EXTENDING A TIME-VARYING MULTIVARIABLE MENDELIAN RANDOMISATION MODEL TO ACCOMMODATE TWO OUTCOME MEASUREMENTS

EXTENDING A TIME-VARYING MULTIVARIABLE MENDELIAN RANDOMISATION MODEL TO ACCOMMODATE TWO OUTCOME MEASUREMENTS

By ALEXANDER PERO, Mathematics and Statistics

A Thesis Submitted to the School of Graduate Studies in Partial
Fulfillment of the Requirements for
the Degree Master of Statistics

McMaster University

MASTER OF STATISTICS (2024)

Hamilton, Ontario, Canada (Mathematics and Statistics)

TITLE:	Extending a time-varying multivariable Mendelian randomisation model to accommodate two outcome measurements
AUTHOR:	Alexander Pero BS (Mathematics and Statistics), McMaster University, Hamilton, Canada
SUPERVISOR:	Dr. Angelo Canty and Dr. Katherine Davies
NUMBER OF PAGES:	x, 60

Abstract

The application of multivariable Mendelian randomisation (MVMR) to analyse time-varying data with multiple measurements of both an exposure and an outcome is unclear. The purpose of this thesis is to develop and examine the properties of a potential model to extend MVMR to handle two measurements of both an outcome and an exposure.

The exposure effect at Time 1 is estimated using univariable Mendelian randomisation (MR), while the exposure effects at Time 2 are estimated using MVMR by using a set of single nucleotide polymorphisms (SNPs) exclusive to the first outcome measurement. Simulations examining the properties of the causal effect estimates in the model under different scenarios were undertaken. The scenarios included different sampling schemes (1, 2, or 4 samples) for summary statistics.

Confidence intervals were too wide, over-coverage was present when following the one-sample scheme, while slight under-coverage in both the two-sample and four-sample schemes was observed. Parameter estimators appeared to be mainly unaffected by increasing instrument strength. Increasing the number of SNPs pertaining to each exposure led to increased biases for the parameters affecting the second outcome measurement. Lastly, parameter estimates maintained acceptable coverage and small biases for different scenarios of overlapping SNPs.

The inclusion of SNPs pertaining to the first outcome measurement in a time-varying

MVMR model with two exposure and two outcome measurements allows for the estimation of exposure effects at both time points. However, the apparent drop in performance when the number of SNPs increases is of concern.

Acknowledgements

I would like to thank my supervisors, Dr. Angelo Canty and Dr. Katherine Davies, for supporting and motivating me throughout the completion of my degree. Their aid, feedback, and flexibility with my research were invaluable and greatly helped. Thank you to Dr. Ben Bolker for agreeing to take part in my defence. Thank you to my family and friends for always supporting me in my endeavours.

Table of Contents

Abstract **iii**

Acknowledgements **v**

1 Introduction **1**

2 Background **3**

2.1 Genetics 3

2.2 Mendelian Randomisation 5

2.3 Multivariable Mendelian Randomisation 15

2.4 Time-Varying Mendelian Randomisation 18

3 Extending the Time-Varying Model for Two Time Points of the Exposure and Outcome **23**

4 Simulations and Results **31**

4.1 Simulation 1 - Causal effect estimates without SNPs G_3 33

4.2 Simulation 2 - Causal effect estimates with SNPs G_3 for different sampling schemes 37

4.3 Simulation 3 - Causal effect estimates with SNPs G_3 for different magnitudes of SNP effects 40

4.4 Simulation 4 - Causal effect estimates with SNPs G_3 for different numbers of SNPs 40

4.5 Simulation 5 - Causal effect estimates with SNPs G_3 and G_{23} 44

5 Conclusion **51**

List of Figures

2.1 Examining the relationship between exposure and outcome without (a) vs with (b) instrumental variable 6

2.2 A comparison of Mendelian randomisation (MR) and Randomised control trial (RCT). 8

2.3 Graphical representation of the IV assumptions with G denoting the IV, X denoting the exposure, Y denoting the outcome, U denoting unobserved confounders. A dashed arrow represents desired variable relationships whereas a dotted arrow represents undesired variable relationships under core instrumental variable (IV) assumptions. 9

3.4 Preliminary model with a single time varying exposure X and outcome Y measured at two different time points 1 and 2. X_1 and X_2 as well as Y_1 do not share any SNPs. 25

3.5 Proposed model with a single time varying exposure X and outcome Y measured at two different time points 1 and 2. Exposures X_1 and X_2 as well as outcome Y_1 do not share any SNPs with the addition of SNPs G_3. 30

4.6 Boxplots of parameter estimates for an empty set of SNPs G_3. True parameter values are indicated by the red lines in each plot. 36

4.7 Boxplots of parameter estimates for different numbers of samples. True parameter values are indicated by the red lines in each plot. 39

4.8 Boxplots of parameter estimates for different minimum absolute SNP effects. True parameter values are indicated by the red lines in each plot. . 42

4.9 Boxplots of parameter estimates for different numbers of SNPs in each G_k. True parameter values are indicated by the red lines in each plot. . . 45

4.10 Boxplots of parameter estimates for different proportions of overlap. True parameter values are indicated by the red lines in each plot. 48

4.11 Boxplots of parameter estimates for different levels of overlap with fixed SNP total. True parameter values are indicated by the red lines in each plot. 50

List of Tables

4.1 Mean estimates, standard deviations, and mean standard errors with estimator bias and coverage when G_3 is empty 35

4.2 Mean estimates, standard deviations, and mean standard errors with estimator bias and coverage by number of samples 38

4.3 Mean estimates, standard deviations, and mean standard errors with estimator bias and coverage for different minimum absolute instrument strengths 41

4.4 Mean estimates, standard deviations, and mean standard errors with estimator bias and coverage for different numbers of SNPs per set G_k . . . 43

4.5 SNP combinations for Scenario 1 - modifying the proportion of overlap . 46

4.6 SNP combinations for Scenario 2 - modifying the proportion of overlap with fixed total SNPs 46

4.7 Mean estimates, standard deviations, and mean standard errors with estimator bias and coverage for different proportions of overlap G_{23} 47

4.8 Mean estimates, standard deviations, and mean standard errors with estimator bias and coverage for differing overlap with fixed SNP total . . . 49

Chapter 1

Introduction

When measuring two variables, for example Body Mass Index (BMI) and Blood Pressure (BP), we are often interested in how BMI affects BP. Through observational studies we may deduce a correlation between BMI and BP. However, to tease apart a potential causal relationship one may conduct a randomized control trial (RCT). Although these studies are the gold standard, they are often time-consuming and expensive. Fortunately, Mendelian randomisation (MR), a method for deducing causal effects from observational data with the use of genetic variants or single nucleotide polymorphisms (SNPs) as instrumental variables (IV), can be performed (Davey Smith and Ebrahim, 2003; Lawlor et al., 2008).

We may also be interested in how the relationship between the two variables changes over time. Knowing how the relationship changes can help guide policy or intervention with the goal of modifying the exposure and the outcome. A recent study employed a specific form of MR called multivariable Mendelian randomisation to analyse the effect of a time-varying exposure on an outcome (Sanderson et al., 2022). One such example is analysing the relationship between adolescent BMI, adult BMI, and adult BP.

However, in that paper they assumed there was a single measure of the outcome and did not consider the implications of adding another measurement of the outcome, for example adolescent BP. In this thesis, we will examine how MR can be used to analyse the relationship between two measurements of exposure and outcome. As such, this thesis aims to analyse a model with an exposure and outcome measured at two distinct time points. It is assumed the first outcome measurement is taken between the two exposure measurements, whereas the second outcome measurement is taken after the second exposure measurement.

The thesis is organized in the following manner. Chapter 2 provides the background in genetics required to understand Mendelian randomisation. The chapter also provides a literature review and background of Mendelian randomisation, multivariable Mendelian randomisation and the application of Mendelian randomisation to time-varying data. Chapter 3 describes a proposed model detailing a time-varying scenario with two measurements of an exposure and an outcome with a proposed methodology to analyse the model. Chapter 4 includes a description of the simulation studies we performed to evaluate the properties of the model and the results of those studies. A discussion as well as future potential directions of research is provided in Chapter 5.

Chapter 2

Background

Since MR requires a certain background knowledge in genetics, this chapter focuses on introducing key concepts in genetics as related to MR followed by a brief background for instrumental variable analysis, on which MR is based, as well as a presentation of important methods and assumptions in MR as related to the work undertaken in this thesis.

2.1 Genetics

A chromosome is a molecule of deoxyribonucleic acid (DNA) carrying genetic information. Humans are born with 23 pairs of chromosomes. One chromosome from each pair is a copy inherited from the mother while the other is inherited from the father. The first 22 pairs of chromosomes are autosomes (or non-sex chromosomes). The 23rd is a pair of sex chromosomes which can either be XX for a female or XY for a male (Davey Smith and Ebrahim, 2003).

A DNA molecule has a double helix structure composed of two linked strands of nucleotide bases wound together. Nucleotide bases are the building blocks of DNA. The four possible bases are adenine (A), cytosine (C), thymine (T), and guanine (G). Within the double helix structure the strands of nucleotides form complementary sequences where A pairs with T, and C pairs with G. Therefore, it is only required to know the sequence of one strand to obtain all the information pertaining to the sequence of the other strand; they contain the same information. For example, if at a particular location on the DNA strand, or locus, the genetic sequence is AATGCT then the complementary strand will read TTACGA (Davey Smith and Ebrahim, 2003).

Individuals possess mutations in their genetic code. A mutation affecting a single nucleotide at a particular locus for at least 1% of the population is known as a single nucleotide polymorphism (SNP). Despite many different forms of mutations, SNPs are the most common. Other mutations concerning single nucleotides at a particular locus affecting less than 1% of the population are classified as rare variants (Wright, 2005). SNP variant forms are known as alleles. If there are only two possible variants of a particular SNP, it is a biallelic SNP. Although SNPs with three or four alleles, they are very rare and the remainder of the thesis will solely focus on biallelic SNPs. The prevalent variant in the population is known as the major allele, while the less prevalent variant is known as the minor allele.

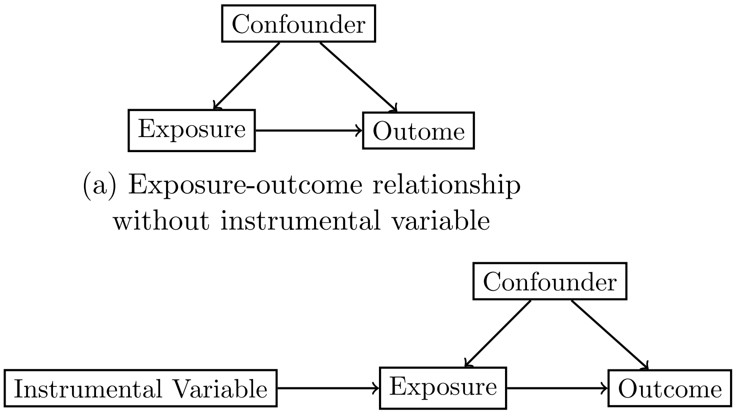
An individual's genotype at a specific locus is represented by their inherited alleles. An individual inherits one allele from each parent. If both alleles are the same the individual is considered homozygous. If they are different the individual is heterozygous. It is often convenient to represent the genotype by the number of minor alleles, either 0, 1, or 2. Alleles inherited at genetic markers far apart on the same chromosome are considered independent since correlation in the genome decreases as the distance between genetic markers increases. The non-random inheritance of alleles caused by SNPs closely located on the chromosome as a result of the recombination of the segregated chromosome leads to a correlation in the inheritance of certain alleles known as linkage disequilibrium (Davey Smith and Ebrahim, 2003).

The principle detailing the distribution of alleles within the population is called Hardy-Weinburg equilibrium. It states that allele and genotype frequencies will remain constant within a population in the absence of evolutionary influences. The frequencies will also remain constant from generation to generation. If the frequency of the minor allele, b, within the population (minor allele frequency, MAF) is denoted p and the frequency of the major allele, B, is denoted q (where $p + q = 1$), then under the assumption of random mating the expected frequencies of the genotypes for SNPs following the Hardy-Weinburg equilibrium are p^2 for homozygote bb, $2pq$ for heterzygote Bb (or bB), and q^2 for homozygote BB. By defining G as the number of minor alleles (with possible values 0, 1, or 2) and with a probability of inheriting a minor allele equivalent to the MAF p, then the distribution of G is clearly Binomial with parameters (2, p) (Hardy, 1908; Weinberg, 1908).

2.2 Mendelian Randomisation

Mendelian randomisation is based on instrumental variable analysis, a method developed in econometrics to examine causal relationships from observational data (Thomas and Conti, 2004). While the term *instrumental variable* appeared in 1945 (Reiersöl, 1945), the first recorded use of instrumental variables in 1928 pertained to the analysis of the effects of tariffs on vegetable and animal oils (Wright, 1928). An instrumental variable

(IV) is a variable strongly associated with an exposure employed to explore a causal relationship between the exposure and outcome (Figure 2.1) (Lawlor et al., 2008). An exposure is a potential causal risk factor for an outcome, and an outcome is a factor or trait that is thought to be affected by the exposure (Burgess et al., 2023).



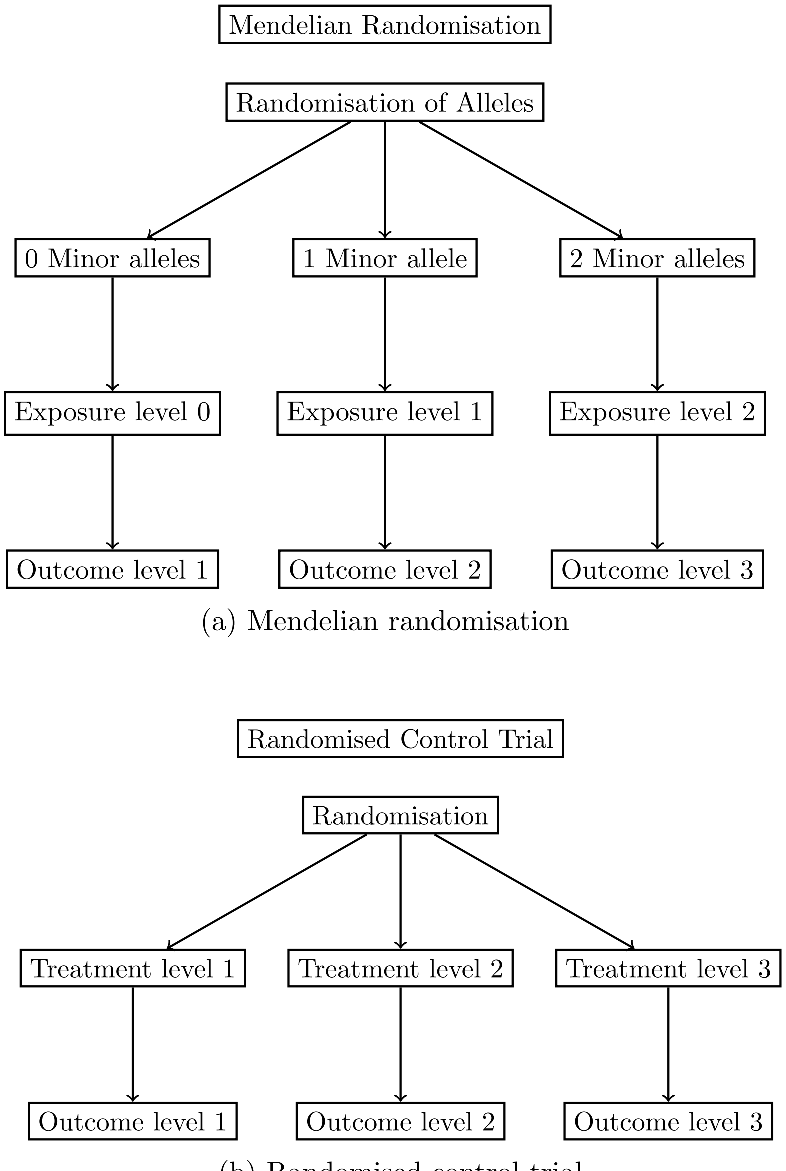
(a) Exposure-outcome relationship without instrumental variable

(b) Exposure-outcome relationship with instrumental variable

Figure 2.1: Examining the relationship between exposure and outcome without (a) vs with (b) instrumental variable

MR is the exploration of the causal effect of an exposure on an outcome using observational data from epidemiological studies via instrumental variable analysis using genetic instruments (Wehby et al., 2008). These genetic instruments are genetic variants or SNPs found in the population (Thomas and Conti, 2004). The first study employing the concept of Mendelian randomisation used the apolipoprotein E (ApoE) gene variants to investigate the effect of serum cholesterol levels on cancer (Katan, 2004). Subsequently, MR has been widely used as a method to overcome shortcomings in observational epidemiology. MR has often been likened to a randomised control trial (RCT). Suppose there is an RCT conducted to explore the effect of a treatment on an outcome. Participants are randomly assigned to two or more treatment groups. Randomisation helps balance the distribution of unobserved confounders between groups (Hingorani and Humphries, 2005; Smith and Ebrahim, 2005; Nitsch et al., 2006). The treatment groups are then subjected to different levels of exposure potentially resulting in different levels of outcome, if there is a causal relationship between the exposure and the outcome. The RCT allows for attribution of causal effect to the treatment since confounders are equal in distribution between groups. Genetic variants within a population are randomly distributed, separating the population into three distinct groups corresponding to the number of minor alleles. If the genetic variant selected is strongly associated with the outcome then we expect that the levels of exposure will differ across the three groups. The differing levels of exposure result in three different levels of outcome allowing for the estimation of the effect of the exposure on the outcome (Figure 2.2). Akin to a natural experiment, it is expected for confounders to be equal in distribution between groups (indicated by the number of minor alleles) through random mating and the random inheritance of alleles within a population provided that the genetic instruments selected are not related to any confounders.

Although the RCT analogy is useful to understanding the basis for Mendelian randomisation, there is a key difference. The level of exposure is usually fixed in an RCT, whereas different individuals may have different levels of exposure within a group categorised by minor allele frequency in MR. Therefore, it is important to select valid IVs to help differentiate the levels of exposure between groups in MR. Genetic variants strongly associated with the exposure of interest allow for clearer separation of groups, whereas weak instruments only slightly separate exposure groups. Weak instruments can also be characterized by large p-values of an F-test for regression for the instrument-exposure



(a) Mendelian randomisation

(b) Randomised control trial

Figure 2.2: A comparison of Mendelian randomisation (MR) and Randomised control trial (RCT).

association. Although weak instruments are valid it is preferred to choose strong instruments as they help maximize the separation between groups reducing biases due to possible confounders (Burgess and Thompson, 2011).

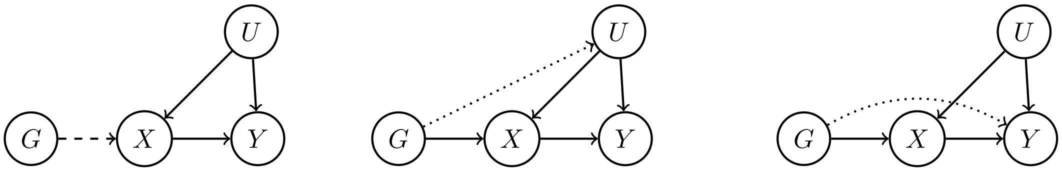
To detect a causal effect of the exposure on the outcome, the IV (or set of IVs) must satisfy the three following core assumptions:

Assumption 1: The IV is associated with the exposure of interest.

Assumption 2: The IV is not associated with any confounders in the exposure-outcome relationship.

Assumption 3: The IV does not directly affect the outcome (Didelez and Sheehan, 2007; Lawlor et al., 2008).

Assumptions 1–3 are represented graphically in Figure 2.3 with G denoting the IV, X denoting the exposure, Y denoting the outcome, and U denoting unobserved confounders.



(a) IV Assumption 1 (b) IV Assumption 2 (c) IV Assumption 3

Figure 2.3: Graphical representation of the IV assumptions with G denoting the IV, X denoting the exposure, Y denoting the outcome, U denoting unobserved confounders. A dashed arrow represents desired variable relationships whereas a dotted arrow represents undesired variable relationships under core instrumental variable (IV) assumptions.

Assumption 1 can be verified empirically using the F-statistic for simple linear regression (Equation 2.1) of the SNP-exposure association to examine the association strength between an SNP and exposure. IVs violating this assumption are known as weak instruments (Staiger and Stock, 1994).

$$F = \frac{\sum_{i=1}^{n} \left(\hat{x}_i - \bar{x}\right)^2}{\frac{\sum_{i=1}^{n} (x_i - \hat{x}_i)^2}{n-2}} \sim F_{1,n-2} \tag{2.1}$$

However, formal tests to confirm Assumption 2 do not exist. In practice, the assumption is typically verified via the use of sensitivity analyses (Glymour et al., 2012).

Instrumental variable assumptions may be violated through a variety of phenomena pertaining to biological mechanisms and non-Mendelian inheritance. The biological mechanism pleiotropy, whereby an IV is associated with the outcome through a pathway other than the exposure of interest, violates Assumption 3 (Davey Smith and Hemani, 2014). This type of violation is specifically called horizontal-pleiotropy and methods such as MR-EGGER and MR-Lasso have been developed to identify and adjust for its effects (Bowden et al., 2015; Rees et al., 2019). Linkage disequilibrium may result in the violation of Assumptions 2 or 3 as chosen IVs may be correlated to genetic variants associated with alternate exposures (Hernán and Robins, 2006).

Deriving a point estimator of the causal effect of the exposure on the outcome often requires an additional assumption. All relationships between IV, exposure, outcome, and confounders must be linear with the absence of statistical interactions (Didelez and Sheehan, 2005).

Denote a single IV by G, the continuous exposure by X, and the continuous outcome by Y. Let θ_X be the effect of G on X, θ_Y be the effect of G on Y, θ_{0_k} with k indexing the exposure or outcome, β_X the effect of X on Y, and ϵ random noise such that:

$$X = \theta_X G + \theta_{0_X} + \epsilon_{\theta_X}$$

$$\begin{aligned} Y &= \beta_X X + \epsilon_Y \\ &= \beta_X \left(\theta_X G + \theta_{0_X} + \epsilon_{\theta_X}\right) + \epsilon_Y \\ &= \beta_X \theta_X G + \beta_X \theta_{0_X} + \beta_X \epsilon_{\theta_X} + \epsilon_Y, \end{aligned}$$

we can then redefine the slope, intercept and error terms to define Y as:

$$Y = \theta_Y G + \theta_{0_Y} + \epsilon_{\theta_Y},$$

such that

$$\begin{aligned} \beta_X \theta_X G &= \theta_Y G \\ \beta_X \theta_X &= \theta_Y \\ \beta_X &= \frac{\theta_Y}{\theta_X}. \end{aligned}$$

Then the estimator for the causal effect of X on Y (Didelez et al., 2010), β_X, is given by

$$\hat{\beta}_X = \frac{\hat{\theta}_Y}{\hat{\theta}_X}, \tag{2.2}$$

otherwise known as the Wald ratio (Wald, 1940). An approximation of the asymptotic standard error for the above ratio estimator can be obtained using the delta method:

$$se\left(\hat{\beta}_X\right) \simeq \sqrt{\frac{se\left(\hat{\theta}_Y\right)^2}{\hat{\theta}_X^2} + \frac{\hat{\theta}_Y^2 se\left(\hat{\theta}_X\right)^2}{\hat{\theta}_X^4} + \frac{\hat{\theta}_Y se\left(\hat{\theta}_Y\right) se\left(\hat{\theta}_X\right)\rho}{\hat{\theta}_X^3}}, \tag{2.3}$$

where ρ is the correlation between $\hat{\theta}_Y$ and $\hat{\theta}_X$ (Thomas et al., 2007; Burgess et al., 2017).

However, in practice it is common to follow the NO Measurement Error (NOME) assumption, $\mathrm{Var}\left(\theta_X\right) \approx 0$. Under the NOME assumption, Equation 2.3 can be approximated by:

$$\begin{aligned} se\left(\hat{\beta}_X\right) &\simeq \sqrt{\frac{se\left(\hat{\theta}_Y\right)^2}{\hat{\theta}_X^2}} \\ &= \sqrt{\frac{1}{\hat{\theta}_X^2 se\left(\hat{\theta}_Y\right)^{-2}}}. \end{aligned} \tag{2.4}$$

(Bowden et al., 2016)

Although the Wald ratio estimate is suitable for a single instrument, in practice it is better to combine the information from multiple instruments. This can be done by employing the two-stage least squares method (TSLS). Define a collection of L IVs (or SNPs) as G, and then the estimate of the causal effect can be found in the following manner. In the first stage, regress the values of X on the instruments in G to then obtain fitted values of X $\left(\hat{X}\right)$ from the regression equation. Then for the second stage, regress the values of Y on $\hat{X}$; the obtained regression slope in the second stage is the causal effect estimate of interest (Angrist et al., 2000). The estimate provided by the TSLS method is exactly equivalent to the Wald ratio estimate when a single instrument is used (Burgess et al., 2017).

TSLS is a method suitable for individual-level data. Such data consist of the SNP, the exposure, and the outcome for each individual. However, with the advent of Genome-Wide Association Studies (GWAS), many studies publish readily available summary-level data of the estimated associations between the l^{th} genetic variant g_l and X $\left(\hat{\theta}_{X_l}\right)$ as well as Y $\left(\hat{\theta}_{Y_l}\right)$, including their standard errors. When there are multiple uncorrelated genetic variants ($L > 1$), the estimators for each genetic variant can be combined using the inverse-variance weighted (IVW) method where squared standard errors are used in place of the variances for practical reasons:

$$\hat{\beta}_{X_{IVW}} = \frac{\sum_{l=1}^{L} \hat{\beta}_{X_l} se\left(\hat{\beta}_{X_l}\right)^{-2}}{\sum_{l=1}^{L} se\left(\hat{\beta}_{X_l}\right)^{-2}} = \frac{\sum_{l=1}^{L} \hat{\theta}_{Y_l}\hat{\theta}_{X_l} se\left(\hat{\theta}_{Y_l}\right)^{-2}}{\sum_{l=1}^{L} \hat{\theta}_{X_l}^{2} se\left(\hat{\theta}_{Y_l}\right)^{-2}} \tag{2.5}$$

with standard error

$$se\left(\hat{\beta}_{X_{IVW}}\right) = \sqrt{\frac{1}{\sum_{l=1}^{L} se\left(\hat{\beta}_{X_l}\right)^{-2}}} = \sqrt{\frac{1}{\sum_{l=1}^{L} \hat{\theta}_{X_l}^{2} se\left(\hat{\theta}_{Y_l}\right)^{-2}}} \tag{2.6}$$

(Burgess et al., 2013).

The method provides greater weight to instruments resulting in estimates with smaller standard errors. When there is a single genetic variant ($L = 1$), the estimate $\hat{\beta}_{X_{IVW}}$ is equivalent to Equation 2.2.

The IVW estimate $\hat{\beta}_{X_{IVW}}$ may equivalently be obtained by performing weighted least squares (WLS) regression without an intercept and with weights $Var\left(\hat{\theta}_{Y_l}\right)^{-1}$ (or $se\left(\hat{\theta}_{Y_l}\right)^{-2}$ in practice) using the following model:

$$\hat{\theta}_{Y_l} = \beta_X \hat{\theta}_{X_l} + \epsilon_l, \epsilon_l \sim \mathcal{N}\left(0, Var\left(\hat{\theta}_{Y_l}\right)\right) \tag{2.7}$$

(Burgess et al., 2016).

Adding an overdispersion parameter ϕ to the distribution of the error in Equation 2.7 leads to a model known as a variable-effects model:

$$\hat{\theta}_{Y_l} = \beta_X \hat{\theta}_{X_l} + \epsilon_l, \epsilon_l \sim \mathcal{N}\left(0, \phi^2 Var\left(\hat{\theta}_{Y_l}\right)\right).$$

The IVW estimator for the variable-effects model is equivalent to Equation 2.5:

$$\begin{aligned}\hat{\beta}_{X_{IVW}} = \frac{\sum_{l=1}^{L} \hat{\theta}_{Y_l}\hat{\theta}_{X_l}\phi^{-2} se\left(\hat{\theta}_{Y_l}\right)^{-2}}{\sum_{l=1}^{L} \hat{\theta}_{X_l}^2 \phi^{-2} se\left(\hat{\theta}_{Y_l}\right)^{-2}} &= \frac{\phi^{-2}\sum_{l=1}^{L} \hat{\theta}_{Y_l}\hat{\theta}_{X_l} se\left(\hat{\theta}_{Y_l}\right)^{-2}}{\phi^{-2}\sum_{l=1}^{L} \hat{\theta}_{X_l}^2 se\left(\hat{\theta}_{Y_l}\right)^{-2}} \\ &= \frac{\sum_{l=1}^{L} \hat{\beta}_{X_l} se\left(\hat{\beta}_{X_l}\right)^{-2}}{\sum_{l=1}^{L} se\left(\hat{\beta}_{X_l}\right)^{-2}};\end{aligned}$$

however, the standard error of the estimator is now given by:

$$\begin{aligned}se\left(\hat{\beta}_{X_{IVW}}\right) = \sqrt{\frac{1}{\sum_{l=1}^{L} \hat{\theta}_{X_l}^2 \phi^{-2} se\left(\hat{\theta}_{Y_l}\right)^{-2}}} &= \sqrt{\frac{\phi^2}{\sum_{l=1}^{L} \hat{\theta}_{X_l}^2 se\left(\hat{\theta}_{Y_l}\right)^{-2}}} \\ &= \frac{\phi}{\sqrt{\sum_{l=1}^{L} \hat{\theta}_{X_l}^2 se\left(\hat{\theta}_{Y_l}\right)^{-2}}};\end{aligned}$$

where in practice $\hat{\phi} = \max(1, RSE)$. The abbreviation RSE stands for Residual Standard Error for the fitted model using Equation 2.7. When fixing $\phi = 1$ the variable-effects model becomes equivalent to the model in Equation 2.7, otherwise known as the fixed-effects model. Although both provide the same point estimate for given data, the variable-effects model allows for larger standard errors (Thompson and Sharp, 1999; Bowden et al., 2017).

Summary-level data has also contributed to the rise of two-sample MR, whereby the summary associations between the SNP and the exposure, and that of the SNP and the outcome are obtained from two separate and partial or non-overlapping samples drawn from the same population. Two-sample MR greatly increased the scope and practicality of MR by allowing the combination of results from different studies resulting in increased sample sizes and statistical power. Additionally, two-sample MR requires the harmonization of data between studies. Harmonization is the process by which instrumental SNPs are properly paired between studies, ensuring the effect estimates correspond to the same alleles for the exposure and outcome (Hartwig et al., 2017). Following data harmonization, two-sample MR can be implemented via the IVW method or its WLS equivalent using the provided summary associations and their standard errors.

2.3 Multivariable Mendelian Randomisation

Cases arise where it is of interest to examine the effect of multiple exposures on a single outcome, or when it is not feasible to obtain SNPs only associated with the exposure of interest resulting in pleiotropy and violating MR Assumptions 2 or 3.

These cases can be handled by multivariable MR (MVMR). The term multivariable refers to the inclusion of at least two exposures with a single outcome. Furthermore, exposures can be correlated or exert direct effects on one another. MVMR requires a slight modification to IV Assumption 1 to account for multiple exposures:

Assumption 1: The IV is associated with one or more exposures

(Burgess and Thompson, 2015).

In an MVMR setting using two-sample summary statistics, Assumption 1 can be assessed with the conditional F-statistic for the exposure-SNP associations. The conditional F-statistic (Equation 2.8) works by assessing instrument strength for an exposure conditional on the other exposures included in the model. Without loss of generality, the conditional F-statistic for an exposure X_1 conditional on the rest of the $K-1$ exposures in a model with L SNPs (where SNPs are indexed by l) is given by:

$$F_{X_1|X_{-1}} = \frac{\sum_{l=1}^{L} \left(\frac{1}{\sigma^2_{X_{1,l}}}\right) \left(\hat{\theta}_{X_{1l}} - \hat{\delta}_1 \hat{\Theta}_{-X_{1,l}}\right)}{L-(K-1)} \sim \frac{\chi^2_{(L-(K-1))}}{L-(K-1)}, \tag{2.8}$$

where, $\hat{\delta}_1$ is a vector of size $K-1$ obtained by performing regression using the model (for $\quad l = 1, \ldots, L$):

$$\hat{\theta}_{X_{1l}} = \delta_1 \hat{\Theta}_{-X_{1,l}} + \epsilon_l$$

The term $\hat{\Theta}$ is the K by L matrix of exposure-SNP summary association estimates between each SNP and exposure:

$$\hat{\Theta} = \begin{bmatrix} \hat{\theta}_{X_{11}} & \ldots & \hat{\theta}_{X_{1L}} \\ \vdots & \ddots & \vdots \\ \hat{\theta}_{X_{K1}} & \ldots & \hat{\theta}_{X_{KL}} \end{bmatrix}.$$

The term $\hat{\Theta}_{-X_1}$ is the matrix $\hat{\Theta}$ without its first row, and $\hat{\Theta}_{-X_{1,l}}$ is l^{th} column of the matrix $\hat{\Theta}_{-X_1}$. Lastly, $\sigma^2_{X_{1,l}}$ is obtained from the equation:

$$\sigma^2_{X_{1,l}} = \begin{bmatrix} -1 & \hat{\delta}_2 & \ldots & \hat{\delta}_K \end{bmatrix} \Sigma_l \begin{bmatrix} -1 \\ \hat{\delta}_2 \\ \vdots \\ \hat{\delta}_K \end{bmatrix},$$

where Σ_l is the covariance matrix for the estimated exposure-SNP associations for a given SNP g_l:

$$\Sigma_l = \begin{bmatrix} \sigma_{11,l} & \dots & \sigma_{1K,l} \\ \vdots & \ddots & \vdots \\ \sigma_{K1,l} & \dots & \sigma_{KK,l} \end{bmatrix}.$$

Larger F values in Equation 2.8 denote strong instruments (Sanderson and Windmeijer, 2016; Sanderson et al., 2021). It is important to note the IV does not have to be associated with all of the exposures included within a model. However, the number of IVs used in the analysis must be at least as large as the number of exposures to ensure linear independence in exposure-SNP association or the predicted exposure values to disentangle the effect of the multiple exposures on the outcome (Sanderson et al., 2018).

Implementation of MVMR is similar to MR. For a model with K exposures and L SNPs, the summary-level data for the exposure-SNP associations $\hat{\theta}_{X_{kl}}$ are obtained by regressing the k^{th} exposure on the l^{th} SNP. Similarly, the outcome-SNP associations $\hat{\theta}_{Y_l}$ are obtained by regressing the outcome on each SNP. In a similar fashion to the univariable MR case, summary associations can be derived from a single sample or by using two non-overlapping samples: one for the exposures and one for the outcome. The IVW method is then performed using a WLS framework with the following model and weights $\phi^{-2}Var\left(\hat{\theta}_{Y_l}\right)^{-1}$ (or $\hat{\phi}^{-2}se\left(\hat{\theta}_{Y_l}\right)^{-2}$ in practice):

$$\hat{\theta}_{Y_l} = \beta_1\hat{\theta}_{X_{1l}} + \beta_2\hat{\theta}_{X_{2l}} + ... + \beta_K\hat{\theta}_{X_{Kl}} + \epsilon_l, \epsilon_l \sim \mathcal{N}\left(0, \phi^2 Var\left(\hat{\theta}_{Y_l}\right)\right) \tag{2.9}$$

(Burgess et al., 2015).

In an MVMR setting, the estimates represent the direct effects of the exposures on the outcome, whereas performing separate MR estimates for each exposure would provide a total effect of the exposure on the outcome including any correlations between exposures in affecting the provided causal estimates (Sanderson et al., 2018).

While MVMR is used to perform Mendelian randomisation in the presence of multiple exposures, the exposures need not be distinct. There may be certain scenarios where a single exposure can vary over time and its relationship with the outcome differs at different time points. Performing Mendelian randomisation with repeated measures of an exposure at different time points is called time-varying MR and is discussed in the following section.

2.4 Time-Varying Mendelian Randomisation

Estimates provided by MR consider effects from exposures acting over the course of a lifetime (Davey Smith and Ebrahim, 2003), however, the effects of an exposure on an outcome may not be identical at each time point (Kivimäki et al., 2007). Furthermore, if the SNP-exposure association varies over time then a single measurement of exposure and outcome may not provide a reliable estimate of the exposure's effect on the outcome (Labrecque and Swanson, 2018). Therefore, it is of importance to measure such exposures at multiple time points to help better understand their relationship with the outcome. When dealing with time-varying exposures, meaning an exposure measured at different time points, the interpretation of MR estimates is not the effect of the exposure at a specific point; it is the effect of changing the latent genetic liability, meaning the unmeasured effects of the SNPs associated with the exposure up until the time the outcome occurs. In other words, it is the effect on the outcome such that the genetic liability for the exposure differs enough for a unit increase in the exposure (Morris et al., 2022; Sanderson et al., 2022).

Recent studies suggested using structural mean models for MR to deal with time varying exposures, a semi-parametric method originally developed for use in randomised control trials with censoring (Burgess et al., 2017). Suppose there is a potential outcome Y measured at time point m for exposure X. Denote the potential outcome if the entire population had the exposure as Y^{X_m} and the potential outcome if none of the population had the exposure as $Y^{X'_m}$. Then the average point effect of the exposure can be calculated as the difference between the mean potential outcome of receiving the exposure compared to not receiving the exposure:

$$E\Big(Y^{X_m}\Big) - E\left(Y^{X'_m}\right).$$

In order to identify the causal estimator, the IV assumptions must hold for the measured time, that is the instrument cannot affect the exposure at other time periods or the outcome directly.

The above expression can be applied when there is a single measurement of the exposure. For multiple measurements, the average period or lifetime effect of the exposure for a period $m - t$ to m by calculating the difference between the mean outcome had the entire population received the exposure at each measured time point and the mean outcome had the entire population not received the exposure at each measured time point:

$$E\Big(Y^{X_{m-t,\dots,m-1,m}}\Big) - E\left(Y^{X'_{m-t,\dots,m-1,m}}\right).$$

For this case, identifying the causal estimator requires the measurement of all relevant exposure time points and at least as many SNPs as the number of time points t. Furthermore, there must be at least one SNP such that the SNP-exposure association differs between time points. Untangling exposure effects at different time points would not be possible without such a SNP (Shi et al., 2021, 2022).

The use of MVMR has also been considered. A study examining the effects of adiposity on disease risk using measures of adiposity levels early and later in life used MVMR to separate the effects of adiposity at both life stages (Richardson et al., 2020). Further work examining the use of MVMR with time-varying exposures and a single outcome showed the viability of the method for estimating causal effects over the course of an individual's life. The authors considered a model with an exposure measured at two different time points and a single outcome all affected by a confounder. The exposure at the earlier time point affected the exposure at the later time point. There were three collections of SNPs used as instrument variables: SNPs exclusively associated with the exposure at the earlier time point, SNPs exclusively associated with the exposure at the later time point, and SNPs associated with the exposure at both time points. Three simulations were conducted with SNP selection. The chosen selection criterion was a p-value smaller than 5×10^{-8} using an F-test for significance of regression for each SNP-exposure association.

The first simulation compared the estimation of the effects of exposures on the outcome when the genetic effects of the SNPs were identical for the exposure at each time point as opposed to differing genetic effects of the SNPs for each time point. Using the IVW MVMR method for estimation, they found having differing genetic effects for each exposure resulted in consistent estimators of the causal effect of the exposure on the outcome when the genetic liability of exposure changes such that the exposure changes by a unit. However, when genetic effects were identical at each time point the causal effect estimates were higher in absolute bias as a result of weak instruments evidenced by small conditional F-statistic values.

The second simulation considered a model identical to the first with a slight modification; the outcome affects the exposure at a later time point and a set of SNPs associated with the outcome were added which may have been selected as instruments for the effect estimates. The effect estimates were calculated using the IVW MVMR method following SNP selection with and without the use of Steiger filtering (a method used to determine SNPs explaining more variance in the exposure than the outcome), filtering out SNPs not meeting this criterion (Hemani et al., 2017). They found direct causal effect estimators were biased without the use of Steiger filtering even if the SNPs were strong instruments since the exposure at time point 2 was affected by both the exposure at the earlier time point and the outcome, becoming a collider (Paternoster et al., 2017). A collider is defined as a variable which is directly affected by two other variables (Pearce and Lawlor, 2016). Conditioning on the exposure at time point 2 resulted in collider bias for all estimators. Collider bias arises from the potential violation of IV assumptions due to the presence of a collider which may create an association between SNPs and the outcome other than through the exposure. However, the collider bias was avoided when Steiger filtering was used by eliminating the use of SNPs affecting the exposure at time point 2 via the outcome (Sanderson et al., 2018).

The final simulation considered a model with the exposure measured at three different time points, whereby earlier exposures affect later exposures, one outcome, an unobserved confounder affecting all measures of the exposure and the outcome, and three different sets of SNPs; two distinct sets affecting all three exposure time points, and a third only affecting the exposure at the third time point. They then estimated the direct effects of the exposure at time points 1 and 2 on the outcome ignoring the exposure at time point 3 and the set of SNPs exclusively affecting it for the two following scenarios: the SNPs were not correlated with the exposure at time point 3; and the distinct groups of SNPs had a correlation with the exposure at time point 3. They found some of the effects from the excluded exposure at time point 3 were included in the effect estimates from the exposures at time points 1 and 2 when there is a correlation. However, this was not the case when there was no correlation between the SNPs and the exposure at time point 3.

The authors then used IVW MVMR to assess the effect of BMI during childhood and adulthood on smoking behaviour (defined as cigarettes per day, smoking cessation, and smoking initiation) and on circulating levels of C-reactive protein (CRP). They found only adult BMI significantly affected circulating levels of CRP. They also concluded childhood BMI was not a predictor for smoking behaviour later in life. However, the authors note they did not further examine their results by exploring biases related to pleitropy, colliders or SNP selection (Sanderson et al., 2022).

Chapter 3

Extending the Time-Varying Model for Two Time Points of the Exposure and Outcome

The previous chapter discussed methods to analyse changing exposure-outcome relationships with the introduction of time-varying exposures measured at multiple time points. However, the implications of introducing additional measurements of outcomes are not yet clear. Returning to the introductory example of using BMI as an exposure and BP as an outcome, if these were both measured at two time points, would there be an additional benefit to this increased information? How should such data be handled? The inclusion of additional measurements of outcome was an area of future research outlined by the authors in the paper discussing the viability of MVMR in a time-varying exposure setting (Sanderson et al., 2022). As such, we will consider a model with two distinct time points including two measurements of an exposure and two measurements of an outcome.

First, a preliminary model is described with the following variables. The continuous exposure X is measured at time point 1 (X_1) and at time point 2 (X_2). The continuous outcome Y is also measured twice. The first measurement Y_1 is measured at any time following the measurement of X_1 but before X_2. The second measurement of Y, denoted Y_2, is measured at the same time as or after X_2. U is an unobserved confounder affecting the exposure and outcome at both time points.

The effects of the exposures and earlier outcomes on later outcomes are as follows:

- β_1: the effect of X_1 on Y_1,
- β_{12}: the effect of X_1 on Y_2,
- β_2: the effect of X_2 on Y_2,
- γ: the effect of Y_1 on Y_2,
- ω: the effect of X_1 on X_2.

We assume exposures or outcomes happening at time point 2 cannot affect those happening at time point 1; an event happening at future times cannot affect the present. Furthermore, it is assumed an outcome cannot affect an exposure, although this may be a reasonable occurrence in practice.

The collections of SNPs included in the model are G_1 (the IVs for X_1) and G_2 (the IVs for X_2). SNPs that affect more than one exposure/outcome are not considered. The model is represented graphically in Figure 3.4.

It is proposed to consider each time point in the model separately. At time point 1, the only relevant variables of the model are the SNPs G_1, the exposure X_1, and

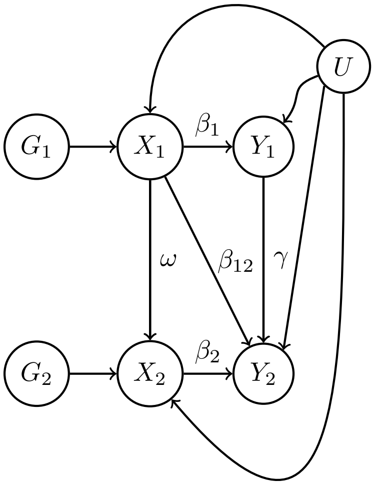


Figure 3.4: Preliminary model with a single time varying exposure X and outcome Y measured at two different time points 1 and 2. X_1 and X_2 as well as Y_1 do not share any SNPs.

the outcome Y_1. The estimation of β_1 can be completed via univariable MR using the following WLS model and weights $\hat{\phi}_1^{-2} se\left(\hat{\theta}_{Y_{1l}}\right)^{-2}$:

$$\hat{\theta}_{Y_{1l}} = \beta_1 \hat{\theta}_{X_{1l}} + \epsilon_{Y_{1l}}, \epsilon_{Y_{1l}} \sim \mathcal{N}\left(0, \phi_1^2 Var\left(\hat{\theta}_{Y_{1l}}\right)\right), \tag{3.10}$$

where $\hat{\phi}_1 = \max(1, RSE)$ and RSE is for the fitted model in Equation 3.10.

At time point 2, all variables in the model are relevant for estimation. The parameters of interest are β_{12}, β_2, and γ, which can be estimated using WLS for MVMR with the following model with weights $\hat{\phi}_2^{-2} se\left(\hat{\theta}_{Y_{2l}}\right)^{-2}$ and all SNPs in G_1 and G_2:

$$\hat{\theta}_{Y_{2l}} = \beta_{12} \hat{\theta}_{X_{1l}} + \beta_2 \hat{\theta}_{X_{2l}} + \gamma \hat{\theta}_{Y_{1l}} + \epsilon_{Y_{2l}}, \epsilon_{Y_{2l}} \sim \mathcal{N}\left(0, \phi_2^2 Var\left(\hat{\theta}_{Y_{2l}}\right)\right), \tag{3.11}$$

where $\hat{\phi}_2 = \max(1, RSE)$ and RSE is for the fitted model in Equation 3.11.

Note, that both Equation 3.10 and Equation 3.11 are based on the variable-effects model. The variable-effects model was chosen as it allows for more conservative standard error compared to the fixed-effects model. It will now be shown that the preliminary model in Figure 3.4 is inadequate to estimate the desired exposure-outcome effects.

Let us define the exposures (X_1, X_2) and outcomes (Y_1, Y_2) in terms of each of the L SNPs g_l with intercept θ_{0_k} where k is one of the exposure or outcome at time points 1 or 2:

$$X_1 = \sum_{l=1}^{L} \theta_{X_{1l}} g_l + \theta_{0_{X_1}} + \epsilon_{\theta_{X_1}}, \tag{3.12}$$

$$X_2 = \sum_{l=1}^{L} \theta_{X_{2l}} g_l + \theta_{0_{X_2}} + \epsilon_{\theta_{X_2}}, \tag{3.13}$$

$$Y_1 = \sum_{l=1}^{L} \theta_{Y_{1l}} g_l + \theta_{0_{Y_1}} + \epsilon_{\theta_{Y_1}}, \tag{3.14}$$

$$Y_2 = \sum_{l=1}^{L} \theta_{Y_{2l}} g_l + \theta_{0_{Y_2}} + \epsilon_{\theta_{Y_2}}. \tag{3.15}$$

Let us define the outcome Y_1 in terms of the exposure X_1

$$Y_1 = \beta_1 X_1 + \epsilon_{Y_1}. \tag{3.16}$$

Let us define the outcome Y_2 in terms of the exposures $(X_1,\ X_2)$ and earlier outcome (Y_1),

$$\begin{aligned} Y_2 &= \beta_{12} X_1 + \beta_2 X_2 + \gamma Y_1 + \epsilon_{Y_2} \\ &= \beta_{12} X_1 + \beta_2 X_2 + \gamma(\beta_1 X_1 + \epsilon_{Y_1}) + \epsilon_{Y_2}. \end{aligned} \tag{3.17}$$

Then by substituting Equations 3.12, 3.13 into Equation 3.17, we obtain:

$$\begin{aligned} Y_2 = {} & \beta_{12} \left(\sum_{l=1}^{L} \theta_{X_{1l}} g_l + \theta_{0_{X_1}} + \epsilon_{\theta_{X_1}} \right) + \beta_2 \left(\sum_{l=1}^{L} \theta_{X_{2l}} g_l + \theta_{0_{X_2}} + \epsilon_{\theta_{X_2}} \right) + \\ & \gamma \left(\beta_1 \left(\sum_{l=1}^{L} \theta_{X_{1l}} g_l + \theta_{0_{X_1}} + \epsilon_{\theta_{X_1}} \right) + \epsilon_{Y_1} \right) + \epsilon_{Y_2} \\ = {} & \beta_{12} \sum_{l=1}^{L} \theta_{X_{1l}} g_l + \beta_2 \sum_{l=1}^{L} \theta_{X_{2l}} g_l + \gamma\beta_1 \sum_{l=1}^{L} \theta_{X_{1l}} g_l + \beta_{12}\theta_{0_{X_1}} + \\ & \beta_2 \theta_{0_{X_2}} + \gamma\beta_1\theta_{0_{X_1}} + \beta_{12}\epsilon_{\theta_{X_1}} + \beta_2\epsilon_{\theta_{X_2}} + \gamma\beta_1\epsilon_{\theta_{X_1}} + \gamma\epsilon_{Y_1} + \epsilon_{Y_2}. \end{aligned} \tag{3.18}$$

Using Equations 3.15 and 3.18, the following equality is apparent:

$$\begin{aligned}\sum_{l=1}^{L}\theta_{Y_{2l}}g_l &= \beta_{12}\sum_{l=1}^{L}\theta_{X_{1l}}g_l + \beta_2\sum_{l=1}^{L}\theta_{X_{2l}}g_l + \gamma\beta_1\sum_{l=1}^{L}\theta_{X_{1l}}g_l \\ &= (\beta_{12}+\gamma\beta_1)\sum_{l=1}^{L}\theta_{X_{1l}}g_l + \beta_2\sum_{l=1}^{L}\theta_{X_{2l}}g_l,\end{aligned} \tag{3.19}$$

demonstrating an identifiability issue when trying to analyse the preliminary model at time point 2 as the collection of SNPs in G_1 represents the total effect of X_1 on Y_2, which includes the effect of X_1 via Y_1. Thus, there is a lack of information making it difficult to estimate β_{12} and γ. It is important to note that this issue does not arise when $\gamma = 0$, since indirect effects of X_1 on Y_2 via Y_1 are not present. This is evidenced by substituting $\gamma = 0$ into Equation 3.19, where the only parameters left to estimate are β_{12} and β_2. Nevertheless, the following model is proposed: with the inclusion of G_3, which is a set of SNPs directly affecting the time point 1 outcome measurement Y_1, the MVMR estimation at time point 2 will avoid the identifiability issue allowing for the estimation of β_{12} and γ (Figure 3.5).

Typically in MR (MVMR included) it is only required to use SNPs for the exposures. However, in a time-varying setting, the first outcome Y_1 will be considered an exposure for Y_2 at time point 2. Considering Y_1 as an exposure provides a theoretical justification for the necessity of the SNPs G_3 to estimate the separate direct causal effects of X_1 and Y_1 on Y_2.

By using a collection of SNPs G_3 exclusive to Y_1, we can keep the outcome Y_2 in terms of X_1, X_2, and Y_1 explicitly (referring to the first equality of Equation 3.17). Then by substituting Equations 3.12, 3.13, 3.14 into Equation 3.17, we obtain:

$$\begin{aligned}
Y_2 = \beta_{12}&\left(\sum_{l=1}^{L}\theta_{X_{1l}}g_l + \theta_{0_{X_1}} + \epsilon_{\theta_{X_1}}\right) + \beta_2\left(\sum_{l=1}^{L}\theta_{X_{2l}}g_l + \theta_{0_{X_2}} + \epsilon_{\theta_{X_2}}\right) + \\
&\gamma\left(\sum_{l=1}^{L}\theta_{Y_{1l}}g_l + \theta_{0_{Y_1}} + \epsilon_{\theta_{Y_1}}\right) + \epsilon_{Y_2} \\
= \beta_{12}&\sum_{l=1}^{L}\theta_{X_{1l}}g_l + \beta_2\sum_{l=1}^{L}\theta_{X_{2l}}g_l + \gamma\sum_{l=1}^{L}\theta_{Y_{1l}}g_l + \beta_{12}\theta_{0_{X_1}} + \\
&\beta_2\theta_{0_{X_2}} + \gamma\theta_{0_{Y_1}} + \beta_{12}\epsilon_{\theta_{X_1}} + \beta_2\epsilon_{\theta_{X_2}} + \gamma\epsilon_{\theta_{Y_1}} + \epsilon_{Y_2}.
\end{aligned} \tag{3.20}$$

Using Equations 3.15 and 3.20, we obtain the following equality:

$$\sum_{l=1}^{L}\theta_{Y_{2l}}g_l = \beta_{12}\sum_{l=1}^{L}\theta_{X_{1l}}g_l + \beta_2\sum_{l=1}^{L}\theta_{X_{2l}}g_l + \gamma\sum_{l=1}^{L}\theta_{Y_{1l}}g_l. \tag{3.21}$$

The equality in Equation 3.21 thus demonstrates estimates for the direct effects of X_1 on Y_2 (β_{12}) and Y_1 on Y_2 (γ) can be estimated when G_3 is included in the model.

Now that the proposed model and method of analysis have been introduced, the following chapter will detail a range of simulations to examine the properties of the model.

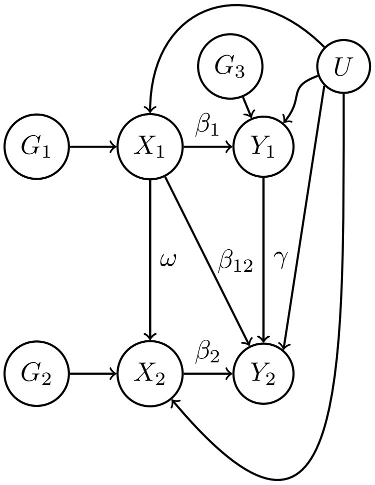


Figure 3.5: Proposed model with a single time varying exposure X and outcome Y measured at two different time points 1 and 2. Exposures X_1 and X_2 as well as outcome Y_1 do not share any SNPs with the addition of SNPs G_3.

Chapter 4

Simulations and Results

This chapter contains several simulations geared towards exploring the properties of the model introduced in Chapter 3 under varying conditions. The first simulation aims to demonstrate the identifiability issue with the initial model described in Chapter 3. The second simulation aims to compare parameter estimation of the model from Figure 3.5 under different sampling conditions. The rest of the simulations examine the performance of the estimators when modifying the strength of the SNP effects on the exposures, the number of SNPs corresponding to X_1, X_2, and Y_1, as well as different scenarios of shared (or overlapping) SNPs between X_2 and Y_1. All simulations follow the same general framework. A total of $L = 30$ SNPs for the first simulation and $L = 45$ SNPs for all subsequent simulations, unless otherwise specified (where g_{klj} is the l^{th} SNP within the set of SNPs G_k for the j^{th} individual), were generated independently from a $\text{Binomial}(2, p_{kl})$. Each MAF p_{kl} was generated independently from a $\text{Beta}(1, 8) \times 0.45 + 0.05$ in order to keep the range of the MAFs between 0.05 and 0.5, while allowing for smaller MAFs to be generated with higher probability. The 45 SNPs were equally divided into the three non-overlapping sets G_1, G_2, and G_3. The exposures for the j^{th} individual (where $j = 1, ..., n$) were constructed as follows:

$$X_{1j} = \sum_{l=1}^{15} \alpha_{1l} g_{1lj} + \kappa_1 U_j + \epsilon_{X_1 j}, \tag{4.22}$$

$$X_{2j} = \sum_{l=1}^{15} \alpha_{2l} g_{2lj} + \omega X_{1j} + \kappa_2 U_j + \epsilon_{X_2 j}, \tag{4.23}$$

where each $\alpha_{kl} \overset{\text{iid}}{\sim} N(0, 1)$ without the interval $(-0.3, 0.3)$. The interval was omitted from the standard normal distribution to avoid weak instruments. The other random variables for the j^{th} individual were distributed as follows:

- $\epsilon_{X_1 j} \overset{\text{iid}}{\sim} N(0, 1)$
- $\epsilon_{X_2 j} \overset{\text{iid}}{\sim} N(0, 1)$
- $U_j \overset{\text{iid}}{\sim} N(0, 1)$,

and the values of κ_1, κ_2, and ω were arbitrarily chosen as 0.4, 0.3, and 0.1 respectively.

The outcomes for the j^{th} individual were then constructed from the exposures as follows:

$$Y_{1j} = \sum_{l=1}^{15} \alpha_{3l} g_{3lj} + \beta_1 X_{1j} + \kappa_3 U_j + \epsilon_{Y_1 j}, \tag{4.24}$$

$$Y_{2j} = \beta_{12} X_{1j} + \beta_2 X_{2j} + \gamma Y_{1j} + \kappa_4 U_j + \epsilon_{Y_2 j}, \tag{4.25}$$

where each $\epsilon_{Y_1 j}, \epsilon_{Y_2 j}, \overset{\text{iid}}{\sim} N(0, 1)$, while κ_3 and κ_4 were arbitrarily set as 0.3 and 0.5 respectively. Lastly, the values of the parameters of interest were $\beta_1 = 0.2$, $\beta_{12} = 0.4$, $\beta_2 = 0.3$, and $\gamma = 0.2$.

The exposure and outcome data were generated for $n = \{20000, 40000\}$ individuals (to be specified in each simulation). The simulation was repeated $N = 1000$ times for each case. Simulations and computations were completed using R 4.3.2 with the aid of the `MendelianRandomization` package (R Core Team, 2023; Yavorska and Staley, 2023). Implementation of the methods described in Chapter 3 pertaining to the model in Figure 3.5 is a two-step process given the availability of summary statistics and their standard errors for the desired SNPs in G_1, G_2, and G_3. First, univariable MR is used to estimate the effect on Y_1 using the SNPs in G_1 and is performed by formatting the summary statistics data with the `mr_input` function, then passing it through the function `mr_ivw` which will provide an estimate for β_1. Second, MVMR is used to estimate the effects on Y_2 using all SNPs by formatting the summary statistics data with `mr_mvinput` and passing it through the function `mr_mvivw` to obtain an estimate of β_{12}, β_2, and γ (Yavorska and Staley, 2023).

4.1 Simulation 1 - Causal effect estimates without SNPs G_3

The purpose of this simulation is to demonstrate the identifiability issue in the estimation of causal effects β_{12} and γ in the time-varying model with exposure and outcome measurements at two time points when the set of SNPs G_3 is empty.

A total of $L = 30$ SNPs (where g_{klj} is the l^{th} SNP within the set of SNPs G_k for the j^{th} individual) were generated independently from a Binomial$(2, p_{kl})$. There were only 30 SNPs since G_3 is empty. The 30 SNPs were equally divided into the two non-overlapping sets G_1 and G_2.

Since G_3 is empty, the outcome Y_1 for the j^{th} individual was then constructed from the exposure at time point 1 as follows:

$$Y_{1j} = \beta_1 X_{1j} + \kappa_3 U_j + \epsilon_{Y_1 j},$$

where each $\epsilon_{Y_1 j} \overset{\text{iid}}{\sim} N(0,1)$, while the effect of the confounder U was still arbitrarily set as $u_3 = 0.3$ and $\beta_1 = 0.2$.

The exposure and outcome data were generated for $n = 20000$ individuals. The data were subsequently split into two equal non-overlapping samples whereby one sample was used to obtain summary statistics for the exposures (θ_{X_1} and θ_{X_2}) and the other was used to obtain the summary statistics of the outcomes (θ_{Y_1} and θ_{Y_2}). Summary statistics were obtained by regressing each of the exposures and outcomes individually on each of the SNPs. The summary statistics were then used to compute the estimates of β_1, β_{12}, β_2, and γ by performing weighted least squares with the models in Equations 3.10 and 3.11, where only the SNPs pertaining to G_1 are used in Equation 3.10, and all SNPs were used in Equation 3.11.

The standard errors for the estimators were then calculated using the following formula:

$$se(\hat{\beta}_c) = \frac{\hat{\phi}}{\sqrt{\sum_l \hat{\theta}^2_{X_d l} se(\hat{\theta}_{Y_f l})^{-2}}},$$

where $\hat{\beta}_c$ was taken to mean the estimate of either β_1, β_{12}, β_2, and γ, with the appropriate summary associations for exposure $\hat{\theta}_{X_d l}$ (where X_d is one of X_1, X_2, or Y_1) and outcome $\hat{\theta}_{Y_f l}$ (where Y_f is one of Y_1 or Y_2) for the l^{th} SNP. When $\hat{\beta}_c$ is taken as $\hat{\beta}_1$, l ranges from 1 to the number of SNPs in G_1. For all other options of $\hat{\beta}_c$, l ranges from 1 to L, the total number of SNPs.

Table 4.1 provides the mean estimate of parameter values across all simulation runs,

Table 4.1: Mean estimates, standard deviations, and mean standard errors with estimator bias and coverage when G_3 is empty

	true value	mean estimate	bias	rel. bias (%)	std. dev.	mean std. error	coverage probability ($N = 1000$)
β_1	0.2000	0.1993	-0.0007	-0.3654	0.0073	0.0074	0.9630
β_{12}	0.4000	0.2955	-0.1045	-26.1234	0.0472	0.0469	0.4140
β_2	0.3000	0.2991	-0.0009	-0.3133	0.0083	0.0073	0.9260
γ	0.2000	0.7151	0.5151	257.5310	0.2349	0.2320	0.4180

the bias of the estimators, their relative bias, the mean standard error across all simulation runs as well as the coverage probability for a 95% confidence interval calculated using a t distribution with $df = 14$ for β_1 and $df = 27$ for β_{12}, β_2, and γ. Relative bias (or rel. bias) for each parameter was calculated as

$$\text{bias}(\%) = \frac{\hat{\beta}_c - \beta_c}{\beta_c} \times 100\%.$$

Boxplots illustrating parameter estimates for each simulation run are provided in Figure 4.6. Although the estimators for both β_1 and β_2 appear unbiased and achieve good coverage probability, the estimators for β_{12} and γ are quite poor in terms of bias and coverage (Table 4.1). Furthermore, the variability of the estimates of β_{12} and especially γ is greater than that of β_1 and β_2, as evidenced by the interquartile range (IQR) for each boxplot in Figure 4.6. The IQRs for each parameter estimates are as follows: 0.3166 for γ; 0.0645 for β_{12}; 0.0101 for β_1; and 0.0111 for β_2. We suspect the poor performance of the estimators is caused by the identifiability issue outlined in Chapter 3; hence the proposed addition of SNPs G_3 affecting Y_1.

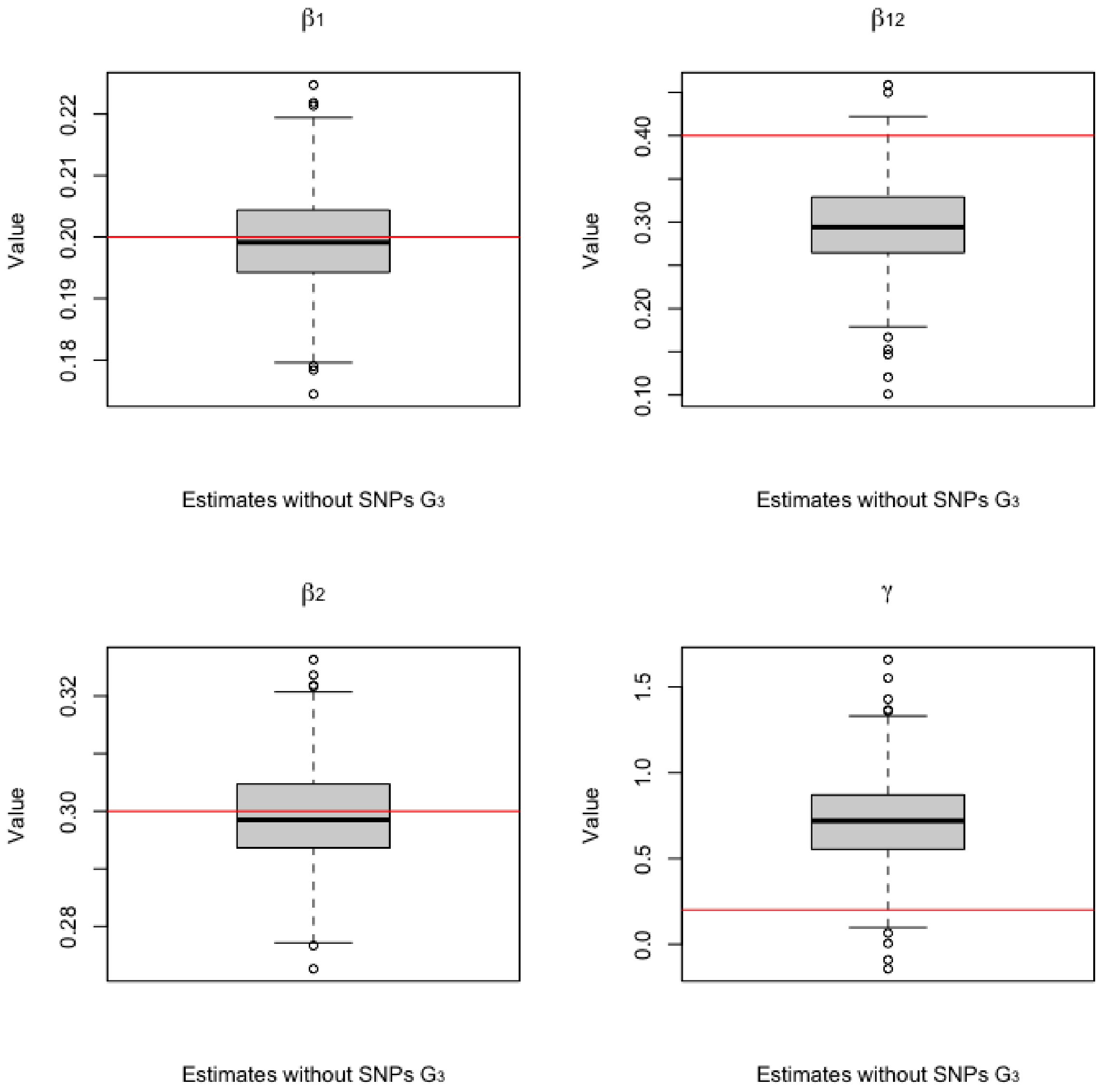


Figure 4.6: Boxplots of parameter estimates for an empty set of SNPs G_3. True parameter values are indicated by the red lines in each plot.

4.2 Simulation 2 - Causal effect estimates with SNPs G_3 for different sampling schemes

The purpose of this simulation was to demonstrate the estimation of causal effects with the inclusion of SNPs G_3 and under three sampling schemes. The three different sampling schemes are the one-, two-, and four-sample schemes. The one-sample scheme involves obtaining all summary statistics for X_1, X_2, Y_1, and Y_2 from a single sample. This sampling scheme is representative of all summary statistics coming from a single study. The two-sample scheme involves obtaining the summary statistics for the exposures X_1 and X_2 from one sample and the summary statistics for the outcomes Y_1 and Y_2 from a separate non-overlapping sample. This scheme is comparable to obtaining summary statistics for the exposures and outcomes from two separate studies. Lastly, the four-sample scheme involves obtaining the summary statistics for each of X_1, X_2, Y_1, and Y_2 from separate samples. This scenario is similar to obtaining the summary statistics from four separate studies. Together, the three schemes capture one middle ground and two extreme scenarios for sampling.

The simulation followed the general simulation framework outlined at the beginning of the chapter. The exposure and outcome data were generated for $n = 40000$ individuals. The individuals were subsequently split into four equal non-overlapping samples containing 10000 individuals called Samples A, B, C, and D. Summary statistics were then obtained using the three sampling schemes outlined above. For the one-sample scheme, only Sample A was used. In the two-sample scheme summary statistics for the exposure were found using Sample A and those for the outcome were found using Sample B. Lastly all Samples A–D were used for X_1, X_2, Y_1, and Y_2 summary statistics respectively, in the four-sample scheme. Parameter estimates and standard errors were obtained in the same manner as Simulation 1.

Table 4.2: Mean estimates, standard deviations, and mean standard errors with estimator bias and coverage by number of samples

	true value	mean estimate	bias	rel. bias (%)	std. dev.	mean std. error	coverage probability ($N = 1000$)	sample(s)
β_1	0.2000	0.2010	0.0010	0.5042	0.0193	0.0209	0.9710	1
β_{12}	0.4000	0.3999	-0.0001	-0.0244	0.0027	0.0109	1.0000	
β_2	0.3000	0.3000	0.0000	0.0070	0.0022	0.0092	1.0000	
γ	0.2000	0.2001	0.0001	0.0444	0.0021	0.0088	1.0000	
β_1	0.2000	0.1998	-0.0002	-0.1148	0.0201	0.0213	0.9660	2
β_{12}	0.4000	0.3955	-0.0045	-1.1350	0.0151	0.0131	0.8970	
β_2	0.3000	0.2972	-0.0028	-0.9390	0.0126	0.0110	0.9020	
γ	0.2000	0.2010	0.0010	0.5224	0.0109	0.0106	0.9410	
β_1	0.2000	0.1979	-0.0021	-1.0508	0.0202	0.0212	0.9680	4
β_{12}	0.4000	0.3964	-0.0036	-0.9062	0.0167	0.0143	0.9040	
β_2	0.3000	0.2961	-0.0039	-1.3034	0.0132	0.0120	0.9110	
γ	0.2000	0.1985	-0.0015	-0.7655	0.0126	0.0115	0.9230	

Firstly, comparing the estimators before and after the inclusion of SNPs G_3 (Table 4.2 and Figure 4.7), we notice improved performance as characterized by smaller biases and higher coverage for the estimators of β_{12}, γ using t distribution 95% confidence intervals with 42 degrees of freedom. The intervals for β_2 used the same degrees of freedom whereas the confidence intervals for β_1 were calculated using 14 degrees of freedom.

Comparing the different sampling schemes we notice the estimator for β_1 provides similar estimates across all three sampling schemes with over-coverage. The story is different for β_{12}, β_2, and γ all of which increase in variability as the number of samples increases. A possible explanation for this pattern is the fact that there is less error within a sample which is using the same individuals for all summary statistics as opposed to obtaining the summary statistics from different samples. We also note the standard

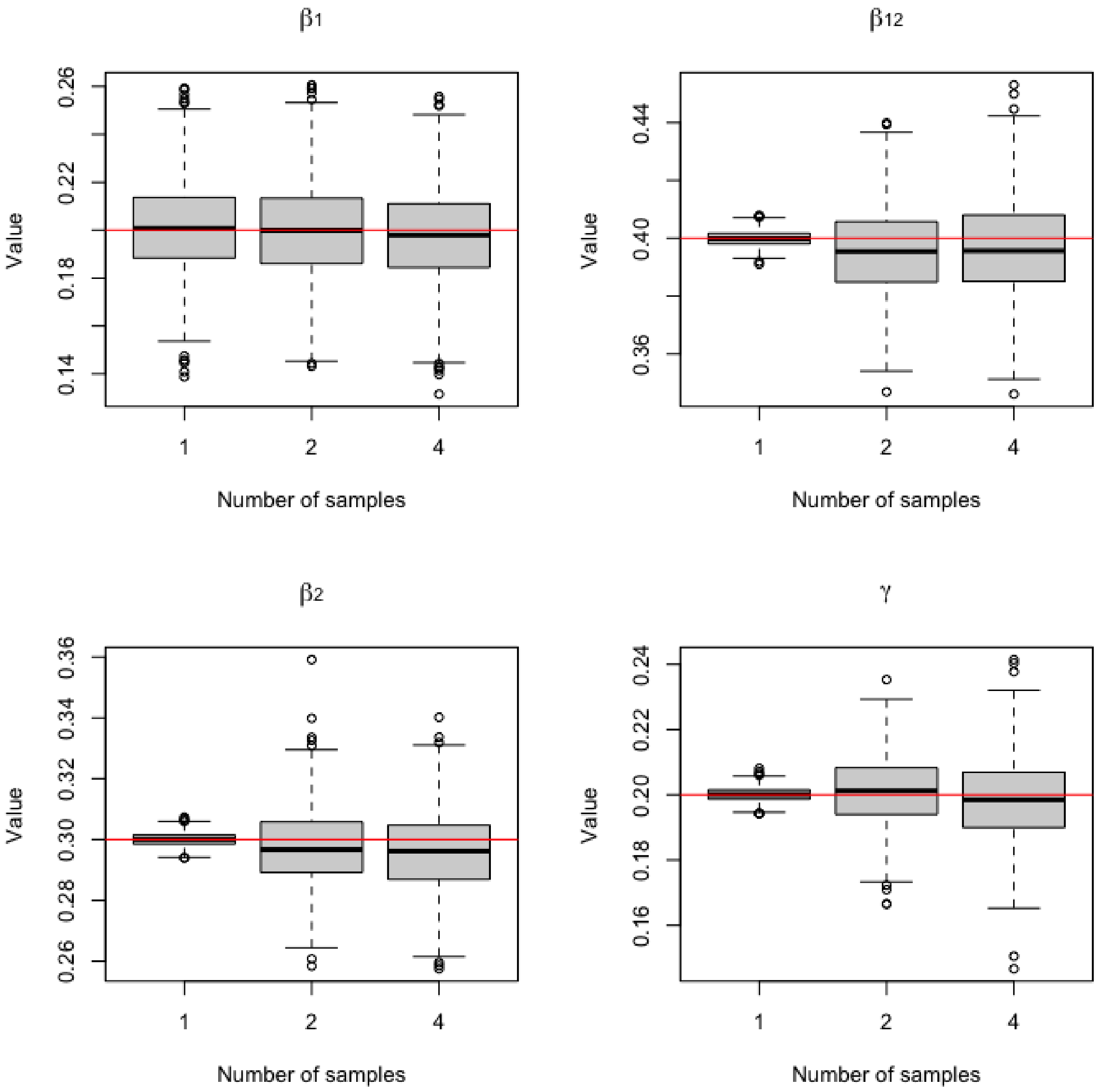


Figure 4.7: Boxplots of parameter estimates for different numbers of samples. True parameter values are indicated by the red lines in each plot.

errors of the estimators are larger on average than their standard deviations within the one-sample scheme.

4.3 Simulation 3 - Causal effect estimates with SNPs G_3 for different magnitudes of SNP effects

The purpose of this simulation was to examine the causal effect estimates under different SNP effects α. The α_{kl}'s from Equations 4.22, 4.23, 4.24, and 4.25 were generated independently from $N(0,1)$ with different and subsequently larger rejection intervals, leading to larger SNP effect magnitudes achieved via an accept-reject algorithm only accepting values outside the rejection intervals which were $|\alpha_{kl}| > \{0, 0.3, 0.6, 0.9, 1.2, 1.5\}$. Following the same procedures as the general simulation framework, the rest of the data was generated and estimates were obtained with the exception of only using $n = 20000$ individuals with the two-sample scheme employed to obtain summary statistics.

It appears as though increasing the minimum generated absolute instrument strength did not significantly influence the performance of the estimators (Figure 4.8). Aside from an initial decrease in variability for the estimates of β_2, the other estimates, namely β_1, β_{12}, and γ, generally remained unchanged regardless of instrument strength.

4.4 Simulation 4 - Causal effect estimates with SNPs G_3 for different numbers of SNPs

Mendelian randomisation is often conducted with large numbers of SNPs as such this simulation sought to examine the causal effect estimates under different numbers of

Table 4.3: Mean estimates, standard deviations, and mean standard errors with estimator bias and coverage for different minimum absolute instrument strengths

	true	mean	bias	rel.	std.	mean std.	coverage	$\lvert\alpha_{kl}\rvert >$
	value	estimate		bias (%)	dev.	error	probability	
							$(N = 1000)$	
β_1	0.2000	0.1992	-0.0008	-0.3918	0.0121	0.0125	0.9620	0
β_{12}	0.4000	0.3982	-0.0018	-0.4428	0.0098	0.0084	0.9100	
β_2	0.3000	0.2965	-0.0035	-1.1654	0.0162	0.0151	0.9250	
γ	0.2000	0.2013	0.0013	0.6721	0.0088	0.0090	0.9470	
β_1	0.2000	0.2001	0.0001	0.0408	0.0120	0.0129	0.9750	0.3
β_{12}	0.4000	0.3981	-0.0019	-0.4808	0.0103	0.0087	0.9060	
β_2	0.3000	0.2978	-0.0022	-0.7221	0.0097	0.0089	0.9210	
γ	0.2000	0.2011	0.0011	0.5632	0.0085	0.0085	0.9500	
β_1	0.2000	0.1995	-0.0005	-0.2668	0.0129	0.0129	0.9670	0.6
β_{12}	0.4000	0.3979	-0.0021	-0.5349	0.0101	0.0087	0.9010	
β_2	0.3000	0.2983	-0.0017	-0.5770	0.0088	0.0076	0.9010	
γ	0.2000	0.2009	0.0009	0.4462	0.0083	0.0082	0.9500	
β_1	0.2000	0.1994	-0.0006	-0.3066	0.0137	0.0136	0.9640	0.9
β_{12}	0.4000	0.3979	-0.0021	-0.5215	0.0103	0.0087	0.8950	
β_2	0.3000	0.2986	-0.0014	-0.4816	0.0087	0.0076	0.9170	
γ	0.2000	0.2009	0.0009	0.4707	0.0076	0.0075	0.9430	
β_1	0.2000	0.1993	-0.0007	-0.3527	0.0119	0.0126	0.9760	1.2
β_{12}	0.4000	0.3972	-0.0028	-0.6954	0.0108	0.0085	0.8580	
β_2	0.3000	0.2986	-0.0014	-0.4613	0.0081	0.0069	0.8980	
γ	0.2000	0.2008	0.0008	0.4197	0.0080	0.0078	0.9460	
β_1	0.2000	0.1996	-0.0004	-0.2230	0.0111	0.0117	0.9710	1.5
β_{12}	0.4000	0.3983	-0.0017	-0.4227	0.0094	0.0081	0.9070	
β_2	0.3000	0.2990	-0.0010	-0.3442	0.0082	0.0072	0.9130	
γ	0.2000	0.2009	0.0009	0.4290	0.0082	0.0080	0.9470	

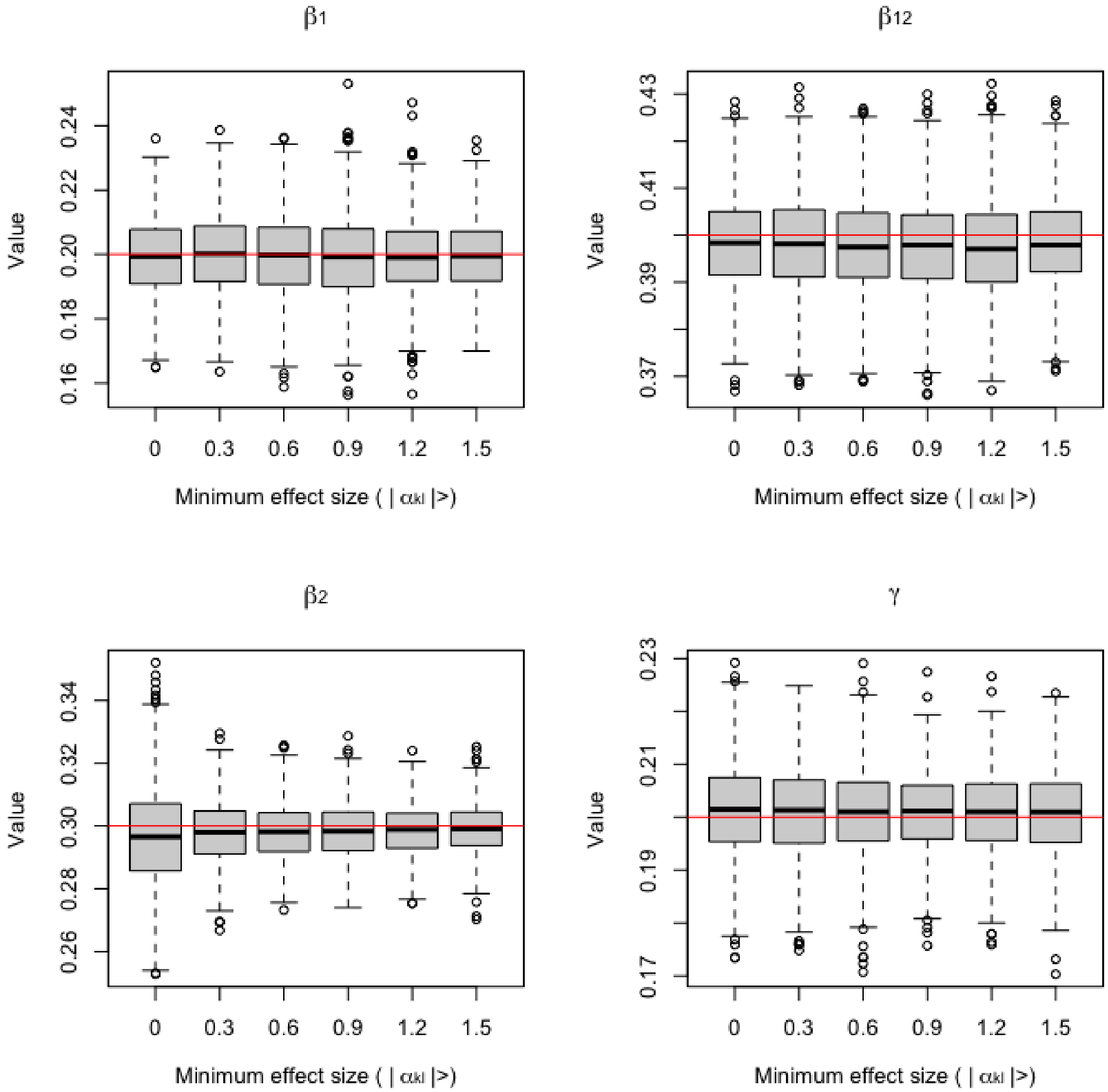


Figure 4.8: Boxplots of parameter estimates for different minimum absolute SNP effects. True parameter values are indicated by the red lines in each plot.

SNPs. The simulation followed the general simulation framework under the two-sample scheme introduced in Simulation 2. Exposures and outcomes were generated for $n = 20000$ individuals and the number of SNPs in G_1, G_2, and G_3 was either 5, 25, 50, 75, or 100, under the restriction each set of SNPs had an equal amount. The simulation was repeated $N = 1000$ for each SNP count.

Table 4.4: Mean estimates, standard deviations, and mean standard errors with estimator bias and coverage for different numbers of SNPs per set G_k

	true value	mean estimate	bias	rel. bias (%)	std. dev.	mean std. error	coverage probability ($N = 1000$)	SNPs per G_k
β_1	0.2000	0.1990	-0.0010	-0.5197	0.0163	0.0177	0.9950	5
β_{12}	0.4000	0.3992	-0.0008	-0.2082	0.0116	0.0101	0.9300	
β_2	0.3000	0.2992	-0.0008	-0.2765	0.0113	0.0104	0.9520	
γ	0.2000	0.2003	0.0003	0.1531	0.0084	0.0079	0.9510	
β_1	0.2000	0.1990	-0.0010	-0.4998	0.0128	0.0135	0.9620	25
β_{12}	0.4000	0.3964	-0.0036	-0.8936	0.0101	0.0084	0.8650	
β_2	0.3000	0.2980	-0.0020	-0.6649	0.0097	0.0086	0.9150	
γ	0.2000	0.2015	0.0015	0.7507	0.0072	0.0073	0.9520	
β_1	0.2000	0.1986	-0.0014	-0.6763	0.0115	0.0112	0.9490	50
β_{12}	0.4000	0.3930	-0.0070	-1.7407	0.0098	0.0078	0.7870	
β_2	0.3000	0.2957	-0.0043	-1.4338	0.0090	0.0079	0.8800	
γ	0.2000	0.2027	0.0027	1.3559	0.0079	0.0077	0.9350	
β_1	0.2000	0.1985	-0.0015	-0.7420	0.0125	0.0124	0.9470	75
β_{12}	0.4000	0.3901	-0.0099	-2.4855	0.0097	0.0079	0.7100	
β_2	0.3000	0.2934	-0.0066	-2.2048	0.0082	0.0072	0.8100	
γ	0.2000	0.2033	0.0033	1.6554	0.0071	0.0068	0.9140	
β_1	0.2000	0.1980	-0.0020	-0.9799	0.0130	0.0123	0.9340	100
β_{12}	0.4000	0.3872	-0.0128	-3.2113	0.0093	0.0077	0.6000	
β_2	0.3000	0.2910	-0.0090	-2.9997	0.0085	0.0072	0.7130	
γ	0.2000	0.2039	0.0039	1.9285	0.0066	0.0066	0.9050	

The estimator for β_1 yielded similar estimates throughout the increase in the number of SNPs, with decent coverage probability. On the other hand, the estimators of effects on Y_2 (β_{12}, β_2, and γ) began performing worse as the number of SNPs increased as characterized by increases in bias and reductions in coverage (Table 4.4). The trend is also evident in Figure 4.9. The exact cause for this drop in performance is unclear. However, it may impact practical applications of the model because it is often preferred to use a large number of SNPs in the estimation when conducting univariable Mendelian randomisation.

4.5 Simulation 5 - Causal effect estimates with SNPs G_3 and G_{23}

The purpose of this simulation was to examine the causal effect estimates under different overlapping SNP conditions. It is important to note the only overlapping SNPs considered are those in the set G_{23}. These SNPs affect the exposure X_2 and the outcome Y_1. Any SNPs affecting Y_1 and X_1 would be invalid SNPs for the estimation of β_1 and as such were not considered.

Two different scenarios of overlap were considered. The first scenario considered changing the proportion of shared SNPs between X_2 and Y_1 by increasing the number of SNPs in G_{23} whilst keeping the number of SNPs affecting X_2 and Y_1 constant. The considered proportions were 0, 0.2, 0.4, 0.6, 0.8, and 1. The exact combinations of SNPs can be seen in Table 4.5. The second scenario once again dealt with changing shared SNPs of X_2 and Y_1. However, the number of SNPs in G_2, G_3, and G_{23} were modified while keeping the total number of SNPs fixed at 45 (see Table 4.6 for exact combinations where the number of SNPs in G_i is denoted by L_i).

In the first scenario, altering the proportion of SNPs shared between X_2 and Y_1 while keeping the same number of SNPs affecting both did not appear to affect the estimation

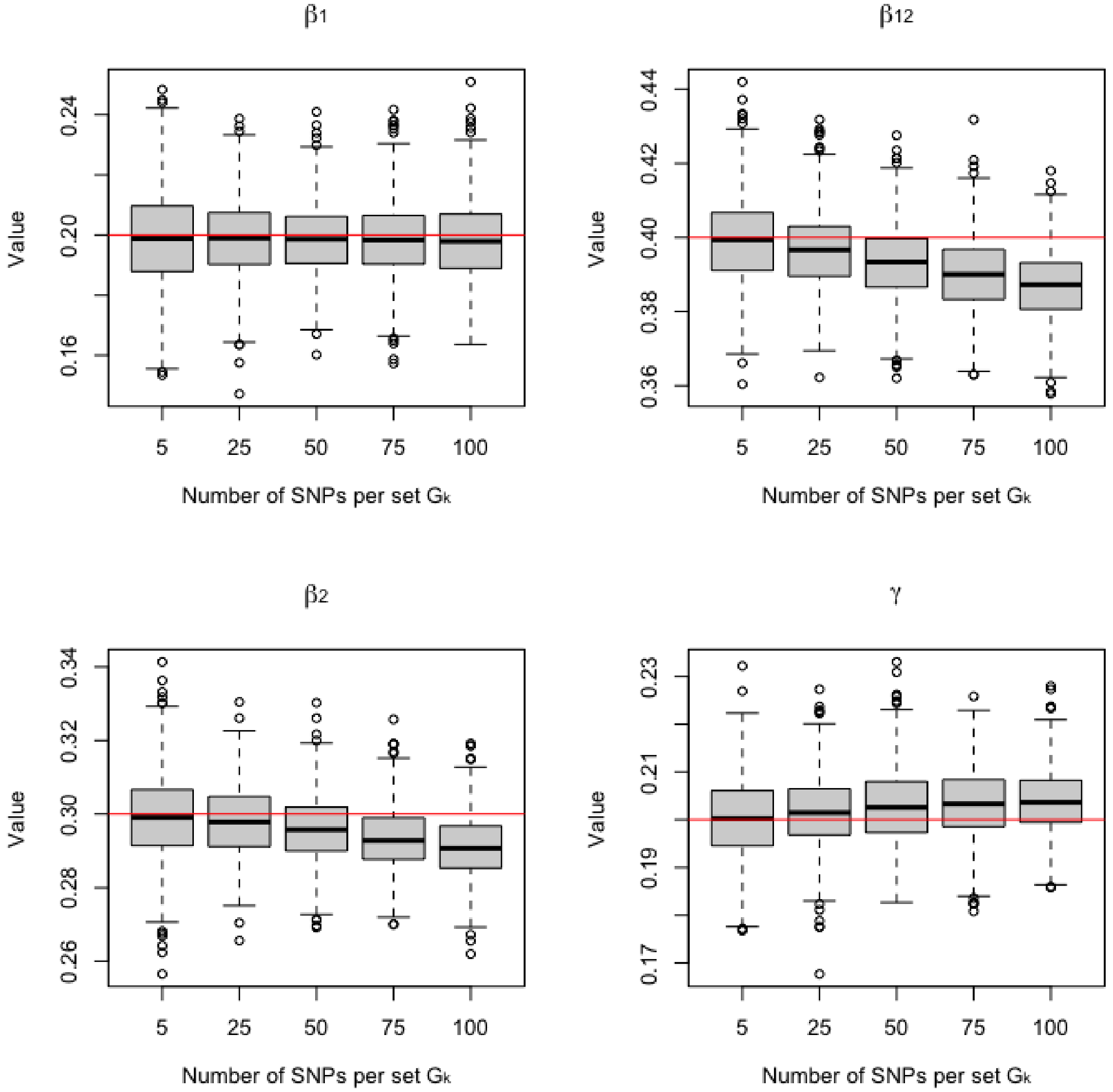


Figure 4.9: Boxplots of parameter estimates for different numbers of SNPs in each G_k. True parameter values are indicated by the red lines in each plot.

Table 4.5: SNP combinations for Scenario 1 - modifying the proportion of overlap

	L_1	L_2	L_3	L_{23}
1	15	15	15	0
2	15	12	12	3
3	15	9	9	6
4	15	6	6	9
5	15	3	3	12
6	15	0	0	15

Table 4.6: SNP combinations for Scenario 2 - modifying the proportion of overlap with fixed total SNPs

	L_1	L_2	L_3	L_{23}
1	15	15	15	0
2	15	12	12	6
3	15	9	9	12
4	15	6	6	18
5	15	3	3	24
6	15	0	0	30

for either time point (as seen in Figure 4.10 and Table 4.7). All estimators yielded good coverage probabilities and small biases with slight over-coverage for β_1 possibly stemming from the estimator's higher mean standard error across the simulation runs (see Table 4.7).

Within the second scenario, altering the proportion of SNPs shared between X_2 and Y_1 while keeping the total number of SNPs in the model constant provided very similar results to the first scenario. Boxplots between numbered of shared SNPs between X_2 and Y_1 in Figure 4.11 appear almost unchanged. Small biases were present and there was slight under-coverage for β_{12} and β_2 (see Table 4.8). These two scenarios provide evidence the model is capable of handling the presence of overlapping SNPs within G_{23}.

Table 4.7: Mean estimates, standard deviations, and mean standard errors with estimator bias and coverage for different proportions of overlap G_{23}

	true value	mean estimate	bias	rel. bias (%)	std. dev.	mean std. error	coverage probability ($N = 1000$)	overlap prop.
β_1	0.2000	0.1992	-0.0008	-0.4110	0.0146	0.0151	0.9690	0
β_{12}	0.4000	0.3974	-0.0026	-0.6433	0.0107	0.0092	0.9030	
β_2	0.3000	0.2985	-0.0015	-0.4876	0.0086	0.0078	0.9250	
γ	0.2000	0.2007	0.0007	0.3557	0.0078	0.0075	0.9370	
β_1	0.2000	0.1993	-0.0007	-0.3560	0.0146	0.0153	0.9690	0.2
β_{12}	0.4000	0.3978	-0.0022	-0.5484	0.0107	0.0092	0.9020	
β_2	0.3000	0.2979	-0.0021	-0.6934	0.0094	0.0083	0.9100	
γ	0.2000	0.2002	0.0002	0.1068	0.0072	0.0073	0.9550	
β_1	0.2000	0.1996	-0.0004	-0.2191	0.0149	0.0153	0.9750	0.4
β_{12}	0.4000	0.3980	-0.0020	-0.4937	0.0101	0.0089	0.9200	
β_2	0.3000	0.2978	-0.0022	-0.7178	0.0104	0.0096	0.9350	
γ	0.2000	0.2001	0.0001	0.0647	0.0072	0.0072	0.9580	
β_1	0.2000	0.1996	-0.0004	-0.2107	0.0155	0.0163	0.9680	0.6
β_{12}	0.4000	0.3984	-0.0016	-0.3971	0.0103	0.0090	0.9080	
β_2	0.3000	0.2978	-0.0022	-0.7498	0.0094	0.0088	0.9230	
γ	0.2000	0.2003	0.0003	0.1512	0.0064	0.0066	0.9540	
β_1	0.2000	0.1996	-0.0004	-0.1875	0.0143	0.0156	0.9760	0.8
β_{12}	0.4000	0.3984	-0.0016	-0.4094	0.0103	0.0088	0.9030	
β_2	0.3000	0.2987	-0.0013	-0.4386	0.0104	0.0095	0.9260	
γ	0.2000	0.2002	0.0002	0.1081	0.0068	0.0068	0.9540	
β_1	0.2000	0.2001	0.0001	0.0665	0.0121	0.0127	0.9800	1
β_{12}	0.4000	0.3985	-0.0015	-0.3720	0.0105	0.0089	0.9100	
β_2	0.3000	0.2985	-0.0015	-0.4942	0.0098	0.0088	0.9310	
γ	0.2000	0.2012	0.0012	0.6165	0.0090	0.0092	0.9610	

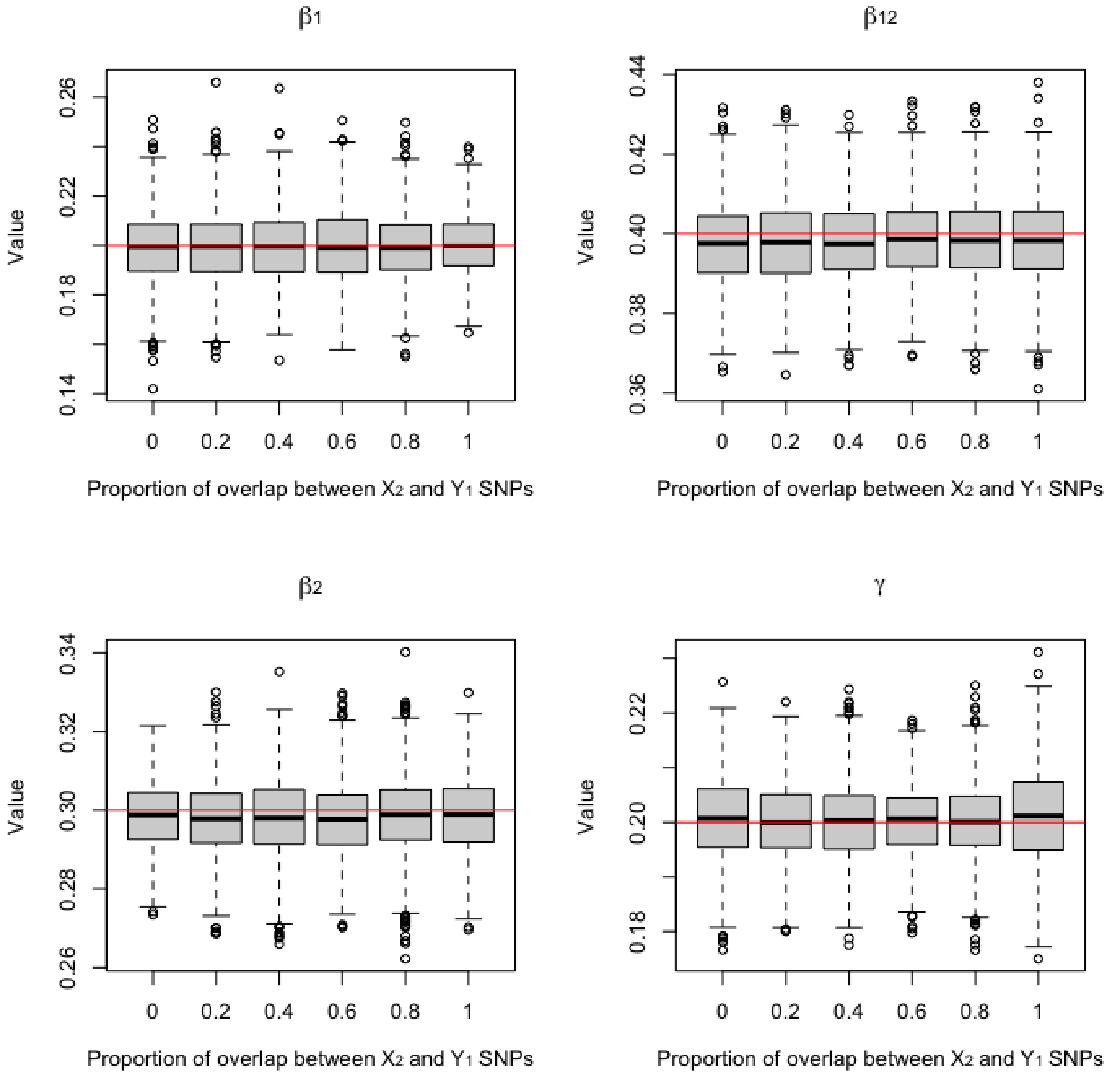


Figure 4.10: Boxplots of parameter estimates for different proportions of overlap. True parameter values are indicated by the red lines in each plot.

Table 4.8: Mean estimates, standard deviations, and mean standard errors with estimator bias and coverage for differing overlap with fixed SNP total

	true value	mean estimate	bias	rel. bias (%)	std. dev.	mean std. error	coverage probability ($N = 1000$)	SNPs in G_{23}
β_1	0.2000	0.1992	-0.0008	-0.4110	0.0146	0.0151	0.9690	0
β_{12}	0.4000	0.3974	-0.0026	-0.6433	0.0107	0.0092	0.9030	
β_2	0.3000	0.2985	-0.0015	-0.4876	0.0086	0.0078	0.9250	
γ	0.2000	0.2007	0.0007	0.3557	0.0078	0.0075	0.9370	
β_1	0.2000	0.1993	-0.0007	-0.3507	0.0151	0.0158	0.9730	6
β_{12}	0.4000	0.3977	-0.0023	-0.5782	0.0107	0.0092	0.8930	
β_2	0.3000	0.2978	-0.0022	-0.7456	0.0093	0.0082	0.9110	
γ	0.2000	0.2001	0.0001	0.0571	0.0071	0.0071	0.9530	
β_1	0.2000	0.1997	-0.0003	-0.1563	0.0162	0.0170	0.9760	12
β_{12}	0.4000	0.3976	-0.0024	-0.5970	0.0103	0.0092	0.9160	
β_2	0.3000	0.2980	-0.0020	-0.6593	0.0092	0.0083	0.9220	
γ	0.2000	0.2003	0.0003	0.1646	0.0062	0.0064	0.9570	
β_1	0.2000	0.1997	-0.0003	-0.1302	0.0166	0.0177	0.9750	18
β_{12}	0.4000	0.3978	-0.0022	-0.5542	0.0108	0.0095	0.9080	
β_2	0.3000	0.2978	-0.0022	-0.7260	0.0084	0.0073	0.9120	
γ	0.2000	0.2004	0.0004	0.1879	0.0062	0.0063	0.9530	
β_1	0.2000	0.1998	-0.0002	-0.1138	0.0161	0.0175	0.9760	24
β_{12}	0.4000	0.3976	-0.0024	-0.6002	0.0110	0.0096	0.9070	
β_2	0.3000	0.2986	-0.0014	-0.4751	0.0083	0.0071	0.9040	
γ	0.2000	0.2003	0.0003	0.1604	0.0064	0.0064	0.9520	
β_1	0.2000	0.2001	0.0001	0.0264	0.0150	0.0159	0.9690	30
β_{12}	0.4000	0.3975	-0.0025	-0.6249	0.0113	0.0100	0.9170	
β_2	0.3000	0.2982	-0.0018	-0.6024	0.0080	0.0068	0.8980	
γ	0.2000	0.2009	0.0009	0.4675	0.0077	0.0075	0.9550	

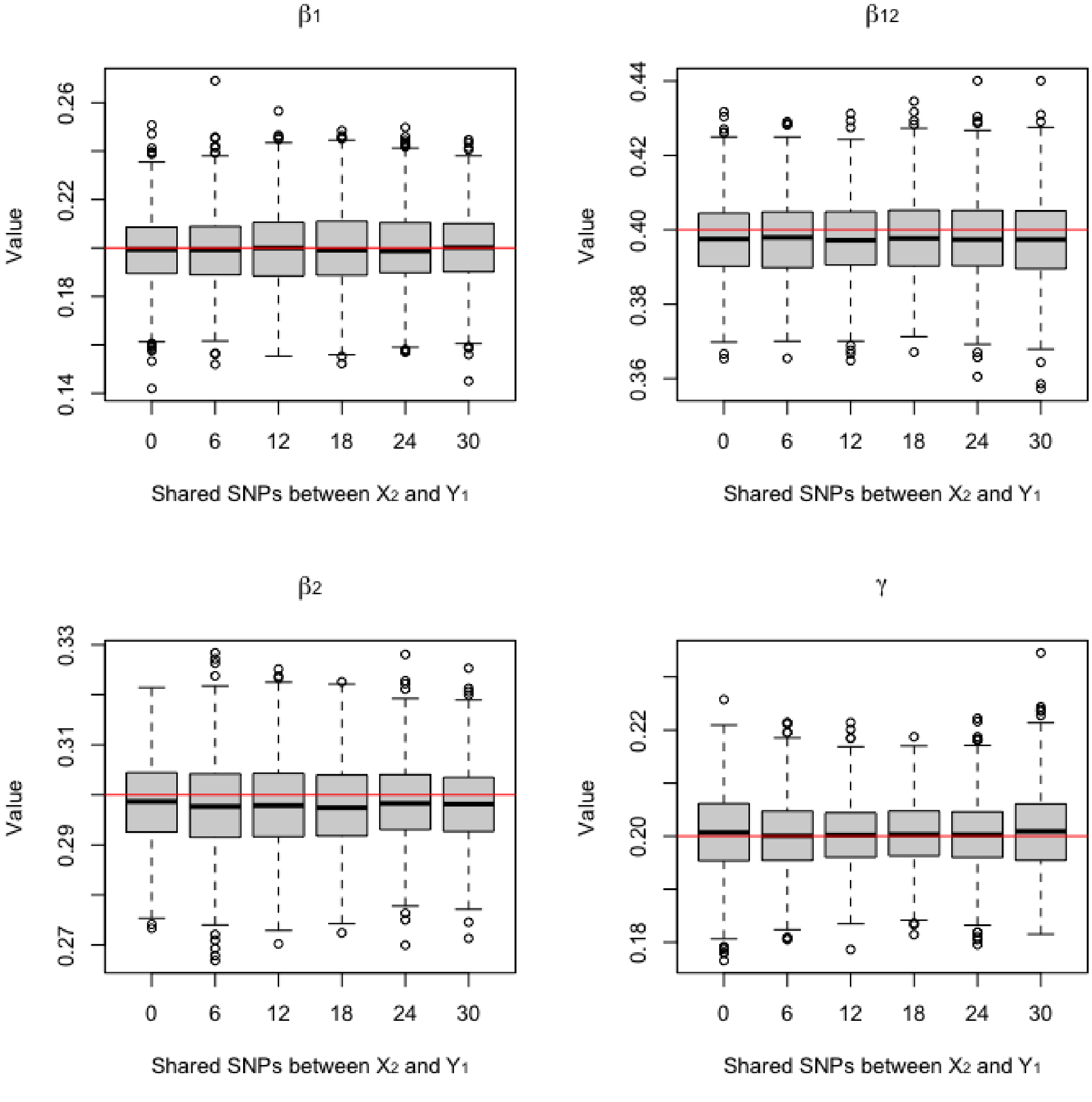


Figure 4.11: Boxplots of parameter estimates for different levels of overlap with fixed SNP total. True parameter values are indicated by the red lines in each plot.

Chapter 5

Conclusion

In this thesis, we explored extending an MVMR framework for a time-varying exposure and outcome measured at two distinct time points. While MVMR is multivariable with respect to the number of exposures (or the number of exposure measurements in this particular scenario), the addition of a second outcome measurement resulted in an identifiability issue which was due to the lack of unique instrumental variables pertaining to the first outcome measurement (Y_1). MVMR was unable to properly estimate and attribute the direct effects of the exposure at time point 1 (X_1) and the outcome at time point 1 (Y_1) on the outcome at time point 2 (Y_2). However, treating Y_1 as an exposure for time point 2 in a time-varying model with an exposure and outcome measured at two time points and having exclusive SNPs as instruments for the exposure at time point 1, X_1, can be used to estimate the variable effects at both time points. Treating the outcome at time point 1 as an exposure requires using valid SNPs for Y_1.

Simulations have shown the estimators perform well under different forms of sampling for the summary-level data. Bias and coverage were similar between the two-sample and four-sample schemes whereas overcoverage was present under the one-sample scheme.

The two-sample and four-sample schemes for summary data may be the more practical scenarios for application since these are comparable to the exposures being measured from one study, the outcomes from another, or all measures coming from four different studies. Under the one-sample scheme, if there was access to the individual-level data then effect estimates can be found via TSLS, a method not explored in this thesis. Furthermore, the one-sample scheme provided the lowest estimator bias and coverage with much smaller standard errors for the estimations at time point 2.

Simulations exploring increasing numbers of SNPs per set G_k demonstrated increases in absolute biases and decreases in coverage for the estimators at time point 2. Although the estimator for time point 1 followed the same pattern, the rate of increase in bias and decrease in coverage was much smaller compared to those of time point 2. This is a major drawback of the model as MR studies may often make use of a large number of SNPs. The reason for this drop in performance for parameter estimators at time point 2, as the number of SNPs increases, is unclear.

Simulations exploring different minimum SNP effect strengths and different types of SNP combinations for the SNP sets for different conditions of overlap demonstrated these parameters did not greatly affect estimators and the model can handle overlap and different instrument strengths.

Although the model introduces an MVMR framework for dealing with a time-varying exposure and outcome measured at two time points did perform well under certain conditions, the framework is not without its drawbacks. Most notably, there must be a unique set of instruments for X_1 not associated with X_2 or Y_1. This method thus relies on the association between the exposure and SNPs changing over time allowing for unique SNPs. Furthermore, if one were to attempt to expand this framework to further include more time points for the exposure and outcome then more sets of unique SNPs may be required for the estimation of the effects at the additional time points. Additionally, this model assumed the outcome cannot affect the exposure. However, in biology, it is possible for feedback loops such that the level of the outcome can affect the level of the exposure. Estimation using this framework under this type of scenario remains to be explored.

Lastly, the estimation is clearly worse as the number of SNPs used increases. Further work should explore why this is the case, and to try and explain the validity of the framework for larger numbers of SNPs which are more representative of real data MR studies.

Bibliography

Angrist, J. D., Graddy, K., and Imbens, G. W. (2000). The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish. *The Review of Economic Studies*, 67(3):499–527.

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, 44(2):512–525.

Bowden, J., Del Greco M, F., Minelli, C., Davey Smith, G., Sheehan, N., and Thompson, J. (2017). A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statistics in Medicine*, 36(11):1783–1802.

Bowden, J., Del Greco M, F., Minelli, C., Davey Smith, G., Sheehan, N. A., and Thompson, J. R. (2016). Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. *International Journal of Epidemiology*, 45(6):1961–1974.

Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*, 37(7):658–665.

Burgess, S., Davey Smith, G., Davies, N., Dudbridge, F., Gill, D., Glymour, M., Hartwig, F., Kutalik, Z., Holmes, M., Minelli, C., Morrison, J., Pan, W., Relton, C., and Theodoratou, E. (2023). Guidelines for performing Mendelian randomization investigations: update for summer 2023.

Burgess, S., Dudbridge, F., and Thompson, S. G. (2015). Re: "Multivariable Mendelian Randomization: The Use of Pleiotropic Genetic Variants to Estimate Causal Effects". *American Journal of Epidemiology*, 181(4):290–291.

Burgess, S., Dudbridge, F., and Thompson, S. G. (2016). Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Statistics in Medicine*, 35(11):1880–1906.

Burgess, S., Small, D. S., and Thompson, S. G. (2017). A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research*, 26(5):2333–2355. PMID: 26282889.

Burgess, S. and Thompson, S. G. (2011). Bias in causal estimates from Mendelian randomization studies with weak instruments. *Statistics in Medicine*, 30(11):1312–1323.

Burgess, S. and Thompson, S. G. (2015). Multivariable Mendelian Randomization: The Use of Pleiotropic Genetic Variants to Estimate Causal Effects. *American Journal of Epidemiology*, 181(4):251–260.

Davey Smith, G. and Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 32(1):1–22.

Davey Smith, G. and Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, 23(R1):R89–R98.

Didelez, V., Meng, S., and Sheehan, N. A. (2010). Assumptions of IV Methods for Observational Epidemiology. *Statistical Science*, 25(1):22 – 40.

Didelez, V. and Sheehan, N. (2005). Mendelian randomisation and instrumental variables: What can and what can't be done. Technical report, University of Leicester, Department of Health Sciences.

Didelez, V. and Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16(4):309–330. PMID: 17715159.

Glymour, M. M., Tchetgen Tchetgen, E. J., and Robins, J. M. (2012). Credible Mendelian Randomization Studies: Approaches for Evaluating the Instrumental Variable Assumptions. *American Journal of Epidemiology*, 175(4):332–339.

Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science*, 28(706):49–50.

Hartwig, F. P., Davies, N. M., Hemani, G., and Davey Smith, G. (2017). Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *International Journal of Epidemiology*, 45(6):1717–1726.

Hemani, G., Tilling, K., and Davey Smith, G. (2017). Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLOS Genetics*, 13(11):1–22.

Hernán, M. A. and Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, 17(4):360–372.

Hingorani, A. and Humphries, S. (2005). Nature's randomised trials. *The Lancet (British edition)*, 366(9501):1906–1908.

Katan, M. B. (2004). Apolipoprotein E isoforms, serum cholesterol, and cancer. *International Journal of Epidemiology*, 33(1):9–9.

Kivimäki, M., Lawlor, D. A., Smith, G. D., Eklund, C., Hurme, M., Lehtimäki, T., Viikari, J. S. A., and Raitakari, O. T. (2007). Variants in the CRP Gene as a Measure of Lifelong Differences in Average C-Reactive Protein Levels: The Cardiovascular Risk in Young Finns Study, 1980–2001. *American Journal of Epidemiology*, 166(7):760–764.

Labrecque, J. A. and Swanson, S. A. (2018). Interpretation and Potential Biases of Mendelian Randomization Estimates With Time-Varying Exposures. *American Journal of Epidemiology*, 188(1):231–238.

Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8):1133–1163.

Morris, T. T., Heron, J., Sanderson, E. C. M., Davey Smith, G., Didelez, V., and Tilling, K. (2022). Interpretation of Mendelian randomization using a single measure of an exposure that varies over time. *International Journal of Epidemiology*, 51(6):1899–1909.

Nitsch, D., Molokhia, M., Smeeth, L., DeStavola, B. L., Whittaker, J. C., and Leon, D. A. (2006). Limits to Causal Inference based on Mendelian Randomization: A

Comparison with Randomized Controlled Trials. *American Journal of Epidemiology*, 163(5):397–403.

Paternoster, L., Tilling, K., and Davey Smith, G. (2017). Genetic epidemiology and Mendelian randomization for informing disease therapeutics: Conceptual and methodological challenges. *PLOS Genetics*, 13(10):1–9.

Pearce, N. and Lawlor, D. A. (2016). Causal inference-so much more than statistics. *International journal of epidemiology*, 45(6):1895–1903.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rees, J. M. B., Wood, A. M., Dudbridge, F., and Burgess, S. (2019). Robust methods in Mendelian randomization via penalization of heterogeneous causal estimates. *PLOS ONE*, 14(9):e0222362.

Reiersöl, O. (1945). *Confluence analysis by means of instrumental sets of variables*. PhD thesis, Stockholm College.

Richardson, T. G., Sanderson, E., Elsworth, B., Tilling, K., and Smith, G. D. (2020). Use of genetic variation to separate the effects of early and later life adiposity on disease risk: Mendelian randomisation study. *BMJ*, 369.

Sanderson, E., Davey Smith, G., Windmeijer, F., and Bowden, J. (2018). An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. *International Journal of Epidemiology*, 48(3):713–727.

Sanderson, E., Richardson, T. G., Morris, T. T., Tilling, K., and Davey Smith, G. (2022).

Estimation of causal effects of a time-varying exposure at multiple time points through multivariable Mendelian randomization. *PLOS Genetics*, 18(7):1–19.

Sanderson, E., Spiller, W., and Bowden, J. (2021). Testing and correcting for weak and pleiotropic instruments in two-sample multivariable Mendelian randomization. *Statistics in medicine*, 40(25):5434–5452.

Sanderson, E. and Windmeijer, F. (2016). A weak instrument F-test in linear IV models with multiple endogenous variables. *Journal of Econometrics*, 190(2):212–221. Endogeneity Problems in Econometrics.

Shi, J., Swanson, S., Kraft, P., Rosner, B., Vivo, I., and Hernán, M. (2021). Instrumental variable estimation for a time-varying treatment and a time-to-event outcome via structural nested cumulative failure time models. *BMC Medical Research Methodology*, 21(258):1–12.

Shi, J., Swanson, S., Kraft, P., Rosner, B., Vivo, I., and Hernán, M. (2022). Mendelian randomization with repeated measures of a time-varying exposure: An application of structural mean models. *Epidemiology*, 33:84–94.

Smith, G. D. and Ebrahim, S. (2005). What can Mendelian randomisation tell us about modifiable behavioural and environmental exposures? *BMJ*, 330(7499):1076–1079.

Staiger, D. and Stock, J. H. (1994). Instrumental variables regression with weak instruments. Working Paper 151, National Bureau of Economic Research.

Thomas, D. C. and Conti, D. V. (2004). Commentary: The concept of 'Mendelian Randomization'. *International Journal of Epidemiology*, 33(1):21–25.

Thomas, D. C., Lawlor, D. A., and Thompson, J. R. (2007). Re: Estimation of bias in nongenetic observational studies using "Mendelian triangulation" by Bautista et .al. *Annals of Epidemiology*, 17(7):511–513.

Thompson, S. G. and Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine*, 18(20):2693–2708.

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics*, 11(3):284–300.

Wehby, G. L., Ohsfeldt, R. L., and Murray, J. C. (2008). 'Mendelian randomization' equals instrumental variable analysis with genetic instruments. *Statistics in Medicine*, 27(15):2745–2749.

Weinberg, W. (1908). Uber den nachweis der vererbung beim Menschen. *Jh. Ver. vaterl. Naturk. Wurttemb.*, 64:369–382.

Wright, A. F. (2005). *Genetic Variation: Polymorphisms and Mutations.* John Wiley & Sons, Ltd.

Wright, P. G. (1928). *The tariff on animal and vegetable oils.* The Macmillan company, New York.

Yavorska, O. and Staley, J. (2023). *MendelianRandomization: Mendelian Randomization Package.* R package version 0.9.0.