.95$ and $w_{i3} = 14.6$ for observations in strata 1, 2 and 3 respectively.

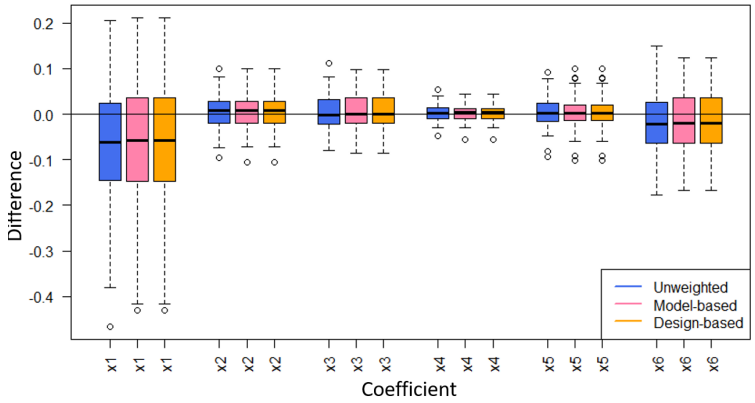

Figure 4.3: The difference between the coefficient estimates and the true values for the stratified sample not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange).

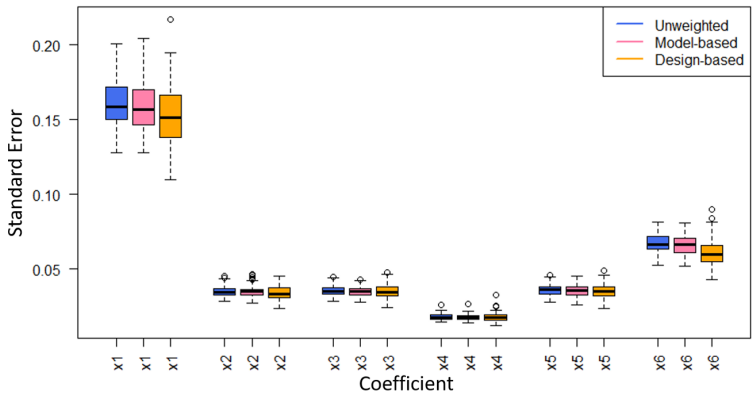

Figure 4.4: The standard errors of the coefficient estimates for the stratified sample not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange).

Figure 4.3 shows that there is a slight bias with the coefficient estimates of the

intercept with the median falling below zero. There is also a slight bias for the coefficient of x_6. The coefficient estimates are the same for model-based regression including weights and the design-based regression but are slightly different to the model-based regression ignoring the weights. The difference however appears to be very small.

Figure 4.4 shows boxplots of the standard errors for the simulations of the stratified sample. Similarly to the results from the simple random sample the range of the standard errors is highest for the intercept term across the three different methods. The median is also highest for this term for all three methods. Across all three methods, the median of the standard errors is lower for the design-based regression. For all of the estimates, with the exception of x_2, the median of the standard errors for the model-based regression including weights is lower than that of the model-based regression ignoring the weights. The difference in medians between these two methods however appears to be small.

F-test

When the F-test was conducted for this scenario, it was found that using sampling weights was recommended 8% of the time. From the results of the above simulations it appears that the coefficient estimates and the standard errors of these estimates are similar across all three methods with any differences found being small. Therefore, it is not unexpected that in this situation, the F-test does not recommend the use of the sampling weights.

However, even though the differences found here are small, it has been shown that there is a difference between the three methods when stratified sampling has been used. When weights are used, the coefficient estimates are the same regardless of whether a model-based or a design-based approach is used. The standard errors of the estimates however, differ across all three methods with, generally, the design-based approach giving the lower standard errors.

4.4.1.2.2 Proportional allocation

The next method of determining the sample size selected in each strata is proportional allocation. This means that a fixed proportion of each strata is chosen using a simple random sample within each strata. For these simulations 10% of each strata will be sampled.

Selecting 10% of each strata means that the sample size in strata 1, 2 and 3 is 20, 36 and 44 respectively. This means that the sampling weights for each strata are $w_{i1} = 10.2$, $w_{i2} = 9.9444$ and $w_{i3} = 9.9545$. Since these three sampling weights are all roughly the same, it is expected that the results of the analysis of this sample will be very similar to the results of the analysis of the simple random sample.

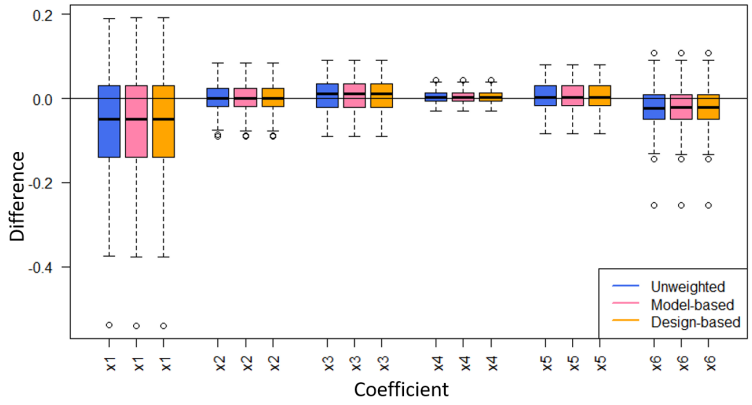

Figure 4.5: The difference between the coefficient estimates and the true values for the stratified sample with proportional allocation not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange).

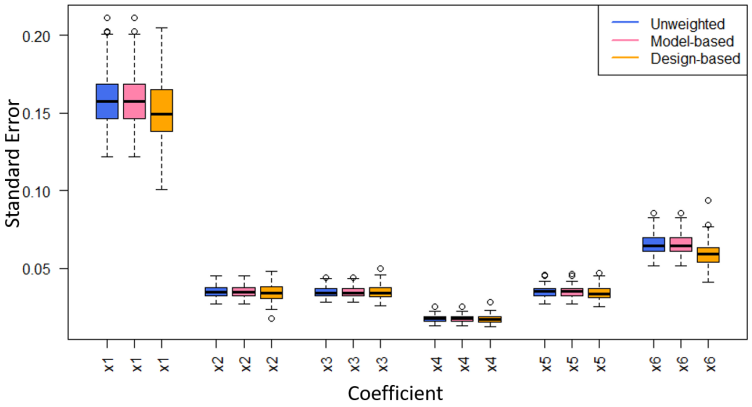

Figure 4.6: The standard errors of the coefficient estimates for the stratified sample with proportional allocation not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange).

It can be seen in Figure 4.5 that the coefficient estimates across the three methods all appear to be equal. This was to be expected since the sampling weights across the three strata are all close to 10. Once again, there is a bias on the intercept term.

The difference between the three methods can be seen in the boxplots of the standard

errors of the coefficient estimates. Both model-based approaches (ignoring and including weights) give similar standard errors. The median of the standard errors across all estimates is lower for the design-based analysis than it is for the model-based analysis. The range of the standard errors, however, is larger across all of the estimates for the design-based regression.

F-test

Since the sampling weights for each strata when using proportional allocation are all similar to each other, it would be expected that there are very few instances, if any, in which it is recommended that sampling weights are included in regression analysis. It was found using the F-test that the use of sampling weights was recommended 7% of the time. This is surprisingly close to the number of times recommended in the previous stratified example.

Once again, despite in general this test not recommending that sampling weights are used, it has been shown that there is a difference in the approaches which can be taken. If a model-based approach is chosen then there appears to be little difference between whether or not weights are used. There has been shown to be a difference between model-based approach and a design-based approach when it comes to calculating the standard errors. The median standard error is lower when taking a design-based approach however the variability in the standard errors is also greater.

4.4.1.2.3 Optimal allocation

The final way in which the sample size within each strata is determined which will be considered is optimal allocation. When a survey is being conducted, there will be certain costs associated with taking an observation for each strata. The purpose of optimal allocation is to obtain the most information possible with the lowest survey cost (Lohr, 2009).

The following formula is used to calculate the optimal sample size in stratum h:

$$n_h = \left(\frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{l=1}^{H} \frac{N_l S_l}{\sqrt{c_l}}} \right) n \tag{4.30}$$

where S_h is the standard deviation of the response variable in stratum h, c_h is the

cost associated with taking an observation in stratum h and H is the number of strata (Neyman, 1934).

Using this formula, keeping equal cost across the strata, the sample size selected from each stratum is 19, 35 and 46. The weights for each stratum are therefore $w_{i1} = 10.73684$, $w_{i2} = 10.22857$ and $w_{i3} = 9.52174$.

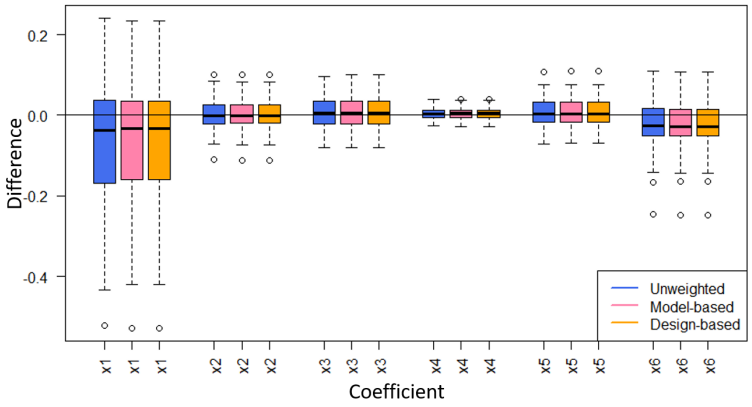

Figure 4.7: The difference between the coefficient estimates and the true values for the stratified sample with optimal allocation not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange).

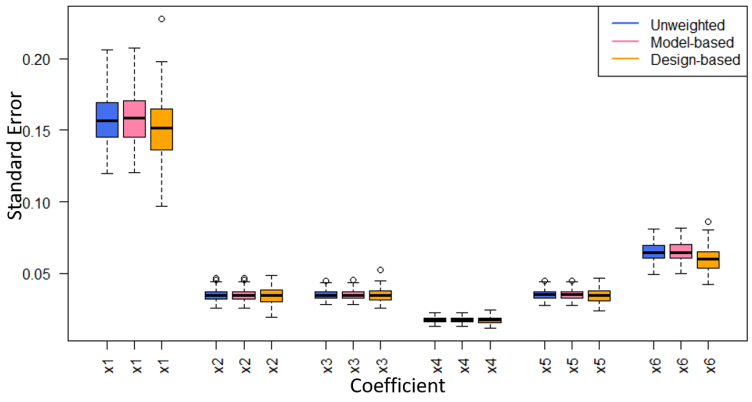

Figure 4.8: The standard errors of the coefficient estimates for the stratified sample with optimal allocation not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange).

Figures 4.7 and 4.8 show that the results across the three methods are similar when optimal allocation is used as when proportional allocation is used. Once again, there is little to no difference in the coefficient estimates between the three methods. There is also still a slight bias in the estimate of the intercept across all three methods.

Similarly, the median standard error for each of the estimates is smallest for the design-based regression than for the model-based regression with the range of the standard errors also being largest when using the design-based methods.

F-test

It was found that using sampling weights was recommended 5% of the time using the F-test described above. This is a similar amount of times to both of the previous scenarios. As before, this is unsurprising since the coefficient estimates calculated both with and without weights are found to be very similar. Again, the main difference between the three approaches is with the standard errors although this difference is small.

4.4.1.2.4 Relationship between sampling scheme and response

Finally the effect of sampling weights when there is an underlying relationship between the response variable and strata is considered. When simulating the response variable, strata has been used as a covariate. The true regression coefficients for strata 2 and strata 3 are 1 and 2 respectively. The number selected for the sample in each strata will be the same as when a fixed number was selected from each strata (30, 40, 30) and therefore the weights of the observations in each strata will also be the same as in these simulations.

Two analyses will be conducted: the first will not include strata as an explanatory variable and the second will include strata.

Analysis not including strata

Initially, a linear regression model will be fitted using only the continuous covariates. It is generally recommended that if the sampling scheme has a relationship with the response variable and not accounted for in the explanatory variables, then weights should be used. Therefore, from the literature it is expected that in this scenario it will be recommended that sampling weights are used the majority of the time. It is expected that the estimated coefficients in the analyses using the sampling weights will be closer to the true values than when weights are ignored.

The coefficient estimates minus the true values can be seen in Figure 4.9. Once again, the coefficient estimates are the same for model-based analysis including weights and the

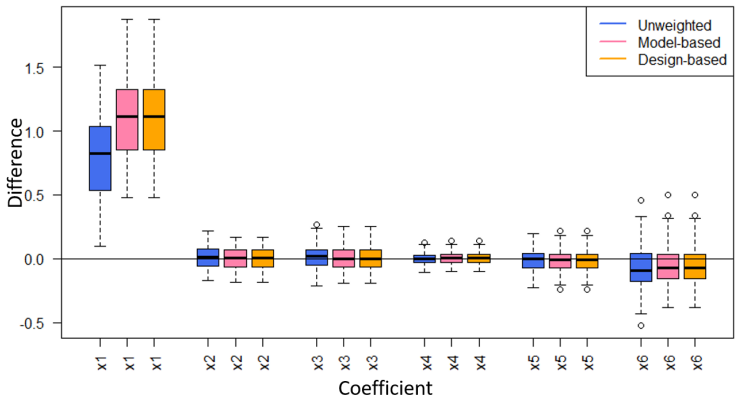

Figure 4.9: The difference between the coefficient estimates and the true values for the stratified sample with underlying relationship (ignoring strata) not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange).

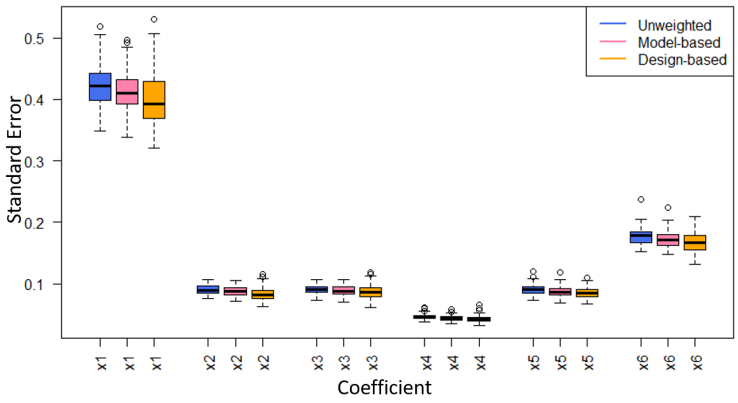

Figure 4.10: The standard errors of the coefficient estimates for the stratified sample with underlying relationship (ignoring strata) not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange).

design based analysis. The biggest difference between the methods can be seen with the intercept term. This difference is the opposite of what was expected. When weights are included, both in a model-based and a design-based regression, the bias of the coefficient estimate for the intercept term is larger than when weights are not included.

For the remaining estimates however, the difference between the coefficient estimates and the true values are as expected and are smaller when weights are included than when they are not. When weights are included, the median of this difference for all of the estimates with the exception of x_6 is very close to zero.

Figure 4.10 shows the standard errors of the estimates. As was seen previously, the range of the standard errors for the design-based analysis is greater than for both of the model-based analyses across most of the coefficients. The median of the standard errors across all of the coefficient estimates are lower for the model-based analysis that includes when weights are not included and then lower again when design-based regression is used.

F-test

When an F-test is used to test the appropriateness of including the sampling weights in this scenario, it is found that it is recommended 100% of the time that weights are used. From the literature this was expected.

The results of the analysis show that including weights tends to improve the coefficient estimates for the continuous covariates however not for the intercept term. Using a design-based analysis lowers the median standard error of the coefficient estimates however the range of these standard errors is generally increased compared to when a model-based approach is used.

Analysis including strata

When the response variable has a relationship with the sampling scheme, it is generally recommended that either sampling weights are used or the sampling scheme is accounted for in the covariates. Next, a linear regression model will be fit including strata as a factor. This will aim to show the difference between the two different recommended approaches.

Figure 4.11 shows the coefficient estimates minus the true values. It can be seen that there is still a bias on the intercept term however it is smaller than when strata is not included. There is also slight bias for the coefficient estimates of strata 2 and strata 3. The bias for all of the coefficient estimates is smaller when sampling weights are used

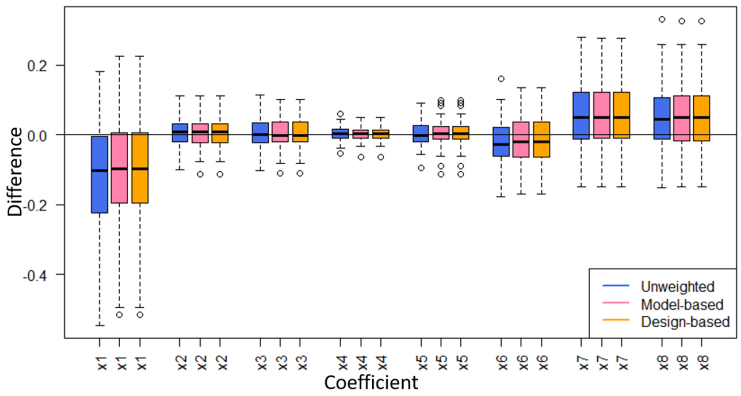

Figure 4.11: The difference between the coefficient estimates and the true values for the stratified sample with underlying relationship (including strata) not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange).

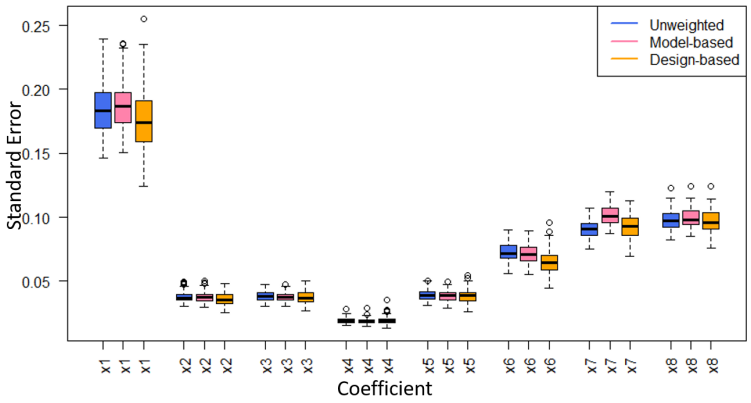

Figure 4.12: The standard errors of the coefficient estimates for the stratified sample with underlying relationship (including strata) not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange).

than when they are not.

Figure 4.12 shows the standard errors of the coefficient estimates. The standard error of the intercept estimate is larger for all three methods than it is for all of the other covariates. The model-based approach ignoring weights has the lowest median and smallest range (excluding a slight outlier) for the intercept than both of the approaches which include the sampling weights. This is also the case for the estimates of the coefficients of x_6 and both of the strata covariates. The range of the standard errors is yet again greater for the design-based estimates than for the model-based estimates

however the difference is generally small.

Since the sampling weight is the same for each strata and strata is being used as a factor in the regression model, it is not possible to fit the extended model in order to use an F-test to determine whether or not it is recommended to use weights in this scenario. Looking at the results from the analysis however it seems that there is only a very small difference in the coefficients between both the coefficient estimates and the standard errors of these estimates between the three methods. Therefore, it is expected both from these results and also the literature that if strata is used as a covariate, then it is not recommended that sampling weights are also used.

Comparing the results of the analysis when strata is included in the regression model with the regression when strata is ignored, it can be seen that the biggest difference is found with the intercept term. When, strata is excluded and sampling weights are used, there is a much larger bias for the estimate of the intercept. When strata is excluded, the median bias for the remaining estimates is close to zero whether sampling weights are used or not. When strata is included as a factor, there is a slight bias on the intercept, one of the continuous covariates and both of the levels of the strata factor.

It has been found that when there is a relationship between the response variable and the sampling scheme that is is important to account for this is some way when fitting a regression model. Assuming in this scenario that relationship of interest is between the response variable and the continuous covariates, then there is very little difference in interpretation between the two recommended methods. Therefore, if incorporating the sampling scheme is possible then sampling weights are not needed. In many surveys, however, the sampling scheme is generally more complicated and may include a lot more strata than what has been used in these simulations. Therefore, including strata as a factor may mean that a large number of coefficients need to be estimated and may lead to a model which is difficult to interpret.

4.4.2 Different sampling size

As the size of the sample increases, the proportion of the population sampled also increases. Therefore the larger the sample, the more representative of the population the sample may be. This section will aim to show how the effect of including sampling weights in analysis changes as the size of the sample increases.

Since it was previously shown that there is little to no difference in the results of the three different methods when looking at a simple random sample and a stratified sample with proportional allocation, from now on we will look only at the latter along with the other scenarios described above.

4.4.2.1 Sampling 10% of the population

In the previous section, the sample size was 10% of the total population. It was found that when the response variable had no relationship with the response variable, more often than not it was not recommended that sampling weights were necessary in regression analysis. The difference between design-based regression and model-based regression (including weights) lay in the standard errors. It was found that, in general, the estimates were less precise when sampling weights were used.

When there was a relationship between the response variable and the sampling scheme it was recommended every time that weights be included. From the literature it was suggested that the sampling scheme could be accounted for either by using the sampling weights or by including the strata used during sampling as covariates in the regression model. It was found that there was little difference in how the relationship of interest would be interpreted between both of these methods. In the simulations conducted, it was simple to compute estimates using either method however if the sampling scheme is more complicated than the one used here, it may be much more difficult to include the strata within the regression model and therefore computing estimates using the sampling weights may be the most logical method to use.

4.4.2.2 Sampling 20% of the population

Next the sample size will be doubled in order to contain 20% of the population. The boxplots produced can be found in the Appendix.

For all of the sampling schemes considered, the results follow a similar pattern. The coefficient estimates are all very similar to when only 10% of the population was sampled. The standard errors are also very similar to what was seen above. As would be expected, the range of the standard errors of the estimates is smaller when a higher proportion of the population has been sampled.

When using the F-test discussed above it was found that in the majority of instances,

with the exception of when there is a relationship between the response variable and sampling scheme, it was not found necessary to include the sampling weights when using regression. For a stratified sample in which a fixed number is selected from each strata, it was found that the use of sampling weights was recommended only 2% of the time. For stratified sampling using proportional allocation and optimal allocation weighting was found necessary 4% and 7% of times respectively. When there is a relationship between the response variable and the sampling scheme, it was found that weights were recommended 91% of the time which is a reduction compared to the previous results.

Comparing these results to when 10% of the population was sampled, it can be seen that with the exception of when optimal allocation is used, when the proportion of the population sampled is increased, it is recommended fewer times that sampling weights are required. This may be because the more of the population sampled, the more representative the sample will be of the population and so sampling weights may not be necessary.

4.4.2.3 Sampling 50% of the population

Finally, the sample size will be increased to contain 50% of the population. It was seen previously that generally with an increase in the proportion of the population sampled the percentage of times that it is recommended that using sampling weights is necessary is decreased. Therefore it is expected that when half of the population is sampled that the sample will be representative enough of the population that the use of weights is recommended very few times.

The boxplots produced can be found in the Appendix. It can be seen that when looking at the coefficient estimates, a similar pattern to that seen when 10% and 20% of the population was sampled can be seen again. The estimates are equal when weights are included regardless of whether a model-based or a design-based approach is taken.

When looking at the standard errors, similarly to the previous simulations, in general the median standard error for each of the estimates is lower when a design-based approach is used than a model based however the range is greater. Once again, the range of the standard errors is smaller when 50% of the population is sampled compared to when 10% or 20% is sampled.

It is when looking at the results of the F-test to determine whether or not sampling

weights should be used that some differences arise from the previous simulations. For the stratified sample in which a fixed number is sampled from each strata, it was found that the use of weights was recommended 3% of the time. This is very similar to when 20% of the population was sampled. When proportional allocation and optimal allocation were used the inclusion of weights was recommended 4% and 2% of the time respectively. For proportional allocation this is the same as when 20% of the population was sampled and for optimal allocation this is a decrease. When there was a relationship between the response variable and the sampling scheme it was found that the use of weights was recommended 0% of the time. This is a large decrease from when both 10% and 20% of the population was sampled. This may be due to the fact that when a large proportion of the population is sampled, the sample is a lot more representative of the population and sampling weights may no longer be required.

4.4.3 Binary response variable

The response variable of interest is not always continuous and so instead of linear regression, logistic regression is used. The majority of the literature is concerned with linear regression and so it will be interesting to see whether or not the use of sampling weights effects analysis when using logistic regression.

Similarly to the linear regression simulations above, when using logistic regression several sampling methods will be considered. Stratified sampling with fixed number in each strata, using proportional allocation and optimal allocation will be considered as well as observing the effect of sampling weights when there is a relationship between the response variable and the sampling scheme.

For each scenario three logistic regression models will be compared. The first will use a model-based approach and ignore the sampling weights. The second will also use a model-based approach but this time the sampling weights will be used in order to estimate the regression coefficients and the variances of these estimates. Finally, the third will take a design-based approach.

The true values of the regression coefficients remain unchanged from the previous simulations. The response variable has been simulated following a binary distribution, taking values of 0 or 1, with $P(Y = 1) = \exp(\eta)/(1 + \exp(\eta))$ where $\eta = X\beta$. The population contains 463 zeros and 537 ones.

It will no longer be possible to use the F-test applied in the above simulations to establish whether or not using weights is necessary. Instead, it will be determined whether a likelihood ratio test (LRT) can be used instead.

In order to conduct a likelihood ratio test the following steps will be taken. Similarly to the F-test, two models will be fit. First, a model including only the covariates of interest (and no weights) and then an extended model which includes the covariates of interest along with these same covariates multiplied by the sampling weights. Define L_0 as the maximum likelihood of the more simple model and L_1 as the maximum likelihood of the extended model. The ratio $\lambda = L_0/L_1$ can then be calculated and will take a value between 0 and 1. The test statistic $\chi^2 = -2\log\lambda$ can now be calculated and compared to the $100(1-\alpha)$ percentile of the Chi-squared distribution with k degrees of freedom where k is the difference in the number of parameters in the two models. If χ^2 is found to be greater than the critical value then it is suggested that weights are necessary in analysis.

4.4.3.1 Stratified Sampling

As before, each observation belongs to one of three strata and the first sampling scheme will select a fixed number of observations from each strata. These values have been chosen to be 30, 40 and 30 from strata 1, 2 and 3 respectively.

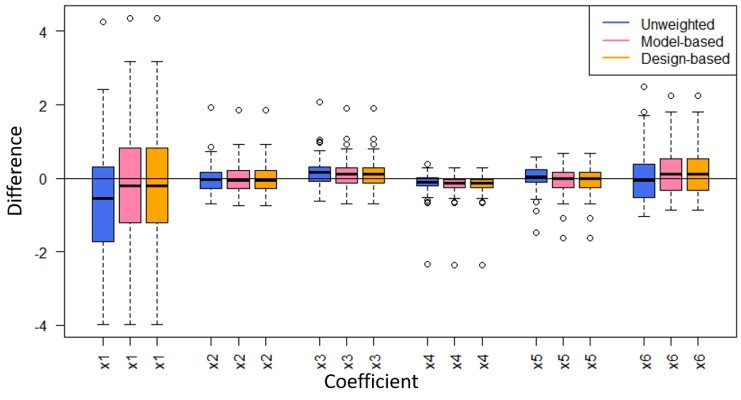

Figure 4.13: The difference between the coefficient estimates and the true values for the stratified sample not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) for the logistic regression model.

It can be seen in Fig 4.13 that, similarly to linear regression, when weights are

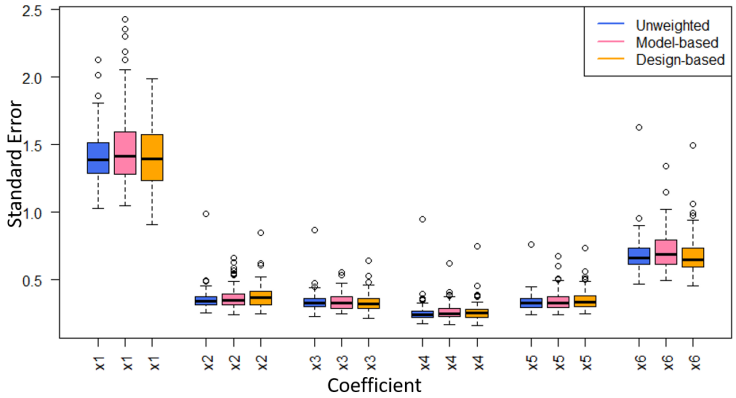

Figure 4.14: The standard errors of the coefficient estimates for the stratified sample not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) for the logistic regression model.

included the coefficient estimates are the same regardless of whether a model-based or a design-based approach is taken. The intercept has a slight bias which is marginally reduced when weights are used. For the remaining covariates, the median difference between the coefficient estimates and true values are all close to zero. In general, other than the coefficient estimates of x_4 and x_6, the median difference is closer to zero when weights are used compared to when they are not.

The differences between the model-based approach and the design-based approach can be seen in the standard errors of the coefficient estimates (Figure 4.14). The median of the standard errors for each of the coefficients is very similar for each of the three methods. Similarly to what was seen in the linear regression examples, the standard error of the intercept is greater than for the coefficients of the continuous covariates. The range of the standard errors appears to be similar for the three methods. This is different to what was found in the linear regression simulations where, generally, the design-based estimates had the largest ranges in standard errors.

The Likelihood Ratio Test

When conducting the likelihood ratio test, it was found that weights were recommended 9% of the time. This is a relatively low number of times which is expected from the results of the analysis since there appears to be very little differences between both the coefficient estimates and the standard errors of the estimates regardless of whether weights are used or not.

4.4.3.1.1 Stratified sampling with proportional allocation

When sampling using proportional allocation the proportion used in these simulations was chosen to be 10%. Once again as expected that under this sampling scheme, there will be very little difference whether weights are used or not. This is because the sampling weights are close to 10 for all strata when the data is sampled in this way.

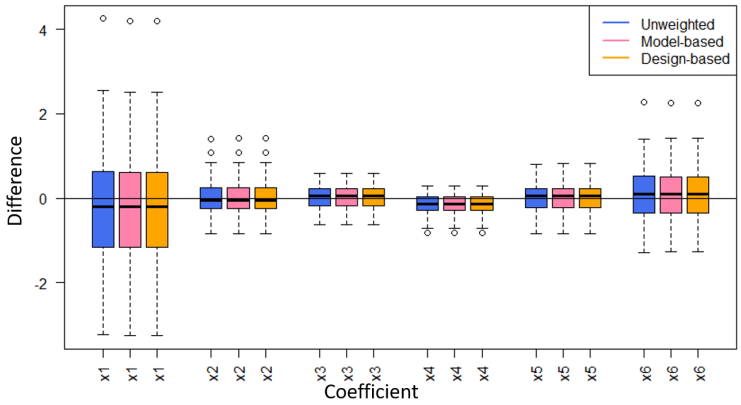

Figure 4.15: The difference between the coefficient estimates and the true values for the stratified sample with proportional allocation not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) for the logistic regression model.

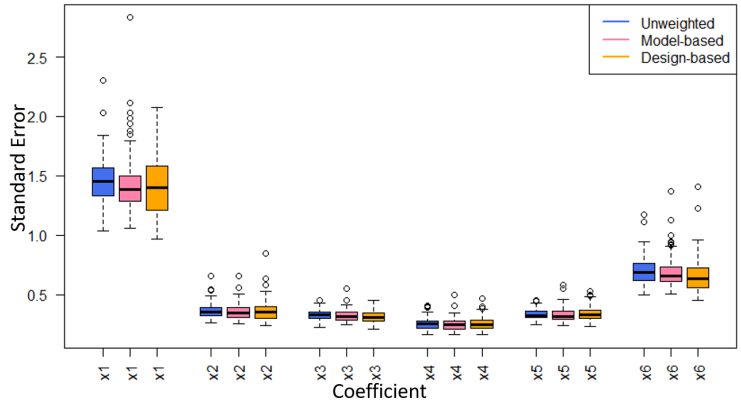

Figure 4.16: The standard errors of the coefficient estimates for the stratified sample with proportional allocation not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) for the logistic regression model.

As expected, there is little to no difference in the coefficient estimates whether weights

are used or not (as seen in Figure 4.15). The median difference is pretty close to zero for all of the coefficients.

The range of the standard errors seen in Figure 4.16 appear to be similar across the three methods. There are also quite a few outliers shown on the boxplots for the three methods. In general, when weights are used, the median standard error is smaller than when weights are not used but are similar whether a model-based or design-based approach is taken.

The Likelihood Ratio Test

Using likelihood ratio tests it is suggested that weights are used 19% of the time. Based on the above analysis this is higher than expected since there is very little difference between the coefficient estimates when weights are used and when they are not. There is a difference however in the standard errors which may account for why it is recommended that weights are used this many times.

4.4.3.1.2 Stratified sampling with optimal allocation

Using Equation 4.30 with a desired sample size of 100, the number of observations selected from strata 1, 2 and 3 are 20, 36 and 44 respectively. This leads to sampling weights of 10.2, 9.9444 and 9.9545. This number is close to 10% of each strata therefore it is expected that the results of this analysis will be similar to that in the previous section when proportional allocation was used to choose the sample size selected from each strata.

As expected, the results seen in Figure 4.17 are very similar to those seen previously when proportional allocation was used. There are only slight differences between the coefficient estimates whether weights are used or not. The median of the differences between the estimates and the true values are all very close to zero.

Also, similarly to when proportional allocation was used, the ranges of the standard errors are similar across the three approaches used and the median standard error is generally lower when weights are used compared to when they are not.

The Likelihood Ratio Test

It was found that 19% of the time it was recommended that the use of sampling weights would be appropriate. This is the same amount found previously when propor-

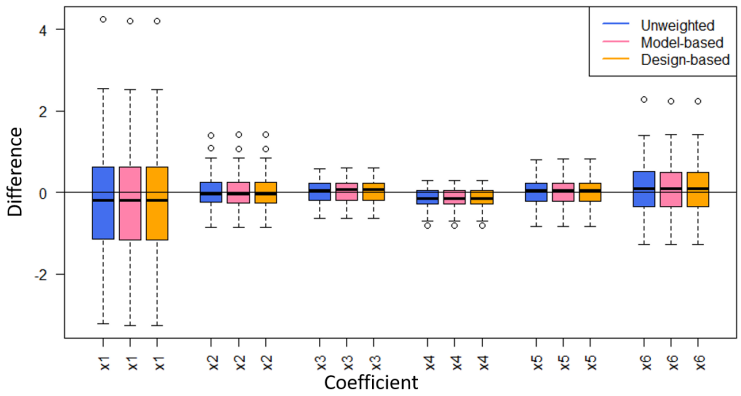

Figure 4.17: The difference between the coefficient estimates and the true values for the stratified sample with optimal allocation not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) for the logistic regression model.

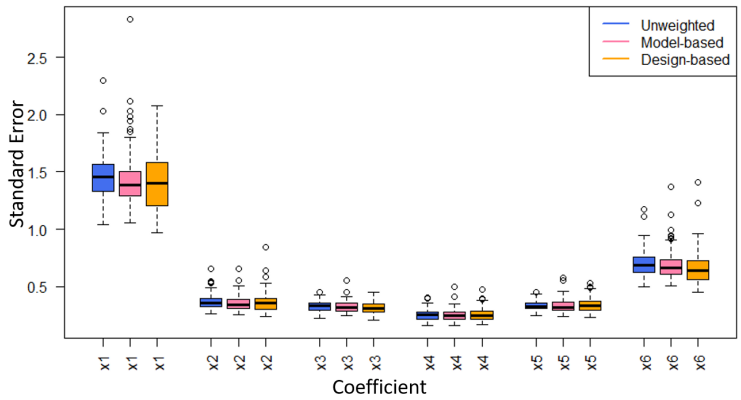

Figure 4.18: The standard errors of the coefficient estimates for the stratified sample with optimal allocation not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) for the logistic regression model.

tional allocation was used. Since the sampling size in each strata was similar in both scenarios this was to be expected.

4.4.3.1.3 Relationship between response variable and sampling scheme

To see the effect of using weights when there is a relationship between the response variable and the sampling scheme, strata was used as covariate when simulating the response variable. The true values of the coefficients of strata were chosen to be 1 and 2 for strata 2 and 3 respectively.

Analysis ignoring strata

For the first analysis, strata was not included as a covariate in the fitted models.

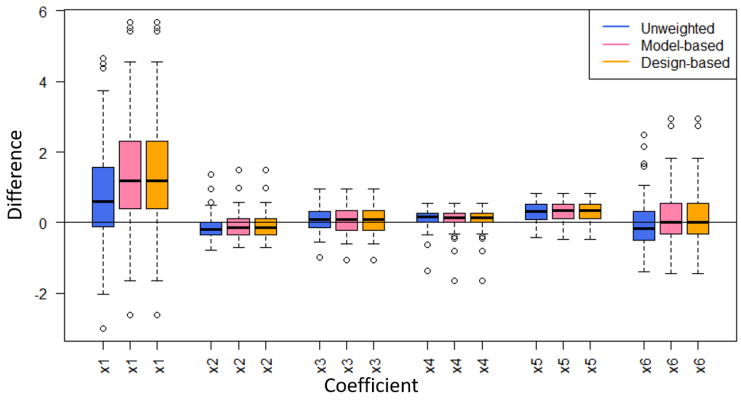

Figure 4.19: The difference between the coefficient estimates and the true values for the stratified sample with underlying relationship (ignoring strata) not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) for the logistic regression model.

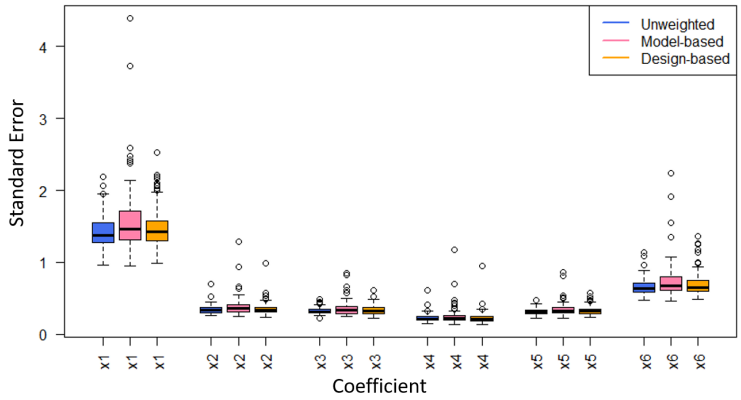

Figure 4.20: The standard errors of the coefficient estimates for the stratified sample with underlying relationship (ignoring strata) not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) for the logistic regression model.

It can be seen in Figure 4.19 that, similarly to the linear regression example, the biggest difference in coefficient estimates between the three methods is with the intercept. There is a bias on the intercept which is reduced when weights are not included. For the coefficients of the remaining covariates, the median difference between the estimates and the true vales is close to zero for all three methods with any bias being slightly reduced by including the sampling weights.

Figure 4.20 shows the standard errors of the coefficient estimates. It can be seen that there are more extreme values for the model-based approach which includes weights than there are for the other two methods. In general, there is little difference between the median standard errors when model-based regression with no weights and design-based regression is used. The range of the standard errors for these two methods also appears to be similar across all of the coefficients. The median of the standard errors when a model-based approach including weights is used is higher than for the other two methods.

The Likelihood Ratio Test

When using a likelihood ratio test it was recommended 58% of the time that the use of sampling weights would be appropriate. This is much lower than the amount of times it was recommended in the linear regression example (100%) however it is still recommended that weights are used more often than not.

Analysis including strata

The second way in which the sampling scheme can be accounted for is by including strata as a covariate in the regression model. Since there are only three strata in these simulations, it is possible to try this method. However, with a more complex sampling structure, especially from large surveys, it may not be possible to account for the sampling scheme in this way as it may lead to a model with a large number of covariates which is difficult to interpret.

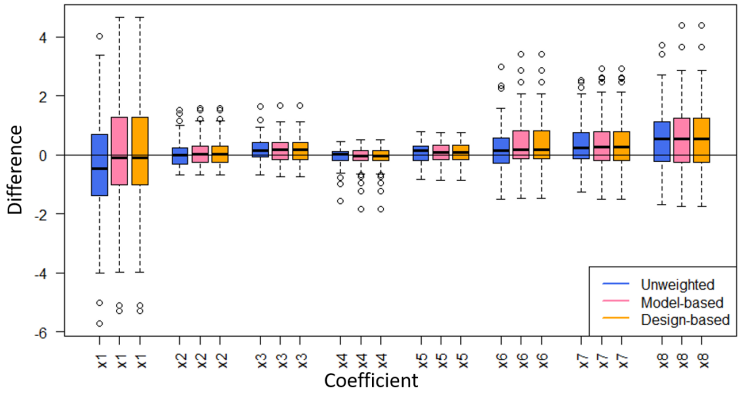

Figure 4.21: The difference between the coefficient estimates and the true values for the stratified sample with underlying relationship (including strata) not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) for the logistic regression model.

Figure 4.21 shows that, other than the intercept, there is very little difference in the

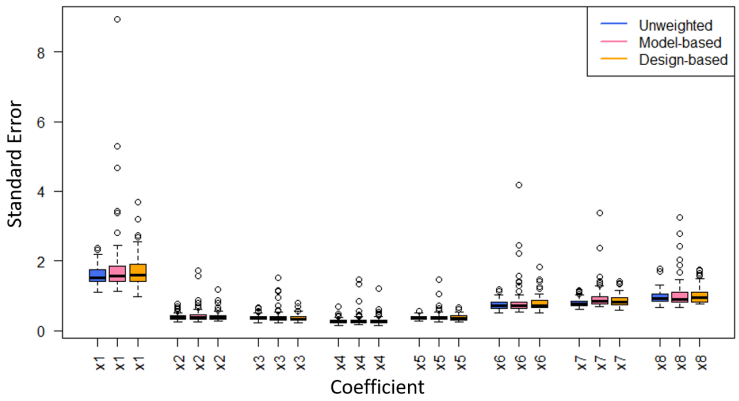

Figure 4.22: The standard errors of the coefficient estimates for the stratified sample with underlying relationship (inclusing strata) not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) for the logistic regression model.

coefficient estimates between the three methods. The intercept has a slight bias when weights are not used which is reduced when weights are included when calculating the estimates.

The standard errors of the estimates seen in Figure 4.22 appear to follow the same pattern as when strata is not included as a covariate. There are more outliers when a model-based approach with weights is used. When a model-based approach without weights and a design based approach is used, the standard errors are similar.

Similarly to the linear regression case, it is not possible to use the likelihood ratio test as the extended model cannot be fit due to singularity. It can be seen however, that there are only very small differences between the coefficient estimates whether weights are used or not. Also, when a model-based approach not including weights and a design-based approach is used, there is little difference in the standard errors. Therefore if strata is included as a covariate then the use of sampling weights as well is not necessary.

4.5 Discussion

The overall aim of this chapter was to determine when and how sampling weights should be used when analysing survey data.

Sampling weights can be used to adjust for unequal sampling probabilities, non-response bias and under-coverage. The sampling weights provided in the PNS and MCS

both account for unequal sampling probabilities and non-response bias. Further details regarding how these weights have been calculated are given in Chapter 2 and Chapter 3.

It has been concluded that when estimating descriptive statistics such as population means and totals, the sampling weights should always be used (De Leeuw et al., 2012). The Horvitz-Thompson estimator described above shows how the sampling weights can be incorporated when computing descriptive statistics.

When using regression models with survey data, it is less clear whether or not weights should be used during the analysis. Tests have been developed for determining whether or not the use of sampling weights is recommended for a linear regression model (DuMouchel and Duncan, 2008). During the simulations section of this chapter, one of these tests was extended for use in a logistic regression model setting.

When it is recommended that weights are used, two ways in which they can be incorporated has been discussed. The first of these is to use a model-based approach with sampling weights. The second is to use a design-based approach. One of the main differences between these approaches is the population to which inference can be made. Using a design-based approach, inference can be made about the specific population from which the sample was selected. Using a model-based approach, inference is made regarding broader sets of populations with similar characteristics (Dorazio, 1999).

Simulations were run to compare the results of these two methods along with the results when ignoring sampling weights. As expected, there was found to be no difference in coefficient estimates when when weights are used regardless of whether a model-based approach or a design-based approach was used. In the majority of instances any bias in the coefficient estimates was found to be lower when weights were used compared to when they were not.

The main difference between the model-based and design-based methods was found to be in the standard errors. Generally the median of the standard errors was found to be lower when a design-based approach was taken compared to when a model-based approach was taken. The range of the standard errors however tended to be greater for the design-based regression.

The simulations using a linear response agree with Lumley's statement that if both weighted and unweighted estimates are valid then the weighted simulations will be less

precise. It was found using the F-test that in general it was not necessary to use the sampling weights during regression suggesting that the unweighted models were valid. For the majority of the scenarios that were considered, the range of the standard errors for the design-based regression was larger than that of the model-based regression.

When comparing the results of the linear regression simulations and the logistic regression simulations it appears that the effect of including sampling weights is similar for both. The main difference found was that when there is a relationship between the response variable and the sampling scheme and strata is not included as a covariate in the model, the F-test recommends that sampling weights are used 100% of the time in the linear regression case whereas in the logistic regression simulations it was only recommended 58% of the time using a likelihood ratio test. Despite the number of times it is recommended that sampling weights are used is lower, it is still recommended more often than not.

When there was a relationship between the response variable and the sampling scheme, two methods for accounting for this were compared. The first did not include strata as a covariate and it was found in this case that it was appropriate to include sampling weights in the analysis. The second included strata as a covariate in the model. It was found that the interpretation of the coefficient estimates for the covariates of interest were similar for both of these methods and so either method would suitably account for the sampling scheme. When the sampling scheme is more complex than the one simulated, it may make more sense to simply use the sampling weights since including strata as a covariate may lead to a model with a lot of different covariates which is difficult to interpret.

When the proportion of the population sampled was increased, it was found that in the majority of cases the results of the F-test indicated that the use of sampling weights was not necessary, especially when there was no relationship between the response variable and the sampling scheme. When there was a relationship, it was found that if the proportion sampled was increased to 20%, it was still recommended to use the sampling weights 91% of the time however this decreased to 0% when the proportion was increased to 50%.

This chapter discusses methods for obtaining parameter estimates for model-based and design-based regression. If there are a lot of potential covariates, however, some

sort of variable selection may need to be implemented. Variable selection methods for an infinite population are discussed in Chapter 5. Variable selection methods for a finite population are discussed in Chapter 6.

4.5.1 Recommendations

When estimating descriptive statistics, the sampling weights should always be used. This chapter aimed to reduce the uncertainty around if sampling weights should be used when conducting regression models. The following recommendations can be used for both linear regression and logistic regression.

When simple random sampling is used there is no reason to use sample weights as all sample weights within the sample should be equal. Therefore, model-based regression is appropriate here.

When stratified sampling is used and a fixed number is chosen from each strata, it was found that the standard errors of the coefficient estimates was smallest when sampling weights were used in a design-based model. This was also found to be case for stratified sampling using proportional allocation and stratified sampling using optimal allocation.

When there is an underlying relationship between the response variable and the sampling scheme, it was found that design-based methods were preferred if the sampling scheme was not accounted for in the covariates of the model. If strata was included as a covariate it was found that there was no need to include sampling weights in the analysis. Therefore the most appropriate method should be used when accounting for the sampling scheme. If there are a lot of strata to be considered, then using the sampling weights would be the recommended method.

It was found that as the proportion of the population sampled increased that the need to include sampling weights in the analysis decreased. This is due to the fact that as the proportion sampled increases, the more representative of the population the sample is. Therefore, if only a small proportion of the population is sampled then the use of sampling weights is recommended.

4.5.2 Unique Contribution

Currently, confusion exists in the literature surrounding sampling weights in a regression model. When researching this subject, conflicting recommendations were found and so simulations were conducted in order to identify if and when sampling weights should be used.

There is existing literature around testing for the appropriateness of sampling weights in a linear regression model. However, little research has been carried out in a logistic regression setting. This chapter extended the existing F-test using a likelihood ratio test for a logistic regression model.

Chapter 5

Methodology and Analysis - Variable Selection

5.1 Introduction

This chapter will aim to answer the research question: Which methods of variable selection are commonly used when analysing survey data? Specifically, how has the lasso been adapted beyond linear regression with continuous covariates for use with more complex data?

The chapter will begin with a review of the current practices used to conduct variable selection including background knowledge, information criteria and penalised likelihoods. Next simulations will be conducted to compare the resulting models when using different methods of variable selection. Finally the current packages available in R to conduct the lasso will be investigated and whether any are suitable to analyse the data from The PNS and the MCS will be discussed.

5.2 Background Knowledge

The use of background knowledge is one method to assist with variable selection. This involves using the results from previous studies or the advice of experts to filter the variables before any statistical models are fit to the data (Heinze and Dunkler, 2017).

This method can be implemented before or after data collection. Given a research question a list can be created containing any information that could possibly be collected

and then cut down based on the factors such as relevance, cost of collection and quality of measurements.

It may possible to sketch a conceptual model using prior knowledge of the subject areas showing any relationships between the independent variables in the data set. During this process certain variables may be found to be redundant and hence can be excluded from either data collection or regression modelling depending on when this step is carried out. The conceptual model may also highlight which of the variables are confounders and hence should be adjusted for when modelling.

Although using background knowledge may start to cut down the number of predictors in a model, it uses little to no statistical theory in order to classify whether or not the resulting model performs better or worse than another model using a different set of predictors. Therefore it is best if this method is used in conjunction with another method of variable selection.

5.3 Variable selection algorithms using information criteria

The most common method of variable selection is hypothesis testing (Heinze et al., 2018). This generally takes the form of comparing two models based on a certain criteria in order to determine which gives the preferred fit to the data.

These tests can be used using a stepwise selection algorithm. This can be done using forward selection, backwards selection or stepwise selection depending on which model is chosen as the initial model.

Forward selection begins with the null model and each of the variables are added one at a time to see which improves the fit of the model most based on a specified criteria. This variable is then added to the model and the process is repeated until it is found that no further variables improve the model fit (Bursac et al., 2008).

Backwards selection begins with the full model and removes each of the variables in turn in order to see which is the least significant based on a specified criteria. This variable is then removed and then the process is repeated until it is indicated that no further variables should be removed from the model (Bursac et al., 2008).

Stepwise selection is a combination of forward selection and backwards selection.

Whereas during forward or backwards selection if a variable is added/removed from the model it remains included/excluded this is not the case during stepwise selection. Each time an independent variable is added to the model, the significance of each of the variables in the model is then checked. A variable found to have significance lower than a pre-specified amount is then removed from the model (Hintze, 2007).

5.3.1 Selection criteria

In order to use any of these algorithms there needs to be a criteria chosen to determine whether or not independent variables are added or removed from the model. In each of the cases described above, at each stage of the algorithm multiple models are fit and need to compared. Analysis of variance (ANOVA) and Akaike Information Criterion (AIC) can be used in order to choose between two or more models.

5.3.1.1 Analysis of variance (ANOVA)

Analysis of variance (ANOVA) can be used to choose between two nested models. First a model containing p covariates is fit. Next, a smaller model in which a subset of the p covariates has been removed from the model is fit. Defining this subset of covariates as $\mathbf{u} \subset \{1, \ldots, p\}$ then ANOVA tests the hypothesis $H_0 : \boldsymbol{\beta}_u = 0$.

The test statistic can then be calculated using:

$$\frac{(RSS_1 - RSS_2)/(p-q)}{RSS_1/(n-p)} \tag{5.1}$$

where RSS_1 and RSS_2 are the residual sum of squares for the full model and the smaller (nested) model respectively, q is the number of covariates in the smaller model and n is the sample size.

This test statistic can then be compared to the critical value $F_{p-q,n-p}$. If the test statistic calculated is greater than the critical value then the the null hypothesis is rejected and it is concluded that the full model is a better fit to the data.

5.3.1.2 The Akaike Information Criterion (AIC)

The AIC can be used to compare the adequacy of two or more models which may or may not be nested. The AIC of a model is calculated using:

$$\text{AIC} = 2p - 2\log L(\hat{\boldsymbol{\beta}}) \tag{5.2}$$

where p is the number of estimated parameters in the model and $L(\hat{\boldsymbol{\beta}}$ is the maximum value of the likelihood function for the model.

When comparing two or more models, the model with the lowest AIC is preferred. Using forward selection, the variable which gives the lowest AIC when included in the model is chosen to be added. The algorithm stops if none of the variables reduce the AIC of the model when added. Using backward selection, the variable which lowers the AIC by the highest value is removed from the model and the algorithm stops when removing any further variables increases the AIC.

5.4 Automatic variable selection

The Least Absolute Shrinkage and Selection Operator (Lasso) is an example of a penalised model selection method. The log likelihood is penalised by subtracting some value, λ, multiplied by the absolute sum of the regression coefficients from it. The Lasso was developed to improve the prediction accuracy and interpretability of regression models by selecting only a subset of the available covariates (Tibshirani, 1996a). The Lasso performs model selection and parameter estimation simultaneously.

5.4.1 Motivation for the use of the Lasso

A linear regression model is of the form

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + e_i, \tag{5.3}$$

where β_0 and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ are unknown parameters and e_i is the error term.

To estimate the parameters using the method of least squares, the following is calculated:

$$\underset{\beta_0,\beta}{\text{minimize}}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_{ij}\right)^2\right\}. \tag{5.4}$$

In general, the use of this method will result in all of the parameter estimates to be non-zero meaning that interpretation of the final model may be difficult and also that the model may overfit the data.

There are two main arguments for using an alternative to the least squares estimate.

1. *Prediction accuracy*: The least squares estimate generally has low bias yet a high variance. Shrinking the values of the regression coefficients (or even having some with a value of zero) can improve prediction accuracy when measured in terms of mean-squared error.

2. *Interpretation*: Instead of using a large number of predictors, it is preferred to use a subset of these predictors which have the strongest effects on the response (Tibshirani, 1996a) .

The lasso uses l_1-regularised regression and obtains parameter estimates by finding solutions to the following:

$$\underset{\beta_0,\beta}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_{ij} \right)^2 \text{ subject to } \|\beta\|_1 \leq \lambda \tag{5.5}$$

where $\|\beta\|_1 = \sum_{i=1}^{p} |\beta_j\|$ is the ℓ_1 norm of β and λ is a parameter that is specified.

The use of the l_1 norm is preferred to that of the l_q since if t is small enough, the lasso gives sparse solution vectors with few non-zero coordinates. This is not the case for l_q-norms with $q > 1$ (Hastie et al., 2015). Therefore a benefit of the lasso is that it conducts variable selection and parameter estimation simultaneously.

5.4.2 Why does the lasso have a model selection property?

To demonstrate the reason behind why the lasso has the model selection property, it can be compared to ridge regression. Ridge regression is another regularisation method which uses a penalty and has the shrinkage property but it does not have the model selection property. The ridge regression method solves a similar optimisation problem to the lasso:

$$\underset{\beta_0,\beta}{\text{minimize}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq \lambda^2. \tag{5.6}$$

The constraint region for ridge regression is a disk $\beta_1^2 + \beta_2^2 \leq \lambda^2$ whereas the constraint region for lasso is a diamond $|\beta_1\| + |\beta_2| \leq \lambda$ (Figure 5.1). Both ridge regression and lasso find the first point where the elliptical contours of the residual sum of squares (centered at the full least squares estimates) intersect with the constraint region. Since unlike a

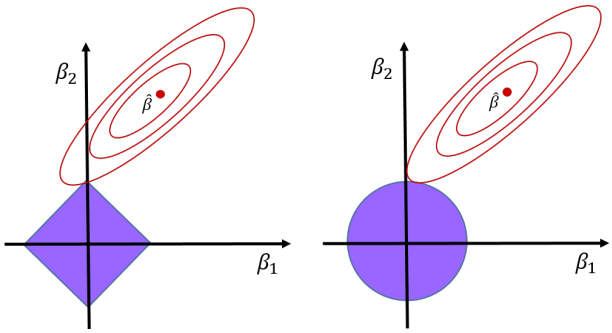

Figure 5.1: Estimation of two parameters for the lasso (left) with the (blue) constraint region given by $|\beta_1| + |\beta_2| \leq \lambda$ and for ridge regression (right) with the constraint region $\beta_1^2 + \beta_2^2 \leq \lambda^2$. The red ellipses are the contours of the residual sum of squares function and are centered at the point $\hat{\beta}$ which is the unconstrained least squares estimate. Adapted from (Hastie et al., 2015).

disk, a diamond has vertices, if the point of intersection is at a vertex then one of the parameters equals zero. When there are more than 2 parameters to be estimated, the constraint region is a rhomboid which gives more chances for the estimates to be equal to zero due to the increased number of flat edges and vertices (Figure 5.2).

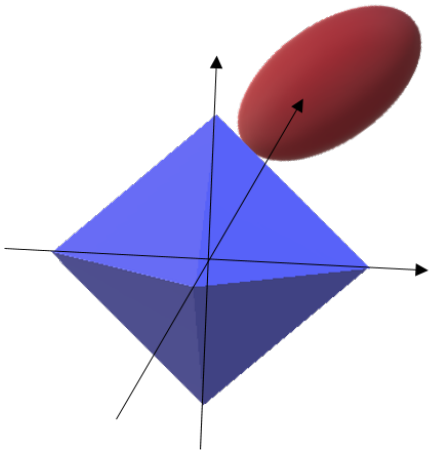

Figure 5.2: An example of a constraint region for the lasso with more than 2 parameters. Adapted from (Hastie et al., 2015).

The lasso has the ability to yield sparse solutions meaning that the solution to the optimisation problem is generally a model with few non-zero coefficients.

5.4.3 The Lasso for Linear Models

With n predictor-response pairs $\{(x_i, y_i)\}_{i=1}^n$, the lasso finds the solution $(\hat{\beta}_0 \hat{\beta})$ to the optimisation problem:

$$\underset{\beta_0,\beta}{\text{minimize}}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2\right\} \text{ subject to } \sum_{j=1}^{p}|\beta_j|\leq\lambda. \tag{5.7}$$

The constraint in (5.7) is equivalent to the ℓ_1-norm constraint in (5.5). The sum of the absolute values of the estimates of the parameters is limited by the bound λ. The magnitude of λ controls how well the model fits to the data with a shrunken parameter estimate corresponding to a more heavily-constrained model.

An alternative way to express Equation 5.7 is in the Lagrangian form:

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}}\left\{\frac{1}{2n}\|\mathbf{y}-\mathbf{X}\beta\|_2^2+\lambda\|\beta\|_1\right\}, \tag{5.8}$$

for some $\lambda\geq 0$. Note, if $\lambda=0$, then this gives the solution to the ordinary least squares problem.

The factor $\frac{1}{2n}$ appearing in Equation 5.7 and Equation 5.8 is often replaced by $\frac{1}{2}$ or 1 to ensure comparability between values of λ across different sample sizes. This has no effect of the results obtained from Equation 5.7 and is simply a re-parametrization of λ in Equation 5.8 (Hastie et al., 2015).

The intercept, β_0, can be omitted from the lasso optimisation during linear regression if the columns of the design matrix and the response variable have been centered (have mean of zero) and the columns of the design matrix have also been standardised (have variance of 1).

5.4.3.1 Cross-validation

The complexity of the model is determined by the value of λ. On one hand, if a large value of λ is chosen, there will be more non-zero coefficients and the the model will fit more closely to the data. This however may lead to a model which over-fits the data. On the other hand, a smaller value of λ will result in fewer non-zero parameter estimates, giving a more sparse model which does not fit to the data as well however is more interpretable.

In general, the aim is to find a value of λ which results in a model with some parameter estimates equal to zero and which also gives a good balance between the two scenarios highlighted above.

Cross-validation is used to estimate the optimal value of λ and be conducted using

the following steps.

1. Randomly divide the full dataset into a K groups, for example 5, 10 or n.
2. Fix one group as the test set and the remaining $K-1$ groups as the training group.
3. Apply the lasso to the training set for a range of λ values.
4. Use each fitted model to predict the responses in the test set.
5. Record the mean squared error for each value of λ.
6. Repeat this process K times with each of the K groups playing the role of the test set once.
7. K different estimates of the prediction error are obtained for each value of λ. Average these K estimates.
8. Plot the cross-validation curve.

When choosing the optimal value of λ there are two methods which can be used. The first selects the value of λ which gives the minimum cross validation error. The second gives the the value of λ which corresponds to a cross validation error of no more than one standard error above it's minimum value.

5.4.3.2 Computation of estimates

For ease of computation, the lasso criterion can be written in the Lagrangian form:

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right\}. \tag{5.9}$$

Assuming that the data has been centered and standardised we can omit the intercept.

5.4.3.3 Soft-Thresholding (Single Predictor)

For a single predictor x_i, based on samples $\{(x_i,y_i)\}_{i=1}^n$, the problem to solve becomes:

$$\underset{\beta}{\text{minimize}}\left\{\frac{1}{2n}\sum_{i=1}^{n}(y_i-x_i\beta)^2+\lambda|\beta|\right\}. \tag{5.10}$$

Since $|\beta|$ does not have a derivative at $\beta = 0$, there is an issue in the use of the standard method of taking the first derivative of the function with respect to β and setting it equal to zero. The problem however can be solved using direct inspection of the function with:

$$\hat{\beta} = \begin{cases} \frac{1}{n}\langle \mathbf{x}, \mathbf{y} \rangle - \lambda & \text{if} \quad \langle \mathbf{x}, \mathbf{y} \rangle > \lambda, \\ 0 & \text{if} \quad \langle \mathbf{x}, \mathbf{y} \rangle \leq \lambda, \\ \frac{1}{n}\langle \mathbf{x}, \mathbf{y} \rangle + \lambda & \text{if} \quad \langle \mathbf{x}, \mathbf{y} \rangle < \lambda \end{cases} \tag{5.11}$$

where $\langle \mathbf{x}, \mathbf{y} \rangle$ is the least squares estimate of β.

This can also be written

$$\hat{\beta} = \mathcal{S}_\lambda(\frac{1}{n}\langle \mathbf{x}, \mathbf{y} \rangle) \quad \text{where} \quad \mathcal{S}_\lambda(z) = \text{sign}(z)(|z| - \lambda)_+. \tag{5.12}$$

$\mathcal{S}_\lambda$ is known as the soft-thresholding operator and moves z towards zero by λ and sets z equal to zero if it's magnitude is less than or equal to λ.

If the data is standardised, then $\hat{\beta} = \mathcal{S}_\lambda(\frac{1}{n}\langle \mathbf{z}, \mathbf{y} \rangle)$ is a soft-thresholded version of the usual least-squares estimate $\tilde{\beta} = \langle \mathbf{x}, \mathbf{y} \rangle$.

5.4.3.4 Cyclic coordinate descent (Multiple Predictors)

In the multivariate case, the predictors are cycled through in a fixed, arbitrary order, updating β_j at the j^{th} step. The objective function is minimised at the corresponding coordinate whilst the other coefficients $(\beta_k, k \neq j)$ are fixed. The objective function (5.8) can be written:

$$\frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \sum_{k \neq j} x_{ik}\beta_k - x_{ij}\beta_j\right)^2 + \lambda \sum_{k \neq j} |\beta_k| + \lambda|\beta_j|. \tag{5.13}$$

The partial residual is expressed as $r_i^{(j)} = y_i - \sum_{k \neq j} x_{ik}\hat{\beta}_k$ and so the update for the j-th coefficient, β_j can be written in terms of it's partial residual:

$$\hat{\beta}_j = \mathcal{S}_\lambda\left(\frac{1}{n}\langle \mathbf{x}_j, \boldsymbol{r}^{(j)} \rangle\right). \tag{5.14}$$

Alternatively, this update can be expressed as:

$$\hat{\beta}_j \leftarrow \mathcal{S}_\lambda\left(\hat{\beta}_j + \frac{1}{n}\langle \mathbf{x}_j, \boldsymbol{r}\rangle\right), \tag{5.15}$$

where $r_i = y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j$ are the full residuals.

The soft-thresholding update (5.14) is applied repeatedly meaning the coordinates of $\hat{\beta}$ and also the residual vectors are updated recurrently (Hastie et al., 2015).

It is often required that a lasso solution is found for varying values of λ. One way in which this can be done is to take the value of λ which gives an all-zeroes vector as the optimal solution and begin with this. This value of λ is given by $\lambda_{max} = \max_j |\frac{1}{n}\langle \mathbf{x}_j, \mathbf{y}\rangle|$. The value of λ can then be decreased by a small amount and coordinate descent can be run until convergence. This value can then be further decreased and the solutions from the previous value can be used as a starting point. Coordinate descent can then be run until convergence. This method is called pathwise coordinate descent and calculates solutions over a range of values of λ.

There are two main advantages to using coordinate descent to obtain solutions to the lasso. The first is that it is quick due to the fact that the coordinate-wise minimisers are available explicitly meaning that searching iteratively for each coordinate is not necessary. The second is that it utilises the sparsity. If λ is large enough, most of the coefficients will be (and remain) equal to zero.

5.4.4 The Lasso for Generalised Linear Models

The lasso was originally created for linear models however it can be extended to generalised linear models. This maximises the likelihood (or minimises the negative log-likelihood) together with an ℓ_1-penalty:

$$\underset{\beta_0,\beta}{\text{minimize}}\left\{ -\frac{1}{n}\mathcal{L}(\beta_0, \beta; \mathbf{y}, \mathbf{X}) + \lambda\|\beta\|_1 \right\} \tag{5.16}$$

where $\mathbf{y}$ is the outcome vector, $\mathbf{X}$ is the matrix of predictors and $\mathcal{L}$ is the specific form of the log-likelihood which changes corresponding to the generalised linear model.

5.4.4.1 Logistic Regression

Given a binary response $y_i \sim \text{Bernoulli}(\mu_i), i = 1, \ldots, n$, consider the logistic model .

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta^T \boldsymbol{x}_i,$$

the negative log-likelihood with ℓ_1-regularisation is given by:

$$-\frac{1}{n}\sum_{i=1}^{n}\{y_i \log(P(Y=1|\boldsymbol{x}_i) + (1-y_i)\log(1-P(Y=1|\boldsymbol{x}_i)\} + \lambda\|\beta\|_1. \tag{5.17}$$

Since $P(Y=1|x_i) = \beta_0 + \beta^T x_i$ this can be written as

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{y_i(\beta_0 + \beta^T \boldsymbol{x}_i) - \log(1+e^{\beta_0+\beta^T \boldsymbol{x}_i})\right\} + \lambda\|\beta\|_1. \tag{5.18}$$

5.4.4.2 Algorithm for computation

When computing the lasso solutions for a logistic model, the proximal-Newton iterative approach is widely used. This involves repeatedly approximating the negative log-likelihood using a quadratic function (Lee et al., 2014). This method is combined with coordinate descent to compute parameter estimates.

More specifically, if the current estimates of the parameters are (β_0^*, β^*) then a quadratic approximation to the unpenalised log-likelihood can be calculated as

$$\ell_Q(\beta_0, \beta) = \frac{1}{2n}\sum_{i=1}^{n} w_i(z_i - \beta_0 - \beta^T \boldsymbol{x_i})^2 + C(\beta_0^*, \beta^*) \tag{5.19}$$

where $C(\beta_0^*, \beta^*)$ is a constant independent from β_0^* and β^* and

$$z_i = \beta_0^* + \beta^{*T}\boldsymbol{x}_i + \frac{y_i - \mu_i^*}{\mu_i^*(1-\mu_i^*)} \quad \text{and} \quad w_i = \mu_i^*(1-\mu_i^*)$$

with μ_i^* defined as $P(Y=1|\boldsymbol{x}_i)$ evaluated at the current parameters. The Newton update is found by minimising ℓ_Q.

In order to minimise the penalised log-likelihood, coordinate descent is used to solve the following:

$$\underset{\beta\in\mathbb{R}^p}{\text{minimize}}\{-\ell_Q(\beta_0, \beta) + \lambda P_\alpha(\beta)\} \tag{5.20}$$

where $P_\alpha(\beta) = (1-\alpha)\frac{1}{2}\|\beta\|_{\ell_2}^2 + \alpha\|\beta\|_{\ell_1}$ is a compromise between ridge regression

when $\alpha = 0$ and the lasso when $\alpha = 1$ and is known as the elastic net penalty (Friedman et al., 2010).

5.4.5 The Group Lasso

The lasso was designed for use with continuous covariates. If there is a categorical variable, the lasso considers each category separately which can result in the final model containing only certain categories from the group. It is desirable however to either keep or remove all of the coefficients of a grouped variable simultaneously. There is an extension called the group lasso which deals with this.

Given a linear regression model:

$$Y = \beta_0 + \sum_{j=1}^{J} X_j \beta_j + \epsilon$$

which contains J groups of covariates and X_j denotes the covariates in group $j \in J$, the group lasso solves

$$\underset{\beta_0 \in \mathbb{R}, \beta_j \in \mathbb{R}^{p_j}}{\text{minimise}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{J} X_{ij} \beta_j)^2 + \lambda \sum_{j=1}^{J} \|\beta_j\|_2 \right\}$$

where $\|\beta_j\|_2$ is the Euclidean norm of the vector β_j.

This adaptation of the lasso has the properties that depending on the penalty, either the entire vector $\hat{\beta}_j$ will be equal to zero or the will be non-zero and that if all groups have only one factor then the problem reduces to the original lasso described above.

In Equation 5.4.5 all groups are penalised by the same amount and so it may mean that groups with more levels are more likely to be selected than groups with fewer levels. This can be overcome by weighting the penalties by a factor of $\sqrt{df_j}$ where df_j is the degrees of freedom of the j-th of the j-th predictor (Yuan and Lin, 2006).

5.4.5.1 Computation of estimates

Computations of estimates for the group lasso are obtained using block coordinate descent. This is similar to cyclic coordinate descent used in the linear regression setting. The difference between the two methods however, is that instead of fixing all but one coefficient, during block coordinate descent all but one of the vectors $\hat{\beta}_j$ is fixed (Yang and Zou, 2013).

5.4.6 The Grouped Logistic Lasso

Combining the two previous methods, the lasso can be further extended to provide estimates for a logistic model with categorical variables.

For a set of n observations $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$ where y takes the form of a binary response variable and $\boldsymbol{x}_i$ contains J groups of predictors, the group lasso for logistic regression solves:

$$\underset{\beta_0 \in \mathbb{R}, \beta_j \in \mathbb{R}^{p_j}}{\text{minimise}} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left\{ y_i(\beta_0 + \beta^T \boldsymbol{x}_i) - \log(1 + e^{\beta_0 + \beta^T \boldsymbol{x}_i}) \right\} + \lambda \sum_{j=1}^{J} s(df_j) \|\beta_j\|_2 \right\} \quad (5.21)$$

where $\|\beta_j\|_2$ is the Euclidean norm of the vector β_j and the function $s(df_j)$ alters the penalty relative to the size of the vector $\hat{\beta}_j$ (Meier et al., 2008). As with in the group lasso for linear regression, generally $s(df_j) = \sqrt{df_j}$.

5.4.7 Computation of estimates

Similarly to grouped linear regression, when computing estimates using grouped logistic lasso, block coordinate descent can be used. One at a time, all but one of the parameter groups is fixed and Equation 5.21 is solved for the remaining parameter group $\boldsymbol{\beta}_j$. In order to find the solution, a Newton iterative approach such as described above can be used (Meier et al., 2008).

5.4.8 R packages for computation of the Lasso

There are a variety of R packages in order use the lasso including `glmnet`, `grpreg`, `grplasso` and `gglasso`.

The `glmnet` package was developed to fit generalized linear models using a penalised maximum likelihood. Therefore, this package can be used for computing estimates using the lasso for linear regression and logistic regression. It can be used to compute estimates using a lasso penalty as well as a ridge penalty or an elastic net penalty.

The `grpreg` package was created to fit generalised linear models with grouped penalties. Therefore, this package can also be used for either a continuous response or a binary response. The penalties that can be applied when using this package include

the group lasso, group minimax concave penalty (MCP) and group smoothly clipped absolute deviation (SCAD).

The `grplasso` package and the `gglasso` were also developed in order to fit a generalised linear model using a grouped lasso penalty.

Since interest lies in survey data, it may be found necessary to include sampling weights during analysis. Two of the packages listed (`grpreg` and `gglasso`) do not allow for the inclusion of sampling weights. Also, many responses to survey questions take the form of a categoric or factor variable and so grouped lasso will be needed. The `glmnet` package does not allow for grouped lasso.

When computing estimates the size of the penalty, λ, needs to be determined. The optimal value of λ is generally found using cross-validation. The `grplasso` does not conduct cross validation.

Table 5.1: The features of the glmnet, grpreg, grplasso and gglasso packages.

	Lasso	Logistic	Group	Weights	Cross validation
glmnet	Yes	Yes	No	Yes	Yes
grpreg	Yes	Yes	Yes	No	Yes
grplasso	Yes	Yes	Yes	Yes	No
gglasso	Yes	Yes	Yes	No	Yes

Table 5.1 summarises whether each of the packages discussed above can include each of the features which will be required when conducting the grouped logistic lasso for survey data. It can be seen that none of the packages include all of the necessary features and so one of them will have to be adapted in order to conduct analysis of the PNS and MCS data.

5.5 Simulations

5.5.1 Comparing variable selection methods

Using step-wise selection with AIC is a very common method of variable selection. This section will compare the results of this method with the results when the lasso is used. When using the lasso, two different values of λ will be compared. The first of these values will be the value of λ which gives the minimum cross validation error. The second of these values will be chosen using the "one standard error" rule which gives the smallest value of λ which corresponds to a cross-validation error of no more than one standard

error away from the minimum cross-validation error.

When choosing the true values of β, some of them will be set to zero. Then when comparing the two methods, the number of false positives and false negatives will be calculated. A false positive occurs when the coefficient estimate is not equal to zero when the true value is equal to zero. A false negative occurs when the coefficient estimate is equal to zero when the true value is not.

The bias and mean squared error will also be calculated. The average of each will be calculated first for the whole model, next for the coefficients whose true values are non-zero and finally for the coefficients whose true values are zero.

The two methods will be compared over a variety of scenarios. In the first, the number of variables will be changed, starting with a small number and increasing until there is a large number of possible variables which could be included in the model.

Next, the sample size will be increased to see whether or not this affects how many false positives and false negatives are given by the two methods.

Finally, the error used when simulating the response variable will be varied and the results of the two methods will be compared when there is low variability in the response as well as a larger variability.

5.5.1.1 Changing number of variables

For the following simulations both the response and the explanatory variables will be continuous and the sample size will remain fixed at 500. When simulating the response variable the variance of the error term will be chosen to be 5% of the variance of $X\beta$, where X is the design matrix containing all of the possible covariates and β is the true values of the coefficients (some of which will be equal to zero). First, a small amount of covariates will be used and then the number will be increased. The number of possible covariates used will be 10, 20, 50 and 100. Although 100 seems like a large number of variables to consider, as technology gets more advanced and surveys become more accessible it is becoming increasingly common for surveys to collect large amounts of information from participants (Winerman, 2018).

Some of the true coefficient values will be set to zero in order to see whether or not this is detected by either of the two methods. For each number of covariates, different amounts (30%, 50% and 70%) of coefficients with true value zero will be used. Each

time 200 repeated simulations will be run and the average number of false positives and negatives will be calculated by comparing the coefficient estimates with the true values. Mean squared error and bias will also be calculated.

When looking at the average number of false positives and negatives in Table 5.2 it can be seen that none of the three methods produce any false negatives with one exception. When using 100 potential covariates, with the true coefficients of 30 of them zero, the average number of false positives when using the lasso with the one standard error rule is 0.025.

When looking at the false positives, the lasso using the one standard error rule to determine λ gives on average the least when starting with 10 or 20 possible covariates. When starting with 50 or 100 covariates, stepwise selection using the AIC gives the least amount of false positives when the coefficients of 30% or 50% of them are zero. When 70% of the true coefficients are equal to zero, the lasso using the one standard error rule has the lowest mean number of false positives.

In all scenarios when changing the number of covariates, the lasso using the value of λ which minimises the cross validation error was found to give the most average number of false positives.

When looking at the bias in Table 5.3 it can be seen that step wise selection using AIC gives the lowest total bias and the lowest bias for the coefficients whose true values are non-zero. When looking at the bias for the coefficients whose true values are equal to zero, it can be seen that the lasso using the one standard error rule to determine λ gives the lowest bias. This is the true for all number of covariates regardless of how many of the true coefficient values are equal to zero.

The mean squared errors calculated whilst varying the number of covariates can be found in Table 5.4. Generally, stepwise selection using AIC results in the lowest total mean squared error. The exceptions to this occur when 70% of the true coefficients are equal to zero and there are 20, 50 or 100 covariates. When this is the case, the lasso using the value of λ which results in the minimum cross validation error gives the lowest total mean squared error.

Stepwise selection using the AIC also results in the lowest mean squared error when considering only the coefficients whose true values are non-zero. The only exception to this is when there are 50 covariates and 70% of the true coefficients are equal to zero. In

this case, the lasso using the value of λ which results in the minimum cross validation error gives the lowest mean squared error.

The lasso using the one standard error rule to select λ gives the lowest mean squared error when looking at the coefficients whose true values are equal to zero. This is the case regardless of how many covariates are used and how many of the true coefficients are equal to zero.

Table 5.2: The average number of false negatives and false positives when AIC and Lasso are used and the number of potential variables is changing

	Average number of false negatives			Average number of false positives		
10 variables						
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
3	0	0	0	0.505	1.870	0.150
5	0	0	0	0.775	2.555	0.085
7	0	0	0	1.150	2.245	0.045
20 variables						
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
6	0	0	0	1.005	4.555	0.920
10	0	0	0	1.575	6.005	0.865
14	0	0	0	2.320	1.845	0.120
50 variables						
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	0	0	0	2.855	11.680	5.030
25	0	0	0	4.550	15.975	5.695
35	0	0	0	6.250	16.220	3.895
100 variables						
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
30	0	0	0.025	7.345	23.925	14.165
50	0	0	0	10.255	32.435	17.655
70	0	0	0	14.380	33.470	13.750

5.5.1.2 Changing sample size

For the following simulations both the response and the explanatory variables will be continuous and the number of variables will remain fixed at 50. When simulating the response variable the variance of the error term will be chosen to be 5% of the variance of $X\beta$, where X is the design matrix containing all of the possible covariates and β is the true values of the coefficients (some of which will be equal to zero). First, a small sample size will be used and then the number will be increased. The sample sizes used will be 100, 200, 500, and 1000.

Some of the true coefficient values will be set to zero in order to see whether or not this is detected by either of the two methods. For each sample size, different amounts (30%, 50% and 70%) of coefficients with true value zero will be used. Once again, 200 repeated simulations will be run each time and the average number of false positives and

Table 5.3: The total model bias, the bias for the non-zero coefficients and for the zero coefficients when AIC and Lasso are used and the number of variables is changing

	Total mean bias			Non-zero coefficient bias			Zero coefficient bias		
10 variables									
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
3	-0.003	-0.018	-0.070	-0.089	-0.470	-1.755	0.029	0.045	0.004
5	-0.002	-0.026	-0.100	-0.051	-0.878	-3.316	-0.009	-0.005	-0.001
7	0.001	-0.014	-0.048	0.002	-0.714	-2.392	0.018	0.011	0.003
20 variables									
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
6	-0.003	-0.042	-0.195	-0.049	-0.541	-2.606	0.033	-0.062	0.023
10	-0.011	-0.054	-0.181	0.128	-0.929	-3.277	-0.079	-0.057	-0.01834
14	0.003	-0.012	-0.046	0.033	-0.372	-1.322	0.027	0.017	-0.001
50 variables									
Number of true zeros	AIC	Lasso	Lasso 1se	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	0.041	-0.183	-0.790	0.212	-1.011	-4.393	0.034	-0.012	0.006
25	0.015	-0.181	-0.537	0.077	-1.458	-4.133	0.044	0.068	0.006
35	0.006	-0.117	-0.284	0.013	-1.520	-3.587	0.029	0.025	0.017
100 variables									
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
30	-0.105	-1.05	-2.737	-0.138	-2.733	-7.567	-0.372	-0.532	-0.338
50	0.002	-0.792	-1.740	-0.066	-3.199	-6.885	0.077	0.094	0.063
70	-0.005	0.537	-1.011	0.107	-3.314	-6.471	-0.061	-0.068	-0.023

Table 5.4: The total mean squared error, the mean squared error for the non-zero coefficients and for the zero coefficients when AIC and Lasso are used and the number of variables is changing

	Total MSE			Non-zero coefficient MSE			Zero coefficient MSE		
10 variables									
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
3	0.00115	0.00123	0.00454	0.00012	0.00014	0.00057	0.00006	0.00049	0.00000
5	0.00057	0.00064	0.00239	0.00007	0.00009	0.00040	0.00004	0.00002	0.00000
7	0.00021	0.00022	0.00088	0.00003	0.00004	0.00022	0.00001	0.00001	0.00000
20 variables									
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
6	0.00855	0.00956	0.02322	0.00049	0.00053	0.00152	0.00028	0.00031	0.00003
10	0.00421	0.00475	0.01209	0.00026	0.00032	0.00109	0.00014	0.00012	0.00001
14	0.00159	0.00147	0.00407	0.00010	0.00015	0.00061	0.00006	0.00004	0.00000
50 variables									
Number of true zeros	AIC	Lasso		AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	0.14571	0.15575	0.23791	0.00320	0.00344	0.00636	0.00203	0.00213	0.00059
25	0.06379	0.06663	0.10683	0.00153	0.00185	0.00395	0.00096	0.00074	0.00017
35	0.02267	0.02045	0.03622	0.00621	0.00090	0.00221	0.00364	0.00017	0.00002
100 variables									
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
30	1.25062	1.33196	1.70577	0.01395	0.01497	0.02237	0.00868	0.00897	0.00392
50	0.55952	0.56748	0.72929	0.00668	0.00794	0.01308	0.00438	0.00325	0.00124
70	0.22971	0.19389	0.27113	0.00295	0.00432	0.00821	0.00198	0.00086	0.00024

negatives will be calculated along with the bias and mean squared errors.

The average numbers of false positives and negatives produced for the three methods when changing the sample size can be found in Table 5.5. It can be seen it is only when the sample size is 100 that any false negatives are produced by any of the three methods. When the sample size is 100, the lasso using the value of λ which results in the lowest cross validation error has the lowest average number of false negatives. When looking at the false positives, when the sample size is 100, stepwise selection using AIC results in the least. When the sample size is 200 or 500 and 30% or 50% of the true coefficients are equal to zero, step wise selection once again results in the lowest average number of false positives. When the sample size is 200 or 500 and 70% of the true coefficient values are equal to zero, the lasso using the one standard error rule results in the lowest number of false positives. When the sample size is 1000 and 30% of the true coefficients are equal to zero, stepwise selection using AIC results in the lowest number of false positives. When 50% or 70% of the true coefficients are equal to zero, the lasso using the one standard error rule results in the lowest average number of false positives.

The bias calculated when varying the sample size can be found in Table 5.6. Similarly to when changing the number of covariates, step wise selection using AIC gives the lowest total bias and the lowest bias for the coefficients whose true values are non-zero. For the coefficients whose true values are equal to zero, it can be seen that the lasso using the one standard error rule to determine λ gives the lowest bias. This is true for all number of covariates regardless of how many of the true coefficient values are equal to zero with one exception. With a sample size of 100 and 50% of the true coefficients equal to zero, stepwise selection using the AIC results in the lowest bias.

The mean squared errors calculated whilst varying the number of covariates can be found in Table 5.7. When the sample size is 100, the lasso using the value of λ which results in the lowest cross validation error gives the lowest total mean squared error. This method also results n the lowest mean squared error when 70% of the true coefficients are equal to zero regardless of the sample size used. When the sample size is 200, 500 or 1000 and 30% or 50% of the true coefficients are equal to zero, stepwise selection using AIC results in the lowest total mean squared error.

Stepwise selection using the AIC also results in the lowest mean squared error when considering only the coefficients whose true values are non-zero. The exceptions to this

are when the sample size is equal to 100 and 30% or 50% of the true coefficients are equal to zero and when the sample size is equal to 500 and 70% of the true coefficients are equal to zero. In these cases, the lasso using the value of λ which results in the minimum cross validation error gives the lowest mean squared error.

The lasso using the one standard error rule to select λ gives the lowest mean squared error when looking at the coefficients whose true values are equal to zero. This is the case regardless of how many covariates are used and how many of the true coefficients are equal to zero.

Table 5.5: The average number of false negatives and false positives when AIC and Lasso are used and the sample size is changing

	Average number of false negatives			Average number of false positives		
sample size 100						
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	0.425	0.075	0.240	5.095	12.245	8.775
25	0.040	0.020	0.004	8.180	16.505	10.785
35	0	0	0	11.355	25.480	17.760
sample size 200						
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	0	0	0	3.345	12.075	6.950
25	0	0	0	5.530	15.850	8.125
35	0	0	0	7.680	16.425	5.935
sample size 500						
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	0	0	0	2.855	11.680	5.030
25	0	0	0	4.550	15.975	5.695
35	0	0	0	6.250	16.220	3.895
sample size 1000						
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	0	0	0	2.460	11.460	3.640
25	0	0	0	3.985	16.120	3.790
35	0	0	0	5.765	15.905	2.465

5.5.1.3 Changing error

For the following simulations both the response and the explanatory variables will be continuous, the number of variables will remain fixed at 50 and the sample size will remain fixed at 500. When simulating the response variable the variance of the error term will be chosen to be $Z\%$ of the variance of $X\beta$, where X is the design matrix containing all of the possible covariates, β is the true values of the coefficients and Z is a value to be varied. The values that Z will take during these simulations is 5, 10, 20 and 50.

Once again, some of the true coefficient values will be set to zero in order to see whether or not this is detected by either of the two methods. For each change in Z,

Table 5.6: The total model bias, the bias for the non-zero coefficients and for the zero coefficients when AIC and Lasso are used and the sample size is changing

	Total mean bias			Non-zero coefficient bias			Zero coefficient mean bias		
sample size 100									
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	-0.007	-0.589	-1.472	-0.663	-3.738	-8.558	1.501	1.115	0.912
25	0.071	-0.430	-0.913	0.359	-3.629	-7.299	0.194	0.334	0.283
35	0.002	-0.321	-0.555	-0.038	-3.884	-6.886	0.026	-0.058	-0.023
sample size 200									
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	0.039	-0.290	-1.035	0.222	-1.664	-5.777	-0.072	0.128	0.070
25	-0.029	-0.349	-0.762	-0.132	-2.640	-5.856	-0.097	-0.049	-0.006
35	0.004	-0.202	-0.403	-0.076	-2.590	-5.118	0.055	0.028	0.000
sample size 500									
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	0.041	-0.183	-0.790	0.212	-1.011	-4.393	0.034	-0.012	0.006
25	0.015	-0.181	-0.537	0.077	-1.458	-4.133	0.044	0.068	0.006
35	0.006	-0.117	-0.284	0.013	-1.520	-3.587	0.029	0.025	0.017
sample size 1000									
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	-0.016	-0.171	-0.692	-0.008	-0.858	-3.799	-0.087	-0.214	-0.112
25	0.015	-0.121	-0.422	0.090	-0.942	-3.242	0.024	0.010	-0.008
35	-0.007	-0.092	-0.234	-0.084	-1.150	-2.930	0.001	-0.002	0.001

Table 5.7: The total mean squared error, the mean squared error for the non-zero coefficients and for the zero coefficients when AIC and Lasso are used and the sample size is changing

	Total MSE			Non-zero coefficient MSE			Zero coefficient MSE		
sample size 100									
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	1.44376	1.36563	1.59161	0.02969	0.02967	0.03935	0.02499	0.01983	0.01166
25	0.61633	0.55830	0.67126	0.01356	0.01535	0.02248	0.01055	0.00637	0.00347
35	0.22801	0.14742	0.19106	0.00502	0.00625	0.01076	0.00414	0.00136	0.00054
sample size 200									
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	0.42856	0.45911	0.58947	0.00935	0.01007	0.01528	0.00612	0.00644	0.00262
25	0.19356	0.19670	0.26625	0.00450	0.00549	0.00949	0.00306	0.00216	0.00079
35	0.06995	0.05792	0.08511	0.00179	0.00255	0.00504	0.00118	0.00049	0.00013
sample size 500									
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	0.14571	0.15575	0.23791	0.00320	0.00344	0.00636	0.00203	0.00213	0.00059
25	0.06379	0.06663	0.10683	0.00153	0.00185	0.00395	0.00096	0.00074	0.00017
35	0.02267	0.02045	0.03622	0.00621	0.00090	0.00221	0.00364	0.00017	0.00002
sample size 1000									
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	0.06493	0.07105	0.13160	0.00144	0.00157	0.00358	0.00086	0.00097	0.00017
25	0.02892	0.03114	0.05984	0.00072	0.00088	0.00226	0.00040	0.00034	0.00004
35	0.01039	0.00981	0.02117	0.00029	0.00043	0.00131	0.00017	0.00008	0.00001

different amounts (30%, 50% and 70%) of coefficients with true value zero will be used. Each time 200 repeated simulations will be run and the average number of false positives and negatives will be calculated as well as the bias and mean squared error.

The average numbers of false negatives and false positives produced when varying the error used to simulate the response variable can be found in Table 5.8. When Z is chosen to be 5% the average number of false negatives for all methods is zero. When $Z = 10$, stepwise selection using AIC and the lasso using the one standard error rule produce a very small number of false negatives. When Z is equal to 20 or 50, the lasso using the value of λ which results in the minimum cross validation error gives the lowest average number of false negatives.

When looking at the false positives, when Z is equal to 5, 10 or 20 and 30% or 50% of the true coefficients are equal to zero, stepwise selection using the AIC results in the lowest number of false positives. When 70% of the true coefficients are equal to zero, the lasso using the one standard error rule results in the lowest false positives. When $Z = 50$, the lasso using the one standard error rule results in the lowest average number of false positives regardless of how many of the coefficients true values are equal to zero.

The bias calculated for each of the three methods when varying the error used when simulating the response variable can be found in Table 5.9. The total bias and the bias for the coefficients whose true values are non-zero is lowest when stepwise selection with AIC is used regardless of the error used or the number of coefficients whose true value is equal to zero. Generally, when looking at the bias for the coefficients whose true values are equal to zero, the lasso using the one standard error rule gives the lowest. The exceptions to this occur when $Z = 10$ and 30% or 50% of the true coefficients are equal to zero and when $Z = 50$ and 30% of the true coefficients are equal to zero. In these three instances, stepwise selection using AIC gives the lowest bias.

The mean squared errors produced for the three different methods when changing the error used when simulating the response variable can be found in Table 5.10. In general, the lasso using the value of λ which gives the minimum cross validation error results in the lowest total mean squared error. The exceptions are when $Z = 5$ and 30% of the coefficients have true value equal to zero and when $Z = 10$ and 30% or 50% of the true coefficients are equal to zero. In these cases, stepwise selection using AIC gives the lowest total mean squared error. A further exception is when $Z = 10$ and 70% of the

true coefficients are equal to 70 with the lasso using the one standard error rule resulting in the lowest total mean squared error.

When looking at the mean squared error of the coefficient whose true values are non-zero, stepwise selection using AIC generally gives the lowest. The exceptions to this occur when $Z = 5$ and 70% of the true coefficients are zero, when $Z = 20$ and 30% of the true coefficients are zero and when $Z = 50$ and 30% or 50% of the true coefficients are zero. In these instances, the lasso using the value of λ which results in the minimum cross validation error gives the lowest mean squared error.

When looking at the mean squared error for the coefficients whose true values are equal to zero, the lasso using the one standard error rule gives the lowest average value for all errors and regardless of how many of the true coefficients are equal to zero.

Table 5.8: The average number of false negatives and false positives when AIC and Lasso are used and the error used when simulating the response variable is changing.

	Average number of false negatives			Average number of false positives		
error 5%						
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	0	0	0	2.86	11.68	5.03
25	0	0	0	4.55	15.98	5.70
35	0	0	0	6.25	16.22	3.90
error 10%						
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	0.03	0	0.01	2.76	11.85	4.99
25	0	0	0	4.95	16.55	5.48
35	0	0	0	6.21	15.93	1.03
error 20%						
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	3.09	0.38	2.33	2.75	11.24	4.47
25	0.35	0.03	0.39	4.54	15.91	5.40
35	0	0	0.01	6.02	15.93	3.65
error 50%						
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	16.22	9.20	26.43	2.90	6.85	0.82
25	7.38	3.41	12.08	4.58	12.26	2.16
35	1.57	0.76	3.01	6.13	14.52	2.45

5.5.2 Centering and standardisation

When conducting the lasso the design matrix can be centered and standardised. When centering, each of the columns of the design matrix will be altered in order to have a mean of 0. When standardising, the design matrix is orthonormalised such that $X^T X = 1$. For group lasso, standardisation orthonormalises the design matrix such that $X_g^T X_G = nI_{df_g}$, where X_g is a matrix made up of the columns of the design matrix corresponding to the gth predictor, df_g is the degrees of freedom of the g-th predictor and n is the sample

Table 5.9: The total bias, the bias for the non-zero coefficients and for the zero coefficients when AIC and Lasso are used and the error used when simulating the response variable is changing

	Total bias			Non-zero coefficient bias			Zero coefficient bias		
error 5%									
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	0.041	-0.183	-0.790	0.212	-1.011	-4.393	0.034	-0.012	0.006
25	0.015	-0.181	-0.537	0.077	-1.458	-4.133	0.044	0.068	0.006
35	0.006	-0.117	-0.284	0.013	-1.520	-3.587	0.029	0.025	0.017
error 10%									
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	0.078	-0.356	-1.591	0.443	-1.923	-8.777	-0.024	-0.135	-0.148
25	0.053	-0.321	-1.060	0.400	-2.539	-8.078	0.006	0.070	-0.076
35	-0.047	-0.276	-0.358	-0.197	-3.287	-4.508	-0.181	-0.073	0.0175
error 20%									
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	-0.035	-0.862	-3.392	-0.230	-4.757	-18.868	0.089	-0.074	0.051
25	0.049	-0.736	-2.192	0.233	-5.907	-16.874	0.149	0.257	0.015
35	0.024	-0.468	-1.138	0.054	-6.063	-14.366	0.115	0.096	0.065
error 50%									
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	-1.455	-4.305	-10.567	-8.126	-24.011	-58.638	0.103	0.230	-0.161
25	-0.437	-2.563	-6.707	-3.796	-20.119	-51.532	0.451	0.417	-0.059
35	-0.029	-1.257	-3.015	-1.353	-16.241	-37.877	0.454	0.243	0.084

Table 5.10: The total mean squared error, the mean squared error for the non-zero coefficients and for the zero coefficients when AIC and Lasso are used and the error used when simulating the response variable is changing

	Total MSE			Non-zero coefficient MSE			Zero-coefficient MSE		
error 5%									
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	0.14571	0.15575	0.23791	0.00320	0.00344	0.00636	0.00203	0.00213	0.00059
25	0.06379	0.00666	0.10683	0.00153	0.00185	0.00395	0.00096	0.00074	0.00017
35	0.02267	0.02045	0.03622	0.00621	0.00090	0.00221	0.00364	0.00017	0.00002
error 10%									
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	0.56820	0.61115	0.94614	0.01246	0.01376	0.02529	0.00797	0.00852	0.00239
25	0.25655	0.26875	0.42807	0.00624	0.00741	0.01584	0.00377	0.00305	0.00064
35	0.09087	0.08179	0.04507	0.00235	0.00353	0.00259	0.00152	0.00072	0.00000
error 20%									
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	2.80056	2.47354	3.87865	0.06435	0.05552	0.10457	0.03227	0.03166	0.00761
25	1.07106	1.06645	1.72915	0.02629	0.02967	0.06408	0.01550	0.01180	0.00253
35	0.36275	0.32718	0.58100	0.00993	0.01432	0.03550	0.00583	0.00280	0.00037
error 50%									
Number of true zeros	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se	AIC	Lasso min	Lasso 1se
15	18.69641	12.56565	19.98806	0.43130	0.30824	0.55311	0.21131	0.09795	0.00508
25	8.00168	5.88259	10.36977	0.21261	0.17663	0.39420	0.09895	0.05161	0.00482
35	2.51192	1.96554	3.59531	0.07631	0.08935	0.22156	0.03689	0.01531	0.00144

size.

Using a simulated data set, the difference in the results of the group logistic lasso can be compared between when centering and standardisation are used and when they are not.

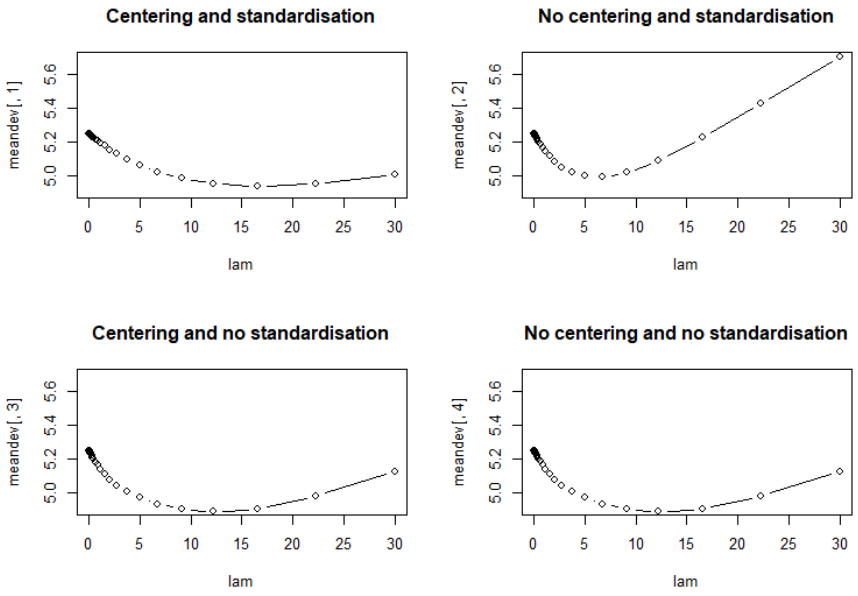

Figure 5.3: Cross validation error curves for the same dataset varying whether the design matrix is centered and standardised.

The default for centering and standardisation using the `grplasso` command is `TRUE` for both. Figure 5.3 shows the cross validation curve for the four different options of whether the design matrix is centered, standardised, both or neither.

The cross validation curve for when both centering and standardisation is set to `FALSE` is the same as when standardising is set to `FALSE` and centering is left as default. It appears that in order to be centered, the data must first be standardised and so the `grplasso` command treats both of these instances the same when computing.

It can also be seen that the main difference between the four plots is when the data is standardised but not centered. Therefore it is implied that centering and standardisation should be used together or neither should be used but not one without the other.

From Figure 5.3 it can be seen that there is very little difference between the curves for when both centering and standardisation is used and when neither is used. The value of λ which gives the minimum cross validation error in each case is 16.52 and 12.26 respectively.

Since it is not clear what the difference in results is between when centering and stan-

dardisation is used and when it is not with one replication, we make repeated simulations to study the effect on the coefficient estimates more systematically.

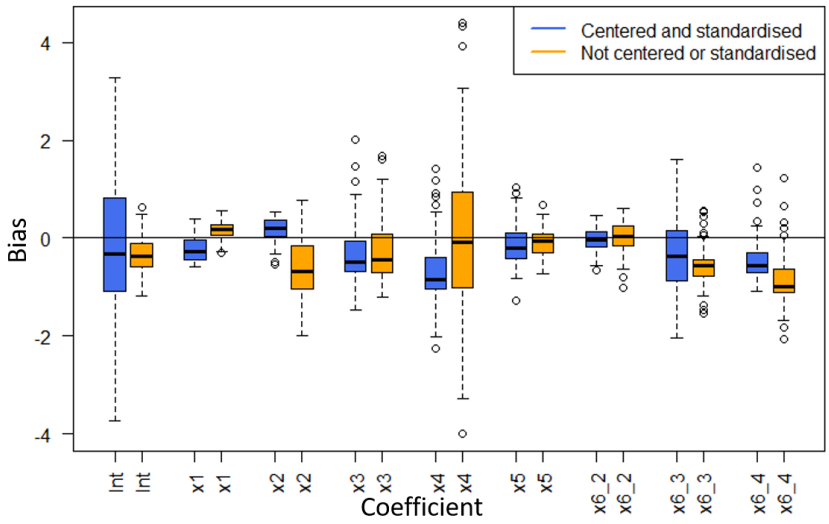

Figure 5.4: The bias of each of the coefficients produced by the lasso when the design matrix has been centered and standardised (blue) and when the design matrix has not been centered or standardised (orange).

Figure 5.4 shows boxplots of the bias of the coefficient estimates obtained when fitting the lasso both when the design matrix has been centered and standardised and when it has not. It can be seen that there is no clear pattern as to which of the methods results in the lowest bias as it varies between the coefficients.

Hastie et al suggest that centering and standardisation is necessary when the covariates are measured using different units. If the covariates have all been measured using the sane units then centering and standardisation may not be required (Hastie et al., 2015).

5.6 Discussion

The main aim of this chapter was to examine the different types of variable selection frequently used. The most common method of variable selection is step-wise selection. This involves adding or removing variables to a model and comparing the results based on a certain criteria. These criteria include ANOVA and AIC.

The lasso is a method which simultaneously selects a model and estimates the coef-

ficients. Stepwise selection may not always lead to the best model globally. The lasso is both computationally efficient and due to the criteria being convex, selects the best global solution. Therefore, the lasso is generally superior to using stepwise selection (McConville et al., 2017).

The lasso was originally developed by Tibshirani in 1996 for use with a continuous response variable and continuous covariates (Tibshirani, 1996b). It has since been developed for use with a binary response variable as well as for use with factor variables. In the group lasso case, all levels of a factor will either be included or excluded from the final model. These two methods have also been combined in order to be able to logistic regression with grouped variables.

There are a variety of R packages that can be used to compute lasso estimates. In order compute estimates using the group logistic lasso, either the `grpreg`, the `grplasso` or the `gglasso` can be used. Since our data comes from a survey, the sampling weights may need to be included. Therefore the `grplasso` package will need to be used. The methods of including sampling weights in the lasso are discussed further in Chapter 6.

Simulations were conducted to examine the differences between stepwise selection and the lasso in a variety of scenarios. For the lasso, when choosing the value of the penalty parameter λ, both the value of λ which minimised the cross validation error and the value of λ which gave a cross validation error of one standard deviation from the minimum were used.

When considering false negatives, very few were produced when the number of covariates and the sample size was varied using any of the three methods. The largest number of false negatives were found when the size of the error term was increased. In this case, the lasso with λ using the one standard error rule was found to give the most false negatives and the lasso using λ which gave the lowest cross-validation error gave the least false negatives. In terms of inference, this implies that as variability increases the the lasso using λ which gives the lowest cross-validation error is less likely to exclude useful information than the other two methods.

The lasso using λ which gave the lowest cross-validation error was found to produce the most false positives on average. This is consistent with findings from Guo (2015) who also concluded that the lasso has a higher false positive rate compared to stepwise selection using AIC (Guo, 2015). Various further extensions to the lasso to control for

the false positive rate (Drysdale et al., 2019) (Sampson et al., 2013) (Javanmard et al., 2019) however these methods are not considered further in this project.

Using the lasso causes the non-zero coefficients selected to be biased towards zero therefore it was expected that the average bias for the lasso models would be larger than for the models obtained using stepwise selection. The simulations confirmed this with stepwise selection generally resulting in the lowest average bias. In order to reduce the bias of the lasso it is recommended that an unrestricted model is fit using the variables whose coefficients are found to be non-zero using the lasso (Hastie et al., 2009).

Across all of the simulations, it was found that in approximately half of the scenarios the minimum average mean squared error was given by stepwise selection and for the other half of the scenarios it was given by the lasso using the value of λ which minimised the cross validation error. The lasso tended to have the lowest mean squared error when there were a larger number of coefficients with a true value of zeros. It also had gave the lowest mean squared error when the error used to calculate the response variable was increased.

Although the lasso using λ which gave the lowest cross-validation error tends to produce more false positives than stepwise selection, it produces less false negatives when the error in simulating the response variable is larger. When looking at the mean squared error there is little difference between the methods and a method has been suggested to deal with the bias resulting from the use of the lasso. Therefore, in terms of inference it appears that although the lasso may include more variables than necessary, the interpretation of coefficients will be similar regardless of the method used.

This chapter has summarised variable selection methods which have been developed for use with infinite populations. When analysing the PNS and MCS inference regarding a finite population is required and sampling weights may be required. Chapter 6 discusses how variable selection can be conducted whilst incorporating the sampling weights.

Chapter 6

Methodology and Analysis – Survey Weighted Variable Selection

This chapter will combine methods from the previous chapters to examine how the lasso can be further extended for use with survey data. The methods of calculating coefficient estimates using the survey lasso will be described. Implementation of survey lasso in R will be discussed. These methods will then be applied to the data from the PNS and the MCS and a comparison between the methods will be made.

6.1 The lasso for survey data

The methods discussed in Chapter 6 assume that the data used is drawn independently from an infinite population. For survey sampling this is not the case as the sample comes from a finite population.

Consider a finite population of size N, denoted by $U = \{1, \ldots, N\}$. Assume that a sample $S \subset U$ is constructed with sampling design $p(.)$. Denoting $s \in U$ as the realisation of S, then $p(s)$ is the probability of selecting a sample index $s \subset U$. The first-order inclusion probability for the kth unit and the second-order inclusion probability for the kth and lth units are defined as

$$\pi_k = \Pr(k \in S) = \sum_{s \subset U : k \in s} p(s) \quad \text{and} \quad \pi_{k,l} = \Pr(k, l \in S) = \sum_{s \subset U : k,l \in s} p(s). \tag{6.1}$$

It is assumed that $\pi_k > 0$ for all $k \in U$.

6.1.1 Survey weighted lasso

As can be seen in Chapter 6, the linear lasso for an infinite population solves the following:

$$\underset{\beta}{\text{minimize}} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\}, \tag{6.2}$$

where y is the response vector, X is the design matrix, $\|\beta\|_1$ is the ℓ_1 norm of β and $\lambda > 0$ is a parameter that is to be specified.

For convenience this can be written as

$$\underset{\beta}{\text{minimize}} \left\{ (y_s - X_s\beta)^T (y_s - X_s\beta) + \lambda \sum_{i=1}^{p} |\beta_i| \right\}. \tag{6.3}$$

McConville et al.(2017) propose a survey weighted linear lasso which solves the following:

$$\underset{\beta}{\text{minimize}} \left\{ (y_s - X_s\beta)^T \Pi_s^{-1} (y_s - X_s\beta) + \lambda \sum_{i=1}^{p} |\beta_i| \right\} \tag{6.4}$$

where $X_s = [x_j^T]_{j \in S}$, $y_s = [y_j]_{j \in s}$ and $\Pi_s = \text{diag}(\pi_j)_{j \in s}$ a diagonal matrix of the inclusion probabilities for the sample (McConville et al., 2017).

6.1.2 Survey weighted logistic lasso

When the response variable of interest is binary then logistic regression should be used. For an infinite population, the logistic lasso minimises the negative log-likelihood subject to an l_1 penalty (Park and Hastie, 2007).

This means solving:

$$\underset{\beta}{\text{minimize}}\left\{ -\sum_{i=1}^{n}\left\{ y_i(\beta_0+\beta^T\boldsymbol{x}_i)-\log(1+e^{\beta_0+\beta^T\boldsymbol{x}_i})\right\}\right\} \text{ subject to } \sum_{j=1}^{p}|\beta_j|\leq\lambda. \quad (6.5)$$

McConville (2011) defines the survey weighted logistic lasso. The coefficient estimator for the survey weighted logistic lasso is given by:

$$\underset{\beta}{\text{minimize}}\left\{ -\sum_{i\in s}\left\{ \frac{1}{\pi_i}y_i(\beta_0+\beta^T\boldsymbol{x}_i)-\log(1+e^{\beta_0+\beta^T\boldsymbol{x}_i})\right\}\right\} \text{ subject to } \sum_{j=1}^{p}|\beta_j|\leq\lambda. \quad (6.6)$$

6.1.3 Survey weighted group lasso

As mentioned in Chapter 6, for factor variables with a group structure, the lasso described above is not appropriate. Therefore, the group lasso which includes all or none of the levels from a group should be used.

As discussed in Chapter 6, For the infinite population case, Yuan and Lin (2006) state that given a linear regression model:

$$Y=\beta_0+\sum_{j=1}^{J}X_j\beta_j+\epsilon$$

which contains J groups of covariates and X_j denotes the covariates in group $j\in J$, the group lasso solves

$$\underset{\beta}{\text{minimize}}\left\{ \sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{J}X_{ij}\beta_j)^2+\lambda\sum_{j=1}^{J}\|\beta_j\|_2\right\}. \quad (6.7)$$

For the finite population survey case, McConville defines the coefficient estimator for the survey weighted group lasso as

$$\underset{\beta}{\text{minimize}}\left\{ (y_s-X_s\beta)^T\Pi_s^{-1}(y_s-X_s\beta)+\lambda\sum_{j=1}^{J}\|\beta_j\|_2\right\}. \quad (6.8)$$

6.1.4 Survey weighted group logistic lasso

For an infinite population, the coefficient estimator for the group logistic lasso was given in Chapter 6 as:

$$\underset{\beta}{\text{minimize}}\left\{ \sum_{i=1}^{n} \left\{ y_i(\beta_0 + \beta^T \boldsymbol{x}_i) - \log(1 + e^{\beta_0 + \beta^T \boldsymbol{x}_i}) \right\} + \lambda \sum_{j=1}^{J} s(df_j)\|\beta_j\|_2 \right\} \tag{6.9}$$

For a survey weighted group logistic lasso coefficient estimator, the formula for the survey weighted logistic lasso given above can be extended to use the group penalty term (McConville, 2011). Therefore, the coefficient estimator is given by:

$$\underset{\beta}{\text{minimize}}\left\{ \sum_{i=1}^{n} \left\{ \frac{1}{\pi_i} y_i(\beta_0 + \beta^T \boldsymbol{x}_i) - \log(1 + e^{\beta_0 + \beta^T \boldsymbol{x}_i}) \right\} + \lambda \sum_{j=1}^{J} s(df_j)\|\beta_j\|_2 \right\}. \tag{6.10}$$

6.2 Selection of the penalty parameter

Since the logistic lasso is needed, many of the covariates in both of the data sets have a grouped structure and we would like to account for the sampling weights, `grplasso` is the only current available package to compute the survey-weighted group lasso. As mentioned in Chapter 6, there is no function to run cross-validation in the `grplasso` package in order to choose the penalty parameter used when fitting the model. Therefore in order to select an optimal value of λ this must computed be separately.

When conducting cross-validation for the grouped logistic lasso some alterations need to be made from the algorithm described in Chapter 6 for the linear lasso.

Since the response variable is binary, it isn't possible to use the mean squared error (MSE) described above directly. Instead a new criteria to minimise needs to be selected.

The first method which can be conducted is using a probability threshold when predicting the values of the response variable. First, the grouped logistic lasso is fit to the training set over a range of λ. Then μ_{pred} can be found for observation in the test data set. Next, a threshold value, τ, is chosen and for all values of $\mu_{pred} > \tau$ the predicted response y_{pred} is 1. For any values of $\mu_{pred} \leq \tau$ the predicted response y_{pred} is 0. The MSE can then be calculated as $\sum(y_{pred,i} - y_{test,i})^2$ where $y_{test,i}$ is the true response for the i-th observation in the test data set. The mean of this error can then be calculated for each value of λ and plotted. The value of λ which minimises this error would be chosen.

A problem with this method is choosing the value τ. Another similar method which doesn't require a choice of threshold calculates $\sum(\mu_{pred,i} - y_{test,i})^2$. The mean can then be calculated for each value of λ and plotted. Once again the value of λ which minimises this error would be chosen.

A further method calculates the negative log likelihood and chooses λ such that this value is minimised. For logistic regression the negative log-likelihood is given by

$$-\sum_{i=1}^{n} \left(y_i \log(\mu_i) + (1 - y_i) \log(1 - \mu_i)\right). \tag{6.11}$$

During cross-validation, the lasso is fit over a range of values of λ and then the negative log-likelihood of the test data-set of size n_2 can be calculated as

$$-\sum_{i=1}^{n_2} \left\{y_{test_i} \log(\mu_{pred_i}) + (1 - y_{test_i}) \log(1 - \mu_{pred_i})\right\}. \tag{6.12}$$

The mean for each value of λ across the k test data-sets can be calculated and the value of λ which gives the lowest mean is then chosen. Using the negative log-likelihood is the most commonly used method when choosing a value of λ using cross-validation (Meier et al., 2008) and therefore will be the method used going forward.

6.2.1 Problems with implementation

Unfortunately, despite working for simulated data, cross validation for group logistic lasso using sampling weights could not be implemented for the PNS or MCS data. When trying to use the method of cross-validation described above, a minimum value of the negative log-likelihood was not reached and therefore no optimal value of λ could be obtained.

Further cross-validation criterion have been considered including weighted log-likelihood and mean-squared error. Similarly to when the negative log-likelihood was used however, a value of λ which minimised these criterion could not be found.

It is currently unclear as to why this method to choose a value of λ does not work and so the implementation of cross validation for group logistic lasso using sampling weights is left as future work.

6.2.2 Using alternative methods to obtain the penalty parameter

The `grpreg` package includes an option to compute cross validation although as mentioned in Chapter 6 it does not contain an option to account for the sampling weights. McConville (2011) shows that despite a model-based estimator of the optimal value of λ, such as obtained using `grpreg,` being slightly less consistent, under similar conditions gives a very similar value to when using a model-assisted estimator (McConville, 2011).

Therefore the model-based estimator of λ will be computed using the `grpreg` and then after adjusting this value for the slight difference in formulation between the equation which the `grpreg` and the `grplasso` packages minimise over, this value will be used in the `grplasso` package both with and without the sampling weights.

A sensitivity analysis will be conducted to examine the impact that varying the value of λ around this chosen value has. The value will be varied slightly and the non-zero coefficients selected using each value will be compared.

6.3 Analysis of the PNS and the MCS

For each of the data sets, the following analysis will be conducted. First, using the cross validation function in the `grpreg` package, a value of λ will be found. This value will then be adjusted to be used in the `grplasso` package. After conducting a sensitivity analysis a value of λ to be used with `grplasso` will be selected. Then, two models will be obtained using the grouped logistic lasso. The first will include the sampling weights and the second will not. This will allow any differences in the non-zero coefficients to be compared when the sampling scheme is accounted for.

There is currently no standard methods for computing standard error estimates for the coefficients obtained using the lasso. Therefore, in order to make an inference about the relationship between intellectual disability and socio-economic and health variables, a design-based generalised linear model will be fit using the variables selected whilst using the group logistic lasso.

Step-wise selection using AIC as described in Chapter 6, will also be conducted. The model chosen using this method will then be compared to the two models selected using the group logistic lasso.

After fitting each of these models, inference will be made regarding the relation-

ship between intellectual disability and socio-economic and health variables in the two countries.

6.3.1 Analysis of the PNS

Using the methods described above, the value of λ chosen was 11.90. After varying λ around this value it was observed that there was no difference in the variables selected and only little difference in the coefficient estimates. Therefore, 11.90 will be used as the penalty parameter when fitting the group logistic lasso both with and without sampling weights. The results from fitting a design-based regression model using the coefficients selected with the weighted and unweighted lasso along with the results from selecting a model using stepwise selection can be found in Tables 6.1 and 6.2.

6.3.2 Analysis of the MCS

Similarly to the analysis of the PNS, using the methods described above, the value of λ chosen was 17.24. After varying λ around this value it was observed that there was no difference in the variables selected and only little difference in the coefficient estimates. Therefore, 17.24 will be used as the penalty parameter when fitting the group logistic lasso both with and without sampling weights. The results from fitting a design-based regression model using the coefficients selected with the weighted and unweighted lasso along with the results from selecting a model using stepwise selection can be found in Table 6.3.

6.3.3 Comparison of methods

It can be seen in both analyses that, similarly to the simulations conducted in Chapter 5, the lasso selects a higher number of variables than when stepwise selection using AIC is used. This is more apparent in the analysis if the MCS. Although this is the case, the inference is similar regardless of the method used.

When looking at the coefficient estimates they appear to be similar across the three methods. When considering the confidence intervals of the coefficient estimates, any variable that would be found to be significant at the 5% level (the 95% confidence interval does not contain zero) using one method is also generally found to be significant using the other methods. There are a small number of exceptions to this. The majority

Table 6.1: Coefficient estimates with 95% confidence intervals and standard errors for the design-based regression models based on the variables selected using unweighted lasso, weighted lasso and stepwise selection using AIC for the PNS data

	Unweighted lasso		Weighted lasso		AIC	
	Estimate (Conf. Int)	Std Error	Estimate (Conf. Int)	Std Error	Estimate (Conf. Int)	Std Error
Intercept	-5.055 (-7.486, -2.624)	1.241	-5.615 (-8.188,-3.042)	1.313	-5.380 (-7.658, -3.102)	1.162
House - Apartment	-0.477 (-1.294, 0.340)	0.417	-0.435 (-1.243, 0.374)	0.413	-	-
- Lodging	-1.468 (-3.667, 0.732)	1.122	-1.344 (-3.540, 0.853)	1.121	-	-
Walls - Inadequate	0.252 (-0.587, 1.092)	0.429	0.136 (-0.705, 0.977)	0.429	-	-
Roof - Inadequate	-	-	0.959 (0.123, 1.794)	0.426	0.931 (0.103, 1.758)	0.422
Floor - Inadequate	0.327 (-0.444, 1.099)	0.394	0.328 (-0.400, 1.057)	0.372	0.367 (-0.259, 0.992)	0.319
Water supply - primary inad, second ad	1.076 (-0.166, 2.318)	0.634	1.117 (-0.153, 2.386)	0.648	1.128 (-0.157, 2.413)	0.656
- Primary and second inad	-1.400 (-2.709, -0.091)	0.668	-1.164 (-2.432, 0.104)	0.647	-1.138 (-2.412, 0.136)	0.650
- Primary inad no second	0.152 (-0.330, 0.634)	0.246	0.210 (-0.233, 0.653)	0.226	0.204 (-0.241 , 0.649)	0.227
Running water - No	0.304 (-0.244, 0.851)	0.279	-	-	-	-
Drinking water - Boiled	-0.544 (-2.335, 1.246)	0.914	-0.489 (-2.298, 1.321)	0.923	-	-
- Treated at home	0.089 (-0.739, 0.917)	0.423	0.142 (-0.655, 0.939)	0.407	-	-
- Industry mineralised	0.203 (-0.288, 0.695)	0.251	0.221 (-0.266, 0.708)	0.249	-	-
- Without treatment	-0.113 (-0.620, 0.393)	0.258	-0.106 (-0.587, 0.375)	0.245	-	-
Number of rooms	-	-	-	-	-0.055 (-0.211, 0.101)	0.080
Number of bedrooms	0.112 (-0.093, 0.317)	0.105	0.108 (-0.098, 0.314)	0.105	0.153 (-0.066, 0.371)	0.111
Kitchen - No	0.557 (-0.054, 1.168)	0.312	0.603 (0.001, 1.204)	0.307	0.528 (-0.092, 1.149)	0.317
Number of bathrooms	0.251 (-0.081, 0.583)	0.170	0.194 (-0.162, 0.549)	0.181	0.286 (-0.143, 0.714)	0.218
Sewage - Inadequate	-	-	-	-	-	-
Waste - Inadequate	-0.361 (-0.866, 0.144)	0.258	-0.441 (-0.948, 0.065)	0.259	-0.406 (-0.908, 0.096)	0.256
Electricity - Inadequate	-0.363 (-1.845, 1.118)	0.756	-0.397 (-1.829, 1.036)	0.731	-0.553 (-1.989, 0.884)	0.733
Basic goods - Two	0.202 (-1.275, 1.679)	0.754	0.189 (-1.312, 1.689)	0.765	-	-
- Three	0.346 (-1.070, 1.762)	0.722	0.281 (-1.209, 1.772)	0.760	-	-
Status goods - One	-	-	0.381 (-0.179, 0.941)	0.286	0.450 (-0.073, 0.972)	0.267
- Two	-	-	-0.113 (-0.819 , 0.592)	0.360	-0.044 (-0.734, 0.647)	0.352
- Three	-	-	0.263 (-0.552, 1.078)	0.416	0.354 (-0.421, 1.129)	0.395
- Four	-	-	0.816 (-0.410, 2.041)	0.625	0.903 (-0.350, 2.157)	0.640
- Five	-	-	1.124 (0.067, 2.180)	0.539	1.184 (0.148, 2.221)	0.529
Comp and Int - Computer	-	-	0.488 (-0.269, 1.245)	0.386	0.480 (-0.277, 1.237)	0.386
- Internet	-	-	0.268 (-0.719, 1.255)	0.503	0.261 (-0.712, 1.233)	0.496
- Neither	-	-	0.624 (-0.034, 1.282)	0.336	0.615 (-0.037, 1.268)	0.333
Health visits - Every 2 months	-0.751 (-1.617, 0.116)	0.442	-0.749 (-1.618, 0.121)	0.443	-0.736 (-1.605, 0.133)	0.443
- From 2 to 4 times	-0.607 (-1.208, 0.005)	0.307	-0.622 (-1.228, -0.015)	0.309	-0.570 (-1.195, 0.055)	0.319
- Once	0.327 (-0.304, 0.958)	0.322	0.356 (-0.268, 0.980)	0.318	0.331 (-0.291, 0.953)	0.317
- Never received	-0.324 (-0.937, 0.289)	0.313	-0.351 (-0.968, 0.266)	0.315	-0.356 (-0.978, 0.266)	0.317
- Not registered	0.095 (-0.416, 0.607)	0.261	0.070 (-0.429, 0.568)	0.254	0.043 (-0.444, 0.530)	0.248
Endemic visits - Every 2 months	-0.024 (-0.561, 0.512)	0.274	-0.044 (-0.568, 0.479)	0.267	-0.044 (-0.566, 0.478)	0.267
- From 2 to 4 times	0.217 (-0.357, 0.791)	0.293	0.154 (-0.418, 0.726)	0.292	0.144 (-0.427, 0.714)	0.291
- Once	-0.126 (-0.683, 0.431)	0.284	-0.177 (-0.747, 0.393)	0.291	-0.200 (-0.774, 0.373)	0.293
- Never received	-0.667 (-1.191, 0.163)	0.262	-0.740 (-1.258, -0.222)	0.264	-0.828 (-1.347, -0.309)	0.265

Table 6.2: Continuation of Table 6.1

	Unweighted lasso		Weighted lasso		AIC	
	Estimate (Conf. Int)	Std Error	Estimate (Conf. Int)	Std Error	Estimate (Conf. Int)	Std Error
Sex - Female	-0.373 (-0.702, -0.044)	0.168	-0.380 (-0.708, -0.052)	0.167	-0.369 (-0.702, -0.037)	0.170
Age	0.339 (0.249, 0.399)	0.031	0.347 (0.287, 0.407)	0.031	0.354 (0.296, 0.412)	0.030
Race - Non-white	0.078 (-0.284, 0.439)	0.185	0.137 (-0.258, 0.532)	0.202	-	-
Can read and write - No	4.042 (3.225, 4.860)	0.417	4.022 (3.214, 4.830)	0.412	3.974 (3.202, 4.747)	0.394
Education level - Completed elementary	-0.042 (-0.753, 0.668)	0.363	-0.077 (-0.749, 0.594)	0.343	-0.097 (-0.769, 0.574)	0.343
- Completed high school	-2.379 (-3.474, -1.285)	0.558	-2.422 (-3.482, -1.362)	0.541	-2.448 (-3.509, -1.387)	0.541
- Graduated/ Masters	-2.310 (-3.957, -0.066)	0.841	-2.325 (-3.954, -0.696)	0.831	-2.295 (-3.920, -0.671)	0.829
Physical disability - Yes	1.536 (0.684, 2.389)	0.435	1.505 (0.638, 2.372)	0.442	1.493 (0.614, 2.371)	0.448
Hearing disability - Yes	0.948 (-0.217, 2.112)	0.594	0.950 (-0.330, 2.231)	0.653	0.947 (-0.341, 2.234)	0.657
Visual disability - Yes	0.268 (-0.538, 1.074)	0.411	0.276 (-0.551, 1.104)	0.422	0.284 (-1.113 0.544)	0.423
Health status - Average	0.849 (0.438, 1.260)	0.210	0.876 (0.468, 1.285)	0.208	0.903 (0.494, 1.313)	0.209
- Below average	1.256 (0.402, 2.111)	0.436	1.248 (0.400, 2.096)	0.433	1.324 (0.493, 2.155)	0.424
Health limits activities - Yes	-0.453 (-1.104, 0.199)	0.333	-0.484 (-1.133, 0.166)	0.331	-0.503 (-1.133, 0.127)	0.322
Chronic illness - Yes	2.388 (1.953, 2.823)	0.222	2.377 (1.944, 2.809)	0.221	2.376 (1.952, 2.800)	0.216
Sees same doctor - Yes	0.524 (0.074, 0.974)	0.230	0.464 (0.034,-0.894)	0.219	0.461 (0.017, 0.905)	0.227
Last consulted doctor - 1 to 2 years	-0.090 (-0.867, 0.687)	0.396	-0.041 (-0.756, 0.674)	0.365	-0.040 (-0.704, 0.624)	0.339
- More than two years	-0.676 (-1.426, 0.074)	0.383	-0.658 (-1.410, 0.093)	0.384	-0.637 (-1.354, 0.079)	0.366
- Never	-2.197 (-3.957, -0.436)	0.898	-2.068 (-3.812, -0.324)	0.890	-2.154 (-3.884, -0.425)	0.882
Times consulted doctor	0.043 (0.015, 0.070)	0.014	0.041 (0.013, 0.069)	0.015	0.040 (0.011, 0.069)	0.015
Last consulted dentist - 1 to 2 years	0.040 (-0.534, 0.614)	0.293	0.052 (-0.484, 0.588)	0.274	-	-
- More than two years	0.197 (-0.325, 0.719)	0.266	0.183 (-0.326, 0.693)	0.260	-	-
- Never	-0.165 (-0.669, 0.370)	0.273	-0.168 (-0.684, 0.348)	0.263	-	-
Health care in last 2 weeks - Yes	0.395 (-0.186, 0.975)	0.296	0.385 (-0.204, 0.975)	0.301	0.361 (-0.220, 0.942)	0.297
Hospitalised - Yes	-	-	0.504 (-0.187, 1.195)	0.353	0.517 (-0.180, 1.213)	0.355
Emergency care - Yes	0.953 (0.173, 1.734)	0.398	0.713 (-0.099, 1.525)	0.414	0.618, (-0.185, -1.421)	0.410
Alternative treatment -Yes	-	-	-	-	-	-

Table 6.3: Coefficient estimates with 95% confidence intervals and standard errors for the design-based regression models based on the variables selected using unweighted lasso, weighted lasso and stepwise selection using AIC for the MCS data

	Unweighted lasso		Weighted lasso		AIC	
	Estimate (Conf. Int)	Std Error	Estimate (Conf. Int)	Std Error	Estimate (Conf. Int)	Std Error
Intercept	-3.761 (-4.456, -3.066)	0.355	-3.949 (-4.676, -3.222)	0.371	-3.598 (-4.149, -3.047)	0.281
Job seekers - Yes	-0.364 (-1.012, 0.285)	0.331	-0.330 (-0.980, 0.320)	0.332	-	-
Income support - Yes	0.435 (0.095, 0.774)	0.173	0.463 (0.147, 0.779)	0.161	0.491 (0.181, 0.800)	0.158
Sickness support - Yes	0.785 (0.491,1.078)	0.150	0.779 (0.487, 1.072)	0.149	0.805 (0.510, 1.101)	0.151
Tax credit - Yes	0.260 (0.015, 0.505)	0.125	0.260 (0.020, 0.501)	0.123	0.302 (0.073, 0.531)	0.117
Family benefits - Yes	0.629 (0.105, 1.153)	0.268	0.604 (0.084, 1.124)	0.266	0.618 (0.089, 1.147)	0.270
Housing benefits - Yes	-	-	-	-	-	-
Other benefits - Yes	1.180 (0.729, 1.631)	0.230	1.173 (0.727, 1.620)	0.228	1.187 (0.736, 1.639)	-
Coping financially - Alright	0.145 (-0.217, 0.507)	0.185	0.148 (-0.214, 0.510)	0.185	-	-
- Getting by	0.184 (-0.195, 0.563)	0.193	0.168 (-0.211, 0.547)	0.193	-	-
- Some difficulty	0.219 (-0.239, 0.676)	0.233	0.203 (-0.255, 0.661)	0.233	-	-
- Very difficult	0.561 (-0.017, 1.140)	0.298	0.541 (-0.040, 1.123)	0.297	-	-
Number of rooms	-0.037 (-0.106, 0.032)	0.035	-0.043 (-0.115, 0.029)	0.037	-0.042 (0.106, 0.022)	0.230
Damp - Small problem	0.042 (-0.402, 0.486)	0.226	-	-	-	-
- Some problem	-0.093 (-0.464, 0.278)	0.189	-	-	-	-
- Large problem	-0.128 (-0.751, 0.495)	0.318	-	-	-	-
Sex - Female	-0.826 (-1.061, -0.592)	0.120	-0.824 (-1.059, -0.590)	0.120	-0.818 (-1.052, -0.584)	0.032
Age - 11	-0.065 (-0.298, 1.67)	0.119	-0.071 (-0.300, 0.159)	0.117	-0.080 (-0.310, 0.149)	0.119
-12	1.449 (0.326, 2.571)	0.573	1.259 (0.218, 2.300)	0.531	1.227 (0.200, 2.255)	0.117
School year - Not in Yr 6	-0.241 (-0.669, 0.218)	0.234	-	-	-	-
Comp and Int - Comp	-	-	-0.108 (-0.782, 0.566)	0.344	-	-
- Internet	-	-	0.192 (-0.566, 0.951)	0.387	-	-
- Neither	-	-	-0.366 (-1.058, 0.326)	0.353	-	-
Health status - Very good	-0.245 (-0.515, 0.025)	0.138	-0.252 (-0.522, 0.018)	0.138	-0.215 (-0.481, 0.052)	0.524
- Good	-0.252 (-0.606, 0.102)	0.181	-0.269 (-0.622, 0.084)	0.180	-0.224 (-0.574, 0.125)	0.136
- Fair	-0.501 (-0.998, -0.013)	0.249	-0.485 (-0.970, 0.000)	0.248	-0.456 (-0.936, 0.023)	0.178
- Poor	0.020 (-0.823, 0.863)	0.430	0.026 (-0.811, 0.863)	0.427	0.102 (-0.727, 0.931)	0.245
Longterm illness - Yes	0.298 (-0.134, 0.731)	0.221	0.313 (-0.119, 0.745)	0.220	-	0.423
Visual disability - Yes	0.431 (-0.081, 0.944)	0.261	0.421 (-0.090, 0.932)	0.261	0.426 (-0.093, 0.946)	0.265
Hearing disability - Yes	0.807 (0.232, 1.382)	0.293	0.784 (0.204, 1.365)	0.296	0.817 (0.247, 1.387)	0.291
Physical disability - Yes	0.885 (0.484, 1.285)	0.204	0.882 (0.482, 1.281)	0.204	0.902 (0.505, 1.300)	0.203
Health limits activities - Yes	2.075 (1.646, 2.504)	0.219	2.067 (1.639, 2.495)	0.219	2.333 (2.057, 2.608)	0.141
Hosp - Less than once a month	0.035 (-0.308, 0.377)	0.175	0.050 (-0.292, 0.392)	0.174	0.035 (-0.307, 0.377)	0.175
- Once or twice a month	0.204 (-0.092, 0.500)	0.151	0.209 (-0.086, 0.504)	0.151	0.190 (-0.104, 0.484)	0.150
- Once or twice a week	0.603 (0.220, 0.985)	0.195	0.616 (0.233, 0.999)	0.195	0.604 (0.223, 0.985)	0.194
- Several times a week	0.300 (-0.151, 0.752)	0.231	0.308 (-0.142, 0.757)	0.229	0.286 (-0.162, 0.735)	0.229
- Every/almost every day	0.817 (0.468, 1.165)	0.178	0.825 (0.479, 1.172)	0.177	0.799 (0.451, 1.146)	0.177
Seen dentist in last year - Yes	-	-	0.082 (-0.308, 0.471)	0.200	-	-
Race - Non-white	-1.118 (-0.418, 0.182)	0.153	-	-	-	-
Number of people at home	-	-	0.039 (-0.042, 0.119)	0.041	-	-
Below poverty line - Yes	0.085 (-0.212, 0.382)	0.152	-	-	-	-

of these exceptions, however, occur for one level of a grouped variable in which the remaining levels are not found to be significant. The only binary variables found to be significant at the 5% level using only one or two of the methods are whether the material of the roof is adequate and whether emergency care has been received at home in the last year in the analysis of the PNS.

The standard errors of each of the estimates also appear to be similar across the three methods.

6.4 Discussion

The aim of this chapter was to combine the methods discussed in the previous chapters and apply them to the data from the PNS and the MCS.

Since the lasso is generally used for non-survey data, a slight adaptation is needed in order to use it with survey data from a finite population. These adaptations have been shown to be suitable for logistic regression, grouped variables and a combination of the two (McConville et al., 2017).

Using these methods, various models were selected for each of the data sets. The first of these was selected using group logistic lasso without the survey weights, the second was selected using the survey-weighted group logistic lasso and the final was selected using stepwise selection based on the AIC. In order to calculate the standard errors of the estimates and based on the suggestion of Hastie et al to reduce the bias of the coefficients (Hastie et al., 2009), a design-based regression model was then fit including the variables found to have non-zero coefficients for the two models selected when using the lasso.

Although the three methods used select different models, when looking at the confidence intervals created for the coefficients in each model, the inference that can be made is very similar for all three models for each of the data sets. There are very few instances in which a variable is found to be significant in only one or two of the models.

Chapter 7

Results

This chapter will answer the research question: How do poverty and health variables interrelate with intellectual disability in Brazil and the UK? It will also answer: Is it possible to profile different types of children that need lower and higher levels of support to aid in identifying subgroups for selective interventions to alleviate inequalities in education?

This chapter will use the knowledge gained in the previous chapters to select a final model to describe the relationship between intellectual disability, poverty and health variables for both Brazil and the UK. First, it will be discussed whether or not there is a need to include an interaction between age and school year. Then after fitting the final models to each of the datasets, inference will be made for each of the countries separately and then finally, an international comparison will be made.

The previous chapter identified that regardless of the method of variable selection used, the inference was very similar. Using the AIC for model selection proved to select the fewer variables than when using the Lasso. Therefore when making the final inference about the relationship between intellectual disability, poverty and health variables, this method will be used in order to select the most parsimonious model. Weights will be used when calculating the coefficient estimates.

7.1 Interaction between age and school year

Since there is a natural relationship between the age of a child and the level of education or school year that they are in, an interaction term between these two variables will be

considered for each of the two countries. The interaction term will be kept in the final model if it is found to be significant at the 5% level.

7.2 Results - Brazil

Stepwise selection using AIC was used as method of variable selection. Coefficient estimates were calculated using design-based methods as described in Chapter 4. The final model for the PNS can be found in Table 7.1.

7.2.1 Inference regarding intellectual disability in Brazil

Looking at the results of the analysis of the PNS it can be seen that nearly all of the variables that were highlighted as potential indicators of poverty in Chapter 2 were found not to be significant at the 5% level. The only exception to this was whether or not the home has a kitchen. If a home does not have a kitchen the odds of a child having an intellectual disability was increased. The odds of a child having an intellectual disability was found to be between 1.02 and 3.25 times higher for a child living in a home with no kitchen compared to a child living in a home with a kitchen.

Age was also found to be significant with an increase in age resulting in an increase of the odds of intellectual disability. This appears to be a logical finding as the older the child the more likely it is that any intellectual disability will be identified and therefore there is a greater potential of diagnosis.

Sex was found to be significant with a male found to have odds of between 1.63 and 2.64 times higher than a female of having an intellectual disability.

Whether a child is able to read and write was found to be significant. If a child is not able to read and write the odds of that child having an intellectual disability was found to be between 22.6 and 82.3 times higher compared to a child who is able to read and write.

When looking at the level of education, it can be seen that only some of the levels of education are found to be significant at the 5% level. The coefficient estimate of graduated/masters was found to be very large. This is probably due to the fact that there is a very small number of children in the sample with an intellectual disability and it was found during the exploratory analysis that there were no children in the sample

Table 7.1: Coefficient estimates, 95% confidence intervals and standard errors for the final model fitted to the PNS data

	Estimate	95% Conf. Int.	Std. Error	z value	p-value	
(Intercept)	-5.77	(-7.77, -3.77)	1.02	-5.66	0.00	***
Roof - Inadequate	0.81	(-0.02, 1.64)	0.42	1.92	0.05	.
Water sup - Primary inad, second ad	0.97	(-0.28, 2.22)	0.64	1.52	0.13	
- Primary and second inad	-1.82	(-3.03, -0.61)	0.62	-2.94	0.00	**
- Primary inad no second	-0.14	(-0.58, 0.30)	0.23	-0.63	0.53	
Kitchen - No	0.60	(0.02, 1.18)	0.30	2.02	0.04	*
Number of bathrooms	0.34	(0.04, 0.64)	0.15	2.21	0.03	*
Sex - Female	-0.31	(-0.65, 0.03)	0.17	-1.79	0.07	.
Age	0.42	(0.36, 0.48)	0.03	13.72	0.00	***
Sex - Female	-0.73	(-0.97, -0.49)	0.12	-7.62	0.00	***
Can read and write - No	3.77	(3.12, 4.41)	0.33	11.44	0.00	***
Education level - Completed elementary	2.78	(1.30, 4.26)	0.76	3.68	0.00	***
- Completed high school	-5.89	(-19.41, 7.64)	6.90	-0.85	0.39	
- Graduated/ Masters	56.65	(9.15, 104.5)	24.24	2.34	0.02	*
Physical disability - Yes	-1.49	(-2.36, -0.62)	0.45	-3.35	0.00	***
Hearing disability - Yes	-1.12	(-2.13, -0.10)	0.52	-2.15	0.03	*
Health status - Average	0.97	(0.58, 1.37)	0.20	4.83	0.00	***
- Below average	1.35	(0.51, 2.19)	0.43	3.16	0.00	**
Health limits activities - Yes	0.29	(-0.26, 0.84)	0.28	1.03	0.30	
Chronic illness - Yes	-2.37	(-2.79, -1.95)	0.21	-11.06	0.00	***
Sees same doc - Yes	-0.56	(-1.04, -0.08)	0.25	-2.27	0.02	*
Last consulted doctor - 1 to 2 years	-0.06	(-0.78, 0.66)	0.37	-0.17	0.87	
- More than 2 years	-0.73	(-1.49, 0.03)	0.39	-1.89	0.06	.
- Never	-2.60	(-4.34, -0.87)	0.88	-2.95	0.00	**
Times consulted doctor	0.05	(0.02, 0.08)	0.01	3.58	0.00	***
Emergency care - Yes	-0.81	(-1.66, 0.03)	0.43	-1.89	0.06	.
Age * Completed elementary	-0.26	(-0.37, -0.14)	0.06	-4.42	0.00	***
Age * Completed high school	0.17	(-0.66, 1.00)	0.42	0.40	0.69	
Age * Graduated/ Masters	-3.57	(-6.44, -0.69)	1.47	-2.43	0.02	*

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 " 1

who had an intellectual disability and had also graduated. Therefore the odds for this variable may be over-stated and hence interpretation may not be accurate.

When considering further disabilities, physical disability and hearing disability were found to have a significant relationship with intellectual disability. For both of these disabilities, the odds of a child having an intellectual disability were found to be decreased if a child had a physical or hearing disability compared to if they didn't.

Many of the health variables were found to be significant at the 5% level. If a child's general health status was reported to be average or below average then the odds of them having an intellectual disability was found to be increased when compared to a child with a reported general health status of above average. The worse the general health of a child was reported to be the more the odds of an intellectual disability were found to increase though there was found to be a slight overlap in the confidence intervals for average health status and below average health status.

If a child has been diagnosed with a long-standing or chronic illness then the odds of them having an intellectual disability was found to be decreased. The odds of a child having an intellectual disability was found to be between 7.02 and 16.3 times higher for a child with no chronic illness than for a child with a chronic illness.

If a child has never consulted a doctor, the odds of them having an intellectual disability are reduced when compared to a child who has consulted a doctor in the last 12 months. There is no significant difference between the odds of a child who has consulted a doctor in the last twelve months and a child who last consulted a doctor more than 12 months ago. The number of times that a child has consulted a doctor in the last 12 months was also found to have a significant relationship with the odds of a child having an intellectual disability. For each additional time that a child has seen a doctor in the last 12 months, the odds of the child having an intellectual disability increase by between 1.08 and 1.07.

The interaction between age and school level was also found to be significant however as previously mentioned, due to the lack of children in the sample with intellectual disability who have reached higher levels of education, the interpretation of these odds may not be accurate.

In summary, it appears that eduction and health variables are more likely to indicate whether a child has an intellectual disability in Brazil than socio-economic variables

which may be indicative of poverty.

7.3 Results - UK

As with the above analysis, stepwise selection using AIC was used as method of variable selection. Coefficient estimates were calculated using design-based methods as described in Chapter 4. The final model for the MCS can be found in Table 7.2.

7.3.1 Inference regarding intellectual disability in the UK

From looking at the analysis of the MCS data it can be seen that there is a relationship between a family receiving a benefit of some sort and a child having an intellectual disability. The only benefits investigated that were not found to be significant were job seekers allowance, housing benefits and family benefits. If a family receives income support, sickness support, tax credit, family benefits or a different unspecified benefit then it is found that the odds of the child in that family having an intellectual disability is increased. The other variables which could be indicative of poverty or hardship along with whether a family is classed as living below the poverty line was not found to be significant.

If a child is female, then the odds of them having an intellectual disability were found to be reduced compared to if a child is male. For a male child the odds of intellectual disability was found to be between 1.93 and 2.83 times higher than for a female.

Age was not found to be significant. This may be due to the majority of the children in the sample being the same age. Similarly, school year was found not to be significant. Therefore, the interaction between age and school year was also not found to be significant at the 5% level.

When looking at further disabilities, all were found to have a relationship with intellectual disability. If a child has any of these disabilities then the odds of intellectual disability are found to be increased. For a child with a visual disability, the odds of intellectual disability are between 1.06 and 2.71 times higher than for a child with no visual disability. For a child with a hearing disability, the odds of intellectual disability are found to be between 1.20 and 3.25 times higher than for a child who does not have a hearing disability. With regards to physical disability, a child who has a physical dis-

Table 7.2: Coefficient estimates, 95% confidence intervals and standard errors for the final model fitted to the MCS data

	Estimate	95% Conf. Int	Std. Error	z value	p-value	
(Intercept)	-3.83	(-4.06, -3.60)	0.12	-32.92	0.00	***
Income support - Yes	0.54	(0.29, 0.77)	0.12	4.35	0.00	***
Sickness support - Yes	0.83	(0.60, 1.05)	0.12	7.08	0.00	***
Tax credit - Yes	0.34	(0.16, 0.53)	0.09	3.64	0.00	***
Family Benefits - Yes	0.43	(-0.06, 0.88)	0.24	1.79	0.07	.
Other Benefits - Yes	1.10	(0.65, 1.52)	0.22	4.96	0.00	***
Sex - Female	-0.85	(-1.04, -0.66)	0.10	-8.64	0.00	***
Visual disability - Yes	0.54	(0.06, 1.00)	0.24	2.26	0.02	*
Hearing Disability - Yes	0.68	(0.18, 1.18)	0.26	2.67	0.01	**
Physical disability - Yes	0.88	(0.53, 1.23)	0.18	4.94	0.00	***
Health limits activities - Yes	2.11	(1.91, 2.32)	0.11	20.10	0.00	***
Hosp - less than once a month	0.05	(-0.23, 0.32)	0.14	0.32	0.75	
-Once or twice a month	0.18	(-0.07, 0.42)	0.13	1.41	0.16	
-Once or twice a week	0.47	(0.14, 0.78)	0.16	2.85	0.00	**
-Several times a week	0.34	(-0.05, 0.71)	0.19	1.74	0.08	.
-Every/ almost every day	0.76	(0.48, 1.05)	0.15	5.23	0.00	***

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 " 1

ability has odds of between 1.70 and 3.42 times higher than a child who does not have an intellectual disability.

If a child's health was reported to limit their daily activities then the odds of intellectual disability was found to be increased to between 6.75 and 10.18 times higher than when compared to a child whose health was not reported to limit their daily activities. The number of times that a child was hospitalised in the last 12 months was found to be significant. Generally, the more common hospitalisation was the greater the odds of intellectual disability.

In summary, a families socio-economic position according to the poverty line was not found to have a significant relationship with intellectual disability. However, if a family is receiving additional support in the form of benefits then the odds of intellectual disability was found to increase suggesting that the socio-economic position of a family may influence whether a child has an intellectual disability in the UK. Some health variables were also found to be significant in terms of their relationship with intellectual disability.

7.3.2 International comparison

There were a number of variables which can be directly compared between Brazil and the UK. The effect of gender is found to be the same in each country with a female being less likely to have an intellectual disability than a male. Race was found not to be significant in either country.

Since the range of ages in the MCS is a lot smaller than that in the PNS, it is difficult to compare the full effect that age has on intellectual disabilities between the two countries. Increased age was found to increase the odds of intellectual disability in Brazil.

Health variables in both countries have a relationship with intellectual disability although the variables found to be significant varied between the countries. In both surveys, the general health status of a child was reported. This variable was only found to be significant in Brazil with a poorer health status increasing the odds of intellectual disability. Similarly, whether a child has a long-term or chronic illness was only found to be significant in Brazil. On the other hand, if health was reported to limit a child's daily activity the odds of them having an intellectual disability was found to increase in

the UK. This variable was not found to be significant in Brazil.

In the UK the presence of a visual, hearing or physical disability was found to increase the odds of having an intellectual disability. This differs to Brazil where visual disability was not found to be significant and the presence of physical disability or hearing disability was found to decrease the odds of intellectual disability.

In regards to socio-economic position, different variables were used in each country. In Brazil, many variables relating to the adequacy of a home were used along with variables relating to access to goods such as a stove, refrigerator and television. It was found that, with the exception of a home having a kitchen, none of these variables had a significant relationship with intellectual disability. In the UK, an indicator of a family living below the poverty line was used along with indicators relating to a variety of benefit schemes being received by a family. Although living below the poverty line was not found to be significant, numerous benefits were. Therefore, there appears to be little to no relationship between socio-economic position in Brazil and some indication of a relationship in the UK.

7.4 Discussion

In Brazil, the only variable that that could be related to poverty found to be significant was whether or not a home has a kitchen. The lack of a kitchen in a home increases the odds of intellectual disability although the lower bound of the 95% confidence interval of the odds is close to one. Therefore, despite this variable being found significant the results of this analysis suggests that there is little to no relationship between intellectual disability and poverty in Brazil.

The largest relationship in Brazil was found to be between intellectual disability and whether a child can read and write. The odds of a child having an intellectual disability if they are unable to read and write are up to 82 times that of a child who is able to read and write. This suggests that if a child is not meeting targets with regards to this it could be a strong indicator that the child has an intellectual disability. It was also found that the odds of a child having an intellectual disability was reduced if they were in education past what is compulsory. This implies that many intellectually disabled children are not in further education.

Despite the majority of poverty variables not being found to be significant with regards to intellectual disability, many of the health variables were found to be significant. A child having a physical disability increased the odds of intellectual disability. This is important to note as in many cases children with physical disabilities often face issues in terms of receiving education due to accessibility issues a lack of resources (França et al., 2008).

The general health status of a child was found to have a relationship with intellectual disability. The worse the health status of a child was reported to be, the greater the risk of intellectual disability. In addition to this, for each additional time a child has visited a doctor in the last 12 months the odds of intellectual disability were found to increase. This suggests that overall, children with poorer general health are more likely to have an intellectual disability in Brazil.

Therefore, in Brazil it seems as though the failure to meet educational targets such as the ability to read and write in addition to a poor general health status and the need to consult a doctor numerous times within a 12 month period may be factors which can be used when attempting to profile a child who may have an intellectual disability.

In the UK, the commonly used method of indicating that a person is living in poverty was available. The poverty line is generally set at 60% of a countries median income and if the income of a family falls below this line then they are defined as living in poverty (Eurostat, 1998). This variable however, was not found to have a significant relationship with intellectual disability. The socio-economic status of a family however may in fact have a relationship with intellectual disability since many of the benefits were found to be significant.

A few of the variables relating to health were found to be significant in the UK. If a child's health was reported as limiting their daily activities the odds of intellectual disability are increased. Also, the more often that a child had been hospitalised within 12 months, the more likely intellectual disability is. The presence of a further disability was also found to increase the odds of intellectual disability.

Therefore, in the UK it seems as though a families need for extra support in the form of benefits along with a poorer general health which results in a limitation of daily activities may be factors which can be used when attempting to profile a child who may have an intellectual disability.

When making a comparison between the two countries it can be seen that health variables were found to be significant in both countries. The poorer the health of a child the more likely they are to have an intellectual disability. The socio-economic position of a family seems to be more significant in the UK than in Brazil.

Chapter 8

Conclusions, Discussion and Future Work

The overall aim of this project was to determine whether or not there is a relationship between intellectual disability and poverty and health variables in Brazil and the UK. Through determining this relationship, the aim was to then try to profile a child who is at greater risk of intellectual disability in order to result in a quicker diagnosis and earlier access to the support and resources available for children with intellectual disabilities. Policies regarding children with intellectual disabilities which have been adopted within the two countries in recent years are similar and hence making a comparison between the two countries will provide further insight into how both systems can be improved.

The need for a study of this kind was highlighted during the literature review. There have been very few studies into intellectual disability in Brazil and any studies which have been conducted tend to focus only on small regions (Mercadante et al., 2009). This study was based on data from a recent national survey and therefore it is possible to make an inference regarding intellectual disability for the whole population of Brazil. Although there have been various studies into intellectual disability in the UK, in order to make an international comparison and confirm findings from such studies, a new analysis has been conducted.

The poverty line is the recommended measure to be used in international comparisons when trying to define a person as living in poverty (Eurostat, 1998). Based on the data available in the PNS however, there is no way to determine whether or not someone is living below the poverty line and so alternative measures had to be used. The general

definition of poverty is "when a person's resources are well below their minimum needs" (Goulden and D'arcy, 2014). Therefore one possible way proposed to determine whether a person is living in poverty in Brazil is to look at the adequacy of the home and an individuals access to basic resources. Based on recommendations from the IBGE, the data regarding the materials which various parts of a home is constructed can be classified as adequate or inadequate and the availability of resources can be grouped into basic goods or status goods (Fundação Instituto Brasileiro de Geografia and Estatística. Departamento de População and Indicadores Sociais, 1998).

An issue related to the comparison aspect of this project was the lack of corresponding variables within the two data sets. The indicators of potential poverty from the PNS had no equivalent in the MCS and hence in order to determine the socio-economic position of an individual, different measures had to be used between the two countries. The MCS contains an indicator of whether a families income falls below the poverty line of the UK and hence this variable was included as an indicator of poverty.

In addition to the poverty line, a families socio-economic status could also be measured by the need of additional support in the form of benefits. Therefore in addition to the poverty line, these benefits may be used to identify if a family is of a lower socio-economic position in the UK.

In order to conduct appropriate analysis of the two data sets two main statistical issues were first addressed. The first of these was to determine how to appropriately account for the complex sampling design of the two surveys. The second was to determine a method to select relevant variables from the large number of available variables in both the PNS and the MCS.

In order to account for the complex survey design both model-based and design based approaches were considered. Simulations showed no difference between the coefficient estimates and little differences in the standard errors between the two methods for the sampling schemes considered. Since the population for which inference is required is the population from which the sample is taken, design-based analysis is used (Dorazio, 1999).

When comparing current variable selection methods simulations show that there is little difference when it comes to making inference from the results of models selected using step-wise selection and the lasso. This was further confirmed during the analysis

of the PNS and MCS data.

Silva and Skinner (1997) highlight the need for variable selection in finite populations (Nascimento Silva and Skinner, 1997) and McConville et al (2017) extend the lasso accounting for sampling weights for use with a finite population (McConville et al., 2017). These methods were considered for use when analysing the data from the PNS and the MCS. There was found to be very little difference in inference when a model was selected using survey weighted lasso and stepwise selection based on AIC. The model found using AIC included fewer variables and hence this method was used in the final analysis of both data sets in order to find the most parsimonious model.

Diagnosis of intellectual disability is not consistent in Brazil and many people never receive a diagnosis (Carvalho and Forrester-Jones, 2016). Therefore if it possible to profile a child who is likely to have an intellectual disability it may lead to a quicker diagnosis and therefore quicker access to appropriate support. The earlier support and resources are available to a child with an intellectual disability, the less likely they are to have poor outcomes in life and hence the financial costs as a result of the disability will be reduced for both the family and also for society (Battiscombe, 1974).

After the analysis of the PNS data, a potential profile of a child likely to have an intellectual disability in Brazil was found to be: a child who is unable to read and write with poor general health and multiple visits to a doctor within a 12 month period.

The ability to read and write was found to be the largest indicator that a child may have an intellectual disability in Brazil. Therefore if a child is struggling to meet targets in regards to this it may be advisable to consider testing for intellectual disability in order to provide adequate support in assisting the child throughout their education. The earlier a child is given support in such areas, the more likely they are to reach their full potential with regards to education and employment. A child who does not pursue education when it is no longer compulsory in Brazil, has higher odds of intellectual disability which could suggest that current practices of educating children with intellectual disability can be improved.

Similarly can be suggested regarding the health of a child. If a child has poor general health then they are more likely to have a reduced education compared to a child with a good general health. If a child is found to have poor general health and requires multiple consultations with a doctor throughout a 12 month period, it may be advisable

to consider testing for intellectual disability in order for the appropriate support to be provided to the child and prevent the gap in education widening further.

In the UK, a potential profile of a child likely to have an intellectual disability was found to be: a child in a family who requires extra financial support in the form of benefits along with a poorer general health which results in a limitation on daily activities.

Despite an indicator of poverty based on the poverty line not being found to have a significant relationship with intellectual disability, many benefits were found to have a relationship. Since many of these benefits are received by a family whose income is low, this suggests that the lower the socio-economic position of a family, the greater the chance of the child having an intellectual disability. Therefore, one suggestion would be to closely monitor the development of children in families who require extra financial support, in order to identify any intellectual disability as soon as possible. The earlier this support is provided the more likely a child is to achieve good outcomes in life and therefore will be less likely to require similar financial support from the government in the future.

Similarly to Brazil, poorer health was found to increase the odds of intellectual disability and so also monitoring children in the UK who have a level of health which restricts their daily activities may lead to a quicker diagnosis of intellectual disability when relevant. With a quicker diagnosis the less likely the gap in educational attainment is to widen.

In both countries physical disability was found to have a relationship with intellectual disability. A child with a physical disability is more likely to have an intellectual disability compared to a child without. Similarly to issues in education for children with poor general health, in many cases education is often limited for children with physical disabilities. This is usually due to access issues and a lack of appropriate support and resources in schools (França et al., 2008). Therefore in order to prevent the gap in educational attainment between disabled children and non-disabled children widening, it may be suggested that the intellectual development of children with physical disabilities should be closely monitored in order to diagnose any intellectual disability as quickly as possible.

8.1 Recommendations and Unique Contribution

This thesis has the potential to have impact on the early intervention of children with intellectual disability. A method has been developed in order to profile children who are likely to have a disability in both the UK and Brazil. From this certain recommendations can be made.

In Brazil, the specific profile was found to be a child who is unable to read and write with poor general health and multiple visits to the doctor each year. Since there is limited diagnosis of intellectual disability in Brazil this may act as a good indicator for schools to highlight if a pupil requires additional support. The earlier a diagnosis is achieved, the less likely the gap in educational attainment is to widen between disabled and non-disabled children.

In the UK, the specific profile was found to be a child in a family which requires additional financial support along with a poorer general health. If this profile is used in order to determine the educational support available to a child it may mean that children who have an underlying disability but are yet to be diagnosed may benefit earlier and are therefore less likely to fall behind their peers. This in turn means that they may be less likely to require financial support and assistance from the government in the future, breaking the cycle.

In this work, it has been determined when and how it is appropriate to use sampling weights when conducting regression modelling, clearing up existing confusion from literature. A test for the appropriateness of the use of sampling weights in a logistic regression setting has also been developed.

There have been no large scale studies into intellectual disability in Brazil and so this is a unique contribution to this field.

8.2 Future work

An additional statistical issue with both datasets, which has not been investigated in this project, is the imbalance in the responses. For both data sets, there is only a small proportion of children who have an intellectual disability. Analysis of imbalanced data may be biased towards the majority group (Wang and Yao, 2009). There are ways in

which the data can be trained in order to balance it including over-sampling the minority class or under-sampling the majority class (Frei, 2019). These methods however have not been researched for survey data. It is thought that the use of either of these methods would confound with the sampling weights. In future it would be useful to examine how to adapt these methods for survey data and investigate whether there are any further methods to deal with imbalanced survey data.

In Chapter 7, a method of cross validation for the survey weighted group logistic lasso was considered. However, it was not possible to implement this method for unknown reasons. A further look into this will allow an optimal value of the penalty parameter to be selected and survey weighted group lasso to be implemented using only the `grplasso` package, without the need of further packages.

When analysing the data from the MCS, only a subset of the available variables were selected in order to allow an international comparison with Brazil to be made. Therefore, it would be interesting to see whether any of the additional variables in the MCS are also found to be related to intellectual disability and allow a more detailed profile of a child who is more likely to have an intellectual disability in the UK to be built.

The MCS is a longitudinal study and so it would also be interesting to see how the relationship between intellectual disability and the variables from the MCS change over time. There is currently no data available in Brazil, however, that would allow a longitudinal study such as this to be completed.

Appendix A

Appendix

A.1 Sampling weights - simulations

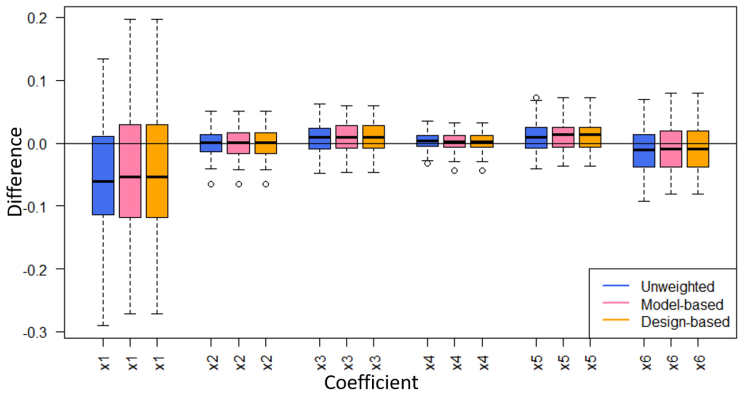

Figure A.1: The difference between the coefficient estimates and the true values for the stratified sample, not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) when sampling 20% of the population.

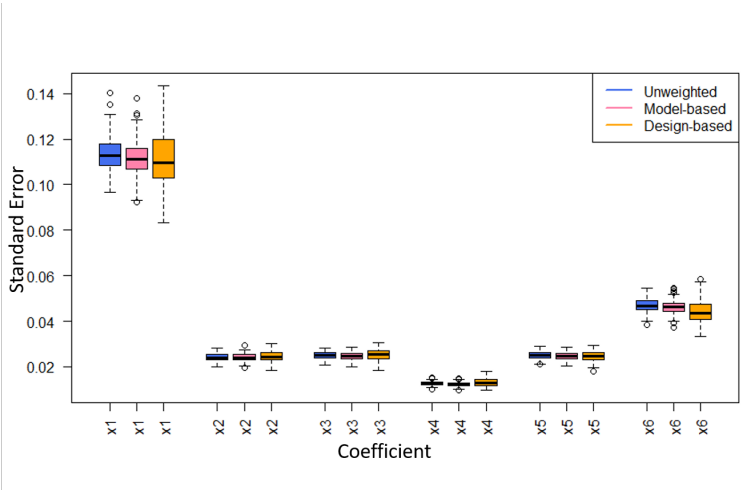

Figure A.2: The standard errors of the coefficient estimates for the stratified sample, not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) when sampling 20% of the population.

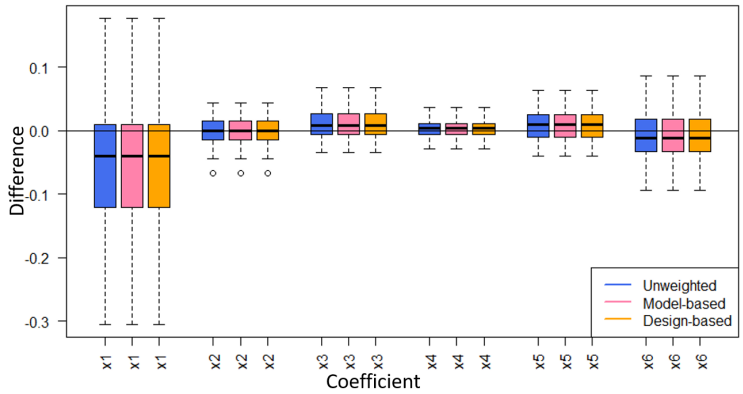

Figure A.3: The difference between the coefficient estimates and the true values for the stratified sample with proportional allocation, not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) when sampling 20% of the population.

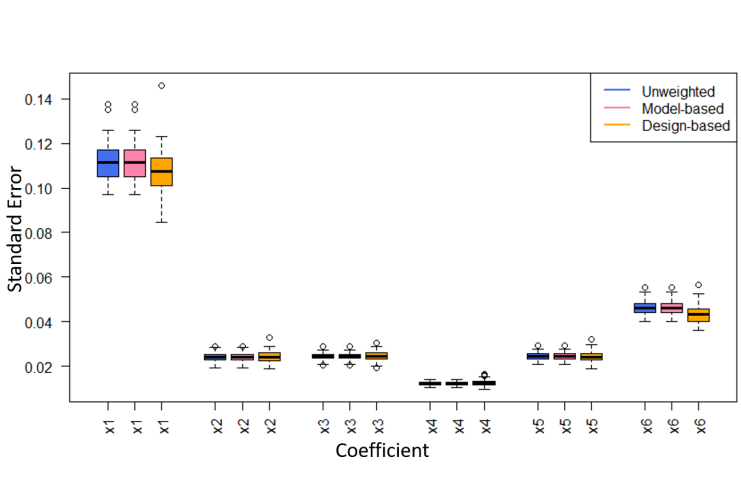

Figure A.4: The standard errors of the coefficient estimates for the stratified sample with proportional allocation, not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) when sampling 20% of the population.

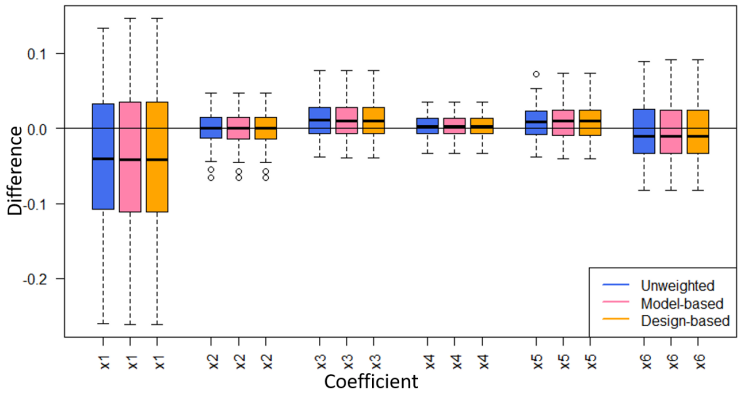

Figure A.5: The difference between the coefficient estimates and the true values for the stratified sample with optimal allocation, not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) when sampling 20% of the population.

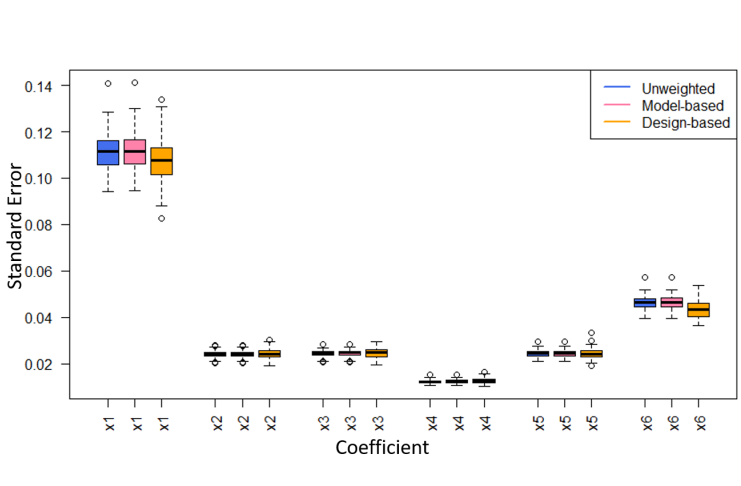

Figure A.6: The standard errors of the coefficient estimates for the stratified sample with optimal allocation, not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) when sampling 20% of the population.

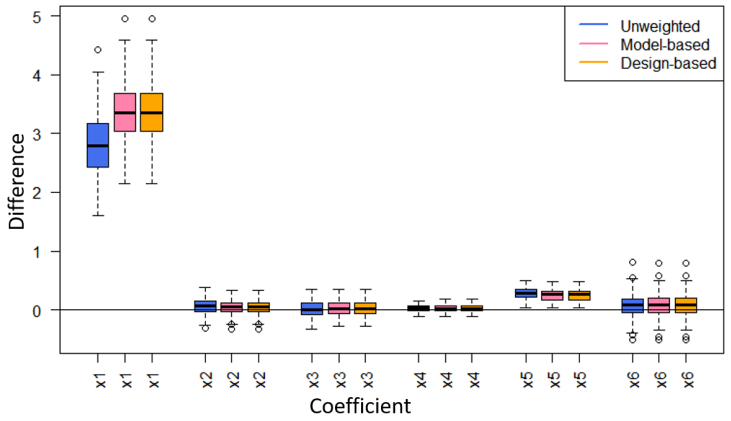

Figure A.7: The difference between the coefficient estimates and the true values for the stratified sample with underlying relationship (ignoring strata), not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) when sampling 20% of the population.

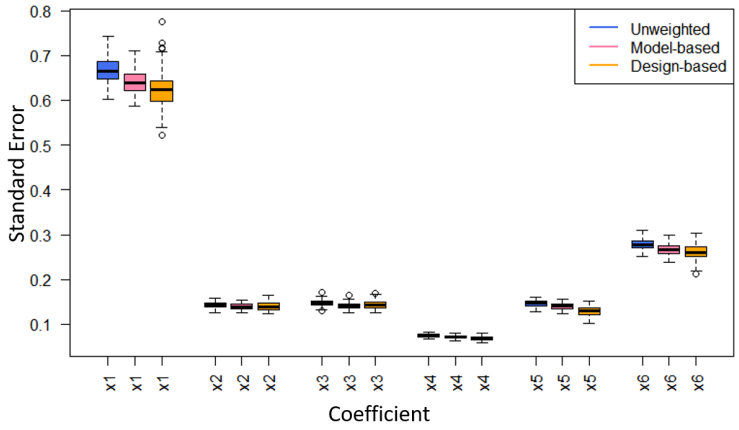

Figure A.8: The standard errors of the coefficient estimates for the stratified sample with underlying relationship (ignoring strata), not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) when sampling 20% of the population.

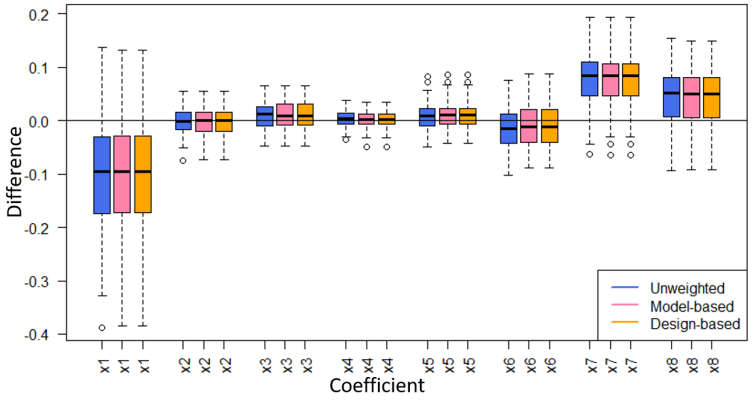

Figure A.9: The difference between the coefficient estimates and the true values for the stratified sample with underlying relationship (including strata), not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) when sampling 20% of the population.

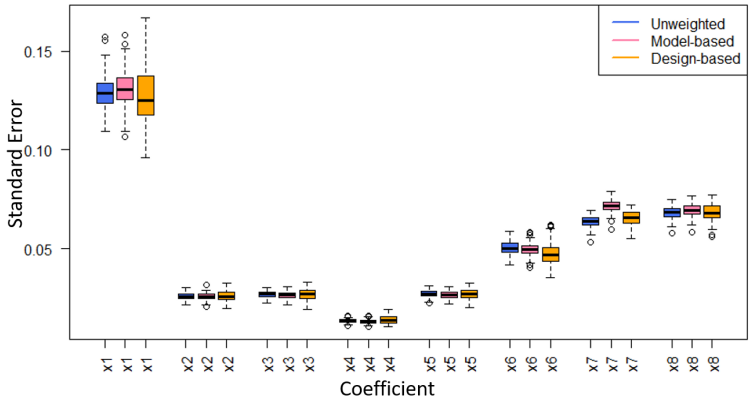

Figure A.10: The standard errors of the coefficient estimates for the stratified sample with underlying relationship (including strata), not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) when sampling 20% of the population.

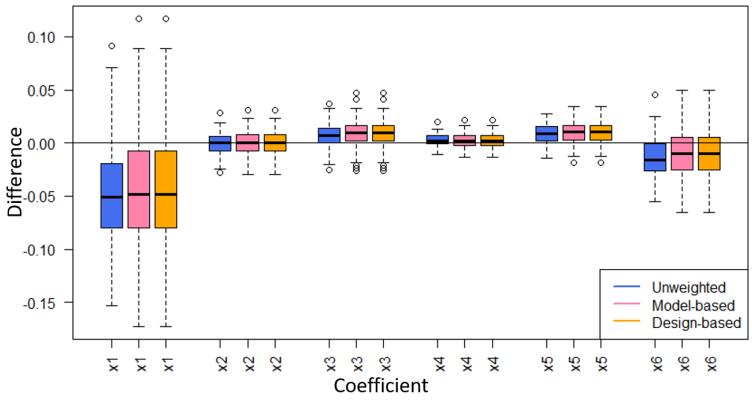

Figure A.11: The difference between the coefficient estimates and the true values for the stratified sample, not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) when sampling 50% of the population.

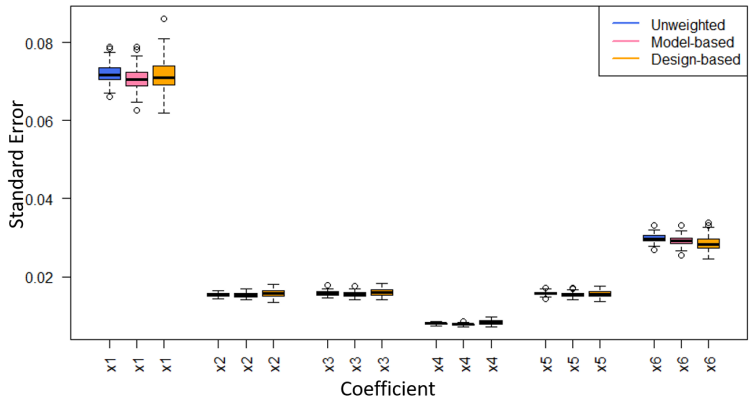

Figure A.12: The standard errors of the coefficient estimates for the stratified sample, not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) when sampling 50% of the population.

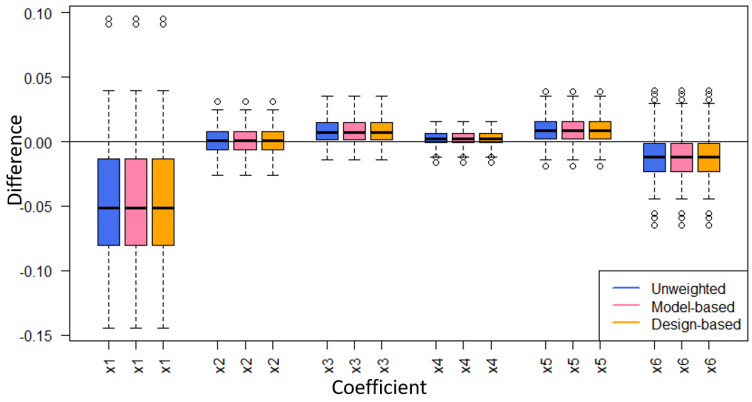

Figure A.13: The difference between the coefficient estimates and the true values for the stratified sample with proportional allocation, not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) when sampling 50% of the population.

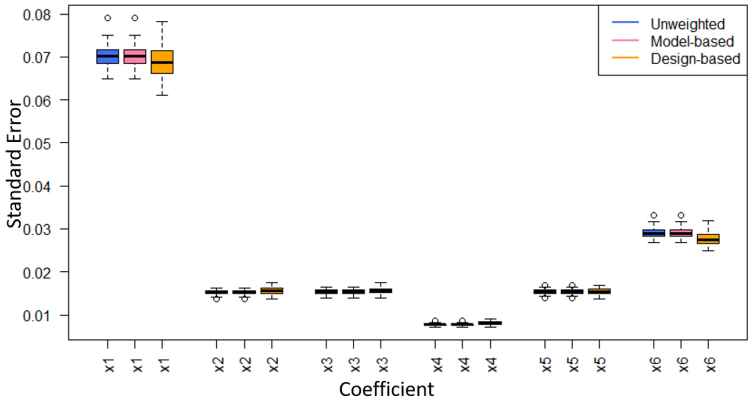

Figure A.14: The standard errors of the coefficient estimates for the stratified sample with proportional allocation, not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) when sampling 50% of the population.

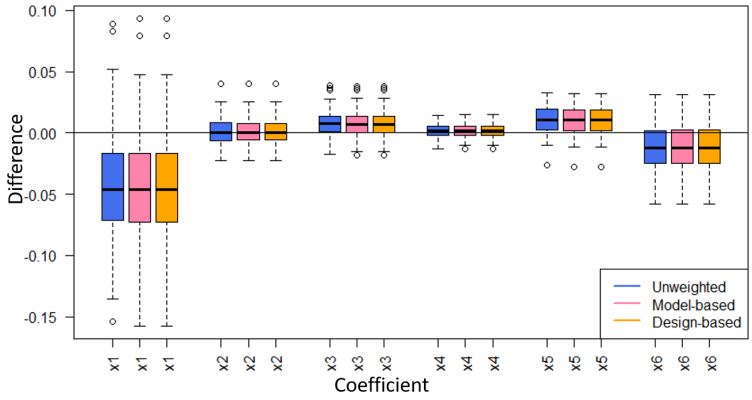

Figure A.15: The difference between the coefficient estimates and the true values for the stratified sample with optimal allocation, not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) when sampling 50% of the population.

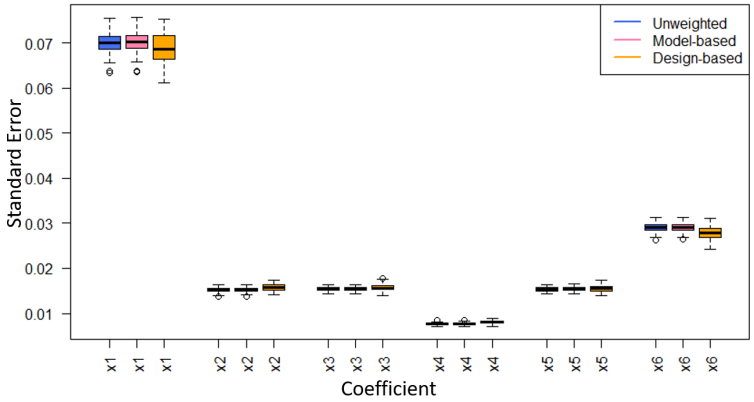

Figure A.16: The standard errors of the coefficient estimates for the stratified sample with optimal allocation, not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) when sampling 50% of the population.

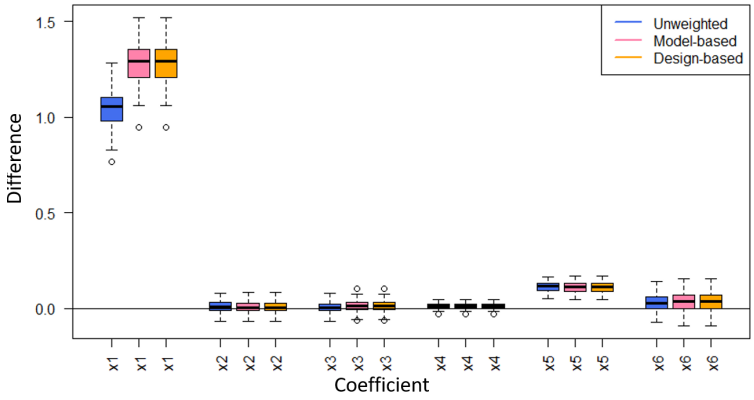

Figure A.17: The difference between the coefficient estimates and the true values for the stratified sample with underlying relationship (ignoring strata), not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) when sampling 50% of the population.

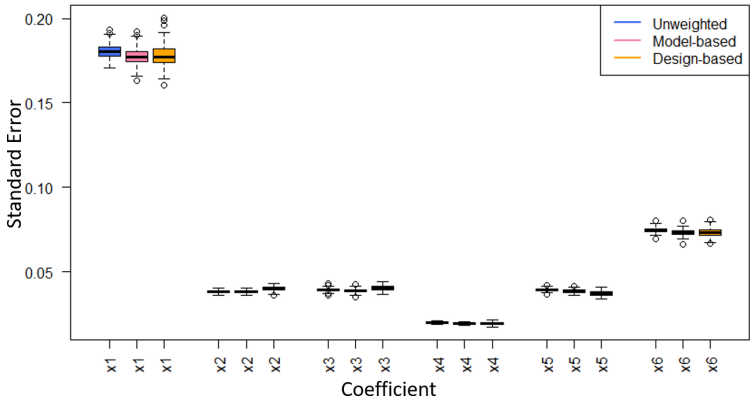

Figure A.18: The standard errors of the coefficient estimates for the stratified sample with underlying relationship (ignoring strata), not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) when sampling 50% of the population.

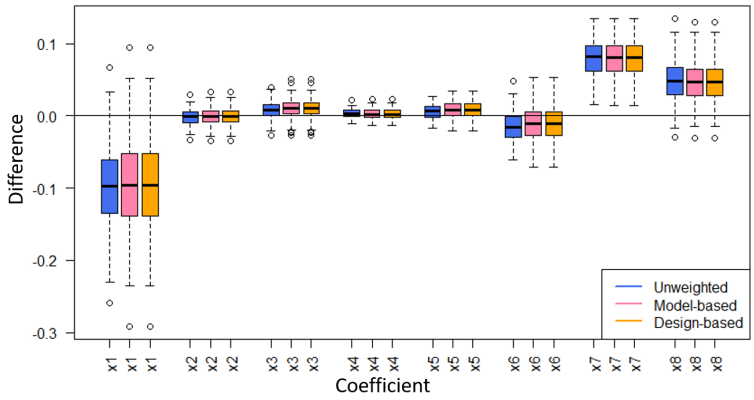

Figure A.19: The difference between the coefficient estimates and the true values for the stratified sample with underlying relationship (including strata), not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) when sampling 50% of the population.

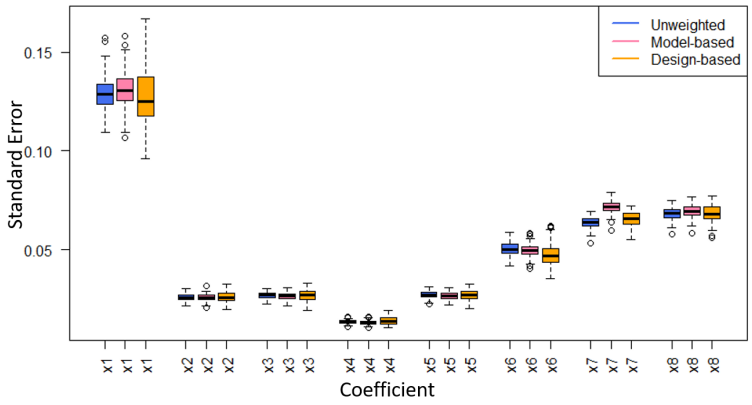

Figure A.20: The standard errors of the coefficient estimates for the stratified sample with underlying relationship (including strata), not including weights (blue), including weights - model-based (pink) and including weights - design-based (orange) when sampling 20% of the population.

References

Asparouhav, T. and Muthen, B. (2007). Testing for informative weights and weights trimming in multivariate modeling with survey data. *Proceedings of the 2007 JSM meeting in Salt Lake City, Utah, Section on Survey Research Methods.*

Assis, S. G. d. (2009). Children and youth with and without disabilities.

Battiscombe, G. (1974). *Shaftesbury: a biography of the seventh Earl, 1801-1885*. London: Constable.

Block, P. (2007). Institutional utopias, eugenics, and intellectual disability in brazil. *History and Anthropology*, 18(2):177–196.

Bollen, K. A., Biemer, P. P., Karr, A. F., Tueller, S., and Berzofsky, M. E. (2016). Are survey weights needed? a review of diagnostic tests in regression analysis. *Annual Review of Statistics and Its Application*, 3(1):375–392.

British Council (2013). The education systems of england & wales, scotland and northern ireland.

British Psychological Society (2000). Learning disability: Definitions and contexts. British Psychological Society Leicester.

Bursac, Z., Gauss, C. H., Williams, D. K., and Hosmer, D. W. (2008). Purposeful selection of variables in logistic regression. *Source code for biology and medicine*, 3(1):17.

Carvalho, E. N. S. d. and Forrester-Jones, R. (2016). Country profile: intellectual disability in brazil. *Tizard Learning Disability Review*, 21(2):65–74.

Chambers, R. L. and Skinner, C. J. (2003). *Analysis of survey data*. John Wiley & Sons.

Chromy, J. R. and Abeyasekera, S. (2005). Statistical analysis of survey data. *Household sample surveys in developing and transition countries, studies in methods. New York: United Nations.*

Cooper, V., Emerson, E., Glover, G., Gore, N. J., Hassiotis, A., Hastings, R., Knapp, M. R. J., McGill, P., Oliver, C., Pinney, A., et al. (2014). Early intervention for children with learning disabilities whose behaviour challenges. briefing paper. *Challenging Behaviour Foundation.*

Cullis, A. (2007). Infant mortality in the millennium cohort study (mcs) sample areas.

Damacena, G. N., Janeiro-rj, R. D., Janeiro-rj, R. D., Federal, U., Gerais, D. M., Enfermagem, E. D., Horizonte-mg, B., Janeiro-rj, R. D., Pesquisas, D. D., Janeiro-rj, R. D., Pereira, C. A., Pesquisas, D. D., and Janeiro-rj, R. D. (2013). The Development of the National Health Survey in Brazil. *National Health Survey Methodology Article*, 24.

Dauncey, M. (2015). Special Education al Needs (SEN)/ Additional Learning Needs (ALN) in Wales. *The National Assembly for Wales.*

De Leeuw, E. D., Hox, J., and Dillman, D. (2012). *International handbook of survey methodology.* Routledge.

Dickens, W. T. (1990). Error components in grouped data: Is it ever worth weighting? *Review of Economics and Statistics*, 72(2):328–333.

Dorazio, R. M. (1999). Design-based and model-based inference in surveys of freshwater mollusks. *Journal of the North American Benthological Society*, 18(1):118–131.

Drysdale, E., Peng, Y., Hanna, T. P., Nguyen, P., and Goldenberg, A. (2019). The false positive control lasso. *arXiv preprint arXiv:1903.12584.*

DuMouchel, W. H. and Duncan, G. J. (2008). Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples. *Journal of the American Statistical Association*, 78.

Durrant, G. B. et al. (2005). Imputation methods for handling item-nonresponse in the social sciences: a methodological review. *ESRC National Centre for Research*

Methods and Southampton Statistical Sciences Research Institute. NCRM Methods Review Papers NCRM/002.

Emerson, E. (2015). The Determinants of Health Inequalities Experienced by Children with Learning Disabilities. *Public Health England.*

Emerson, E., Hatton, C., Robertson, J., Roberts, H., Baines, S., Evison, F., and Glover, G. (2012). People with learning disabilities in england 2011. *Durham: Improving Health & Lives: Learning Disabilities Observatory.*

Emerson, E. and Heslop, P. (2010). A working definition of learning disabilities. *Improving Health and Lives: Learning Disabilities Observatory*, pages 1–4.

Emerson, E., Robertson, J., Baines, S., and Hatton, C. (2016). Obesity in british children with and without intellectual disability: cohort study. *BMC Public Health*, 16(1):644.

Emerson, E., Shahtahmasebi, S., Lancaster, G., and Berridge, D. (2010). Poverty transitions among families supporting a child with intellectual disability. *Journal of Intellectual and Developmental Disability*, 35:224–234.

Eurostat (1998). Recommendations on social exclusion and poverty statistics. *Luxembourg: Statistical Office of the European Communities.*

Finn, D. and Goodship, J. (2014). Take-up of benefits and poverty: an evidence and policy review. *JRF/CESI Report.*

FPLD (2017). Learning disabilities.

França, I. S. X. d., Pagliuca, L. M. F., and Baptista, R. S. (2008). Policies for the inclusion of disabled people: limits and possibilities. *Acta Paulista de Enfermagem*, 21(1):112–116.

Frei, L. (2019). A deep dive into imbalanced data: Over-sampling.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Fuller, W. (2009). *Sampling Statistics.* Wiley.

Fundação Instituto Brasileiro de Geografia and Estatística. Departamento de População and Indicadores Sociais (1998). *Pesquisa sobre padrões de vida, 1996-1997*. Ibge.

Goulden, C. and D'arcy, C. (2014). A Definition of Poverty. *JRF Programme Paper*, pages 1–10.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2011). *Survey methodology*, volume 561. John Wiley & Sons.

Guo, P. (2015). Improved variable selection algorithm using a lasso-type penalty, with an application to assessing hepatitis b infection relevant factors in community residents. *PLoS ONE*, 10.

Hahs-Vaughn, D. L. and Lomax, R. G. (2006). Utilization of sample weights in single-level structural equation modeling. *The Journal of Experimental Education*, 74(2):163–190.

Hansen, K. (2012). Millennium Cohort Study: First, second, third and fourth surveys. A guide to the datasets (Seventh Edition). *Centre for Longitudinal Studies*, pages 1–114.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity - The Lasso and Generalisations*. CRC Press.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46(6):1251–1271.

Heinze, G. and Dunkler, D. (2017). Five myths about variable selection. *Transplant International*, 30(1):6–10.

Heinze, G., Wallisch, C., and Dunkler, D. (2018). Variable selection–a review and recommendations for the practicing statistician. *Biometrical Journal*, 60(3):431–449.

Hintze, J. (2007). Ncss statistical system user's guide iii: Regression and curve fitting.

Jannuzzi, G. d. M. (2005). A educação do deficiente no brasil: dos primórdios ao início do século xxi. *Cadernos de Pesquisa*, 35(124):255–256.

Javanmard, A., Javadi, H., et al. (2019). False discovery rate control via debiased lasso. *Electronic Journal of Statistics*, 13(1):1212–1253.

Joshi, H., Plewis, I., Cullis, A., Sadigh, M., Dodd, K., Woods., and Chapman, A. (2002). CLS Cohort Studies Centre for Longitudinal Studies Study.

Ke, X. and Liu, J. (2015). Deficiência intelectual. *Textbook of Child and Adolescent Mental Health.*, pages 1–27.

Kott, P. S. (1991). A model-based look at linear regression with survey data. *The American Statistician*, 45(2):107–112.

Lai, D.-C., Tseng, Y.-C., Hou, Y.-M., and Guo, H.-R. (2012). Gender and geographic differences in the prevalence of autism spectrum disorders in children: Analysis of data from the national disability registry of taiwan. *Research in developmental disabilities*, 33(3):909–915.

Lee, J., Sun, Y., and Saunders, M. (2014). Proximal newton-type methods for minimising composite functions. *SIAM Journal on Optimization*.

Lehtonen, R. and Pahkinen, E. (2004). *Practical methods for design and analysis of complex surveys*. John Wiley & Sons.

Lohr, S. L. (2009). *Sampling: design and analysis*. Nelson Education.

Lumley, T. (2010). *Complex Surveys - A Guide to Analysis Using R*. Wiley.

Mantoan, M. T. E. and Valente, J. A. (1998). Special education reform in brazil: an historical analysis of educational policies. *European Journal of Special Needs Education*, 13(1):10–28.

Marinho, A. S. d. N. (2009). Poverty, disability and violence. *Ciencia & saude coletiva*, 14:21–23.

Martin, E. A. and De La Puente, M. (1993). *Research on sources of undercoverage within households*. US Bureau of the Census.

Maulik, P. K., Mascarenhas, M. N., Mathers, C. D., Dua, T., and Saxena, S. (2011). Prevalence of intellectual disability: a meta-analysis of population-based studies. *Research in developmental disabilities*, 32(2):419–436.

McConville, K. (2011). *Improved estimation for complex surveys using modern regression techniques*. PhD thesis, Colorado State University. Libraries.

McConville, K. S., Breidt, F. J., Lee, T. C., and Moisen, G. G. (2017). Model-assisted survey regression estimation with the lasso. *Journal of Survey Statistics and Methodology*, 5(2):131–158.

McKenzie, K. (2013). Searching for a diagnosis.

McKenzie, K., Murray, G., Murray, A., Delahunty, L., Hutton, L., Murray, K., and O'Hare, A. (2018). Child and adolescent intellectual disability screening questionnaire to identify children with intellectual disability. *Developmental Medicine & Child Neurology*, 61(4):444–450.

McKenzie, K., Paxton, D., Murray, G., Milanesi, P., and Murray, A. L. (2012). The evaluation of a screening tool for children with an intellectual disability: The child and adolescent intellectual disability screening questionnaire. *Research in developmental disabilities*, 33(4):1068–1075.

Meier, L., Van De Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71.

Mercadante, M. T., Evans-Lacko, S., and Paula, C. S. (2009). Perspectives of intellectual disability in latin american countries: epidemiology, policy, and services for children and adults. *Current Opinion in Psychiatry*, 22(5):469–474.

Meyer, A. (2010). Brazil education. `https://www.brazil.org.za/brazil-education.html`.

Mostafa, T., Platt, L., Rosenberg, R., and Smith, K. (2014). Millennium Cohort Study. *Centre for Longitudinal Studies*.

Nascimento Silva, P. and Skinner, C. J. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, 23(1):23–32.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97:558–606.

Norris, N. (2007). Raising the school leaving age. *Cambridge Journal of Education*, 37(4):471–472.

Oakland, T. (2004). Learning disabilities internationally and in brazil: issues to consider in developing services for brazilian students with learning disabilities. *Avaliação Psicológica*, 3(2):115–120.

Park, M. Y. and Hastie, T. (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677.

Paulo, M. and Freitas, S. D. (2014a). Coordenação de Métodos e Qualidade - COMEQ Sumário. pages 1–22.

Paulo, M. and Freitas, S. D. (2014b). Coordenação de Métodos e Qualidade - COMEQ Sumário. pages 1–22.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review / Revue Internationale de Statistique*, 61(2):317–337.

Pfeffermann, D. and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, 61(1):166–186.

Plewis, I., Calderwood, L., Hawkes, D., Hughes, G., and Joshi, H. (2007). Millennium cohort study: technical report on sampling. *London: Centre for Longitudinal Study, Institute of Education*.

Razzouk, D., Zorzetto, R., Dubugras, M. T., Gerolin, J., and Mari, J. d. J. (2006). Mental health and psychiatry research in brazil: scientific production from 1999 to 2003. *Revista de Saúde Pública*, 40:93–100.

Roberts, G., Rao, N., and Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74(1):1–12.

Roser, M. and Ortiz-Ospina, E. (2016). Global rise of education. *Our World in Data*.

Sampson, J. N., Chatterjee, N., Carroll, R. J., and Müller, S. (2013). Controlling the local false discovery rate in the adaptive lasso. *Biostatistics*, 14(4):653–666.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

Scior, K. and Werner, S. (2015). Changing attitudes to learning disability. *London: Mencap*.

Shahtahmasebi, S., Emerson, E., Berridge, D., and Lancaster, G. (2011). Child Disability and the Dynamics of Family Poverty, Hardship and Financial Strain: Evidence from the UK. *Journal of Social Policy*, 40(04):653–673.

Solon, G., Haider, S. J., and Wooldridge, J. (2013). What are we weighting for? NBER Working Papers 18859, National Bureau of Economic Research, Inc.

Souza-Júnior, P. R. B. D., Freitas, M. P. S. D., Antonaci, G. D. A., and Szwarcwald, C. L. (2015). Desenho da amostra da Pesquisa Nacional de Saúde 2013. *Epidemiologia e Serviços de Saúde*, 24(2):207–216.

Stanek, C. (2013). The educational system of brazil. *IEM Spotlight*, 10(1):1–6.

Tibshirani, R. (1996a). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. (1996b). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, 58:267–288.

Tramontina, S., Martins, S., Michalowski, M. B., Ketzer, C. R., Eizirik, M., Biederman, J., and Rohde, L. A. (2002). Estimated mental retardation and school dropout in a sample of students from state public schools in porto alegre, brazil. *Brazilian Journal of Psychiatry*, 24(4):177–181.

Wang, S. and Yao, X. (2009). Diversity analysis on imbalanced data sets by using ensemble models. In *2009 IEEE Symposium on Computational Intelligence and Data Mining*, pages 324–331. IEEE.

Wears, R. L. and Lewis, R. J. (1999). Statistical models and occam's razor. *Academic Emergency Medicine*, 6(2):93–94.

Wheeler, D. C., VanHorn, J. E., and Paskett, E. (2008). A comparison of design-based and model-based analysis of sample surveys in geography. *The Professional Geographer*, 60(4):466–477.

WHO (1992). *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines*. Geneva: World Health Organization.

Winerman, L. (2018). Big data gets bigger.

Winship, C. and Radbill, L. (1994). Sampling weights and regression analysis. *Sociological Methods & Research*, 23(2):230–257.

Wood, D. (2006). What kind of school shall we be? *The Guardian*.

Yang, Y. and Zou, H. (2013). A fast unified algorithm for computing group-lasso penalized learning problems. *Statistics and Computing, Accepted*.

Yansaneh, I. S. (2003). Construction and use of sample weights. *Designing Household Surveys Samples: Practical Guidelines*.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.