Mendelian Randomisation with Many Dependent Instruments

Thesis submitted for the degree of

Doctor of Philosophy

At the University of Leicester.

by

Chin Yang Shapland BSc MSc

Department of Health Sciences

University of Leicester.

1st September 2016.

Mendelian Randomisation with Many Dependent Instruments

Thesis submitted for the degree of

Doctor of Philosophy

At the University of Leicester.

by

Chin Yang Shapland BSc MSc

Department of Health Sciences

University of Leicester.

1st September 2016

Abstract

Mendelian Randomisation with Many Dependent Instruments

CY. Shapland

Mendelian randomisation is well known for its weak instrument bias, which is caused by the weak genetic association with the exposure. Usually, the most significant SNP within a gene region is selected to represent the association with the exposure of interest. However, if the causal variant was not genotyped then this proxy with weaker association will be a worse instrument. Moreover, a GWAS significant proxy may not show significance in another population. In the human genome, there are many SNPs in linkage equilibrium (LD) within a gene. The correlation between the SNPs and causal variant may increase power to detect the underlying association with the exposure. My thesis will investigate whether many SNPs in LD within a gene region can provide a stronger instrument than a single proxy for the causal variant(s). The thesis will first establish the gains from having multiple SNPs in LD as instruments. Simulation of realistic LD patterns will be used to assess both classical and Bayesian approaches to the estimation of the causal effect with many dependent SNPs. A Bayesian approach to Mendelian randomisation is preferable to the classical estimation with many dependent SNPs.

Acknowledgements

I would first like to express my gratitude to my two main supervisors, Professor John Thompson and Professor Nuala Sheehan, both have been not only excellent supervisors and mentors, but also friends.

Thanks also go to other members of the Department of Health Sciences for their input and support, including fellow PhD students and researchers. In particular, I'd like to thank Dr. Vicki Jackson for taking the time to read through a draft of this thesis.

I would also like to thank Dr. Michael Crowther, for being the ballast throughout the calm and rough seas to finishing this thesis.

Finally, I would like to thank my parents, my stepfather, and the rest of my friends, for their support over the past few years. Especially to my stepfather, whom without the encouragement I would not even be on the academic path.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This chapter gives a brief overview of Mendelian randomisation, discusses possible advantages of extracting full information from a gene, lists the objectives of the thesis and finally presents an overview of the thesis. See Chapter 2 for a fuller description of Mendelian randomisation and an explanation of genetic terms.

1.1 Mendelian Randomisation

Conventional epidemiology has made important contributions to knowledge about disease aetiology. However, due to factors such as selection bias, reverse causation and unmeasured confounding, most of the correlation between exposure and disease outcome cannot be interpreted as causal, but only as association. Therefore some findings from observational epidemiological studies cannot be replicated in randomised control trials (RCT). Unmeasured confounding is one of the most likely explanations for false associations in observational studies; confounding has been defined, as an important influence on the outcome that differs systematically between the comparison groups [68]. For example, Vitamin C has been found to have a protective effect on cardiovascular disease in observational studies but the protective effect was not found in an RCT, most likely because the relationship is confounded by lifestyle [181]. People who take Vitamin C are more likely to be health conscious than the general population, less likely to smoke and more likely to exercise regularly, and consequently this group has lower risk of cardiovascular disease.

Instrumental variable analysis has been developed in the field of econometrics as a method to infer causality between an endogenous variable (X) and an outcome (Y),

free from the effect of confounding (U). The word "endogenous" is an econometric term for an explanatory variable that is correlated with the confounding variable. A variable is a valid instrument (Z) only if it is (1) associated with the endogenous variable, (2) independent of the confounding variable and (3) independent of the outcome of interest given the endogenous and confounding variables. Figure 1.1 shows the diagram of instrumental variable analysis [117]. Provided these core assumptions are satisfied and the relationship is linear and without interactions, the X-Y association is the ratio of the Z-Y and Z-X associations. If the three core assumptions do not hold then the estimation of the X-Y association is subject to bias [135].

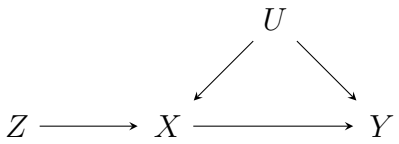


Figure 1.1: Instrumental variable analysis; Z is the instrument associated with the endogenous variable (X), Y is outcome of interest and U is the confounding effect between X and Y.

Mendelian randomisation is instrumental variable analysis with genetic variants as instruments [286]. The concept of Mendelian randomisation was first introduced by Katan [162] and popularised by Davey Smith and Ebrahim [79]. Mendelian randomisation has been described as the "natural" RCT [81]; the exposure can sometimes be a phenotype or a behaviour influenced by a genetic variant. By Mendel's law of independent assortment [130], the inheritance of a genetic variant is random, similar to patients being randomised to different treatments in an RCT. In cases where exposures are considered as unethical or impractical to randomise, for example smoking and alcohol consumption, Mendelian randomisation can be the alternative [80]. The genetic instruments are usually in the form of single nucleotide polymorphisms (SNPs) [84]. SNPs are positions where a single nucleotide in the deoxyribonucleic acid (DNA) is altered. DNA is a molecule that carries genetic information. A SNP could change the protein product, or affect gene expression, which alters the exposure and in turn the disease risk. Genome-wide association studies (GWAS) measure thousands of SNPs in a large number of individuals to identify their association with complex traits [216]. Due to the development of GWAS, there is a substantial number of SNP associations with phenotypes and disease outcomes that are currently available, which means there are plenty of SNPs to use as genetic

instruments.

As well as Mendel's law of independent assortment and the large number of SNPs available, there are other reasons that make SNPs desirable candidates as instruments; common SNPs have been well-characterised in their biological function and a SNP can measure lifelong exposure, since exposures are generally affected by genetic variants from birth to adulthood [80]. Another key point is that a SNP is unlikely to be related to confounders, such as socioeconomic status and behaviour, typical of observational epidemiological studies [79]. Even though the three core assumptions can be considered biologically plausible, there is no procedure that fully tests these assumptions. There are three major genetic conditions that can invalidate the core assumptions; direct pleiotropy, pleiotropy via linkage disequilibrium and population stratification. Pleiotropy is defined as an association with more than one phenotype. If a SNP is associated with another exposure that also affects the outcome of interest, then this SNP is not a valid instrument, as it violates assumption 3 (or 2 if the additional exposure affects the confounder). Pleiotropy via linkage disequilibrium can also violate assumptions 3 and 2, which will be discussed later. Population stratification occurs when different populations show differences in disease rates and allele frequencies, which confounds the association between genetic instrument and outcome of interest.

For the estimation of the causal effect, the most popular estimators, thanks to their simplicity, are the Wald (ratio) estimator for a single instrument and two-stage least squares (2SLS) for multiple instruments [26]. The Wald estimator is the ratio of the coefficients from the Z-Y and Z-X associations. The 2SLS algorithm predicts X from its genetic association in the first-stage, then derives the causal effect from the regression of Y on the predicted X in the second stage. For multiple instruments, there are also other estimators available, such as Limited Information Maximum Likelihood (LIML), Generalised Method of Moments (GMM), semi-parametric and Bayesian methods.

There is limited methodological research in Bayesian approaches to Mendelian randomisation. Consequently, the usage of Bayesian methods is sparse in comparison to classical instrumental variable methods (see Chapter 8). The fundamental idea of the Bayesian approach is that it derives statistical inferences through combining prior information with the information from one's data. If informative priors are available then the precision of the causal effect estimate can be greater than that obtained from classical approaches [153]. Due to the ever-increasing number of

GWAS, the use of informative priors is now possible. For example, when there are multiple instruments and one of them has been previously published as being genome-wide significant for the exposure of interest, then more weight can be put on this specific SNP within the Bayesian analysis.

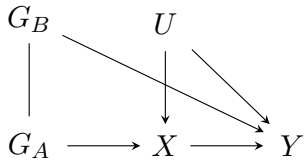
Even if a genetic variant does satisfy the core assumptions, bias can come from weak instruments. A genetic variant typically explains a small proportion of the variation in an exposure and with finite samples, the combination will introduce bias to the causal effect estimate, commonly known as weak instruments bias [253]. An instrument is considered weak if its F-statistic from the association with exposure is less than 10;

$$F = \frac{R^2(n-1-k)}{(1-R^2)k} \tag{1.1}$$

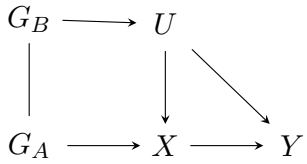
where R^2 is the proportion of the variability in exposure that is explained by the instrument(s), n is the size of the sample and k is the number of instruments. Weak instrument bias can be reduced by including information from multiple genotype-exposure and genotype-outcome association studies and combining them to estimate the causal effect via meta-analysis [266]. The lack of power to detect an effect between X and Y is also caused by the weak association between genotype and exposure. Statistical power can be increased by having a larger sample size [79], however this can be expensive. The use of multiple instruments can increase power and precision in the estimation of causal effect [219]. For the increase of precision with multiple instruments, each additional instrument must explain additional variation in the exposure of interest [219]; Equation 1.1 shows that as k increases, the F-statistic will decline if R^2 remains the same. However, having multiple SNPs with an F-statistic > 10 is not always possible and the threshold does not always guarantee less than 10% bias in the estimation of causal effect. Alternatively, the LIML estimator and allele score can be implemented; LIML is median unbiased with many weak instruments [61, 85]. Allele scores can increase the F-statistic of a single instrument by combining multiple SNPs [227]. The scoring for each SNP is based on the presence of the risk allele and weighted by the risk allele's association with the exposure of interest [49].

1.2 Extracting full information from a gene

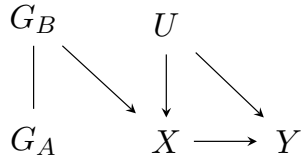
The instruments used in the multiple SNPs Mendelian randomisation are usually independent, i.e. each instrument has its own independent effect on exposure and is not in linkage disequilibrium (LD) with another SNP. LD is described as the correlation between alleles at different loci [230]. It is not surprising that most Mendelian randomisation studies exclude dependent SNPs as potential instruments, as LD can lead to a violation of the three core assumptions of a instrumental variable. A genetic instrument (G_A) does not satisfy assumption 2 or 3 when it is in LD with another variant (G_B) that is directly associated with Y (Figure 1.2a) or associated with U (Figure 1.2b) respectively. The assumptions are satisfied if G_A is in LD with G_B the functional variant of X, Figure 1.2c. There are many cases where Mendelian randomisation studies only use the most significant SNP as an instrument even though they have genotyped thousands of SNPs [76, 87, 243, 306]; for example Cuellar-Partida et al. [76] have used only 10% of the 17,749 SNPs as instruments to infer causality between short-sightedness and educational attainment.



(a) Violation 1



(b) Violation 2



(c) No violation

Figure 1.2: Linkage Disequilibrium

The use of multiple instruments, where each instrument explains the variation of the exposure independently and satisfies the core assumptions will increase precision and power for the estimation of causal effect. The question that I would like to

address in this thesis is **can more information can be extracted from a gene region than is provided by a single proxy SNP for the unknown causal variant?**

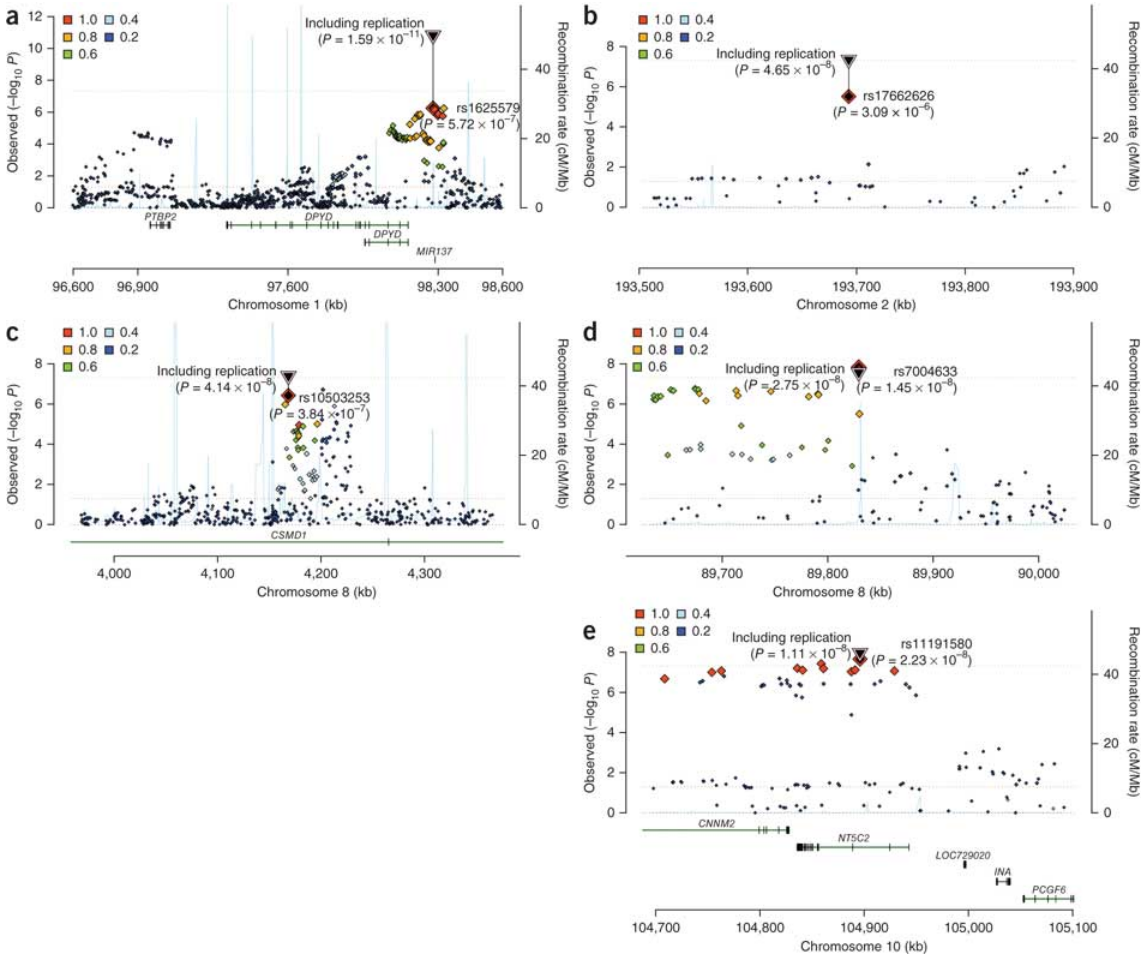


Figure 1.3: Regional plot taken from Schizophrenia Psychiatric Genome-Wide Association Study Consortium et al. [247]. On the x axis is the genomic position, and on the y axis is log_{10} P (p-value) for the SNP association with schizophrenia. The triangle is the most significant SNP and the LD of each SNP is based on the correlation with this most significant SNP. Colour coding (from red to blue) denotes LD information; see also the legend within the plot. The bottom of each plot gives the gene name and its region on the chromosome.

If the causal variant is unknown or unmeasured, due to its rarity in the population, then the joint information from all available variants in LD with the causal variant can predict the causal association [284]. Genetic association studies are usually split between direct and indirect associations [216]. A direct association is where the identified genetic variant has a causal role, whereas indirect associations

are where the associated variant has no causal role but is associated with the causal variant via LD. Johnson et al. [152] and Clayton et al. [69] have suggested having multiple "tag" SNPs to capture all the information in one gene for indirect association studies. In addition, the complexity of identifying the causal variant for a specific disease means not many direct association studies are available in the literature [197]. For example, Figure 1.3 shows the regional plots from Schizophrenia Psychiatric Genome-Wide Association Study Consortium et al. [247], where they have identified five SNPs associated with schizophrenia; in plot (e) there are SNPs with almost identical p-values and hence it is difficult to differentiate between the functional variant and the SNPs in LD with the functional variant. Moreover, a disease outcome may have multiple causal variants [144, 193, 203]. A single variant is unlikely to account for all of the LD between the unknown causal and genotyped SNPs in the region [301] and may underestimate the variance. Therefore, many SNPs in LD have the potential to be strong instruments that explain the total variance within the region.

Mendelian randomisation studies have been utilising genome-wide significant SNPs as instruments. However these SNPs may not be significant in another population; a population can have a different effect size for the genetic association to another population [239, 260]. For example, Imamura et al. [149] have found two loci with type II diabetes risk allele specific to the Japanese population. The same loci could not be replicated in East Asian (non-Japanese), South Asian, European and Mexican/latino population, when replication was performed with a SNP from the same region that is in high LD with the lead (i.e. most significant) SNP from the Japanese population. They have argued that it could be that the SNPs in LD with the causal variant only exist in Japanese population or LD between SNPs and causal variant is the same for all five populations but the causal variant only has an effect in the Japanese population. A study that collected data in the United States of America has found APOE$-\epsilon 4$ allele to have significant association with Alzheimer's disease in people of European but not African or Hispanic ancestry [261].

Realistically, genetic markers will be in LD; due to population genetic factors, such as natural selection, genetic drift and mutation [251], LD will continue to exist under many conditions. These factors affect the local recombination rate and in turn the level of LD [15]. Recombination is the process that generates new gene or chromosome combinations. Strong LD usually exists in a block-like structure between two recombination hotspots, known as haplotype blocks [108]. The blue

vertical lines in Figure 1.3 are the recombination rates, the taller the line the higher the recombination rate. Plot (a) shows the continuous decline of LD from the lead SNP with the increase in physical proximity within a haplotype block, between two relatively high recombination points (at 97,600 and 98,400 kilobases). Plot (e) has SNPs that are almost in complete LD with the lead SNP in the absence of recombination. The average length of a haplotype block is 16 kilobases (kb) and consists of 70 SNPs on average for European populations and the size of haplotype blocks is population dependent [150]. An average length gene is 53.6kb [258], thus there are approximately 3 haplotype blocks within a gene.

Using the same motivation as genetic association studies with SNPs from a single gene region, Burgess et al. [52] have investigated whether multiple variants in a single gene region from multiple studies can provide stronger instruments than a single variant from multiple studies. They compared allele scores with meta-analysis and other summarised data methods, where they modified these algorithms to incorporate the correlation between SNPs. They have simulated 15 correlated SNPs and each SNP has its own genetic effect on the exposure. Another simulation study by Wang et al. [283], showed that if the functional variant of the exposure was unmeasured or unknown, the surrogate marker could be a weak instrument as it does not explain the true amount of variation, therefore the LD information between typed and the causal variant could potentially explain the full variation within a gene region and have higher statistical power for the gene-exposure association in the first-stage of 2SLS. A haplotype is a set of alleles at linked loci in a single chromosome and an individual has two haplotypes, one inherited maternally and the other paternally. Two approaches were compared by Wang et al. [283], both based on 2SLS. SNP-IV selects SNPs as instruments via stepwise regression and haplotype-IV sorts haplotypes with similar effect sizes into the same group. They compared the two approaches by simulating 11 SNPs, only two of the SNPs were the functional variants of the exposure. Burgess et al. [52] and Wang et al. [283] have both concluded that the precision and power for the estimation of causal effect have increased with multiple correlated variants from a single gene as instruments rather than a single variant. Their findings raise the question of whether it would be possible to obtain the same benefit with multiple dependent SNPs in individual-level data.

There are several Mendelian randomisation studies that have utilised SNPs from the same gene region as instruments, for example, a study of plasma level on car-

diovascular disease using SNPs from the *AHSG* gene [103], adiponectin level and type 2 diabetes with SNPs from *ADIPOQ* [299], adiposity level and cardiovascular disease where instruments are SNPs from *TRIB1* [78]. However, these studies have given individual causal effect estimates from each SNP which are affected by weak instrument bias. Recent Mendelain randomisation studies have combined their dependent SNPs into a single instrument, using allele scores [49]. However, due to lack of external data with the same genetic-exposure association and population, they either derived the weights for the allele score from the dataset under analysis or did not weight the risk allele. The former approach is subject to bias [53] and the latter to sampling error [95]. Therefore, there are few guidelines on using SNPs from the same region as instruments in Mendelian randomisation.

1.3 Objectives

To summarise, the reasons for using multiple dependent SNPs are the complexity of knowing the causal variant(s) and the potential for SNPs in LD with the causal variant(s) to be stronger instruments than a single proxy. Therefore in this thesis I propose to investigate the advantage of having multiple dependent SNPs from the same gene region and whether a single most significant SNP is sufficient as an instrument for the estimation of causal effect in Mendelian randomisation. I will consider scenarios where the causal variant for the exposure is unknown or unmeasured and the dependent SNPs are associated with the exposure through their correlation with the causal variant. The exposure and outcome of interest will be assumed to be continuous in an individual-level dataset with information on SNP genotypes, exposure and outcome of interest. The 3 main questions of this thesis are;

1. Are there any gains from using many dependent SNPs from the same gene?
2. What is the most efficient estimator for many dependent SNPs?
3. In comparison to the classical approaches, are Bayesian approaches more efficient with many dependent SNPs?

For question 1, I will investigate the effect of LD on the amount of variation explained by a non-causal SNP through algebra and simulations.

For question 2, I will first review the Mendelian randomisation and instrumental variable analysis literature to find the recommended estimator for many instruments. Random LD patterns and MAFs will be simulated to find the most efficient estimator with many dependent SNPs, and the estimators compared to find any gains in excluding weak instruments within the gene region, by implementing instrument selection criteria.

Finally, the Bayesian approach to Mendelian randomisation and instrumental variable analysis will be reviewed for Question 3. Simulation of realistic genetic data will aim to find whether there are any gains for the Bayesian in comparison to classical approaches with many dependent SNPs as instruments.

1.4 Overview of Thesis

The background to Mendelian randomisation is in Chapter 2. This includes, the genetic and statistical limitations of Mendelian randomisation, a summary of its methodological development, and review of the applied literature on studies that used dependent instruments and their approach to causal estimation.

Chapter 3 is the statistical background that consists of a review of algorithms that have aimed to specifically reduce weak instruments bias, from the literature on Mendelian randomisation and instrumental variable analysis. Then follows the mathematics of 2SLS and recommended efficient estimators for many instruments.

Chapter 4 describes the general simulation method for the following chapters, which includes procedures to generate the genotype, exposure and outcome of interest for individuals, evaluation criteria to monitor performance and number of simulations to obtain the optimal accuracy level for the evaluation criteria.

Chapter 5 aims to find when using 2SLS, how much of the variation in exposure can be explained by SNPs in LD with the causal SNP. This will be derived mathematically but also demonstrated by simulation. Each scenario will have a different minor allele frequency (MAF), variation explained by the causal SNP, sample size or, where appropriate, correlation of SNPs with the causal SNP.

The performance of 2SLS with multiple dependent instruments will be examined in Chapter 6, where the potential instruments are non-causal SNPs but in LD with the causal variant. The first simulation experiment aims to investigate the effect of

LD in 2SLS, where the SNPs are included as instruments according to the ranking of p-value from their association with the exposure. The second experiment compares the performance of 2SLS with a single significant SNP or multiple non-causal SNPs and whether the instrument selection policy changes with different sample sizes and number of potential instruments available in the dataset.

As previous research has demonstrated that many weak instruments bias the causal effect estimate in 2SLS, Chapter 7 will compare the performance of 2SLS with the efficient estimators described in Chapter 3. Hence, this chapter aims to find the most efficient estimator with many dependent instruments.

To offer a Bayesian alternative to the classical approaches, Chapter 8 will review the methodological literature on Mendelian randomisation and instrumental variable analysis, and describe a Bayesian approach for the estimation of the causal effect. The objective here is to find a Bayesian algorithm that shows potential in improving the performance from the classical approaches, to Mendelian randomisation with many dependent instruments.

Chapter 9 will describe and demonstrate the Bayesian algorithm found in Chapter 8. The comparison of the classical and Bayesian approaches to Mendelian randomisation is described in Chapter 10. This chapter consists of 3 experiments; 2 experiments where the simulated datasets have SNPs that vary in MAF and LD patterns. The third experiment will use the genome simulator package to generate realistic patterns for the genetic instruments.

Finally, this thesis finishes with Chapter 11, which comprises a discussion of the findings from each chapter, their limitations and future work.

Chapter 2

Overview of Mendelian Randomisation

2.1 Introduction

This chapter reviews the problem of confounding in epidemiology and describes how instrumental variables can be used to adjust for confounding. After a brief review of some genetic terminology, the use of genetic variants as instruments is described.

2.2 Confounding in observational studies

2.2.1 Definition of confounding

Confounding is one of the major systematic errors in an epidemiological studies. Clayton and Hills [68] stated that;

> there is always a possibility that an important influence on the outcome, ..., differs systematically between the comparison groups. It is then possible that part of an apparent effect of exposure is due to these differences, and the comparison of the exposure group is said to be *confounded*.

Confounding can also be found in a clinical trials setting due to the systematic differences between the comparison groups [122].

The aim of a medical study is usually to measure the effect of an exposure on an outcome, but in practice the size of this effect is impossible to observe directly

as one individual cannot be both exposed and unexposed at the same time. For example, in a population of smokers, the mortality rate is I_1 and had those people not smoked the rate would have been I_0. A measure of the effect would be $I_1 - I_0$. However, I_0 is not observable, thus, I_0 is substituted by the mortality rate observed from another population of non-smokers. If the counterfactual rate (I_0) does not equal the mortality rate of non-smokers in the other population, then it is said that the measure of association is confounded [242].

A confounder must be associated with both the disease and exposure, as shown on Figure 2.1. However, association alone is not sufficient as Figure 2.2 shows. Intermediate factors, have often been mistaken for confounders. An intermediate factor lies between the exposure and the disease on the causal pathway, hence a confounder has to be an extraneous risk factor [121].

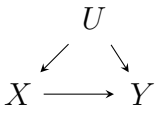


Figure 2.1: Confounder

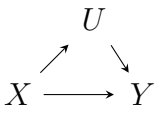


Figure 2.2: Correlation with exposure

To summarise, the following three conditions must hold for a variable to be a confounder are [242];

1. A confounder must be an extraneous risk factor for the disease

2. A confounder must be associated with the exposure under study in the population that are at risk

3. A confounder must not be affected by the exposure or the disease. In particular, it cannot be an intermediate step in the causal path between the exposure and the disease

These three conditions are necessary but not sufficient characteristics of a confounder. Further details are discussed by Rothman et al. [242].

2.2.2 Methods for Controlling Confounding

Measured Confounding

Measured confounding can be controlled by design or analysis. In the design based method, the systemic error created by confounding can be reduced by matching the exposed and unexposed cohort. There are two types of matching, individual and group matching. Individual matching is where individuals are paired up by their similarity in factors. Group matching assembles individuals with similar factors into groups. However, this can be difficult to achieve, since it is not always possible to match every person with more than a few factors. In addition matching introduces selection bias as it is selecting individuals based on their similarity in the factors.

Stratification is an analysis based method for controlling confounders. This divides the data according to a particular confounder, statistical analyses are then performed within strata. Thus, the chosen confounder is no longer affecting the exposure. In consequence, stratification can cause a sparse data problem; each stratum may contain little data and the more confounders that are controlled for, the worse the problem of sparse data.

A solution to the sparse data problem is to stratify in a way that balances the probability of allocation as estimated by the propensity score. The propensity score is a conditional probability of treatment given the observed covariates [238]. The propensity score allows data to be stratified on a single variable, instead of multivariable adjustment, which reduces the sparse data problem.

Regression models are often used to avoid sparse data problems [120]. Regression models allow the outcome to be dependent on both the exposure and the confounders; the true effect of exposure is estimated by adjusting for the confounders. However, regression models do not always solve the sparse data problem, as the number of degrees of freedom available for covariate adjustments being dependent on the sample size. Hierarchical regression model could potentially resolve this issue by giving flexibility to the degrees of freedom [117].

Unmeasured Confounding

In observational studies, unmeasured confounders are often ignored [155]. In consequence, the result produced is subject to systematic error and may be unreliable. Randomisation is the most popular design based method for reducing bias from unmeasured confounding. Randomisation assigns individuals into groups at random

and hence tends to balance the unmeasured confounders [241].

Sensitivity analysis can be used to investigate unmeasured confounding, by forming an 'educated guess' for the unmeasured confounder in the exposed group with the disease, and the effect of the unmeasured confounder in the unexposed group [242]. Then the 'educated guess' is included into the statistical analysis and the changes in result are examined.

The multiple regression analysis (extension to regression modelling) can be used to adjust for unmeasured confounding, this method is known as 'bias modelling', where its theory contains a combination of sensitivity analysis, Bayesian and Monte Carlo sensitivity analysis approaches. See Greenland [118] for more details on bias modelling.

In Econometrics, instrumental variable (IV) analysis is the method most commonly used to avoid unmeasured confounding.

2.3 The assumptions of IV analysis

There are three assumptions which must be satisfied for a variable to be used as an instrument [90]. Let Z be an instrumental variable, X a risk factor or an exposure, Y the disease-outcome and U all of the unmeasured confounders as illustrated in Figure 2.3. The assumptions are;

1. Z is associated with X,
2. Z is independent of U,
3. Z is independent of Y given X and U.

In the estimation of causal effect we make the further assumption that,

4. All associations in are linear and without interactions [264].

2.4 Genetic Background

The human genome is made up of long strands of deoxyribonuleic acid (DNA), which contains nucleotides bases of four possible types, adenine (A), cytosine (C), guanine (G) and thymine (T). The nucleotide bases pair up with one another, A with

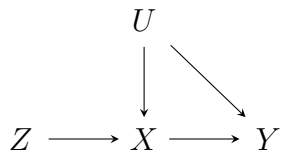


Figure 2.3: Directed acyclic graph (DAG) for the instrumental variable model. Z is the instrumental variable, X is the exposure of interest, Y is the outcome of interest and U is the confounding effect.

T and G with C to form base pairs (bp) of a double-stranded DNA. The human genome consists of 3.2×10^9 bp arranged into 23 pairs of chromosomes. There are 22 homologous pairs of autosomes and two sex chromosomes, XX and XY in females and males respectively. For each chromosomal pair, one copy is inherited from individual's mother and the other from the father. In each chromosome certain regions of DNA are known as genes, which encode the instructions for assembling amino acids into proteins. Chromosome 1 is the largest human chromosome with 3,168 genes and the Y chromosome is the smallest with 344 genes. The average length of protein-coding genes is 53.6×10^3 bp. However the length of a gene ranges from a few hundred to 2.4×10^6 bp long [258].

$$CTCGATCA\underline{G}AGCTGACT$$
$$CTCGATCA\underline{A}AGCTGACT$$

Figure 2.4: The bases of one strand of DNA are shown from a particular locus on two copies of a chromosome, the underlined pair of nucleotides is a SNP.

99.9% of the human genome is identical between two unrelated individuals and the variations in DNA sequence can be categorised into single nucleotide polymorphisms (SNPs), there are approximately 10 million SNPs [112]. A SNP is where a single nucleotide found at a chromosomal position (locus) is replaced by a different nucleotide, Figure 2.4. In each locus, the different possible nucleotides are known as alleles; the underlined nucleotides bases in Figure 2.4, G and A, are the two alleles for the SNP at this particular locus. The less common allele at a particular locus in a given population is labelled as the minor allele. The proportion of minor alleles in a SNP, within a given population, is known as the minor allele frequency (MAF). The distribution of MAF has an exponential decline from 0.01 to 0.5 [175]. 5 million

SNPs have MAF of greater than 10% and approximately 10 million with frequency greater than 1% [234].

The genotype of an individual consists of two alleles, one from each chromosome pair. Using the SNP in Figure 2.4 as an example, where the alleles are A and G, the genotypes that an individual can have are AA, AG or GG. A genotype is termed homozygous if an individual has the same allele on both chromosomes, or heterozygous if they are different. The allelic code on a single strand is called a haplotype.

Table 2.1: A glossary of genetic terms.

Alleles	variant at a given locus that varies between chromosomes in the population
Chromosome	a structure that carries a collection of genes located on a long string of DNA
Gene	a section of DNA sequence which codes for a protein or functional RNA molecule
Genotype	is the combination of alleles at a particular locus
Haplotype	is the set of alleles at linked loci on a single chromosome, an individual has two haplotypes at any loci, one inherited from mother and the other from father
Linkage Disequilibrium (LD)	the correlation between alleles at different loci
Locus	a unique chromosomal location
Phenotype	is a measurable characteristic of the subject
Pleiotropy	a genetic effect on more than one phenotype
Population Stratification	confounding by ethnicity
Recombination	is the process that occurs during meiosis, where sections of DNA are broken and recombined to produced new chromosomes
Minor Allele Frequency (MAF)	is the frequency of the less common allele within the population
Single Nucleotide Polymorphisms (SNPs)	are positions where a single nucleotide in the DNA is altered

Table 2.1 defines the genetic terms that will be used in this thesis.

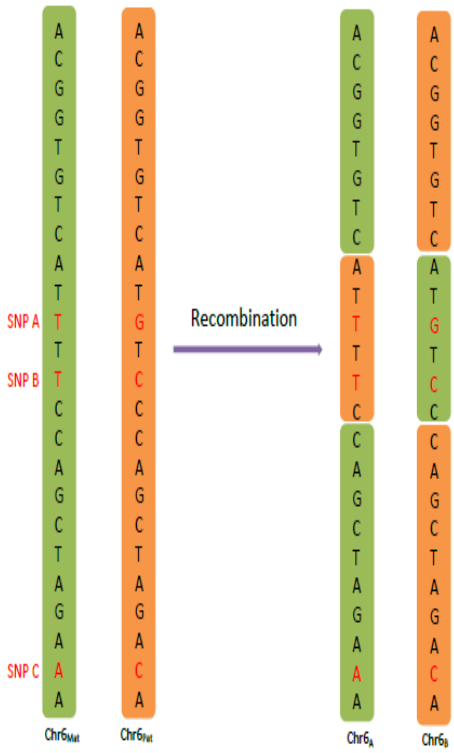


Figure 2.5: Illustration of recombination using a pair of chromosome (Chr 6) of an individual with 3 SNPs (SNP_A, SNP_B and SNP_C) taken from Jackson [151]; $Chr6_{Mat}$ and $Chr6_{Pat}$ is inherited maternally and paternally respectively. Two new chromosomes are formed, $Chr6_A$ and $Chr6_B$.

2.4.1 Linkage disequilibrium

Linkage disequilibrium (LD) is the correlation between a pair of SNPs and is commonly measured by the square of Pearson's correlation coefficient, r^2. LD extends over 1-100 kb, and is often discontinuous rather than declining smoothly with distance. LD is influenced by factors such as natural selection, genetic drift, population subdivisions and mutation [251]. Local recombination rates is the major determinant on how these factors affect LD between a pair of loci or in a region [15]. Recombination involves the exchange of genetic material either between multiple chromosomes or between different regions of the same chromosome, see Figure 2.5. The average recombination rate over 500 kb across the human genome is about 0.5cM, with the

range of 0.19cM to 1.25cM. A centimorgan (cM) is the unit of linkage that refers to the distance between two genetic loci determined by the frequency in which recombination occurs between them. There are also recombination hotspots, which are regions of chromosome where recombinations occur most frequently. These hotspots on average appear every 50 kb [18] and create a block-like structure of LD, known as haplotype blocks [108]. The average block size for European ancestry is 16 kb with 70 SNPs but only 5 of them have MAF greater than 5% [150]. Very common SNPs with MAF greater than 0.25 have mean maximum r^2 of 0.97 and for rarer SNPs (MAF < 0.05) of 0.76.

2.5 Genetic Instruments

Mendelian randomisation is IV analysis with genetic instruments [286]. In 1985, Katan [162] came up with the earliest concept of Mendelian randomisation. Observational studies had reported that low cholesterol increased the risk of cancer, which he believed to be reverse causation rather than causal. To distinguish the relationship, he thought of using Apolipoprotein E (apoE) polymorphism as an instrument, since the apoE-2/3/4 alleles control the phenotype of cholesterol. As the cholesterol level increases, apoE increases from 2 to 4. Therefore, to check for a causal relationship between low cholesterol and cancer, one could use a study of apoE on cancer.

This section will give a summary for the estimation of the causal effect (further details of these estimators can be found in the review by Burgess et al. [53]).

2.5.1 Estimation of causal effect

Suppose that an instrument satisfies all of the assumptions in Section 2.3. Then, for the simple case of a single instrument and a continuous outcome, the causal regression coefficient of disease outcome (Y) on risk factor (X), β_{XY}, is;

$$\beta_{XY} = \frac{\beta_{ZY}}{\beta_{ZX}} \tag{2.1}$$

where β_{ZY} is the regression coefficient of disease outcome (Y) on genetic instrument (Z) and β_{ZX} is the regression coefficient of risk factor (X) on genetic instrument (Z). This is also known as the Wald (ratio) estimator. The interpretation of Equation 2.1,

is that β_{ZY} accounts for the variation in Y explained by Z and by assumption 2 avoids the variation explained by the confounder, that affects both Y and X.

Genotypic data are becoming more readily available due to the decrease in genotyping costs and the increase in number of genome-wide association studies (GWAS). However, due to data protection and confidentiality laws, individual-level data are not always accessible. With this issue in mind, Thompson et al. [266] and Burgess and Thompson [44] have published statistical approaches to Mendelian randomisation with a single instrument and multiple instruments using summarized data. The former combines studies with information on genotype-phenotype and genotype-outcome associations to give an estimate of the causal effect via Bayesian meta-analysis, which will be outlined in Chapter 8. The latter has focused on two methods; inverse-variance weighted (IVW) estimator, a technique inspired from a GWAS [72], and the other is a likelihood-based method. The IVW calculates the Wald estimator from each instrument and its standard error is derived using the delta method. Then, the estimates from all of the instruments, are combined into a fixed-effect meta-analysis model, to obtain a single causal estimate. The likelihood-based method assumes a linear relationship between X and Y, and the genetic association estimates have a bivariate normal distribution. The causal effect can be estimated by maximum likelihood or the Bayesian method, both algorithm assume all the instruments give the same causal effect. If β_{ZX} and β_{ZY} are from a meta-analysis of different studies, then the covariance matrix of the bivariate distribution can account for their correlations, i.e. heterogeneity between the studies.

The assumption that an outcome is continuous can limit Mendelian randomisation, as epidemiological studies often express causal effects as an risk ratio or odds ratio. The main problem for a binary outcome is that the regression of Y on Z and X on Z with the compatible parameter is no longer a consistent estimator of the causal effect and will result in a biased estimate [90]. There are several IV analysis methods available for binary outcomes described in detail by Greene [116]. However, some of the econometric methods makes assumptions that are not reasonable for Mendelian randomisation. Palmer et al. [218] have provided an overview of IV methods that are suitable for Mendelian randomisation with binary outcomes. For multiple instruments, Bowden and Vansteelandt [30] suggested applying these instruments to a structural mean model (SMM) for a case-control study and Clarke et al. [67] proposed estimating the causal relationship using a generalised method of moments estimator. Bayesian meta-analysis can be implemented to combine in-

formation from multiple studies with binary outcomes [55]. In summary, there are a few approximation methods available for incorporating multiple instruments in Mendelian randomisation with binary outcomes. However, it is still unclear which algorithm should be used to estimate efficiently the causal effect with multiple instruments.

In addition, there are further developments with estimation of causal effect in survival data [262], data with non-linear relationship between X and Y [50], estimation with multiple pleiotropic instruments with multiple outcomes [46] and missing data [48]. Even unbiased estimators with invalid instruments were developed; Egger regression estimates a causal effect with instruments that violate IV assumptions 2 and 3 [31], but instead the instruments have to satisfy the InSIDE (Instrument Strength Independent of Direct Effect) assumption. The InSIDE assumption is violated if several instruments are associated with the same confounder. The weighted median estimator is unbiased, if less than 50% of the instruments violate assumption 2 and 3 [32]. Bidirectional Mendelian randomisation investigates the direction of the causal effect [273, 287]. Mendelian randomisation analysis with mediator, where the mediator gives another pathway from exposure to the outcome of interest [51]. A 2×2 factorial Mendelian randomisation was purposed to understand the effect of multiple treatments on the risk of coronary heart disease [102]. To determine the environmental effect on the gene expression and in turn the disease outcome, two-step epigenetic Mendelian randomization have been advocated [235].

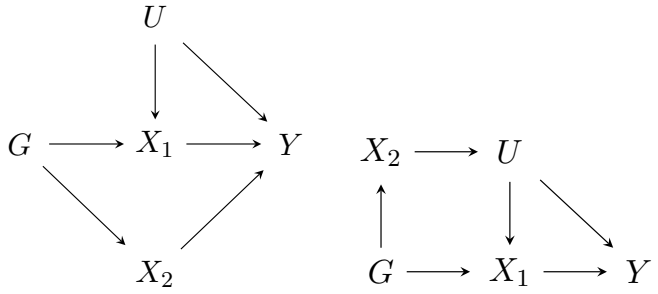
2.5.2 The violations of the IV assumptions

Davey Smith and Ebrahim [79] described several practical limitations to Mendelian randomisation that derive from dependence on the three assumptions and linearity without interactions.

Pleiotropy

Pleiotropy is a genetic variant's potential for having more than one specific phenotypic effect. Let X_1 and X_2 be the two different phenotypes affected by the genetic variant (G). In example 1 shown in Figure 2.6a, G would not be a valid instrument unless X_2 was identified and adjusted for, since the presence of X_2 violates assumption 3. Example 2, Figure 2.6b, shows a violation of assumption 2 due to pleiotropy, since G is not independent of the confounder. The bias from pleiotropy

can be avoided if only variants with a well understood genetic function are used as instruments [82]. Also, having multiple independent genetic variants can help to identify the possible bias. If each genetic variant that affects the risk factor (X) produces a similar estimate of β_{XY}, then the chance of pleiotropy being present is low although they could, in theory, all be biased in the same way.



(a) Example 1 (b) Example 2

Figure 2.6: Illustration of the violations of the IV assumptions via Pleiotropy

Linkage disequilibrium

Linkage disequilibrium (LD) is the statistical association between a pair of alleles at different loci within the population. Such associations exist, because alleles are physically close together and consequently tend to be co-inherited. As described in Section 1.2, pleiotropy via LD can invalidate the three assumptions, but if the instrument is acting as a proxy for the functional variant of the exposure then no assumptions are violated. Dealing with LD is the same as for pleiotropy; the increased knowledge of the genetic function and the use of multiple independent instruments can help identify potential problems.

Population stratification

Population stratification is a type of confounding, whereby allele frequencies and disease outcome or allele frequencies and risk factor, vary between different sub-groups of the population, and thus giving an illusion of association in the overall population. Cardon and Palmer [58] suggested a "classic" example of the violation of assumption 3 via population stratification; Knowler et al. [171] found an association between haplotype Gm3;5,13,14 with type II diabetes, in a study of people with White European and Pima Indian origin, however, the association disappeared when

the analysis was performed separately for different ethnicities. This violation of assumption 3 is illustrated in Figure 2.7, where P is the population stratification. Population stratification can be solved by adjusting for population structure in the analysis or by performing family-based studies [182].

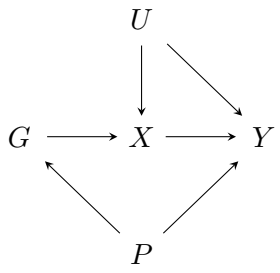


Figure 2.7: Illustration of the violation of IV assumption due to population stratification

Canalization

Canalization, also known as developmental compensation, is the adaptation of an individual to a genetic change where the effect of the genetic change is reduced or removed. For example, if a person has the genetic variant that is associated with high blood pressure and does not experience adverse events induced by high pressure because the person's arteries have become resistant to high blood pressure [281]. Hence, canalization can cause bias in Mendelian randomisation estimation as it is affecting the genotype-outcome association but not the genotype-exposure association. There is currently no strategy known to identify violation via canalization, except through biological knowledge [98].

Statistical assessment of IV assumptions

It is difficult to validate the instrumental variable assumptions, as there are no statistical tests of the assumptions. However, there are procedures that will give confidence that an instrument does satisfy the core assumptions.

The simplest procedure is to test the association between the genetic variant and measured confounders, but there may still be violations through unmeasured confounders. Glymour et al. [113] have suggested four approaches to verify the core assumptions; leverage prior causal assumptions, identify factors that modify the genotype-phenotype association, apply instrumental inequality tests and the

use of multiple instruments to conduct overidentification tests (overidentification is when there are more instruments than endogenous variables). Using heterogeneity in a meta-analysis is a summarised data version of the overidentification test [115]. However, Burgess and Thompson [42] and Palmer et al. [219] have found counter examples to the Glymour et al. [113] approaches, and concluded that the validity of these assumptions must be supported by biological knowledge and not just by empirical tests. Others have proposed biological guidelines based on Bradford Hill criteria [136] to test the plausibility of IV assumptions [54].

2.5.3 Weak genotype-phenotype associations

The genotype-phenotype associations are usually weak in the sense that only a small proportion of the phenotype is explained by the genetic variant and does not provide enough statistical evidence for the association with the phenotype [182]. For example, in the GWAS of Horikoshi et al. [143], the genetic variants that are significantly associated with birth weight, each explained below 2% of the variance in birth weight. This weak association causes two main statistical problems in the estimation of causal effects; weak instrument and finite sample bias, and a lack of power to reject the null hypothesis of no causal effect between X and Y.

Weak instrument and finite sample bias

The magnitude of bias depends on both sample size and variability in phenotype that is explained by the instrument [257]. IV estimators are asymptotically unbiased but biased in finite samples. The strength of a instrument, or set of instruments, is defined by the F-statistics from the association between instrument(s) and exposure of interest ;

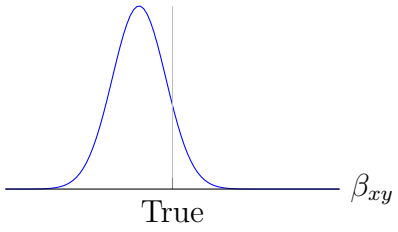
$$F = \frac{R^2(n-1-k)}{(1-R^2)k} \tag{2.2}$$

where R^2 is the proportion of the variability in phenotype that is explained by the genotype(s), n is the size of the sample and k is the number of genetic instrument(s). Staiger and Stock [253] suggest that F-statistics of less than 10 will cause weak instrument bias. However, this definition is misleading, as the term "weak instrument bias" seems to suggest that there is a lack of statistical evidence from the instrument but an instrument can be made stronger by increasing the sample size, as shown by Equation 2.2. The bias of 2SLS to the ordinary least squares (OLS) estimator,

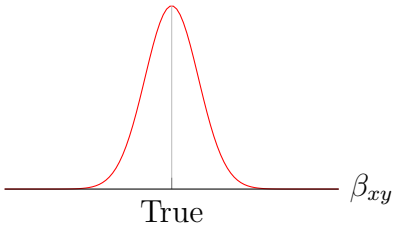
referred to as relative bias, is

$$\frac{\text{bias in 2SLS}}{\text{bias in OLS}} = \frac{\sigma_{\hat{x},\epsilon}/\sigma_{x,\epsilon}}{R^2_{x,z}} \tag{2.3}$$

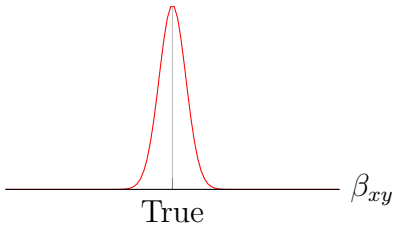
where $\sigma_{i,j}$ is the covariance of i and j, $\hat{x}$ is predicted X by instrument Z and ϵ is error including the confounding effect. Therefore, as variation explained by Z ($R^2_{x,z}$) decreases, the relative bias will increase. Finite sample bias comes from the OLS in the first-stage regression to estimate the coefficient of X on Z. In finite samples, the non-zero correlation between error and instrument is less certain, which creates uncertainty into predicting X and in turn biases the causal effect estimate [29].



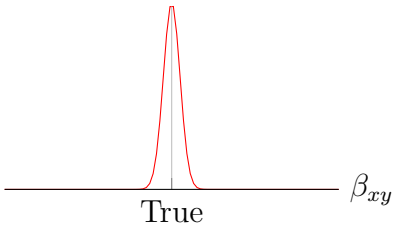
(a) Small R^2 and sample size



(b) Larger R^2 and small sample size



(c) Larger sample size and small R^2



(d) Larger sample size and R^2

Figure 2.8: Illustration of weak instrument and finite sample bias

The effect of weak instrument and sample size are illustrated in Figure 2.8. Consider datasets that have weak instruments (i.e. small R^2) and small sample size, then the distribution for the causal effect estimates take the shape of Figure 2.8a, most of the datasets inaccurately estimate the causal effect, and variation is large (i.e. uncertainty carried from first-stage regression). If the total variation explained by the genetic instruments is increased but the sample size remains small, then estimates will centre at the true causal effect but the variation is still large, Figure 2.8b. Figure 2.8c shows more datasets will estimate the true causal effect with smaller variation if sample size of each dataset is increased even where R^2 is small. Increasing both sample size and R^2 in each dataset induces all datasets to estimate the true causal effect, centred at true causal effect and smaller variation, Figure 2.8d.

Statistical power

The weak association between genetic variant and exposure causes a test for a causal effect to lack power. Pierce et al. [227] have performed simulation studies in order to estimate the power for single and multiple instruments and Freeman et al. [105] have given a formula to calculate power for a single genetic instrument and Brion et al. [36] have given power calculations for multiple instruments. For a binary outcome, power calculation is only available for a single instrument [41]. All of these power calculations demonstrate that power is heavily dependent on sample size, variation explained by the genetic instrument and the effect size between X and Y. Figures 8.1 and 8.2 in Burgess and Thompson [45] are the power curves of these relationships.

2.5.4 Improving instrument strength

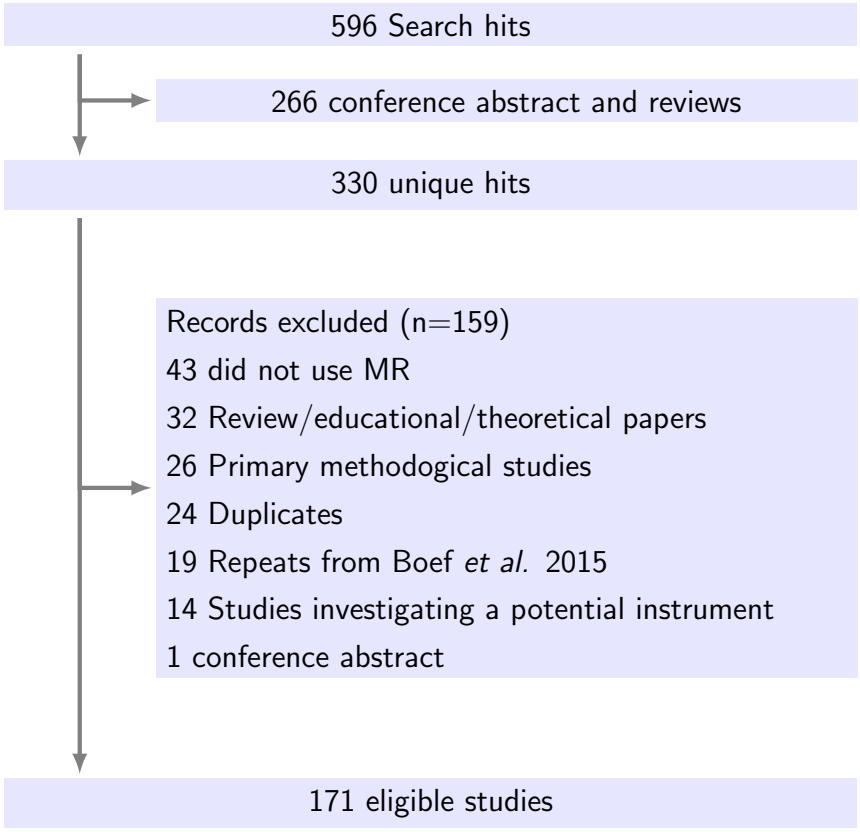
Weak instrument bias and low power in a single instrument can be resolved by having a large sample size. Davey Smith and Ebrahim [79] have advised performing sample size calculations before embarking on a Mendelian randomisation study. However, obtaining a large sample size is not always realistic, due to the expense. Instead, information can be borrowed from other studies, Thompson et al. [266] have proposed combining multiple genotype-phenotype and genotype-outcome associations to estimate the causal effect from a single genetic instrument via meta-analysis and other methods of borrowing information from other studies have been proposed. (See Section 2.5.1)

A commonly used approach is to include multiple instruments to increase power and precision [220]. However, this would not necessarily remove weak instrument bias [227] as seen in Equation 2.2; as k increases, the F-statistic decreases. Thus, this is known as many weak instruments bias. Achieving both unbiased and well-powered estimates of causal effects with multiple instruments, will depend on the additional variance in risk factor explained by each instrument i.e. each instrument included in the first stage regression must contribute to the overall R^2. For the estimation of causal effect with many weak instruments, LIML is recommended as it is median unbiased [85]. Instead of implementing LIML to reduce weak instrument bias, multiple SNPs can be combined into a single instrument [44], a procedure more commonly known in Mendelian randomisation as allele scoring. This method has also been found to increase power [227].

2.6 Review of multiple dependent instruments

Th objective of this section is to describe Mendelian randomisation studies that have used many dependent instruments within a genetic region. Recently, a review of Mendelian randomisation studies between May 2005 and December 2013 was published by Boef et al. [26]. Therefore, the same search and exclusion criteria described in Boef et al. [26] will be adopted to update this review. After completing the update, the next section will include studies that have more than one SNP from a single genetic region as instruments, the information of the statistical method implemented to estimate the causal effect and the type of instrument used in the estimation will be extracted.

Figure 2.9: Summary of the literature search 2014-2016



2.6.1 Review of Mendelian randomisation studies

Employing the same search terms as Boef et al. [26] in EMBASE and restricting publication between January 2014 to June 2016, returned 596 hits. Figure 2.9 gives the summary of the updated literature search, 266 conference abstracts and review

articles were removed by reading titles and abstracts. 159 records were removed by following Boef et al. [26]'s exclusion criteria. In total, there were 171 and 179 eligible studies from the update and Boef et al. [26] respectively.

Table 2.2 gives the number of studies that had independent, dependent and single instruments for the estimation of the causal effect in both Boef et al. [26] and the update. Notice there are only 176 studies from Boef et al. [26] even though their paper reported 179 eligible studies, as there was 1 study that had a non-genetic instrument [17] and 2 studies that were review articles of Mendelian randomisation [63, 91]. From both the published review and update there were 80 studies that had multiple dependent instruments in the form of SNPs in a single genetic region, approximately 23% of the total number of Mendelian randomisation studies. Even though 40% of the studies had independent instruments, there are cases where the genotypes of thousands of SNPs were available but only the most significant SNP in the region was used [76, 86, 243, 306]. 4 of the studies with a single instrument followed the same procedure [189, 221, 222, 265].

Table 2.2: Form of genetic instrument in Mendelian randomisation studies from Boef et al. [26] and my update. The total number of studies is 346

Form of Genetic Instrument(s)	Boef et al. [26]	Update	Total
Independent	47	91	138
Dependent	52	28	80
Single	77	52	129

2.6.2 Mendelian randomisation studies with dependent instruments

Tables 2.3 and 2.4 show the types of genetic instrument and statistical method implemented in 80 studies. Some studies appear in both Tables as these studies have used two types or methods within their analysis.

The types of genetic instrument included in Table 2.3 are as follows: multiple SNPs analysed separately so each SNP has estimated their own causal effect; multiple SNPs included additively into the estimation model; multiple SNPs combined

into a single instrument via genotype combination (or haplotype); weighted allele scores; and unweighted allele scores. Most studies included the dependent instruments additively into the estimator, without accounting for the correlation between SNPs, Table 2.3. The next most popular procedure, used in 17 studies, was combining dependent SNPs into a single instrument via a weighted allele score, however 10 of the studies derived weights for the allele score from the same dataset under analysis [88, 110, 154, 166, 208, 209, 246, 277, 300, 302]. As mentioned in the Chapter 1 this form of weighting is subject to selection bias.

Table 2.4 shows that two-stage least squares (2SLS) is the most popular estimator among studies with multiple dependent instruments. Genotype-outcome association refers to methods that did not derive an effect size for the causal relationship, but monitored the association between genotype that had been established to have an effect on the exposure and outcome of interest. Cox proportional hazards model are usually used in the context of survival analysis. There are a series of methods called two-stage, which have the same procedure as 2SLS but instead of using linear regression in both stage, logistic, cox model or mixed regression were implemented in the second stage.

In comparison to Boef et al. [26], the findings in Mendelian randomisation studies with dependent instruments are not similar. In terms of types of instruments, Boef et al. [26] found that most studies with a single population used multiple SNPs in a single analysis (11 out of 31 studies) and then allele score, not specified as weighted or unweighted, was the next most popular type of instrument. For dependent instruments, there are approximately the same number of studies that used allele score and multiple SNPs in a single analyses. Genotype-outcome association, without further estimation is the most popular statistical method in studies from Boef et al. [26] and 2SLS is the next. Contrary to studies with dependent instruments, the order is reversed, 2SLS is most popular.

2.7 Conclusion

Mendelian randomisation infers causality between exposure and outcome of interest adjusting for unmeasured confounding, which observational epidemiology cannot. Even though there is limited methodological development and lack of guidelines for using SNPs from the same genetic region as instruments, 23% of the 346 Mendelian randomisation studies have used instruments in this form. There are also studies

that only used the most significant SNP among the many SNPs in a single gene and exclude SNPs that are in LD with the lead SNP. The use of multiple dependent SNPs from the same genetic region has the potential to increase power and precision of the causal effect estimate [52, 283]. The next chapter will examine the estimators for multiple dependent instruments but with individual-level datasets, where information on genotypes, exposure and outcome of interest are available from each participant.

Table 2.3: Type of instruments in Mendelian randomisation with dependent instruments for single and multiple population (Pop.).

Type	1 Study Pop.	Refs	> 1 Study Pop.	Refs
Multiple SNPs in separate analyses	14	[1, 4, 27, 34, 75, 78, 103, 173, 180, 180, 188, 191, 275, 296]	3	[59, 101, 231]
Multiple SNPs in a single analysis	26	[28, 35, 37, 38, 65, 73, 93, 109, 123, 134, 156, 157, 165, 170, 187, 199, 212, 255, 259, 271, 276, 278, 280, 291, 299, 308]	2	[72, 148]
Combination of SNPs	8	[156, 157, 173, 180, 236, 291–293]	0	
Weighted allele score	17	[75, 88, 110, 154, 166, 179, 208, 211, 223, 246, 250, 277, 297, 298, 300, 302, 307]	2	[209, 269]
Unweighted allele score	10	[2, 37, 71, 176, 208, 208, 214, 268, 282, 303]	0	

Table 2.4: Statistical methods used Mendelian randomisation studies with single or multiple study population (1 and > 1 Pop. respectively)

Method	1 Pop.	Refs	> 1 Pop.[a]	
Two-stage least squares	25	[37, 38, 73, 75, 110, 123, 154, 165, 176, 179, 188, 208, 214, 259, 271, 276, 277, 280, 282, 291, 292, 298, 299, 307, 308]	0	
Genotype-outcome association	20	[4, 27, 28, 35, 65, 71, 87, 93, 109, 134, 156, 187, 191, 199, 236, 255, 275, 278, 296, 300]	2	[59] [212][b]
Ratio/Wald estimator	11	[1, 2, 34, 78, 103, 167, 211, 297, 302, 303, 306]	3	[101, 147, 231]
Cox proportional hazards model	3	[14, 246, 250]	0	
Generalised least squares regression	4	[5, 77, 196, 304]	0	
Two-stage: linear first, logistic regression in second	4	[209, 223, 268, 293]	3	[72][c] [209, 270]
Two-stage: linear first, cox regression in second	1	[180]	0	
Two-stage: mixed linear first, linear regression in second	1	[228]	0	
IV probit regression	1	[133]	0	
Multiplicative Generalised Method of Moments	1	[157]	0	
Generalised Method of Moments	1	[75]	0	
Limited Information Maximum Likelihood	1	[75]	0	
Variance Component model	1	[173]	0	
Unclear	1	[170]	0	

[a]First performed IV analysis in single study, then meta-analysis to pool the results from each study.
[b]Analysis was performed on multiple populations but did not pool the results.
[c]hierarchical Bayesian Meta-analysis was implemented.

Chapter 3

Statistical Approaches to Mendelian Randomisation with Multiple Dependent Instruments

3.1 Introduction

The review of Mendelian randomisation studies in the previous chapter has shown that dependent instruments are widely used, although some studies only use the most significant SNP in the region. However, there is a lack of methodological development and few guidelines to using many dependent SNPs as instruments. As discussed in Chapter 1, there is evidence that having multiple SNPs in a single region can explain more of the variation in the exposure, than a single most significant SNP. Therefore, the aim of this chapter is to review the potential estimators for multiple dependent instruments in Mendelian randomisation. I will only focus on the most widely used estimators examined by Davies et al. [85], in their study of estimators for many weak *independent* instruments. The main focus of this chapter is the statistical method for many instruments with individual-level data, for other and the general discussion of approaches in Mendelian randomisation see Section 2.5.1 and Burgess et al. [53] for details.

3.2 Instrumental variable analysis

Much of the development of the theory of Mendelian randomisation estimators is reported in the econometrics literature, so, before introducing these instrumental variable estimators, the next section will give the definitions of econometric terms and equivalent terms in epidemiology, and present an example to explain some of the terms in detail.

3.2.1 Econometric Terminology

The definitions in Table 3.1 come from Wooldridge [295].

Table 3.1: A Glossary for econometrics with equivalent epidemiological terms

Econometrics term	Epidemiological term	Definition
Endogenous explanatory variables	Confounded explanatory variables	a variable in a multiple regression that is correlated with the error term, caused by unmeasured confounding or measurement error.
Exogenous explanatory variables	Unconfounded explanatory variables	a variable that is uncorrelated with the error term.
Error term / disturbance	random error	The variable in regression equation that contains unobserved factors that affect the dependent variables, which may also include measurement error.
Error in variables	measurement error	situation where a dependent or an independent variable is measured with error.
Instrument relevance	Assumption 1	an instrument must be relevant for explaining the variation in X.
Concentration parameter	-	is a measure of the magnitude of the instrument relevance.
Exclusion restrictions	Assumption 2 and 3	an instrumental variable does not have a direct effect on the outcome and is not associated with the error.
Structural Equation	-	the equation that measures a causal relationship. See example below.

Reduced form equation -	a linear equation where an endogenous variable is a function of exogenous variables and error terms. See example below.

An example

Consider this equation;

$$X = \alpha_0 + \alpha_1 Z_1 + \alpha_2 U + \epsilon_x$$
$$Y = \beta_0 + \beta_1 X + \beta_2 U + \epsilon_y$$

where X is the exposure, Y the outcome, Z_1 the instrument, U the unmeasured confounding effect, and ϵ_x and ϵ_y are the random error for X and Y respectively. As confounding is unmeasured, this equation can be re-parametrised into the **structural form**,

$$X = \alpha_0 + \alpha_1 Z_1 + e_x$$
$$Y = \beta_0 + \beta_1 X + e_y$$

The structural form in econometrics is usually referred to as the equation that measures a causal relationship. e_x and e_y are now the error term for both confounding and random effect, and are correlated. In econometrics, X is known as the **endogenous** variable, as it is correlated to the error term from the two equations' joint dependence on U. Whereas, Z_1 is an **exogenous** variable, independent of the error term, as it is generated externally from the two equations.

We can substitute X into the second equation to give the **reduced form**,

$$X = \alpha_0 + \alpha_1 Z_1 + e_x$$
$$Y = \beta_0 + \beta_1(\alpha_0 + \alpha_1 Z_1 + e_x) + e_y$$

As coefficients of Z_1 has been structured into α_1 and $\alpha_1\beta_1$, this parametrisation is known as **restricted reduced form**. The equations could be written as,

$$X = \alpha_0 + \alpha_1 Z_1 + e_x$$
$$Y = \beta_0 + \beta_1\alpha_0 + \phi Z_1 + \beta_1 e_x + e_y$$

where ϕ is not restricted to equal to $\alpha_1\beta_1$, this equation is therefore known as the **unrestricted reduced form**.

3.2.2 Generalised Method of Moments

Suppose that we have a model with one outcome (Y), L continuous exposures (X), J instrumental variables (Z) and i denotes the individual and n is the sample size;

$$Y_i = X_i'\beta + e_{yi}, \tag{3.1}$$

where Y_i is a scalar, X_i is $L \times 1$ vector of exposures, β is also $L \times 1$ vector of the regression coefficient and e_{yi} is the error term as a scalar. For an unbiased estimator of the causal effect, the generalised method of moments (GMM) estimator must satisfy the population moment conditions

$$E[g_i(\beta)] = E[Z_i(Y_i - X_i'\beta)] = 0. \tag{3.2}$$

where the Z_i is $J \times 1$ vector of instruments and $J \geq L$. i.e. each instrument must satisfy this condition. This is a necessary condition, if $J < L$, the single structural equation cannot be identified, the equation is known as unidentified in econometrics [294]. As part of the IV assumptions, each instrument Z_i must be independent of error term $(Y_i - X_i'\beta)$, hence the expected association between instrument and error term must equal to 0, Equation 3.2. First consider the case where $J = L$, then we can solve

$$\hat{\beta} = D^{-1}s \tag{3.3}$$

where $D = n^{-1}\sum_{i=1}^{n} Z_i X_i'$ and $s = n^{-1}\sum_{i=1}^{n} Z_i Y_i$. In full matrix notation, $\hat{\beta} = (Z'X)^{-1}Z'Y$, which is the Wald estimator seen in Section 2.5.1. However, for $J > L$, there might not be a solution that satisfies Equation 3.2. Instead, the estimator aims to derive a β that gives the minimum criterion function $Q(\beta)$. The general form of GMM criterion function is

$$\hat{\beta} = arg\min_{\beta} Q(\beta) \tag{3.4}$$

$$Q(\beta) = \bar{g}(\beta)'W\bar{g}(\beta), \tag{3.5}$$

where $\bar{g}(\beta) = n^{-1}\sum_{i=1}^{n} g_i(\beta) = n^{-1}\sum_{i=1}^{n} Z_i(Y_i - X_i'\beta)$ and W is a weight matrix, this is customised for different GMM estimators. The solution to minimum criteria function is not always unique.

Two-stage least squares

Two-stage least squares (2SLS) is a special case of GMM. 2SLS has weighting matrix, $W_z = (n^{-1}\sum_{i=1}^{n} Z_i Z_i')^{-1}$. Provided the inverse exists, then the GMM criterion function for 2SLS becomes

$$\hat{\beta}_{2SLS} = arg\min_{\beta} Q_{2SLS}(\beta) \tag{3.6}$$

$$Q_{2SLS}(\beta) = \bar{g}(\beta)' W_z \bar{g}(\beta), \tag{3.7}$$

Setting $\bar{g}(\beta)' W_z \bar{g}(\beta)$ to zero and rearrange terms gives

$$\hat{\beta}_{2SLS} = (D' W_z D)^{-1} D W_z s, \tag{3.8}$$

where $D = n^{-1}\sum_{i=1}^{n} Z_i X_i'$ and $s = n^{-1}\sum_{i=1}^{n} Z_i Y_i$. The 2SLS estimator is efficient if the population moment conditions are satisfied and when the variance of e_{yi} from each instrument is the same, i.e. $E[e_{yi}^2|Z_i] = \sigma^2$, condition known as homoskedasticity.

Two-step GMM

The difference between two-step GMM and 2SLS is the weighting that is put on $g(\beta)$, two-step GMM derives its weighting from the residual of 2SLS, i.e. weighted by how well the 2SLS estimate predicts Y. As shown by the minimising function Q_{GMM} and weighting matrix W_{GMM};

$$Q_{GMM}(\beta) = \bar{g}(\beta)' W_{GMM} \bar{g}(\beta), \tag{3.9}$$

$$W_{GMM} = (n^{-1}\sum_{i=1}^{n} g_i(\beta_{2SLS}) g_i(\beta_{2SLS})')^{-1} \tag{3.10}$$

where $g_i(\beta_{2SLS}) = \sum_{i=1}^{n} Z_i(Y_i - X_i'\beta_{2SLS})$. $W_z = W_{GMM}$ when $E[e_{yi}^2|Z_i] = \sigma^2$ [207], hence two-step GMM and 2SLS will give the same causal estimate if condition of homoskedasticity is satisfied.

Continuously updating estimator

The continuously updating estimator (CUE) has a similar algorithm to two-step GMM but it is more of an iterative approach. The process begins with a random β with identity matrix as weights. Then, with a new β, CUE re-calculate the weights from the previous β for the current and continues until the estimates converge.

$$Q_{CUE}(\beta) = \bar{g}(\beta)' W_{CUE} \bar{g}(\beta), \tag{3.11}$$

$$W_{CUE} = (n^{-1} \sum_{i=1}^{n} g_i(\beta) g_i(\beta)')^{-1} \tag{3.12}$$

where $g_i(\beta) = \sum_{i=1}^{n} Z_i(Y_i - X_i'\beta)$. As CUE is an iterative process and as mentioned above the GMM criterion function does not always have an unique solution, which causes CUE not able to converge in some cases.

3.2.3 Limited Information maximum likelihood

LIML is also known as least variance ratio estimator. LIML is constructed for each equation individually which is similar to least squares; generating one equation at a time. Least squares estimation does not assume any distribution for the error terms, whereas LIML assumes the error to have a normal distribution. LIML has the same asymptotic distribution as 2SLS, this means that with large enough sample size the two estimators will give the same answers. The LIML estimate is equivalent to minimising this function;

$$Q_{LIML}(\beta) = \frac{\bar{g}(\beta)' W_{LIML} \bar{g}(\beta)}{(\sigma(\beta))^2}, \tag{3.13}$$

$$W_{LIML} = (n^{-1} \sum_{i=1}^{n} Z_i Z_i')^{-1} \tag{3.14}$$

where $(\sigma(\beta))^2 = n^{-1} \sum_{i=1}^{n} (Y_i - X_i'\beta)^2$. Under homoskedasticity, $E[e_{yi}^2 | Z_i] = \sigma^2$, $Q_{CUE} = Q_{LIML}$. Notice, the numerator of Q_{LIML} is the same as the minimising function in 2SLS but the denominator is the variance of the OLS estimate.

3.2.4 An illustration of the estimators

Design

Consider an example where there are two instruments Z_1 and Z_2, exposure (X), outcome (Y), the confounding effect between X and Y (U) and random error for X and Y, ϵ_x and ϵ_y respectively. The true causal effect between X and Y is 0.5, and Z_1 and Z_2, each explain 2% of the variation in X.

To demonstrate the minimising functions and weighting algorithm of 2SLS, two-step GMM, CUE and LIML, I have restricted the estimator to only find a solution from β ranged between -5 and 5 in steps of 0.01. For each estimator, estimates for $g(\beta)$ from each SNP and their interaction will be calculated;

$$g_1(\beta) = n^{-1}\sum_{i=1}^{n} SNP_{1i}(Y_i - X_i'\beta)$$
$$g_2(\beta) = n^{-1}\sum_{i=1}^{n} SNP_{2i}(Y_i - X_i'\beta)$$
$$g_{12}(\beta) = n^{-1}\sum_{i=1}^{n} SNP_{1i}(Y_i - X_i'\beta) * n^{-1}\sum_{i=1}^{n} SNP_{2i}(Y_i - X_i'\beta)$$

For LIML, $g(\beta)$ will be divided by residual of Y, $n^{-1}\sum_{i=1}^{n}(Y_i - X_i'\beta)^2$. These estimates are then combined and weighted according to each algorithm to give their value of Q. As there are 2 instruments in this example the weighting matrix is a square 2×2 matrix (W). Therefore Q is

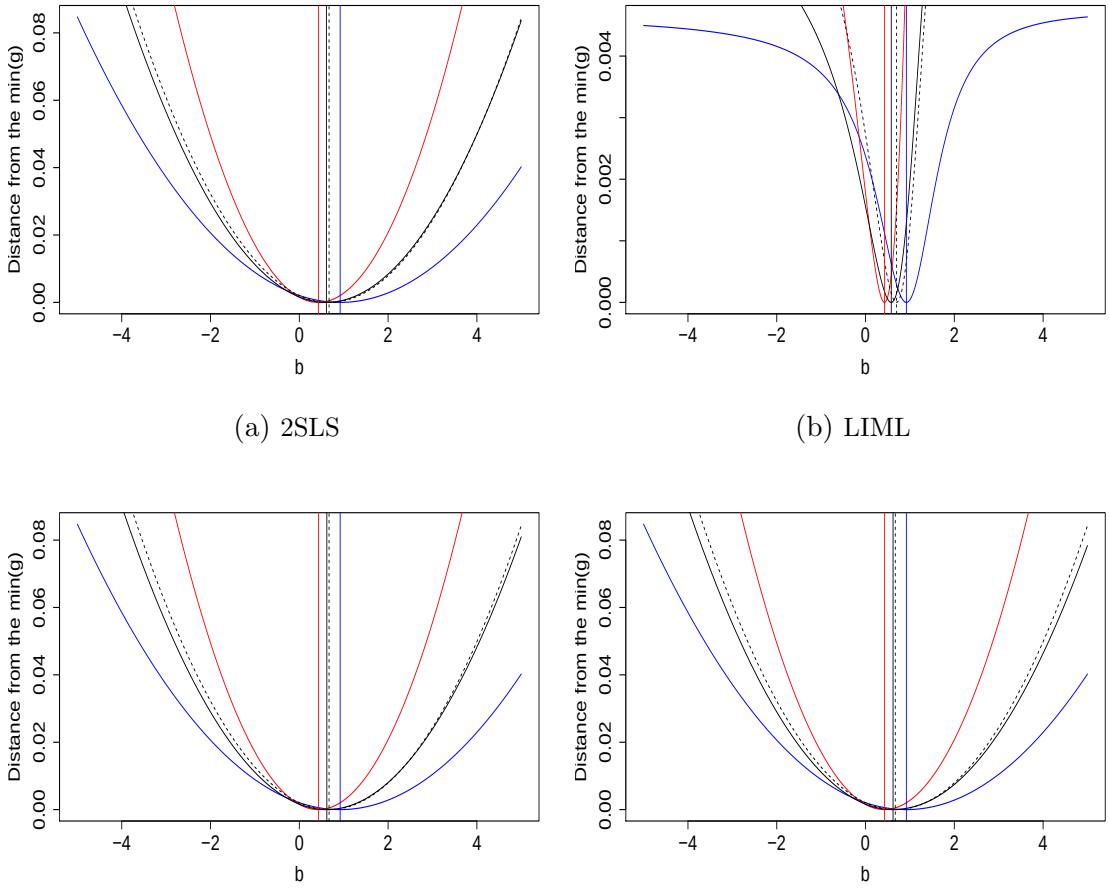
$$Q(\beta) = \frac{g_1 W[1,1] + g_2 W[2,2] + g_{12} W[1,2]}{W[1,1]W[2,2]W[2,1]}$$

Note that the calculation of weighting for each estimator is as described in the previous section.

Results

Figure 3.1 shows all of the algorithms go through different values of β to estimate $g(\beta)$ from Z_1, Z_2, interaction of Z_1 and Z_2 and final $Q(\beta)$, which are the red, blue, dotted and black lines respectively. The weighting matrix in 2SLS is the variance and covariance for each instrument and their interactions. Two-step GMM weighs

each variable based on residual of 2SLS estimate, i.e. weighted by how well 2SLS predicts Y. CUE is an iterative process, it goes through different values of β and re-calculates the weighting matrix base on the residual at every iteration, which is repeated until the estimates converge. The weightings for LIML is exactly the same as 2SLS but the estimate is divided by variance from the β in OLS regression. Figure 3.1 demonstrates the similarity between 2SLS, two-step GMM and CUE as they are based on moments equations. Whereas, Figure 3.1b is slightly different, as Econometricians referred to as a Cauchy-type distribution.



(a) 2SLS

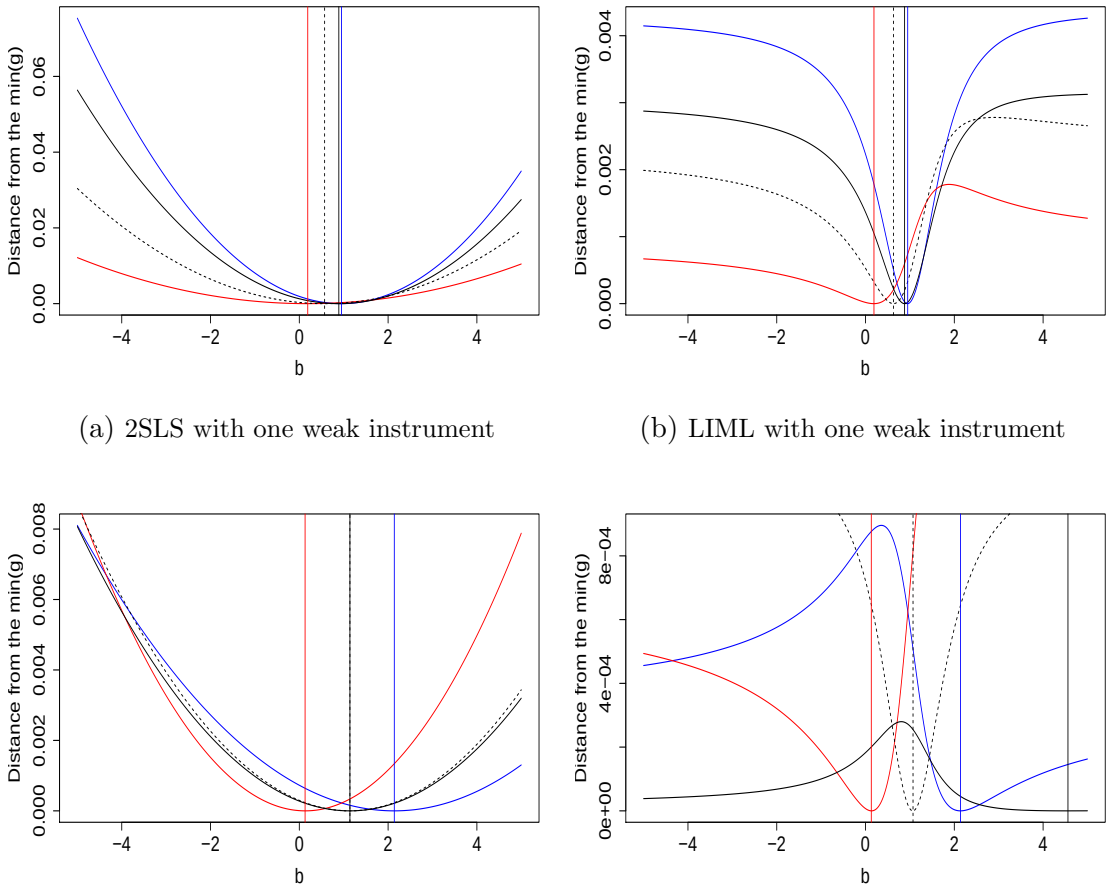
(b) LIML

(c) Two-step GMM

(d) CUE

Figure 3.1: Comparing minimising functions for 2SLS, two-step GMM, CUE and LIML, where Z_1 and Z_2 each explained 2% of variation in X. The y-axis is the distance from the minimum $g(\beta)$ and x-axis range of β. Red, blue, dotted and black lines is $g_1(\beta)$, $g_2(\beta)$, $g_{12}(\beta)$ and final $Q(\beta)$ respectively. The horizontal lines are the minimum $g(\beta)$ for all the steps.

In theory, the LIML algorithm has an advantage in the estimation of the causal effect. As 2SLS with weak instruments is biased towards OLS then the ratio will tend to 1, therefore LIML will minimise its function by ensuring the ratio does not tend to 1. However, there are cases where LIML fails; in the presence of one weak instrument, Figure 3.2b, the 2SLS moves closer to the OLS estimate, the curve of $g_1(\beta)$ for Z_1 (red line) becomes more cubic, this suggests there is a value of β that will give large enough variation in OLS for Q_{LIML} ratio not to be 1. When there is still one strong instrument, the bias in 2SLS is not large enough for LIML to find a β to give large variation in OLS. Figure 3.2d show LIML does fail with weak instruments, notice the Cauchy curve has turned upright for the final Q_{LIML}, the bias is large from both instruments, showing disagreement in causal effect estimate the minimum for $g_1(\beta)$ (red line) is the maximum for $g_2(\beta)$ (blue line). To ensure Q_{LIML} ratio does not reach 1, LIML finds a β that gives large enough variation from the OLS for the bias in 2SLS seem relatively small in comparison.



(a) 2SLS with one weak instrument

(b) LIML with one weak instrument

(c) 2SLS with weak instruments

(d) LIML with weak instruments

Figure 3.2: Comparing minimising functions for 2SLS and LIML, where for "one weak instrument" Z_1 explains 2% of the variation in X and Z_2 explains 0.1% and "weak instruments" is Z_1 and Z_2 each explains 0.1% of variation in X. The y-axis is the distance from the minimum $g(\beta)$ and x-axis range of β. Red, blue, dotted and black lines is $g_1(\beta)$, $g_2(\beta)$, $g_{12}(\beta)$ and final $Q(\beta)$ respectively. The horizontal lines are the minimum Q for all the steps

3.3 Literature review of IV estimators for many weak instruments

The econometrics literature contains extensive research on weak instruments [29, 253] and on many instruments in instrumental variable analysis [24, 60]. The many instruments literature considers instruments that are relatively strong in comparison to the weak instruments literature. In order to narrow down the vast number

of approaches available from both areas, the literature review here will only focus on approaches to "many weak instruments" and find the instrumental variable estimator that is still consistent with many weak instruments. The literature search was performed in Scopus database on May 2015, with keywords as "instrumental variables" and "many weak instruments". There were 24 articles from the search hit, 3 conference papers and 1 paper that did not study the many weak instruments problem were excluded, and 4 papers [44, 85, 227, 248] were based on Mendelian randomisation which will be discussed in the next section.

Chao and Swanson [61] examined the consistency of 2SLS, LIML and the k-class version of 2SLS (B2SLS, commonly known as the biased-corrected 2SLS [206]). Their conclusions were; a robust estimator can improve the precision of point estimation with a large number of weak instruments, and that LIML and B2SLS are still consistent estimators even when the rate of growth in the strength of instruments is slower than that of the number of instruments. Han and Phillips [126] published the limiting distribution for GMM under many weak instruments and conducted a simulation to compare unweighted, two-step GMM and CUE. They concluded that the consistency of GMM estimation is dependent on the sample size, number and strength of the moment conditions (i.e. strength of the instruments) and their simulation demonstrated that CUE gave the less biased and less variable IV estimates compared to unweighted and two-step GMM. In the case of finite samples, Anderson et al. [8] have found LIML to be more attractive than 2SLS, GMM and empirical likelihood (EL) for many weak instruments.

In hypothesis testing, the t-test is the standard test for 2SLS, to check if there is a significant difference between null and the estimated causal effect. However, the t-test lacks power with weak instruments, this reason motivated Anderson and Rubin [7], Kleibergen [168] and Moreira [202] to develop Anderson and Rubin (AR), Lagrange multiplier (LM) and conditional likelihood ratio (CLR) tests respectively. These are alternative testing procedures that are robust to the strength of the instruments. Andrews and Stock [11] found CLR test to be more powerful under many weak instruments than AR or LM tests. Andrews et al. [12] wanted to investigate whether the power from 2SLS, LIML, Fuller (modified version of LIML for weak instruments) and B2SLS would improve on the Wald test, however they found that the power of the CLR test is still superior. The conditioning on Wald and likelihood ratio test was introduced by Moreira [202] as robust test statistics even with weak instruments, where the critical value is dependent on the concentration parameter.

It is well known among econometricians that 2SLS is biased with many instruments, so LIML is commonly used to correct the bias. However LIML's estimated standard error is too small. Hansen et al. [128] derived a corrected standard error (CSE), it is an extension to the method of moments which allows for non-Gaussian distributions. Their results suggest that CSE would still make accurate approximations under many weak instruments.

Instead of assuming the asymptotic distribution of 2SLS is normal, Poskitt and Skeels [229] have derived a multivariate-t approximation. They have provided different scenarios for different concentration coefficients and number of instruments. One particular scenario was very similar to Staiger and Stock [253], where the concentration parameter remains fixed while the number of instruments increases with sample size, they show that the multivariate-t approximation provides a less biased estimate from 2SLS than the normal approximation. However, the authors have cautioned that the investigators need to distinguish a large and a small concentration parameter as the multivariate-t approximation is not appropriate for all cases.

Kapetanios and Marcellino [158, 159] have derived several modified versions of instrumental variables estimators, naming them cross-sectional averaging (CSA) 2SLS, factor 2SLS and factor GMM. CSA is when the first-stage estimates are weighted by 1/number of instruments. Factor estimators reduce a large number of instrumental variables to a smaller number using factor analysis. The authors claim that with many weak instruments, CSA 2SLS performances better than standard 2SLS, but factor 2SLS is worse, and the factor GMM is more effective than the standard GMM. They have also suggested that pre-selection of instruments based on their correlation with the endogenous variable, would further improve the efficiency of their estimators.

LIML, B2SLS and 2SLS are inconsistent estimators under the condition of heteroscedasticity and many weak instruments. Hence the JIV (jackknife instrumental variable) estimator has been suggested as an alternative Chao et al. [62]. Heteroscedasticity arises when the variance of the error term, given the explanatory variable, is not constant. The JIV procedure uses the fitted value from "delete one" observation in the first-stage instead of the fitted value from OLS. Hausman et al. [132] have developed LIML (HLIML) and Fuller versions of JIV (HFUL), where their simulations suggested these estimators' finite properties gave lower bias with narrower confidence interval than LIML, Fuller, JIVE and CUE. Later, Kunitomo [177] have theoretically and numerically proved that their estimation in finite sam-

ple distribution is more accurate HLIML, their method is called asymptotic optimal modification (AOM) LIML estimator. As the name suggests it modifies the asymptotic properties of LIML estimation for many weak instruments and heteroscedasticity. In order to reduce the correlation between instrument and error term, Hansen and Kozbur [127] applied ridge regression in the first-stage to shrink the coefficient of the instruments and the jackknife algorithm. In a many instruments setting, this method is able to effectively capture enough instruments to give an unbiased estimate with a smaller variance, in comparison to 2SLS and JIVE. For cases where heteroskedasticity is unknown, a symmetrical jackknife estimator can be used, it is a jackknife version of GMM developed by Bekker and Crudu [25].

3.3.1 Conclusion from the review

From the many approaches in econometrics, LIML has been identified as the most efficient estimator [61] with many weak instruments, even though there are multiple extensions. Greene [116] summarised two reasons that the maximum likelihood approach is more efficient than the least squares approach with many weak instruments; the finite sample distribution of LIML means that the estimation is less sensitive to weak instruments and LIML is invariant to normalisation of the equation.

The tail behaviour of finite sample distribution of $\hat{\beta}_{LIML}$ does not depend on the degree of over-identification (d=k-m+1, where k is number of instruments and m is number of endogenous variables), $\hat{\beta}_{LIML}$ has Cauchy-type tails (see Figure 3.1b) and has no finite moments. Therefore, the finite sample density of $\hat{\beta}_{LIML}$ is much less sensitive to the addition of superfluous instruments. Note, that if strong instruments do not give the true causal effect estimate then weaker instruments would not be able to correct the estimation.

The invariance property, is a mathematical result, which enables LIML to substitute and exclude parameters out of the equation to then only maximise the likelihood of β. For example, a parameter can appear in the likelihood function as $1/\theta_j$, to simplify, the model can be re-parametrised in terms of $\gamma_j = 1/\theta_j$. See Greene [116] for more detailed explanation.

However, LIML assumes the errors to have a normal distribution. Therefore when the condition of homoskedasticity is not satisfied, LIML is a biased estimator of the causal effect [207]. In contrast, the least squares method does not assume any distribution on the error term. In addition, LIML has been proved to be a biased estimator of the causal effect with weak instruments [124].

3.4 Literature review of Mendelian randomisation with multiple instruments in individual-level data

Multiple instruments have been proposed as a solution to weak instrument bias and low power caused by the weak association between genotype and exposure [220, 227]. Lawlor et al. [182] and Pierce et al. [227] stated that for multiple instruments and continuous outcomes, two-stage least squares (2SLS) is one of the simplest and popular algorithms used to estimate the causal effects. This is evident from most of the existing simulation studies designed for multiple instruments [43, 47, 48, 220, 227] and a recent review of Mendelian randomisation studies [26]. As the name suggests, there are two stages to this method; first stage, regress X on the instrument(s). Second stage, regress Y on the predicted X from the first stage. If the three assumptions (Section 2.3), then predicting X from the genetic instruments is free from the effects of the confounder and consequently the variation in the second stage, Y is only explained by the unconfounded X.

One of the major limitations to using multiple instruments in 2SLS is many weak instruments bias. The first example of this problem was found in a study of effects of length of schooling on wages. Angrist and Krueger [13] used with season of birth as an instrument to begin with. However this instrument only explained 0.012% of the variation in length of school. Therefore they included the interaction of length of schooling and year of birth, and the 180 instruments explained 0.043% of the variation and standard error was decreased by a half. Bound et al. [29] and Hansen et al. [128] have found including multiple instruments decreases the standard error, but at the price of bias, as the 2SLS estimator is biased towards the ordinary least squares (OLS) estimate.

To decrease the weak instrument bias in 2SLS, Pierce et al. [227] have proposed combining the allele scores of the multiple genetic instruments into a single instrument, where the weights are determined by the magnitude of association between the genetic variants and X. However, the weights of the allele score must be pre-determined before analysing the current dataset and the estimate can still be severely biased if the variants included in the allele scores are not valid instruments [49]. The allele score could be unweighted, but then an unweighted allele score will have lower power than estimating with multiple instruments [220].

The limited information maximum likelihood (LIML) [22], the generalised method of moments (GMM) [83] and the continuously updating estimator (CUE) [128] have been suggested by Davies et al. [85] as alternative algorithms to avoid many weak instruments bias. Burgess et al. [53] have also published a review of these estimators. Burgess et al. [48] suggested a modified version of LIML for multiple instruments [124]. Burgess et al. [47] designed a Bayesian approach to Mendelian randomisation with multiple instruments, which claims to be similar to 2SLS, but offers more flexibility in model assumption, e.g. not limited to assumptions of linearity and normality. For the situation when the instrument is weak or the sample size is small, Jones et al. [153] have devised a Bayesian approach to Mendelian randomisation with the use of an informative prior, where the method incorporates current knowledge, see Chapter 8.

3.5 Conclusion

This chapter has given a description and reviews of the instrumental variable estimators for many weak instruments. Even though 2SLS is the most popular method used in Mendelian randomisation studies [26], 2SLS is severely biased with many weak instruments. The econometrics and Mendelian randomisation literatures both agree that LIML is the most efficient estimator with many weak instruments. The open question is whether the same conclusion can be reached with many dependent instruments. The answer to this question will be delayed to Chapter 7. The next chapter will first describe the simulation method for this thesis.

Chapter 4

The Simulation of Mendelian Randomisation with Dependent Instruments

4.1 Introduction

Before beginning with the simulation studies for Mendelian randomisation with many dependent instruments, this chapter will describe the simulation methods and issues to consider when designing simulation studies. A dataset for Mendelian randomisation usually consists of the genotypes of single-nucleotide polymorphism (SNPs) as potential instruments, a risk factor (X) and an outcome of interest (Y). In this thesis, I am interested in assessing the instruments that are correlated with each other. In genetic terms, the SNPs are in linkage disequilibrium (LD) with each other. Furthermore, I will concentrate on the case where only one or two of the SNP(s) is the functional variant(s) for X, thus the other SNPs are associated with X through their correlation with the functional variant(s).

In order to monitor the effect of different LD patterns and minor allele frequency (MAF) on the performance of the estimators, the first section will give the simulation method I have designed to control the SNPs' correlation with the functional variant(s) and MAF. For more random LD patterns and MAF to reflect realistic genome, the implementation of a human genome simulator will be introduced. The next section will give the procedure of simulating the relationship between the functional variant(s), X and Y. These simulation methods may have some minor changes depending on the scenarios in the following chapters, which will be reported in the

relevant design section.

A list of evaluation criteria will also be given, that are to be utilised as tools to assess the performance of the estimators. Finally, to ensure the efficiency of the simulation study, the number of simulations required to provide sufficient accuracy level from each of the evaluation criteria will be calculated. The consideration of these issues for designing simulation studies were based on the guidelines by Burton et al. [56].

4.2 The Genotypes

This section will first describe my method of simulating SNPs in order to control their minor allele frequency (MAF) and the linkage disequilibrium (LD) between them. The LD of the two SNPs is measured by the square of Pearson correlation coefficient (r^2) [137]. The second part will describe a human genome simulator which simulates more realistic LD patterns in comparison to my algorithm.

4.2.1 Genotype of two SNPs

For the haplotypes of each SNP, the presence of the minor allele will be coded as 0 and 1. 10,000 haplotypes will be simulated to generate the genotype. The genotype of each individual is the combination of a pair of haplotypes, i.e. the parent haplotypes. Hence, an individual's SNP genotype will be coded 0,1 or 2 for the presence of the minor allele. The following steps will be taken to simulate the genotypes of 2 SNPs for n individuals;

1. For SNP_c and SNP_1 to be in LD for the desired r^2, MAF of SNP_c and SNP_1, f_c and f_1 respectively. F, the proportion of haplotypes which SNP_{ci} and SNP_{1i} both equal 1, where $i = 1 \ldots 10,000$, can be derived from below;

$$r^2 = \frac{(F - f_c f_1)^2}{f_c(1 - f_c)f_1(1 - f_1)} \tag{4.1}$$

   Note, when $r^2 = 0$ means there is no correlation between the two SNPs and $r^2 = 1$ is perfect correlation (i.e. the two SNPs are completely identical).

2. With f_1, f_c and F, the coding of 0 and 1 in SNP_c and SNP_1 will follow the proportions in the table below;

		SNP_1		
		0	1	
SNP_c	0	$1-f_1-f_c+F$	f_1-F	$1-f_c$
	1	f_c-F	F	f_c
		$1-f_1$	f_1	

There is $1-f_1-f_c+F$ proportion of haplotypes where $SNP_{ci}=0$ and $SNP_{1i}=0$ at the ith sample and F proportion of haplotypes in which $SNP_{ci}=1$ and $SNP_{1i}=1$. There is a constraint on F; $F < min(f_c, f_1)$, since different combinations of MAFs and r^2 are not always possible. For example if $f_c=0.5$, $f_1=0.1$ and the desired r^2 between the SNPs is 0.9, then, from Equation 4.1 F is equal to 0.19, which is not possible for SNP_1, as it only has 0.1 proportion of 1s and F requires both SNP_1 and SNP_c to equal to 1 at the same haplotype positions.

3. One randomly selected set of haplotypes of 2 SNPs is added to another set, to make the genotypes of 2 SNPs for an individual. In other words for 1 SNP, if the pair of haplotypes both coded 1 than the genotype of an individual will be coded 2 and if pair of haplotypes are both 0 than an individual's genotype will be 0.

4. Step 5 is repeated until there are genotypes of 2 SNPs for n individuals.

4.2.2 Genotype of multiple SNPs

The simulation above is only capable of generating two SNPs at a time, thus a more efficient method to simulate multiple SNPs in LD is required. As we can not be certain of the identity of the functional variant, the following simulations will examine the situation where the functional variant was not measured and for a SNP to be unmeasured it is usually because it is rare within the population.

A Scopus search at January 2015 for papers describing studies that had used simulation was based on the keywords "genome-wide association", "linkage disequilibrium", "simulation" and "rare variants" and it identifies 78 studies. 67 papers were excluded as they did not have a description of their simulation method for genotypes. 8 out of 11 studies cited either (or both) Basu and Pan [21] and Wang and Elston [284]. There were 3 studies with original simulation methods for multiple

SNPs in LD [21, 183, 285]. They have all used a multivariate normal distribution to generate correlated latent variables and transform them into haplotypes of SNPs in LD by applying the specified MAF of each SNP. This procedure will be adopted in my simulation of multiple SNPs in LD.

Now I will demonstrate how to simulate different genetic patterns for multiple SNPs by specifying LD with the functional variant(s), SNP_c. The functional variant is defined as having a direct effect on exposure (X), which is described in Section 4.3. The suggestion of multivariate normal distribution for all the SNPs from Wang and Elston [284] requires values in the correlation matrix to be positive definite, this can be difficult for specific LD pattern. In other words, it would be difficult to have the correlations between the SNPs and causal SNP set at specific values that also give positive definite values for the correlation between non-causal SNPs. Apart from the use of the multivariate distribution, the rest of the simulation method will follow Wang and Elston [284].

Patterns I, II and III will assume there is only one functional variant and its genetic position is in the middle of all the SNPs. Pattern IV will have two causal SNPs, but these are also positioned in the middle of their region of SNPs. In order to control the MAF and correlation structure between SNPs, the LD patterns are simplified here and may not be representative of the real data. However, realistic LD patterns will be produce by utilising GENOME simulator [185], the detailed description will be postponed to the next section. Finally, these SNPs are coded into 0, 1 and 2 for the presence of the minor allele according to their MAF.

Before the description of the simulation of multiple genotypes, first several parameters must be defined, apart from which genetic pattern to produce;

k is the number of instruments,

n is the number of individuals,

$\boldsymbol{\rho_{max}}$ is the maximum correlation (or LD) with SNP_c a SNP can have,

f is the MAF of the SNPs.

It is important to note, Z_i and Z_c in "Specifying linkage disequilibrium" are latent variables and will be changed into the genotypes of SNP_i and SNP_c respectively in the following section "Specifying minor allele frequency", where $i = 1 \dots k$.

Specifying linkage disequilibrium

1. Randomly generate n values of Z_c from a Normal distribution $\sim (0, 1)$.

2. The following equation derives the correlation (ρ) between a SNP and the causal SNP, ρ of a SNP will be determined by its proximity to the causal SNP, i.e. the further away they are the weaker their correlation with the causal SNP will be;

$$\rho_i = \rho_{max}(1 - 2\frac{|i - c|}{k}) \tag{4.2}$$

   where position $i = 1 \dots k$ and c is the position of the causal SNP among k SNPs.

   Four different correlation patterns will be simulated for most of the chapters; Pattern I, III and IV will have ρ_{max} of 0.9 and Pattern II have 0.5. Pattern III is designed to produce a haplotype block within the region, hence Equation 4.2 is conditioned if $\rho_i > 0.5$ then ρ_i will be modified to ρ_{max}. Pattern IV splits the SNPs into two group $(k/2)$ for two causal SNPs. The SNPs in each group obtain their correlations through Equation 4.2 according to the position of their causal SNP, $c1$ and $c2$.

3. Simulate Z_i using ρ_i and Z_c

$$Z_i = \frac{\rho_i Z_c + (1 - \rho_i)\varepsilon}{\sqrt{(\rho_i^2 + (1 - \rho_i)^2)}} \tag{4.3}$$

   where ε is the residue of SNP_i after the correlation with SNP_c, in which there are n number of ε with $N(0, 1)$. See Appendix A.1 for the derivation of Equation 4.3.

Specifying minor allele frequency

1. Sort the values of Z_c. Find Z_{cj} at position $(1-f_c)^2$ and f_c^2, where $j = 1, \dots, n$. Labelling these values as a and b accordingly.

2. Then the genotype of SNP_{cj} is coded as 0 when $Z_{cj} < a$, 1 when $a < Z_{cj} < b$ and 2 when $Z_{cj} > b$.

3. Repeat step 1 and 2 to obtain the genotype of k number of SNP_i from Z_i.

4.2.3 GENOME

The previous simulation methods for SNP genotypes, the amount of LD was dependent on their physical proximity from the causal SNP. However, the pattern of LD is also affected by other factors, such as natural selection, genetic drift, mutation and population subdivision [251]. Thus, a simulator for more realistic LD pattern between the SNPs will be required.

In January 2015 *Genetic Epidemiology* published a special issue on the simulation of genetic data, which included an overview of genetic data simulation by Peng et al. [226]. This review summarised 93 simulators and described a website, called the genetic simulation resource (GSR). The website tailors individual requirements by including different options and compares the resulting simulators.

The aim was to simulate haplotypes of the SNPs to monitor their MAF and LD, therefore I have selected the following options in the GSR website;

- Target: Type of simulated data: Haploid DNA sequence
- Output: Data type: Linkage disequilibrium
- Evolutionary Features: Recombination: Varying recombination rates
- Interface: Command-line and script based
- Development: Tested Platforms: Window
- Development: Language: R

As a result of these options, 3 simulators were suggested; GENOME, SimuPop and QMsim. GENOME [185] was chosen as the most suitable simulator, its LD was scaled by the genetic distance which is similar to my simulation method and user friendly. SimuPop [225] requires implementation in Python. QMsim [245] was designed to imitate genetic data from livestock, whereas the objective of my simulation study is to focus on the human genome.

GENOME [185] applies the coalescent-based approach to simulate genome data, where the approach follows the Wright-Fisher neutral model. The Wright-Fisher model is a stochastic evolutionary model where it takes account of mutation, selection, geographical factors, changes in population size, and so on [99]. The standard coalescent-based approach simulates genealogical events backward in time [164], thus GENOME is a backward-time simulator. The simulator starts off with a sample of sequences and simulated generations backwards until one common ancestor is

reached. The backward algorithm is conditional on population size, recombination rates and migration between subpopulations. Since GENOME simulates every generation, it is able to incorporate features such as multiple populations, population stratification and geographic relationships between subpopulations. The inclusion of these parameters is an advantage, as these parameters are only usually seen in forward-time simulators, which are computationally less efficient than backward-time simulators.

Implementation of GENOME simulator

This subsection begins with the options in GENOME that is relevant to my simulation aim, then the steps of generating the genotype data from GENOME;

-pop gives the number of sub-populations and size of sub-samples, -pop 1 10000 extracts 10,000 haplotypes from one population.

-pieces specifies the number of fragments per independent region (chromosome). This controls the number of recombinations that occur within a region.

-len specifies the length in base pair(bp) per fragment. The product of -len and -pieces gives the total length in bp of a whole independent region.

-s fixes the number of SNPs per independent region.

The default of GENOME is to simulate one independent region, which represents one chromosome and therefore is simulated independently from one region to another. It is important to note GENOME does allow for options to control MAF and the proportion of SNPs with specified MAF but this only controls the data output and does not have an effect on the mechanism of the simulator. For detailed instructions and other options in GENOME, see Liang et al. [185].

The following steps are to create haplotype of SNPs from an averaged length protein-coding gene (53.6×10^3bp) with 200 SNPs and 5 recombination points; an average length gene should have approximately 200 SNPs [258]. Recombination usually occurs between haplotype blocks and the average length of a haplotype block is 1000bp for European population [108]. Then the GENOME simulated haplotypes will be transformed into genotypes;

1. The GENOME command-line for simulating 10,000 haplotypes of 200 SNPs from an average length gene;

genome -pop 1 10000 -pieces 5 -s 200 -len 10000

The combination of -pieces 5 and -len 10000 induces GENOME to simulate gene length of 5×10^4bp.

2. GENOME codes presence of an allele as 0 and 1, the MAF of each SNP is therefore the mean of the 10,000 haplotypes.

3. The pearson correlation coefficients (r^2) [137] between the SNPs are obtained;

$$r^2 = \frac{(F - f_i f_j)^2}{f_i(1 - f_i)f_j(1 - f_j)} \tag{4.4}$$

where f_i and f_j are the MAF of ith and jth SNP. F is the proportion of haplotypes of the ith and jth SNP both having 1.

4. 2 out of the 10,000 haplotypes are randomly selected and combined to make the genotype of 200 SNPs for one subject.

5. Step 4 is repeated until there are genotypes of 200 SNPs for n subjects, where n is number of individuals.

4.3 The Exposure and Outcome of Interest

The aim will be to simulate realistic parameter values for exposure (X) and outcome of interest (Y). I have taken as my example the relationship between birth weight and fasting glucose as a simulation model. In 1991, Hales et al. [125] have found type II diabetes in adulthood is associated with low birth weight, and have hypothesized that impaired development of β cell function is a consequence of intrauterine malnutrition which causes type II diabetes in the individual when their diet becomes nutritionally rich in later life. However, recent meta-analysis of GWAS have proven 'the fetal insulin hypothesis' [145], where the hypothesis states that the association between low birth weight and the increased risk of type II diabetes are mediated by genetic factors [131]. The inspiration came from the most recently published GWAS 2013 [143], where most of the significant SNPs explained up to 2% of the variation in birth weight. The population distribution of birth weight and fasting glucose were obtained from the summary statistics of the Office for National Statistics and the Whitehall II cohort study [97] respectively.

4.3.1 One causal variant

Each simulated dataset consists of the genotype of a causal SNP (SNP_c), risk factor (X), disease outcome (Y), and unmeasured confounding (U). The causal SNP will explain 2% of the variation in X. X will explain 6% of the variation in Y. The X and Y will have distribution of $N(3.3, 0.59^2)$ and $N(5.47, 1.32^2)$ respectively. U was drawn randomly from distribution $N(0, 1)$, this will be the unmeasured confounding affecting both X and Y. The following equation describes the relationship between SNP_c, X, Y and U;

$$X_i = \alpha_0 + \alpha_1 SNP_{ci} + \alpha_2 U_i + \varepsilon_{xi} \tag{4.5}$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 U_i + \varepsilon_{yi} \tag{4.6}$$

where ε_x and ε_y are independent random errors with distributions of $N(0, 1)$, $i = 1, \ldots, n$ and n is number of individuals.

The βs and αs were derived empirically through the definition of variance, since we know the variance explained (R^2) and the variance of SNP_c, X and Y. The variance of SNP_c is simply $2f_c(1 - f_c)$, where f_c is the MAF of SNP_c. For example;

$$\alpha_1^2 = R_c^2 \frac{Var(X)}{Var(SNP_c)} \tag{4.7}$$

R_c^2 is 2%, as SNP_c explains 2% of the variation in X. The remaining 98% from X will be split equally between U and ε and their coefficient is calculated in a similar fashion as for SNP_c. The coefficient for the linear regression of Y will have the same derivation as X, where the derivation will start with X explaining 6% of the variation in Y;

$$\beta_1^2 = R_x^2 \frac{Var(Y)}{Var(X)} \tag{4.8}$$

The true coefficient of β_1 will differ between chapters, however this will not affect the estimation of causal effect, as the variation in Y explained by X will remain the same [227].

4.3.2 Two causal variants

For the simulation of two causal SNPs, SNP_{c1} and SNP_{c2}, known as Pattern IV in Section 4.2, is similar to the method above but with slight changes to the equation

of X;

$$X_i = \alpha_0 + \alpha_1 SNP_{c1i} + \alpha_2 SNP_{c2i} + \alpha_3 U_i + \varepsilon_{xi} \tag{4.9}$$

where each causal SNP will explain 1% of the variation in X. The derivation of α coefficients will be the same as above.

4.4 Evaluation Criteria

The purpose of these simulations is to evaluate the performance and the precision of estimators in different scenarios. The causal effect estimate ($\hat{\beta}_1$), its standard error and p-value will be stored after each replication of the simulated data. The summary statistics and evaluation criteria will be calculated from B simulations. The summary statistics will include mean and median of $\hat{\beta}_1$ ($\bar{\hat{\beta}}_1$ and $\tilde{\hat{\beta}}_1$), standard deviation of $\hat{\beta}_1$ (σ) and standard error of $\hat{\beta}_1$ ($\sigma/\sqrt{B}$). As recommended by Burton et al. [56], the following evaluation criteria will be monitored;

Bias

Bias will be derived from the difference between the mean estimates and the true value β_1, $\bar{\hat{\beta}}_1 - \beta_1$. The median bias will also calculated.

Root Mean Squared Error

Root Mean Squared Error (RMSE) measures the overall accuracy by merging bias and variability together; $\sqrt{\frac{1}{B}\sum_{i=1}^{B}(\hat{\beta}_{1i} - \beta_1)^2}$.

Coverage

Coverage is the percentage of simulations that had the true β_1 within the confidence interval of $\hat{\beta}_1$. The coverage should reach approximately 95% (the nominal coverage level).

Power and Type I error

Type I error is the proportion of simulations that rejected the null hypothesis of no effect (p-value < 0.05), when the null hypothesis is true. Power is the probability of not committing type II error (failure to reject the false null hypothesis) and the target power is often taken to be 0.8. Power will be calculated by the proportion

of $\hat{\beta}_1$ that had a p-value$<$ 0.05. To ensure the validity of 0.05 critical threshold, 5% significance level, the empirical type I error will be obtained from a separate simulation.

Outliers

The causal effect estimates from limited information maximum likelihood (LIML) and continuously updating estimator (CUE) are well known to have occasional extreme outliers in the presence of many weak instruments [61, 207]. However, monitoring outliers was not within the guidelines of Burton et al. [56]. Thus to compare the performance of LIML and CUE without the effect of outliers, bias and RMSE will be winsorised and the percentage of outliers will also be reported.

Winsorisation is a transformation that reduces the effect of possibly spurious outliers by limiting extreme values in the data [289]. Winsorisation

1. orders the $\hat{\beta}_1$ from B number of simulations,
2. finds the 20th and 80th value within the ordered list, and
3. Any $\hat{\beta}_1$ that is smaller than the 20th will be set equal to the 20th value and $\hat{\beta}_1$ greater than the 80th value will changed to the 80th value.

The 60% threshold has been advised by Wilcox [289] in their review of the different thresholds.

The percentage of outliers, the definition of an outlier will be the same as for box plots, i.e. $\hat{\beta}_1$ will be labelled as an outlier when it is smaller or greater than median $\mp 1.5 \times IQR$, where the interquartile range (IQR) is the 75th minus the 25th value.

4.5 Number of Simulations

The number of simulations will be selected to ensure the stability of the estimate of interest, while reducing unnecessary computational time. These accuracy level of bias, RMSE, coverage and power will be calculated from the following formula given by Burton et al. [56];

$$\delta = \frac{Z_{1-(\alpha/2)}\sigma}{\sqrt{B}} \tag{4.10}$$

where δ is the specified difference from the true value. B is the number of simulations, σ^2 is the variance for the estimate of interest and $Z_{1-(\alpha/2)}$ is the $1-(\alpha/2)$ quantile of the normal distribution.

Bias

As the true bias and variance were unknown, a small simulation study was performed; the causal effects of 20 datasets were estimated from two-stage least squares (2SLS) with SNP_c as the instrument. Each dataset had the genotype of SNP_c, X and Y for 2,000 individuals, simulation method is as described in Section 4.2 and 4.3. The mean bias and standard deviation from this small simulation study was -0.0227 and 0.0007 respectively. Then, using Equation 4.10, 10,000 simulations will produce a δ of 0.0001, which is within 0.6% accuracy of the bias of -0.0227.

RMSE

The same simulation study as above returned standard deviation of the RMSE of 0.0064 and a mean of 0.2280. 10,000 simulations gave δ of 0.0001, which is within 0.06% accuracy of the RMSE of 0.0064.

Coverage

The nominal coverage is 0.95 and its empirical standard deviation would be 0.0022 (by the definition of standard deviation for proportions). For nominal coverage, its δ equals to 0.0043 from 10,000 simulations. i.e. 0.45% accuracy for 95% coverage.

Power

For the desirable power of 0.8, also a proportion, gave empirical standard deviation of 0.0040. Then through Equation 4.10, the δ from 10,000 simulations is 0.0078.

10,000 simulations will give bias, RMSE, coverage and power, the accuracy level of approximately 0.6%, 0.1%, 0.5% and 1% respectively. The accuracy level for 10,000 simulations seems high, however as mentioned before LIML and CUE are prone to outliers with many weak instruments, which will reduce the accuracy levels of bias and RMSE. Hence, each simulation in the following chapters will be repeated 10,000 to obtain the desirable accuracy from each of the evaluation criteria.

Chapter 5

One and Two SNPs Dependent on the Causal SNP in Two-Stage Least Squares

5.1 Introduction

So far, in the Mendelian randomisation literature, there has been no information on how much variation a non-functional variant can explain in the exposure and on its impact on the estimation of the causal effect. Therefore, this chapter aims to answer this question using two-stage least squares (2SLS), to act as a foundation for the use of many dependent instruments in Mendelian randomisation. 2SLS will be implemented in this primary investigation, as it is the most popular and the simplest method used among all of the algorithms for instrumental variable analysis for continuous outcome [84].

The chapter begins with the scenario where the causal SNP, SNP_c is known and measured and examines the effect of sample size and variation explained by SNP_c on the performance of 2SLS. The second scenario considers the case where the causal SNP is unknown or unmeasured, and a single non-causal SNP, SNP_1 whose association with X is driven by the correlation with SNP_c is measured. This scenario evaluates 2SLS with changes in the correlation between SNP_1 and SNP_c. The next scenario investigates whether there are any gains by including a non-causal SNP as an instrument if SNP_c is measured. The chapter ends with investigating the effect of having two non-causal SNPs as instruments in 2SLS.

5.2 Single causal SNP: Variation explained and sample size

5.2.1 Aims

Section 2.5.3 explained that the weak association between gene and exposure of interest causes weak instrument bias and low power in detecting a causal effect. This section will demonstrate the impact of genetic variation explained and sample size on these statistical properties. Figure 5.1 gives the graphical representation of the relationship between causal SNP (SNP_c), risk factor (X), disease outcome (Y) and unknown confounder (U).

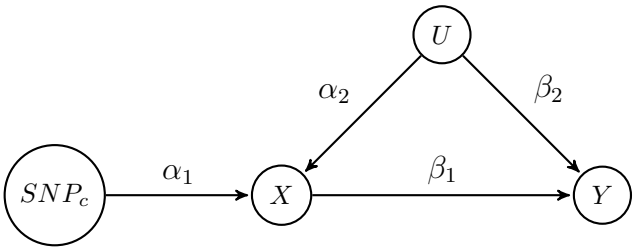


Figure 5.1: Illustration of Mendelian randomisation with one causal SNP

5.2.2 Design

Adopting the method of simulating SNP_c from section 4.2.1, the minor allele frequency (MAF) for SNP_c will be 0.25 (note that the simulation method was designed so that MAF will not effect the variation explained by SNP_c). The variance in X explained by SNP_c, ranged from 0.5% to 3%. X will explain 6% of the variation in Y, simulated as in Section 4.3. Sample sizes of 10,000 and 15,000 will be incorporated into the simulations. SNP_c with different variations explained will be introduced into the 2SLS as a single instrument.

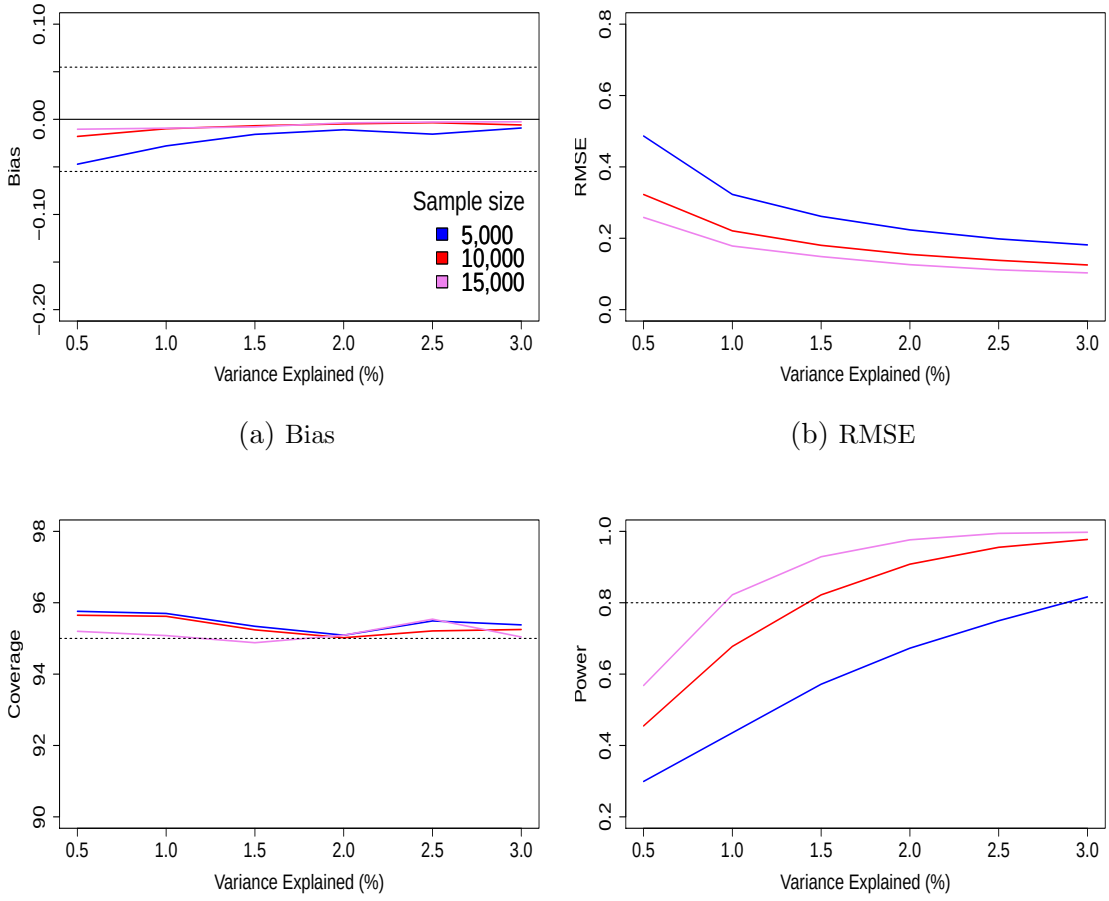
As reasoned in Section 4.5, simulations will be repeated 10,000 times. Bias, RMSE, coverage and power will be monitored to evaluate the performance of 2SLS. Type I error will also be part of the evaluation criteria but derived from a separate simulation where X does not directly influence Y. Winsorisation will not be used throughout this chapter as 2SLS does not give extreme outliers, unlike limited information maximum likelihood (LIML) and continuously updating estimator (CUE). For more detailed explanation see Section 4.4.

5.2.3 Results

Figure 5.2a to 5.2d shows each evaluation criterion plotted against the variation explained by SNP_c. The general pattern for all of the evaluation criteria indicate an improvement on the performance of 2SLS as variation explained and sample size increases; the bias decreases as variation explained increases. The bias for all the variance explained declines towards zero as sample size increases. RMSE reduces as variation explained rises. A sample size of 5,000 had the greatest RMSE for all of the variation explained, compared to other sample sizes. As the sample size and variation explained increases, the standard error of the simulations decreases (Appendix Table B.1). Furthermore, Figure 5.2a and Figure 5.2b illustrated, that the variance explained does not affect the variation between the estimates. The bias (Figure 5.2a) from the three sample sizes will eventually be the same, as variance explained increases. Whereas, the RMSE (Figure 5.2b) remains separate between the three sample sizes, even with increasing variance explained.

From looking at Figure 5.2c, it is difficult to determine which sample size or variation explained has the optimum coverage, in the presence of overlap between the different sample sizes. However, the coverage for variation explained are all approximately around the nominal level. When considering the sample sizes, 15,000 had the highest power for all variation explained. Power also increases as variance explained increases. The critical threshold of 0.05 was valid for power, shown by Appendix Figure B.1, as the type I error for all the sample size and variance explained were approximately 5% (represented by the dotted line).

To illustrate the point made by Staiger and Stock [253] about the F-statistics, Appendix Figure B.2a to B.2e show bias, RMSE, coverage, power and type I error respectively. The performance of 2SLS benefited from the increase of the F-statistics; all of the evaluation criteria reaches its optimum level.



(a) Bias

(b) RMSE

(c) Coverage

(d) Power

Figure 5.2: Evaluation criteria of 2SLS against variance explained by SNP_c with different sample size. The different coloured lines represents varies sample sizes; blue, red, violet represents 5,000, 10,000 and 15,000 individuals within a simulation. The dotted and solid lines in Figure 5.2a is the 10% and zero bias respectively. The nominal 95% coverage level is the dotted line in Figure 5.2c. The dotted line in Figure 5.2d is the 0.8 nominal power.

5.2.4 Conclusion

2SLS benefited from increasing variation explained, as the α_1 is further away from zero. When variation explained is small, 2SLS becomes more precise with increasing sample size, as the variation for the estimates of β_1 becomes smaller. However the variation only decreases with greater sample size, and not with variation explained.

2SLS have been known to suffer from weak instrument bias [29, 39], i.e. F-statistics can influence the amount of bias, as reflected in the results; as the F-

statistic increases, the bias decreases (Figure B.2a). The overlaying of sample sizes in Figure B.2a showed that bias is balanced by both sample size and variation explained; the bias from small variation explained and large sample size is equivalent to bias from large variation explained and small sample size. F-statistics of greater than 10, gave estimates of less than 10% bias (Figure B.2a), which supported Staiger and Stock [253]'s guideline of having F-statistics greater than 10, to avoid weak instrument bias. However, Palmer et al. [220] have found cases where this guideline does not always apply.

After understanding the mechanics of 2SLS with SNP_c, now in view of the rarity of cases where the causal SNP is known, the next scenario will consider SNPs that are not causal but are associated with the causal SNP through linkage disequilibrium.

5.3 Single non-causal SNP in linkage disequilibrium

The aim of this section is to understand the effect on 2SLS using a single SNP (SNP_1) with different levels of linkage disequilibrium (LD) from the causal SNP (SNP_c). LD is commonly measured by the square of Pearson's correlation coefficient (r^2). The correlation allows SNP_1 to be associated with X without being causal itself. The effect of variation explained by SNP_c will also be examined.

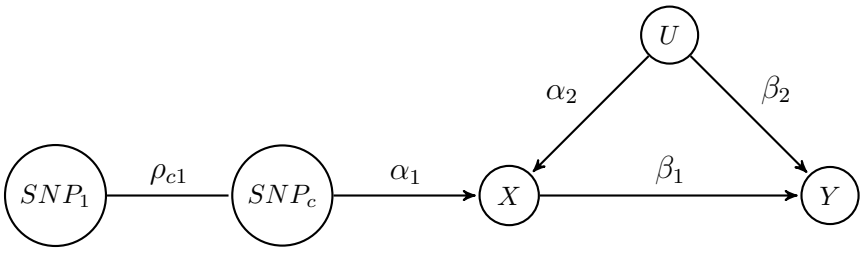


Figure 5.3: Illustration of Mendelian randomisation with causal (SNP_c) and non-causal SNP (SNP_1). Solid line without and with arrow represent correlation and causal relationship respectively.

5.3.1 Mathematics of the variance explained by non-causal SNP

Figure 5.3 show the relationship between SNP_c, SNP_1, X, Y and unknown confounder (U). Notice the solid line without arrow between SNP_c and SNP_1, this is

to emphasis their relationship is correlation not causal.

Regress X on SNP_c

$$X = \alpha_0 + \alpha_1 SNP_c + \varepsilon \tag{5.1}$$

where α_0 is the intercept, ε is the random error and by the definition of the simple linear regression coefficient, the expectation of α_1 is,

$$E(\alpha_1) = \rho_{xc}\frac{\sigma_x}{\sigma_c} \tag{5.2}$$

where ρ_{xc} is the correlation between X and SNP_c, σ_x is the standard deviation of X and σ_c is the standard deviation of SNP_c.

By the definition of variance, the variance in X explained by SNP_c is,

$$\begin{aligned} Var.Explained &= \alpha_1 Var(SNP_c) \\ &= \alpha_1^2 \sigma_c^2 \end{aligned}$$

Substitute 5.2 into variance explained, to yield the expected variance explained;

$$\begin{aligned} E(Var.Explained) &= \rho_{xc}^2 \frac{\sigma_x^2}{\sigma_c^2}\sigma_c^2 \\ &= \rho_{xc}^2 \sigma_x^2 \end{aligned} \tag{5.3}$$

Then, the expected percentage of variance explained by SNP_c is;

$$\begin{aligned} E(\%Explained) &= 100\rho_{xc}^2 \frac{\sigma_x^2}{\sigma_x^2} \\ &= 100\rho_{xc}^2 \end{aligned} \tag{5.4}$$

Regress SNP_c on SNP_1

$$SNP_c = \alpha_0' + \alpha_1' SNP_1 + \varepsilon' \tag{5.5}$$

By the definition of simple linear regression coefficient,

$$E(\alpha_1') = \rho_{c1}\frac{\sigma_c}{\sigma_1} \tag{5.6}$$

where ρ_{c1} is the correlation between SNP_1 and SNP_c, σ_c is the standard deviation

of SNP_c and σ_1 is the standard deviation of SNP_1.

Replace SNP_c in 5.1 by 5.5,

$$\begin{aligned} X &= \alpha_0 + \alpha_1(\alpha_0' + \alpha_1' SNP_1 + \varepsilon') + \varepsilon \\ &= \alpha_0'' + \alpha_1\alpha_1' SNP_1 + \varepsilon'' \end{aligned}$$

Then, the expectation of $\alpha\alpha'$ is,

$$\begin{aligned} E(\alpha_1\alpha_1') &= \rho_{cx}\frac{\sigma_x}{\sigma_c}\rho_{c1}\frac{\sigma_c}{\sigma_1} \\ &= \rho_{cx}\rho_{c1}\frac{\sigma_x}{\sigma_1} \end{aligned} \tag{5.7}$$

By the definition of variance, the expected variance explained in X by SNP_1 is,

$$\begin{aligned} E(Var.Explained) &= (\alpha_1\alpha_1')^2 Var(SNP_1) \\ &= \rho_{cx}^2\rho_{c1}^2\frac{\sigma_x^2}{\sigma_1^2}\sigma_1^2 \\ &= \rho_{cx}^2\rho_{c1}^2\sigma_x^2 \end{aligned}$$

The percentage of variance explained by SNP_1 is then,

$$\begin{aligned} E(\%Explained) &= 100\rho_{cx}^2\rho_{c1}^2\frac{\sigma_x^2}{\sigma_x^2} \\ &= 100\rho_{cx}^2\rho_{c1}^2 \end{aligned} \tag{5.8}$$

which is the percentage of variance explained by SNP_c from equation 5.4 $\times\rho_{c1}^2$, where ρ_{c1}^2 is r^2 of SNP_c and SNP_1.

5.3.2 Simulations

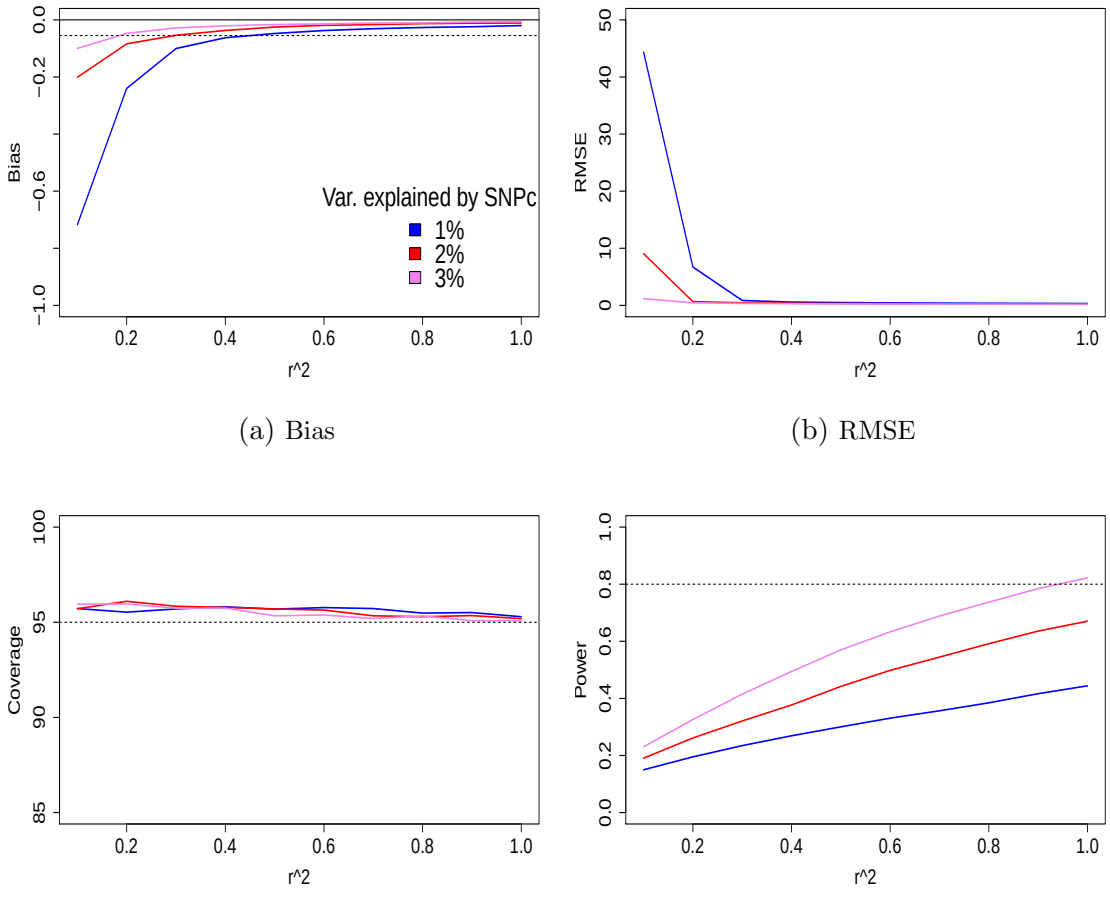
The simulations were applied to be in agreement with the mathematics to comprehend the mechanics of 2SLS. There are two aspects covered in the simulation; the r^2 between SNP_1 and SNP_c, and the variance in X explained by SNP_1.

Design

The MAF for SNP_c and SNP_1 will be 0.5 and the r^2 between them will vary between 0.1 to 1. The genotype of the SNPs will be simulated in section 4.2. The

simulation method in Section 4.3 sets the variation of X explained by SNP_c to 2%, this simulation will also include 1% and 3%. The generation of X and Y for 5,000 individuals is the same as the previous section.

The simulation were repeated 10,000 times for each variance explained by SNP_c. SNP_c will be discarded, the performance of 2SLS will be based on the non-causal SNP as instrument. The same evaluation criteria will be derived as in the previous section.



(a) Bias

(b) RMSE

(c) Coverage

(d) Power

Figure 5.4: Evaluation criteria of 2SLS with SNP_1 as instrument against correlation (r^2) between SNP_c and SNP_1 for when SNP_c explains 1%, 2% and 3% variation in X, shown by the blue, red and violet coloured lines accordingly. In bias, the dotted line is the 10% bias and the solid line is the zero bias. The 95% nominal level is defined by the dotted line in coverage. The 0.8 nominal power is labelled with the dotted line.

Results

Figure 5.4a to 5.4d are the bias, RMSE, coverage and power from 2SLS respectively, for different correlations (r^2) between SNP_c and SNP_1, and levels of variation explained by SNP_c. The allele frequency for both SNP_c and SNP_1 was 0.5.

For all three levels of variation explained, 2SLS have estimated more than 10% bias when r^2 is less than 0.2. RMSE decreases as correlation between SNP_c and SNP_1 increases. Coverage is approximately at 95% nominal level, with different levels of r^2 and variation explained. Power increases with the correlation between SNP_c and SNP_1. The type I error (Appendix Figure B.3) is approximately 5%, for all values of r^2.

The three evaluation criteria demonstrated that the performance of 2SLS deteriorates much faster when SNP_c only explained 1% of variation in X. At r^2 of 0.4, 1% variation explained has more than 10% bias, whereas the other variation explained have less than 10% bias. RMSE for 1% variation explained is over 40 at r^2 of 0.1, while 2% and 3% variation explained have RMSE of less than 10. The highest power for all levels of r^2 is when SNP_c explained 3% of the variation in X. However, 1% and 2% variation explained does not reach 0.8 nominal power, even with perfect correlation with SNP_c.

See Appendix Table B.2 for the tabular representation of these figures.

5.3.3 Conclusion

As r^2 decreases, the performance of 2SLS worsens, as expected since the association between X and SNP_1 is through its correlation with SNP_c. In other words, the weaker the correlation, the weaker the association between X and SNP_1. This relationship can be reflected algebraically; the variance explained in X by SNP_1 equals r^2 times by variance explained in X by SNP_c, proven by Section 5.3.1. This was also confirmed by simulation. Taking the bias as an example, SNP_1 with r^2 of 0.5, where SNP_c explained 2% of the variance, produced an approximate bias of -0.03 for sample size of 5,000 (Appendix B.2), which is the same as the bias from SNP_c with 1% variation explained and a sample size of 5,000 (Appendix B.1).

The correlation between SNP_c and SNP_1 did not affect the coverage or the type I error, this is consistent with results for a single causal SNP (Section 5.2); the change in variation explained by SNP_c did not alter the coverage or the type I error, and the variation explained by SNP_1 is dependent on the variation explained

by SNP_c and r^2. This resulted in the lack of movement in coverage and type I error with the changing r^2.

The next section will investigate whether there are any gains in including a non-causal SNP when a causal SNP is already an instrument in 2SLS.

5.4 Causal SNP plus another

As evident from the previous scenario, 2SLS was under powered when the variation explained by SNP_c is small. This section investigates whether an increasing number of instruments, would improve the power of the estimate in 2SLS. If adding instruments does improve 2SLS, what level of LD for the additional SNP would be best.

5.4.1 Design

From the discussion of two genotypes in Section 4.2, the desired allele frequency for both SNP_c and SNP_1 will be defined as 0.5, since having the same allele frequency generates more values of r^2 for comparison; r^2 will range from 0.1 to 0.9. SNP_c and SNP_1 will be included as instruments in 2SLS. To measure the benefit of including one non-causal SNP as instrument, 2SLS with SNP_c and SNP_1 was compared to 2SLS with only SNP_c as the instrument. The rest of the simulation follows Section 4.3.

The simulation will be repeated 10,000 times for a dataset of 5,000 individuals. The evaluation criteria are the same as the previous sections.

5.4.2 Results

As one instrument (the blue dot-dashed line) increased to two (the violet line) instruments in the first stage of 2SLS, the bias, RMSE and power have improved, as shown in Figures 5.5a to 5.5d. However, there is a minor decrease in coverage from the increase of instruments. The bias for two instruments is closer to zero, than one instrument. RMSE is smaller with two instruments in the first stage of 2SLS. Coverage is further away from 95% nominal level (the dotted line) with two instruments than one, this meant that less than 95% of the simulations had the true value within its confidence interval, Figure 5.5c. The dotted line in Figure 5.5d is the 0.8 nominal power. Power increased from 0.68 to 0.70 with one to two instruments

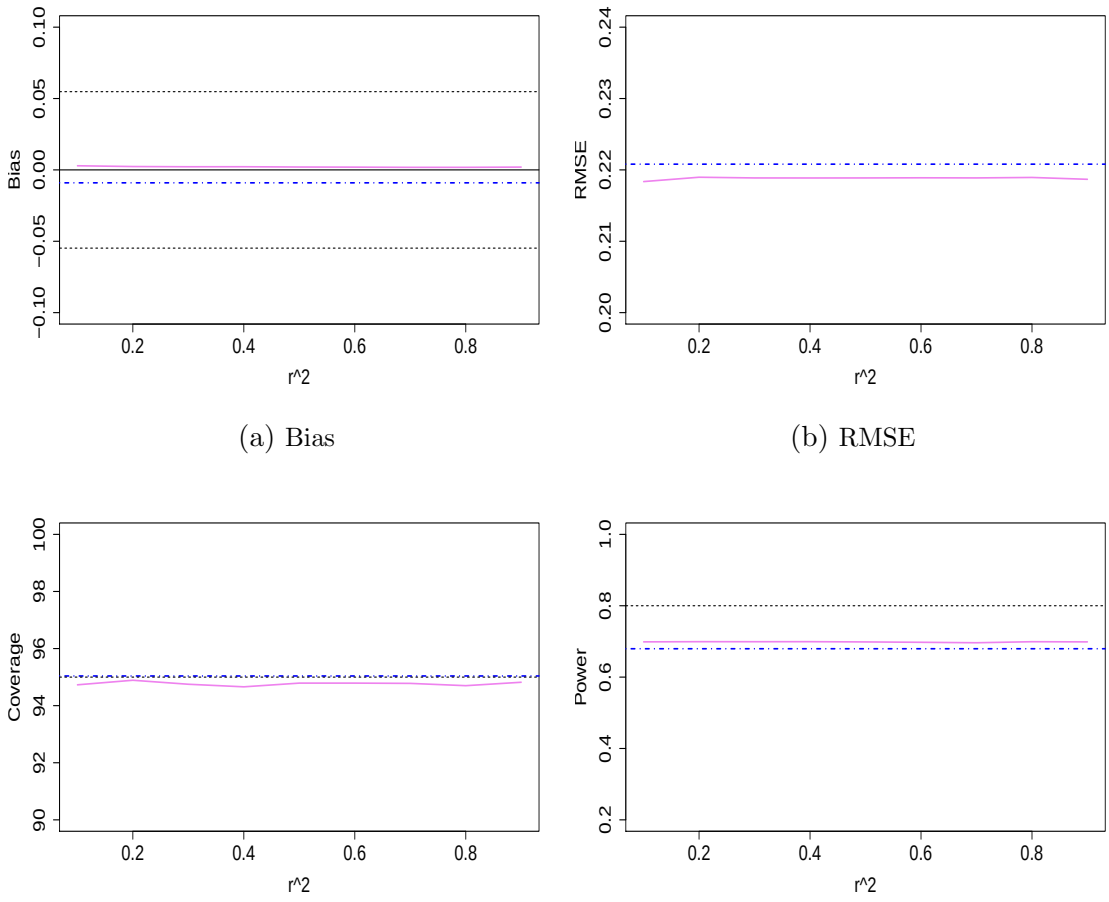
respectively. The type I error is the same for one and two instruments, they are both close to 5% significance level (the dotted line), Appendix Figure B.4.

The evaluation criteria were not affected by the difference in r^2 between SNP_c and SNP_1; this can be seen by Figures 5.5a to 5.5d, all the lines are moderately straight for two instruments, with standard errors of 0.0043 (Appendix Table B.3).

See Appendix table B.3 for the table representation of these figures.

5.4.3 Conclusions

There was evidence of a benefit in including more then one instrument in the 2SLS. However, the correlation between SNP_c and SNP_1 did not affect the 2SLS which suggests that adding another instrument improves the measurement error; since the change in r^2 will affect the variance explained by SNP_1, seen in section 5.3.1. Therefore, the is no benefit in adding another SNP as instrument if the causal SNP is known, in the estimation of causal effect using 2SLS. Otherwise adding a non-causal SNP will subject to unpredictable correction in the measurement error that is dependent on random noise. This section showed results for 2SLS with two instruments, when one is causal. Now consider when causal SNP was not measured, the ability of 2SLS to accurately estimate β_1 with two non-causal SNPs.



(a) Bias

(b) RMSE

(c) Coverage

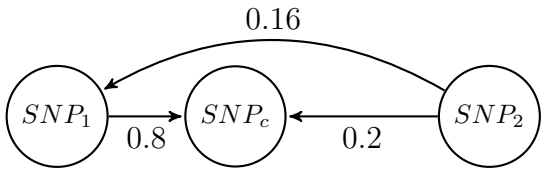
(d) Power

Figure 5.5: Evaluation criteria of 2SLS with instruments of SNP_c and SNP_1, and 2SLS with only SNP_c, against the correlation (r^2) between SNP_c and SNP_1. 2SLS with only SNP_c is represented by the blue dashed and 2SLS with SNP_c and SNP_1 is the violet line. In bias, The black dotted and solid lines represents the 10% and zero bias respectively. The black dotted line is 95% nominal coverage. The 0.8 nominal power is the black dotted line.

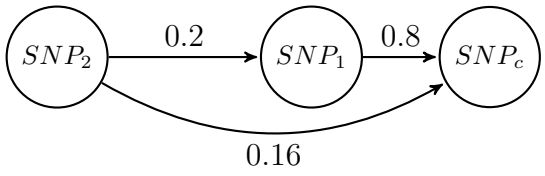
5.5 Two non-causal SNPs

Now consider the situation, when SNP_c is not measured. Would having two non-causal SNPs prove to be beneficial to the performance of 2SLS? In theory, adding an extra SNP to 2SLS should be dependent on which SNP was already chosen as an instrument. If the chosen SNP has a strong correlation with SNP_c then any extra SNP added would only improve the measurement error, the correlation of the extra SNP would not make a considerable difference as is evident from Section 5.4. If the correlation of the chosen SNP is weak, then adding another weak instrument would cause an inaccurate estimate of β_1 from 2SLS.

Another question is, now there are two non-causal SNPs, would the correlation between SNP_1 and SNP_2 effect the performance of 2SLS? Let the correlation between the SNPs be examined in terms of distance, as one of the reasons for the high correlation between the two SNPs is due to their close proximity in genome. For example, SNP_1 is close to SNP_c (i.e. highly correlated), and there are two scenarios; scenario one, when SNP_2 is far away from both SNP_c and SNP_1, shown by Figure 5.6a. And scenario 2, when SNP_2 is far away from SNP_c but closer SNP_1, Figure 5.6b. The question is would the differences in these two scenarios affect the performance of 2SLS.



(a) Scenario 1



(b) Scenario 2

Figure 5.6: Distance between the causal (SNP_c) and two non-causal SNPs (SNP_1 and SNP_2)

5.5.1 Mathematics of the variance explained by two non-causal SNPs

Figure 5.7 displays the relationship between the causal SNP, risk factor, disease outcome, unknown confounder and the correlation between the SNPs.

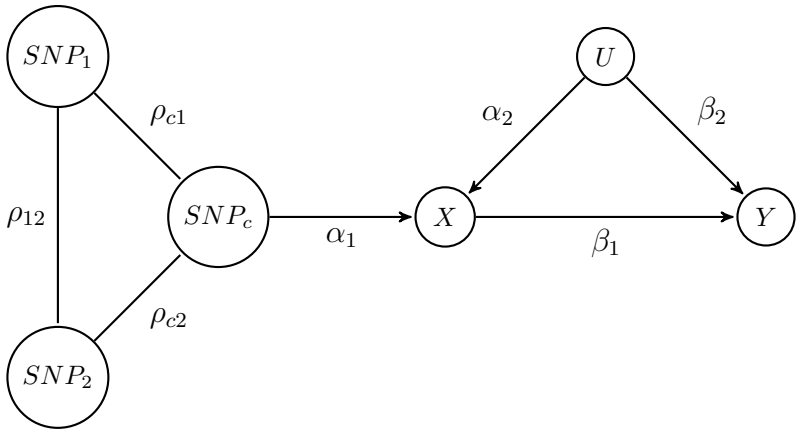


Figure 5.7: Illustration of Mendelian randomisation with a causal SNP and two non-causal SNPs. Solid line without and with arrow represent correlation and causal relationship respectively.

Let the covariance matrix of SNP_c, SNP_1 and SNP_2 be;

$$\Sigma = \begin{bmatrix} \sigma_c^2 & \rho_{c1}\sigma_c\sigma_1 & \rho_{c2}\sigma_c\sigma_2 \\ \rho_{c1}\sigma_c\sigma_1 & \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{c2}\sigma_c\sigma_2 & \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

where σ_c^2, σ_1^2 and σ_2^2 are the variances of SNP_c, SNP_1 and SNP_2 respectively. ρ_{c1}, ρ_{c2} and ρ_{12} are the correlations between SNP_c and SNP_1, SNP_c and SNP_2, and the correlation between SNP_1 and SNP_2 accordingly.

Regress SNP_c on SNP_1 and SNP_2,

$$SNP_c = \alpha_0' + \alpha_1' SNP_1 + \alpha_2' SNP_2 + \varepsilon \tag{5.9}$$

where α_0 is the intercept, ε is the random error and the covariance matrix are

split into four blocks to obtain the regression coefficient;

$$\Sigma = \left[\begin{array}{c|cc} \sigma_c^2 & \rho_{c1}\sigma_c\sigma_1 & \rho_{c2}\sigma_c\sigma_2 \\ \hline \rho_{c1}\sigma_c\sigma_1 & \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{c2}\sigma_c\sigma_2 & \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{array}\right]$$

Then the regression coefficient is [9];

$$\alpha = \Sigma_{..}^{-1}\Sigma_{.c} \tag{5.10}$$

where $\alpha = \begin{bmatrix}\alpha_1' \\ \alpha_2'\end{bmatrix}$, $\Sigma_{..} = \begin{bmatrix}\sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}$ and $\Sigma_{.c} = \begin{bmatrix}\rho_{c1}\sigma_c\sigma_1 \\ \rho_{c2}\sigma_c\sigma_2\end{bmatrix}$. $\Sigma_{..}$ is the covariance matrix of SNP_1 and SNP_2 and $\Sigma_{.c}$ is covariance matrix between SNP_c with SNP_1 and SNP_2.

The variance explained is $\alpha'\Sigma_{..}\alpha$ [9];

$$\begin{aligned} Var.Explained &= \alpha'\Sigma_{..}\alpha \\ &= \Sigma_{.c}'\Sigma_{..}^{-1}\Sigma_{..}\Sigma_{..}^{-1}\Sigma_{.c} \\ &= \Sigma_{c.}\Sigma_{..}^{-1}\Sigma_{.c} \end{aligned}$$

Then variance explained by SNP_1 and SNP_2 in matrix form;

$$\begin{aligned} \Sigma_{c.}\Sigma_{..}^{-1}\Sigma_{.c} &= \begin{bmatrix}\rho_{c1}\sigma_c\sigma_1 & \rho_{c2}\sigma_c\sigma_2\end{bmatrix}\begin{bmatrix}\sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}^{-1}\begin{bmatrix}\rho_{c1}\sigma_c\sigma_1 \\ \rho_{c2}\sigma_c\sigma_2\end{bmatrix} \\ &= \frac{1}{\sigma_1^2\sigma_2^2(1-\rho_{12}^2)}\begin{bmatrix}\rho_{c1}\sigma_c\sigma_1 & \rho_{c2}\sigma_c\sigma_2\end{bmatrix}\begin{bmatrix}\sigma_2^2 & -\rho_{12}\sigma_1\sigma_2 \\ -\rho_{12}\sigma_1\sigma_2 & \sigma_1^2\end{bmatrix}\begin{bmatrix}\rho_{c1}\sigma_c\sigma_1 \\ \rho_{c2}\sigma_c\sigma_2\end{bmatrix} \\ &= \frac{1}{\sigma_1^2\sigma_2^2(1-\rho_{12}^2)}\begin{bmatrix}\rho_{c1}\sigma_c\sigma_1\sigma_2^2 - \rho_{c1}\rho_{12}\sigma_c\sigma_1\sigma_2^2 & -\rho_{c1}\rho_{12}\sigma_c\sigma_1^2\sigma_2 - \rho_{c2}\sigma_c\sigma_1^2\sigma_2\end{bmatrix}\begin{bmatrix}\rho_{c1}\sigma_c\sigma_1 \\ \rho_{c2}\sigma_c\sigma_2\end{bmatrix} \\ &= \frac{1}{\sigma_1^2\sigma_2^2(1-\rho_{12}^2)}(\rho_{c1}^2\sigma_c^2\sigma_1^2\sigma_2^2 - 2\rho_{c1}\rho_{c2}\rho_{12}\sigma_c^2\sigma_1^2\sigma_2^2 + \rho_{c2}^2\sigma_c^2\sigma_1^2\sigma_2^2) \\ &= \frac{\sigma_c^2}{1-\rho_{12}^2}(\rho_{c1}^2 - 2\rho_{c1}\rho_{c2}\rho_{12} + \rho_{c2}^2) \end{aligned} \tag{5.11}$$

Regress X on SNP_c,

$$X = \alpha_0 + \alpha_1 SNP_c + \varepsilon' \tag{5.12}$$

Replace SNP_c in 5.12 with 5.9,

$$X = \alpha_0'' + \alpha_1(\alpha_1' SNP_1 + \alpha_2' SNP_2) + \varepsilon'' \tag{5.13}$$

The variance in X explained by SNP_c is,

$$Var.Explained = \alpha_1^2 \sigma_c^2$$

In terms of the variance in SNP_c explained by SNP_1 and SNP_2, then the variance in X explained by SNP_1 and SNP_2 is,

$$Var.Explained = \alpha_1^2 \alpha' \Sigma_{..} \alpha \tag{5.14}$$

Writing 5.14 in another form

$$= \alpha_1^2 \sigma_c^2 \left[\frac{\alpha' \Sigma_{..} \alpha}{\sigma_c^2} \right] \tag{5.15}$$

Using the expected α_1, $\rho_{xc}\dfrac{\sigma_x}{\sigma_c}$. Then, the expected variance explained is

$$\begin{aligned} E(Var.Explained) &= \rho_{xc}^2 \frac{\sigma_x^2}{\sigma_c^2} \sigma_c^2 \left[\frac{\alpha' \Sigma_{..} \alpha}{\sigma_c^2} \right] \\ &= \rho_{xc}^2 \sigma_x^2 \left[\frac{\alpha' \Sigma_{..} \alpha}{\sigma_c^2} \right] \end{aligned} \tag{5.16}$$

Substituting 5.11 into 5.16, the expected percentage explained is,

$$\begin{aligned} E(\%Explained) &= 100 \rho_{xc}^2 \frac{\sigma_x^2}{\sigma_x^2} \left[\frac{\rho_{c1}^2 - 2\rho_{c1}\rho_{c2}\rho_{12} + \rho_{c2}^2}{1 - \rho_{12}^2} \right] \\ &= 100 \rho_{xc}^2 \left[\frac{\rho_{c1}^2 - 2\rho_{c1}\rho_{c2}\rho_{12} + \rho_{c2}^2}{1 - \rho_{12}^2} \right] \end{aligned} \tag{5.17}$$

which is the percentage of variance in X explained by SNP_c $\times$ the proportion of the variance in SNP_c explained by SNP_1 and SNP_2. When there is no correlation between SNP_1 and SNP_2, i.e. $\rho_{12}^2 = 0$, then the percentage of variance explained is the percentage of variance explained by SNP_c $\times$ the sum of the correlation of the two SNPs with SNP_c.

5.5.2 Simulations

This simulation will demonstrate the mathematical proof above.

Design

Suppose there are three SNPs, and one of the SNPs is causal for X, called SNP_c. SNP_1 and SNP_2 are not causal for X but are correlated with SNP_c and also with each other. The genotype simulation method of Section 4.2 does not include the option to change the correlation between two non-causal SNPs, therefore the following steps will be used to generate the three SNP genotypes for 5,000 individuals;

1. The MAF for SNP_c, SNP_1 and SNP_2 were 0.5, 0.45 and 0.45 respectively. The similarity in MAF was to provide some flexibility in setting the values of correlation.

2. Using the command mvrnorm() from R package, "MASS" [279] to draw from the distribution of a random multivariate normal of SNP_c, SNP_1 and SNP_2 with 5,000 samples(i);

$$\begin{bmatrix} Z_{ic} \\ Z_{i1} \\ Z_{i2} \end{bmatrix} \sim N(0, \Sigma)$$

where

$$\Sigma = \begin{bmatrix} \sigma_c^2 & \rho_{c1}\sigma_c\sigma_1 & \rho_{c2}\sigma_c\sigma_2 \\ \rho_{c1}\sigma_c\sigma_1 & \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{c2}\sigma_c\sigma_2 & \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix},$$

σ_c^2, σ_1^2 and σ_2^2 is the variance of SNP_c, SNP_1 and SNP_2 respectively. ρ_{c1}, ρ_{c2} and ρ_{12} are the desired correlation between SNP_c and SNP_1, SNP_c and SNP_2, and the correlation between SNP_1 and SNP_2 respectively.

3. Sort the values of Z_c. Find Z_{ic} at $(1 - f_c)^2$ and at f_c^2, naming these values as a and b accordingly.

4. To form the genotype of SNP_c for 5,000 samples, each individual has 0 when $Z_{ic} < a$, 1 when $Z_{ic} > a$ and 2 when $Z_{ic} > b$.

5. Repeat step 3 and 4 for Z_{i1} and Z_{i2} to get genotypes of SNP_1 and SNP_2.

Each scenario will be defined by different combinations of correlation, demonstrated by Figure 5.8. ρ_{c1} values were 0.2, 0.5 and 0.8. ρ_{c2}, included 0.2, 0.5, 0.8 and the multiples of these values with values in ρ_{c1}. ρ_{12} has exactly the same range of correlation as ρ_{c2}. For example, ρ_{12} of 0.16 is the multiple of 0.2 and 0.8, i.e. $\rho_{c2} \times \rho_{c1}$. Then 0.16 swaps with ρ_{c2} to 0.2, and ρ_{c1} remains unchanged.

The simulation of the relationships between SNP_c, X and Y follows Section 4.3. 2SLS with one instrument, SNP_c and two instruments, SNP_1 and SNP_2 will be monitored. Each scenario, will be repeated 10,000 times. The correlation with SNP_c shown an impact on the evaluation criteria; previous simulations have found mean bias to be affected by extreme estimates of β_1, due to the unstable Wald ratio caused by near zero α_1. Therefore, median bias will also be calculated.

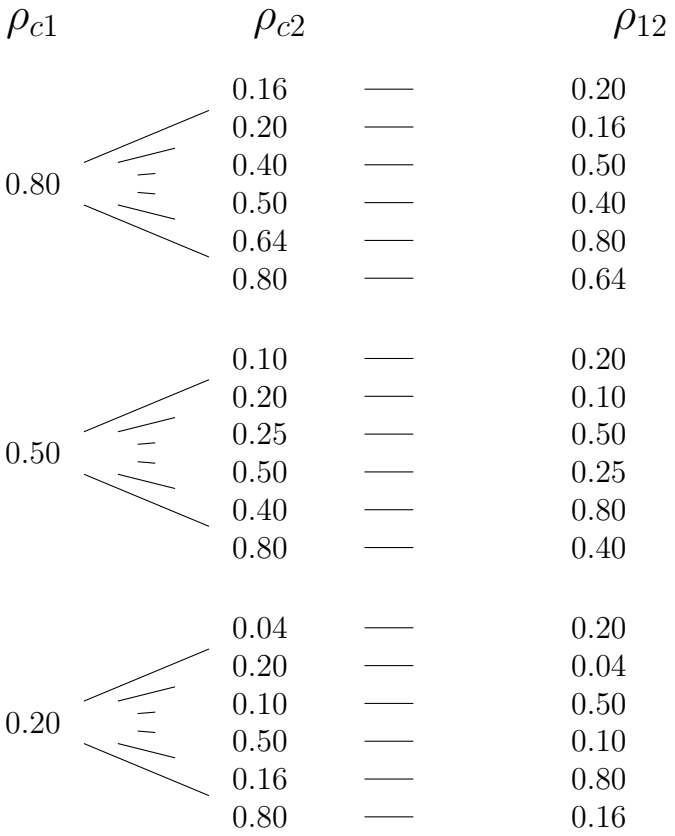


Figure 5.8: Design of the correlation combinations between the three SNPs. ρ_{c1} is the correlation between SNP_c and SNP_1, ρ_{c2} is SNP_c and SNP_2, and ρ_{12} is between SNP_1 and SNP_2.

Results

Table 5.1 shows the difference in bias is greater with two weak instruments, i.e. both ρ_{c1} and ρ_{c2} are less than 0.2. The stronger the correlation in one of the two instruments, the closer the difference in bias is to zero. The difference in RMSE is further away from zero when one instrument does not have a high correlation with SNP_c (i.e. $\rho < 0.8$). There is no distinct difference in coverage (a percentage of simulation that have the true causal effect within the estimation of 95% confidence interval), when comparing the strength in correlations between the instruments. There is a decrease in power with two instruments; the decrease is at its peak when both instruments have $\rho < 0.8$. Same as coverage, there are no recognisable trends from the difference in type I error.

The evaluation criteria seem to be fairly constant as correlation between SNP_1 and SNP_2 increases, ρ_{12}. For clarity, consider when ρ_{c1} is 0.5 and 0.8, the difference seems similar as ρ_{12} increases. There is more of a spread with ρ_{c1} of 0.2, this is because of the different values of ρ_{c2}. There are no distinct patterns in the difference of RMSE, as ρ_{12} increases, the values are only affect by either high ρ_{c2} or high ρ_{c1}. Again for power, the differences were dispersed are not affected by ρ_{12}, but by ρ_{c2} or ρ_{c1}. The differences in coverage and type I error are closer to zero, i.e. the estimates from SNP_1 and SNP_2 are the same as the estimates from SNP_c.

Table 5.1: The difference in evaluation criteria between 2SLS with instruments of SNP_1 and SNP_2, and with only SNP_c. The difference in evaluation criteria is calculated, for example Bias with SNP_1 and SNP_2 - Bias with SNP_c. ρ_{c1}, ρ_{c2} and ρ_{12} is the correlation between, SNP_c and SNP_1, SNP_c and SNP_2, SNP_1 and SNP_2 respectively.

ρ_{c2}, ρ_{12}	Mean Bias	Median Bias	RMSE	Coverage	Power	Type I Error
$\rho_{c1} = 0.8$						
0.80,0.64	0.0114	0.0170	0.0365	0.0000	-0.0849	0.0001
0.64,0.80	0.0097	0.0183	0.0676	-0.3900	-0.1580	0.0040
0.50,0.40	0.0105	0.0202	0.0632	0.0200	-0.1387	-0.0003
0.40,0.50	0.0101	0.0141	0.0653	-0.2300	-0.1521	0.0023
0.20,0.16	0.0111	0.0162	0.0647	-0.3200	-0.1473	0.0033
0.16,0.20	0.0108	0.0176	0.0661	-0.4100	-0.1589	0.0041
$\rho_{c1} = 0.5$						
0.80,0.40	0.0108	0.0195	0.0571	-0.3900	-0.1332	0.0040
0.40,0.80	0.0052	0.0497	0.2561	-0.6600	-0.3506	0.0067
0.50,0.25	0.0117	0.0310	0.1397	-0.0600	-0.2596	0.0006
0.25,0.50	0.0092	0.0437	0.2487	-0.1700	-0.3431	0.0015
0.20,0.10	0.0155	0.0366	0.2268	-0.4500	-0.3338	0.0044
0.10,0.20	0.0096	0.0390	0.2380	-0.1800	-0.3462	0.0017
$\rho_{c1} = 0.2$						
0.80,0.16	0.0083	0.0166	0.0661	-0.4100	-0.1496	0.0040
0.16,0.80	0.1404	0.2610	1.4455	-0.8500	-0.4844	0.0086
0.50,0.10	0.0103	0.0380	0.2213	-0.4800	-0.3249	0.0048
0.10,0.50	0.1090	0.2255	1.7517	-1.4900	-0.4816	0.0150
0.20,0.04	0.0231	0.1441	0.8511	-1.6300	-0.4466	0.0162
0.04,0.20	0.1578	0.2869	2.0236	-1.2000	-0.4749	0.0121

5.5.3 Conclusion

To illustrate the mathematics, Figure 5.9 was produced from equation 5.17, for the correlation between the SNPs against the variance explained. Figure 5.9 shows when

the correlation between SNP_c and the non-causal SNPs becomes stronger, then the variance explained gets closer to the variance explained by SNP_c, therefore the difference in evaluation criteria reduces (Table 5.1), i.e. the estimates from two non-causal SNPs were approximately the estimates from 2SLS with SNP_c. The mathematics were reflected on the simulations; taking median bias as an example, when $\rho_{c1} = \rho_{c2}$ in Table 5.1, where they both have correlations of 0.2, 0.5 and 0.8, the difference in bias declines towards zero as 0.2 increased to 0.8. This is because the variance explained by the two instruments was increasing. Suppose we match the correlations described here to the correlations in Figure 5.9 then, Figure 5.9 shows when ρ_{c1} is 0.2 and ρ_{12} is approximately 0.2 (the black curve)would produce variance explained of under 0.5%. With ρ_{c1} as 0.5 and ρ_{12} as 0.2 (the black curve) yielded variance explained of approximately 1%. When ρ_{c1} is 0.8 and ρ_{12} is approximately 0.5 (the red curve), gives just above 1.5% of variance explained.

The simulations (Table 5.1) showed that 2SLS with one strong instrument within the two, performs approximately just as well as with SNP_c as a single instrument. This is because having a strong SNP, explains a similar amount of variance as SNP_c, shown by section 5.3. It therefore acts as a causal SNP, and adding any other SNP would only correct for measurement error, as shown in Section 5.4. Table 5.1 also suggested two weak instruments would have an adverse effect on the performance of 2SLS. The reason is shown by equation 5.17; the weaker the correlation with SNP_c, the lower the variance explained by both instruments, and therefore greater the difference in estimates, when comparing to SNP_c as an instrument.

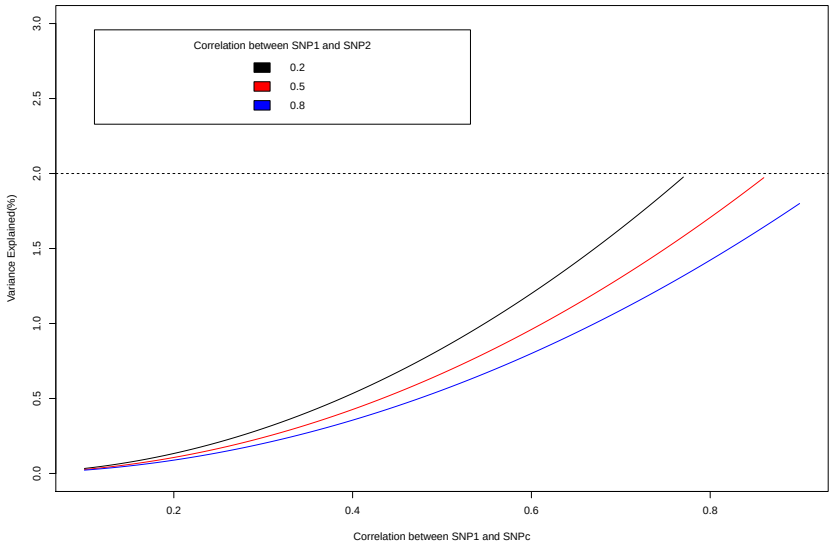


Figure 5.9: Variance explained by SNP_1 and SNP_2 with correlations. The x-axis is the correlation between the non-causal SNP and SNP_c, where $\rho_{c1} \approx \rho_{c2}$, and the y-axis is the variance explained by the two non-causal SNPs. The three different colours represents the correlation between the two non-causal SNPs, ρ_{12}. Black, red and blue represents 0.2, 0.5, 0.8 respectively. The dotted horizontal line is the variance explained by SNP_c but also a cut-off, since it is not possible for the two non-causal SNPs to have complete independence of each other, when they are both highly correlated with the causal SNP.

5.6 Discussion

2SLS was proficient in estimating the causal effect of a risk factor on a disease outcome, using dependent SNPs as instruments. But, this is conditional on the proportion of variance in the risk factor explained by the causal SNP and the sample size of the individual-level data. When the sample size is small and the variance explained is close to zero, there is a higher chance of estimating α_1 as zero, and therefore creating an unstable estimate of β_1.

2SLS with a non-causal instrument is still efficient in estimating the relationship between risk factor and disease outcome. As the variance in the risk factor explained by these dependent SNPs, relies on the correlation with and the variance explained by the causal SNP; the variance explained by the dependent SNP equals r^2 times the variance explained by the causal SNP. Only the measurement error was corrected, when another SNP was included with the causal SNP as instruments in

2SLS. When one of the dependent SNPs is strongly correlated with the causal SNP, then having two dependent SNPs as instruments can still produce accurate estimate of β_1. However, there was no benefit in adding another strongly correlated SNP, unless the two SNPs are independent to a degree (as complete independence is not possible).

As seen from the simulations in this chapter using a single proxy SNP as instrument gave an unbiased estimate of the causal effect if this proxy is highly correlated with the causal SNP. However, if this is not the case then using this proxy SNP will still give a biased estimate, as it does not account for all of the true variation. Similar conclusion have also been drawn in GWAS; due to the incomplete coverage of common genetic variant in the contemporary marker panels available for genotyping [19]. In order to capture the true association of the causal variant(s) that have not been directly genotyped, multiple variants in LD have been utilised. This approach have also found to result in modest power increase [89, 138, 217, 224].

The simulation makes a simplifying assumption that the variance explained by the causal SNP remains the same despite its MAF. However, in reality this may not always be the case; a review by Manolio et al. [194] have suggested that the reason that GWAS have only found modest genetic association with disease risk is because the large effects lies within rare variants. Currently there are not many examples to prove this theory, due to insufficiently large sample size and difficulty in genotyping the rare variants. In the context of Mendelian Randomisation, consider if we have a causal SNP with low MAF as the instrument, which has a large effect size on the exposure of interest, this could still cause weak instrument bias and low precision, as the genotype of a rare MAF SNP consists of mostly zero and will require larger sample size than an instrument with a common MAF.

Conclusions have been drawn from using one and two dependent instruments for 2SLS. The next investigation will be to examine when there are multiple dependent instruments, whether having more than two instruments would improve the performance of 2SLS, from having just the causal SNP.

Chapter 6

Multiple Dependent SNPs in Two-stage Least Squares

6.1 Introduction

The previous chapter has examined the performance of two-stage least squares (2SLS) with two non-causal SNPs in linkage disequilibrium (LD) as instruments. The conclusion was that the variation in X explained by a non-causal SNP is dependent on its correlation with the causal SNP and the variation in X that is explained by the causal SNP. This chapter aims to find an instrument selection criterion that is able to identify SNPs that are highly correlated with the causal SNP and to reduce possible bias from weak instruments in 2SLS.

The SNPs from Experiment 1 will be assumed to have patterned LD as the aim is to quantify the effects of LD on 2SLS. So far, the simulations have generated artificial LD, to control the amount of LD and identify different levels of impact from LD. To find whether the same impact can be achieved with realistic LD, the rest of the experiments in this chapter will use GENOME [185], to form realistic LD patterns. The aim of Experiment 2 will be to find any potential differences between artificial and realistic patterns of LD. Experiment 3 will answer the question whether reasonable performance from 2SLS can still be observed with various selection policies for multiple dependent SNPs.

Every experiment will be presented with the aims, design, results and conclusions. The aims section gives the scenario which the experiment will simulate and the question which the experiment will answer. The design section will follow the simulation methods in Chapter 4, but with some minor changes to tailor for each

experiment. The results section will have graphical illustrations and descriptions of the results. The conclusion section will summarise the results and answer the question in the aims section. This chapter will end with a discussion, where I will describe the findings from each experiment, their limitations and outline the plans for next chapter with the results from this chapter in mind.

6.2 Experiment 1: p-value ranking under patterned linkage disequilibrium

6.2.1 Aims

This experiment will investigate the situation in which data are available on 6 possible correlated instruments in the form of SNPs taken from the same gene. It will consider the case in which none of the potential instruments is the causal SNP. The SNPs will be assumed to have pattern linkage disequilibrium (LD) as defined in Section 4.2. SNPs will be selected for inclusion in the Mendelian randomisation analysis (MR) based on their ranking according to the p-values of their unadjusted associations with the risk factor, X. The aims of the experiment will be to determine the optimal number of instruments for use in a two-stage least squares (2SLS) MR and to understand the factors that influence this number.

6.2.2 Design

The experiment will take the form of a simulation in which 7 SNPs are generated with patterned LD in Section 5.5.2. The maximum ρ will be set to 0.1 till 0.9 in steps of 0.1 to simulate 9 different LD patterns, referred to as simply as ρ. In patterned LD the SNP closest to the causal SNP has a correlation of ρ, while the next SNP has a correlation ρ^2 with the causal SNP and the furthest away has a correlation ρ^3 with the causal SNP. The middle SNP will be assumed to be causal and to explain 2% of the variance in the risk factor, X. X will explain 6% of the variance in the outcome Y. X and Y will be normally distributed and the sample size will be 5,000. The causal SNP will be discarded and the 2SLS will be based on the selected non-causal SNPs. The simulation will be repeated 10,000 times.

In each simulation X will be regressed on each SNP in turn and the SNPs will be ranked according to their p-value. 2SLS analyses will be conducted on the same

data using the best 1, 2,...6 SNPs and the bias, median bias, root mean square error (RMSE), coverage and power will be calculated. A separate simulation in which X is not directly related to Y will be used to estimate the type I error.

A further simulation will identify the first SNP to be chosen in terms of its LD with the causal SNP. In this simulation we will only consider $\rho = 0.5$ and $\rho = 0.9$.

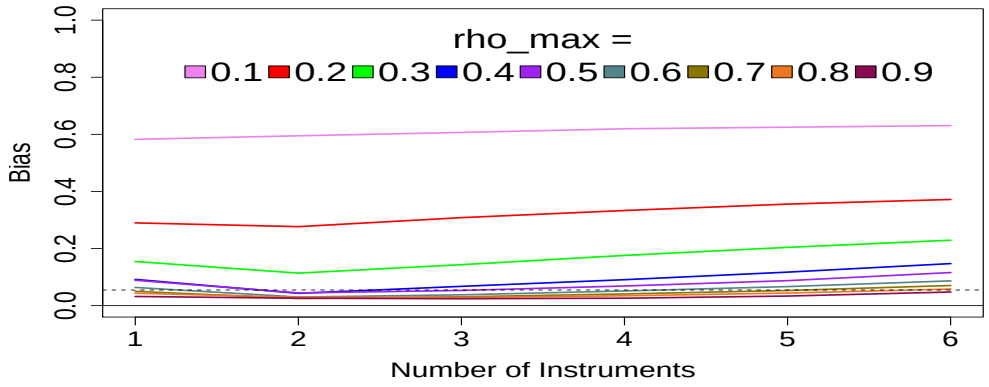
6.2.3 Results

Selection based on p-value ranking

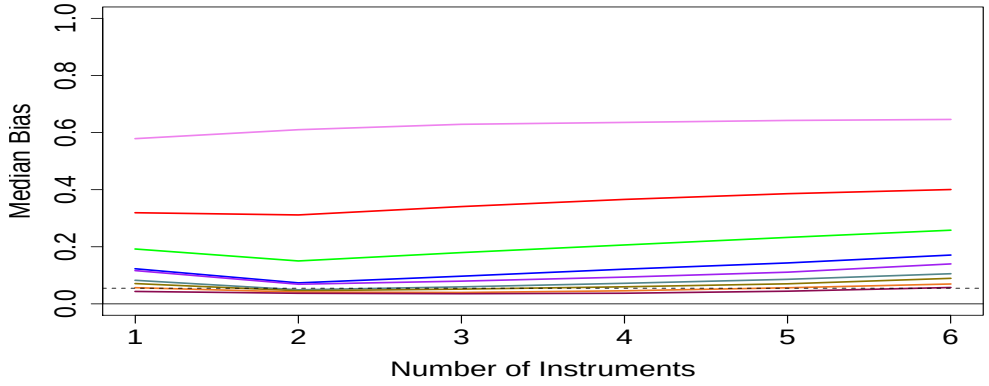
The bias (Figure 6.1a) and median bias (Figure 6.1b) of the MR estimates are greater when the LD is weak and suggests that the best performance is obtained from taking the two most highly ranked SNPs. All number of selected SNPs gave bias below 10% provided ρ was greater than 0.7. The difference in mean (Figure 6.1a) and median (Figure 6.1b) bias demonstrates the existence of a relatively large negative bias in the simulation which pulled the mean towards zero. Taking ρ of 0.4 with two instruments as an example, the bias is below the 10% (the dotted line) but the median bias is above 10%. The RMSE (Figure 6.1c) reflects both the bias and the variance of the MR estimate. It too is highest when the LD is weak but tends to decrease with increasing numbers of selected SNPs implying that extra SNPs improve the variance of the estimate more than they increase the bias.

The coverage of a 95% confidence interval is shown in Figure 6.1d. Low levels of LD can reduce the coverage to around 80% and for all strengths of LD the coverage falls as the number of selected SNPs increases. Using two SNPs improves coverage very slightly when the correlation is high but makes coverage worse when the correlation is low. The reverse pattern is seen in the plot of Type I error (Figure 6.1e) which is poorly controlled when the LD is weak. Figure 6.1f shows the power and appears to show that SNPs in weak LD have the greatest power but this is a misleading impression caused by the failure to control Type I error.

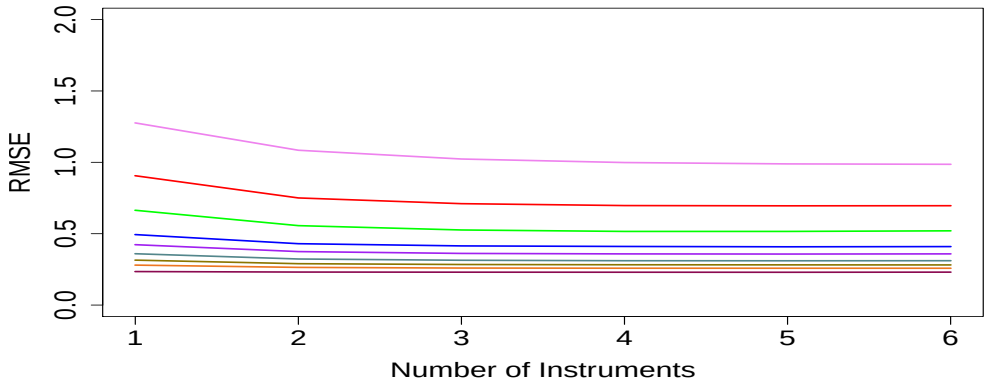
See appendix table C.1 for the table representation of the figures.



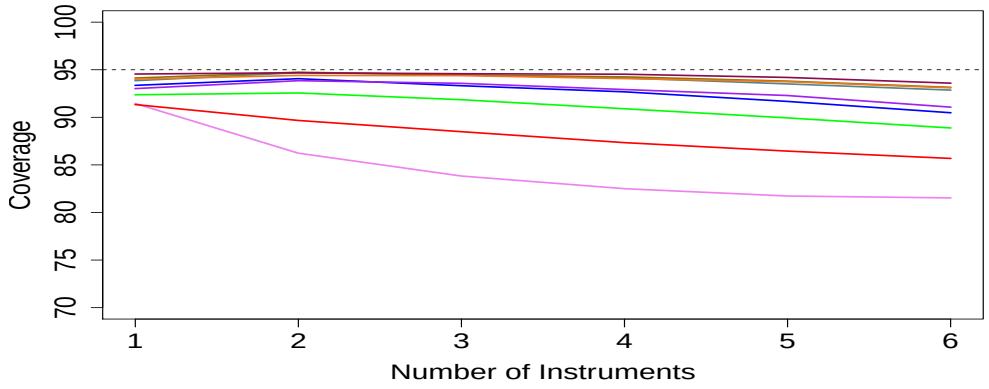
(a) Bias



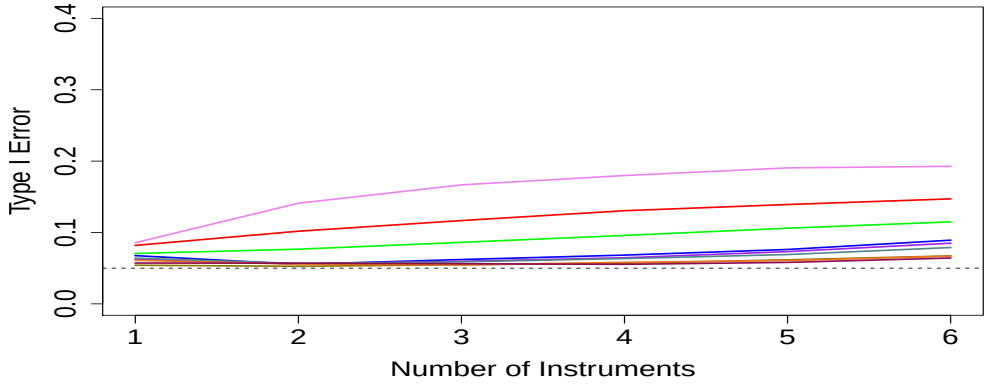
(b) Median Bias



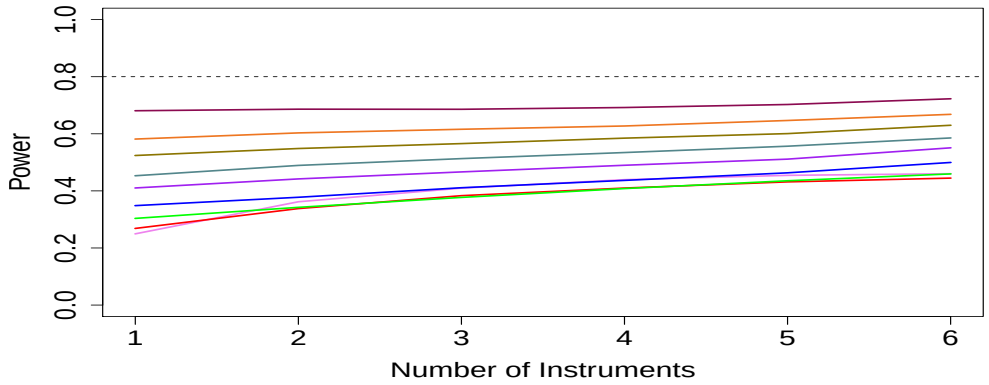
(c) RMSE



(d) Coverage

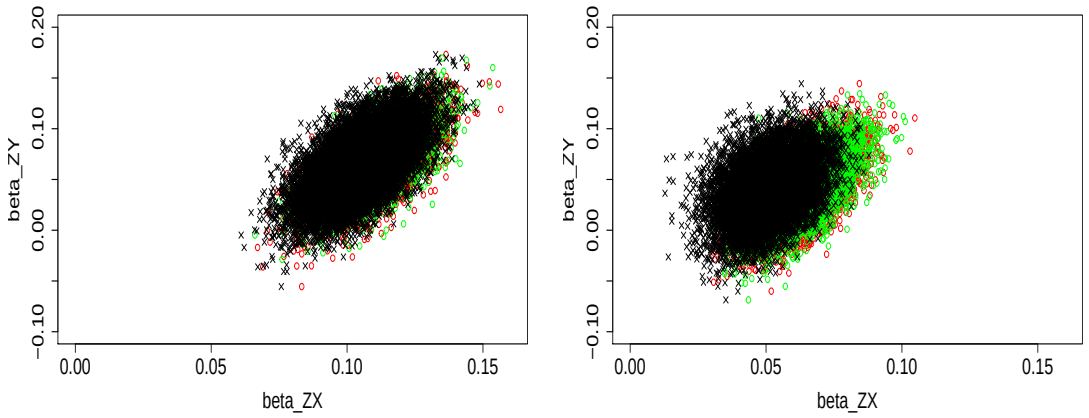


(e) Type I Error



(f) Power

Figure 6.1: Evaluation criteria to measure the performance of 2SLS for different strengths of LD, maximum ρ, based on selecting the best 1,2,...6 SNPs and using them jointly in a 2SLS MR. The colours represents range of maximum ρ, the correlations between the SNPs within a gene, see legend. For bias, the dotted and solid lines represent 10% and 0 bias respectively. The dotted line in coverage, type I error and power is nominal level of 95%, 5% and 0.8 respectively.



(a) $\rho = 0.9$ (b) $\rho = 0.5$

Figure 6.2: The $\hat{\beta}_{ZX}$ and $\hat{\beta}_{ZY}$ from SNPs with the lowest p-value of regression of X on SNP. Green and red coloured dots are the two SNPs closest to the causal SNP, and the black dots are the other SNPs

Identity of the first SNP to be chosen

Table 6.1 shows for $\rho = 0.5$ and $\rho = 0.9$ which of the 6 SNPs was chosen first.

Figure 6.2a and 6.2b are the estimates from the regressions of X on SNP and Y on SNP where the SNP had the minimum p-value. When LD is strong (Figure 6.2a), the estimates between the two closest and the others SNPs are similar. While for weaker LD, the estimates in more distant SNPs are shifted away from those for the two closest SNPs.

When the correlation is low the SNPs furthest from the causal SNP are only weakly associated with X ($\rho = 0.5$, $\rho^2 = 0.25$, $\rho^3 = 0.125$) and so are very rarely selected while when the correlation is higher ($\rho = 0.9$, $\rho^2 = 0.81$, $\rho^3 = 0.729$) all six SNPs remain associated with X and may sometimes be selected. In that case the high correlation ensures that the chosen instruments are never very weak and so there is little bias in the MR estimate. When the correlation is low the MR can be heavily biased especially when by chance a SNP is chosen that has a very low true association with X.

6.2.4 Conclusions

When ρ is low the LD between SNPs is weak and the six available SNPs will be poor substitutes for the causal SNP, the strength of their association with X will be

Table 6.1: The SNP with the lowest p-value from regression of X on the SNP in each simulation. The estimate of the coefficients for regression of X on SNP ($\hat{\beta}_{ZX}$), Y on SNP ($\hat{\beta}_{ZY}$) and the MR estimate of Y on X ($\hat{\beta}_{XY}$) of 10,000 simulations, where k represents the different SNP number, where k=1,...,6. The top row gives the true value of these coefficients.

	No. of simulation with lowest p-value	Mean $\hat{\beta}_{ZX}$	Mean $\hat{\beta}_{ZY}$	Mean $\hat{\beta}_{XY}$
True		0.1180	0.0647	0.5480
$\rho = 0.5$				
SNP_1	7	0.0476	0.0509	1.0443
SNP_2	86	0.0549	0.0528	0.9596
SNP_3	5021	0.0632	0.0405	0.6126
SNP_4	4832	0.0630	0.0408	0.6198
SNP_5	50	0.0524	0.0498	0.9321
SNP_6	4	0.0495	0.0532	1.0511
$\rho = 0.9$				
SNP_1	493	0.1079	0.0665	0.6044
SNP_2	1512	0.1094	0.0656	0.5890
SNP_3	3307	0.1103	0.0642	0.5698
SNP_4	2739	0.1102	0.0655	0.5830
SNP_5	1183	0.1092	0.0658	0.5917
SNP_6	766	0.1082	0.0650	0.5900

weak and we observe the characteristic of the weak instrument bias [253], namely increased bias and RMSE, poorer coverage and type I error control, and misleading powers.

When the observed SNPs are poor surrogates for causal SNPs there is some gain in using the two best SNPs as this reduces bias and RMSE, but adding further SNPs always leads to poorer performance.

It appears that the optimal performance depends on the pattern of LD which means that it is very difficult to suggest a SNP selection policy that will be optimal in all situations. The best policy is likely to depend on the pattern of LD within the gene and strength of the association between the unmeasured causal SNP and X. Physical proximity is not the only factor that affects LD [251], therefore a genome simulator will be utilised to simulate a more realistic LD between the SNPs in the next experiment.

6.3 Experiment 2: p-value ranking under real linkage disequilibrium

6.3.1 Aims

This experiment is similar to Experiment 1, the only difference between Experiment 1 and 2 is the construction of the linkage disequilibrium (LD) between the SNPs; Experiment 2 simulates a more realistic LD, using GENOME [185], whereas Experiment 1 assumed the SNPs had patterned LD. The aim of this experiment is to see whether the conclusions about optimal numbers of instruments for 2SLS MR differs from Experiment 1, under a more realistic genetic structure.

6.3.2 Design

The design of the simulation will follow Experiment 1, except for generating the genotypes of SNPs. GENOME will simulate the genotypes for 5,000 individuals, as described in Section 4.2.3. 7 SNPs will be randomly selected from a section of GENOME simulated strand, i.e. a section of DNA. The middle SNP will be assumed to be causal, SNP_c, and it explains 2% of the variation in risk factor, X.

To extract different patterns of linkage disequilibrium, GENOME will be run 5 times to obtain 5 sections of DNA. Then for each section of DNA, simulation of an MR dataset of 5,000 individuals will be repeated 10,000 times.

6.3.3 Results

Table 6.2 gives information on the 7 SNPs from each section of DNA. Within a section of DNA, usually the SNPs having strongest correlation with SNP_c are selected within the simulations. *DNA 1*, *2* and *5* contained SNPs of the weakest correlations with SNP_c, *DNA 3* and *4* have SNPs that were identical to SNP_c. There were also SNPs identical to each other. i.e. same allele frequency and correlation with SNP_c. The R command for 2SLS, ivreg(), deals with identical instruments by automatically removing one of them from the first-stage regression. Therefore, the estimates from 2SLS will remain unchanged with supposedly an additional instrument.

From Figure 6.3a, all numbers of selected SNPs have less than 10% bias (the dotted line), except for *DNA 5* as it contained SNPs that are weakly correlated with SNP_c. The bias for *DNA 1* to *4* switches from negative to positive, but the

same trend is not seen in median bias (Figure 6.3b), therefore this trend is caused by the extremely negative estimates within the simulations. The RMSE (Figure 6.3c) is the combination of bias and variance of the MR estimate. RMSE is high when the correlation with SNP_c is weak, i.e. *DNA 1* and *5*. The similarity in RMSE for *DNA 2* to *4* suggests that by adding extra SNPs improves the variance but it is balanced by the increase in bias.

The coverage (Figure 6.3d) for all the sections of DNA are approximately 95% nominal level (the dotted line). *DNA 5* contains SNPs in low LD with SNP_c and has the lowest coverage. With additional instruments, there is a decrease in coverage for all the sections of DNA. The type I error increases from the 5% significance level (the dotted line) when including more SNPs and when LD is weak. Figure 6.3f shows a modest improvement in power for all sections of DNA, as the number of selected SNPs increases, however power will need to be adjusted for the rise in type I error.

As expected, the performance of 2SLS from *DNA 1* and *5* are less efficient, in comparison to other sections of DNA; this is due to the weak correlation of these SNPs with SNP_c. For each section of DNA, the standard error for the simulations are fairly similar for the different number of instruments. In Appendix Table C.2 from 1 to 6 instruments, *DNA 3* has the smallest contrast (0.0043 to 0.0042) and *DNA 5* has the greatest (0.0114 to 0.0088).

6.3.4 Conclusions

For the SNPs that were strongly correlated with SNP_c, the bias increased with additional instruments, since all of variation is explained by the SNPs that were already selected as instruments. For small r^2 between the SNPs, the bias decreased from one to two instruments for *DNA 5*. This supports the evidence from Section 5.5.1; when r^2 is small (i.e. close to independent), adding another SNP will increase the variation explained. However, with even greater numbers of instruments, this advantage disappears. The RMSE shows that even though adding more instruments decreases the variation, this does not outweigh the increase in bias. Coverage decreases with additional instruments and the increase in power was balanced by the increase in type I error.

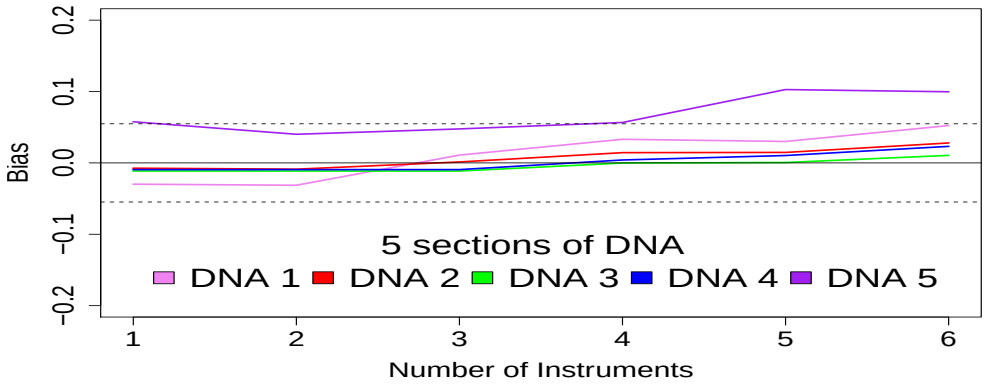
For sections without identical SNPs, including 2 SNPs with the smallest p-value does improve bias, RMSE and type I error. Adding further instruments always leads to the deterioration in performance from 2SLS. This conclusion is similar

to Experiment 1. However patterned LD did not generate sections with identical SNPs, therefore GENOME simulated DNA sections have an additional conclusion for optimal numbers of instruments for 2SLS; when identical SNPs are present, there are no benefits in including more than one instrument.

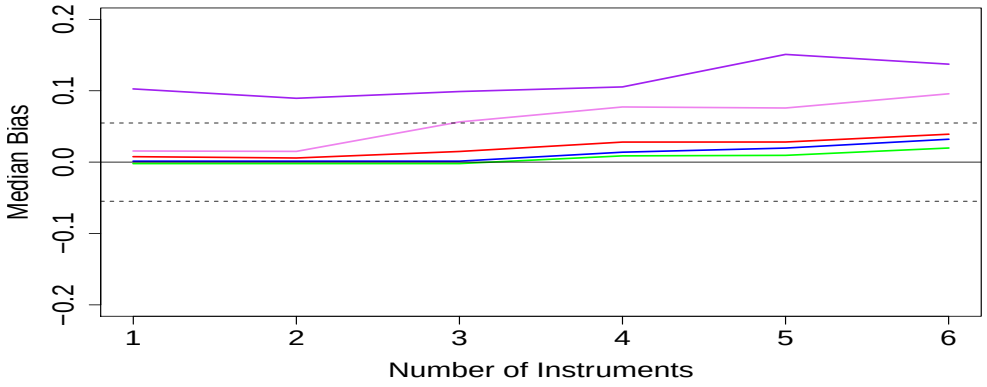
As there were identical SNPs and such variety of LD between the SNPs within the 5 sections of DNA, it is difficult to judge the optimal number of instrument for any one section. Hence, the next section will increase the number of DNA sections examined and investigate whether restricting the instruments according to their properties or applying popular guidelines could prevent extreme estimates, and hopefully give clues about how instruments should be chosen in any one sample to obtain the least bias and variation.

Table 6.2: The GENOME genetic structure for 5 sections of DNA. The correlations with SNP_c, r and r^2, the allele frequency and the number of simulations when the SNP had the lowest p-value (10,000 simulations in total).

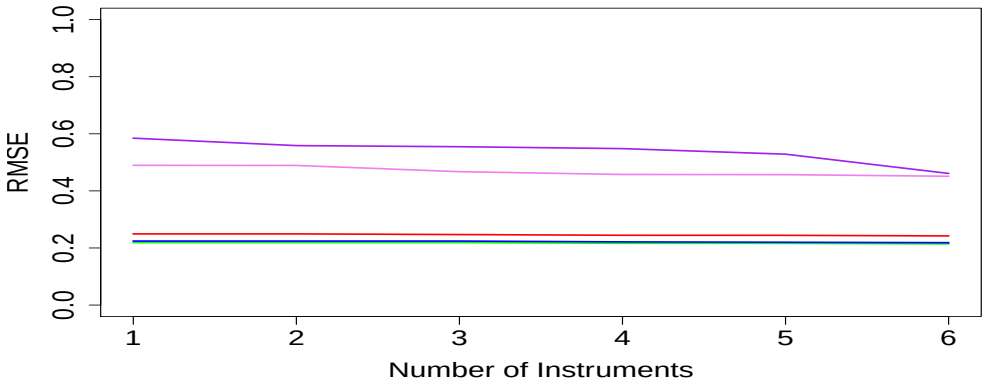
	SNP_1	SNP_2	SNP_3	SNP_c	SNP_4	SNP_5	SNP_6
DNA 1							
r	0.3363	0.3363	0.3363	1.0000	0.3465	0.4914	0.4914
r^2	0.1131	0.1131	0.1131	1.0000	0.1201	0.2414	0.2414
Allele frequency	0.5590	0.5590	0.5590	0.1254	0.5442	0.3726	0.3726
Lowest p-value	148	0	0	-	204	9648	0
DNA 2							
r	0.4917	0.4917	0.8618	1.0000	0.8979	0.8979	0.1380
r^2	0.2418	0.2418	0.7428	1.0000	0.8062	0.8062	0.0191
Allele frequency	0.0982	0.0982	0.2398	0.2040	0.2412	0.2412	0.9308
Lowest p-value	0	0	910	-	9090	0	0
DNA 3							
r	1.0000	0.4384	0.4384	1.0000	1.0000	1.0000	0.1479
r^2	1.0000	0.1922	0.1922	1.0000	1.0000	1.0000	0.0219
Allele frequency	0.8950	0.6210	0.6210	0.8950	0.8950	0.8950	0.1572
Lowest p-value	5612	0	0	-	4388	0	0
DNA 4							
r	0.2870	0.3326	0.3665	1.0000	1.0000	1.0000	1.0000
r^2	0.0824	0.1106	0.1343	1.0000	1.0000	1.0000	1.0000
Allele frequency	0.0802	0.9054	0.8874	0.5142	0.5142	0.5142	0.5142
Lowest p-value	0	0	0	-	4986	0	5014
DNA 5							
r	0.3769	0.3769	0.3769	1.0000	0.1209	0.3248	0.3811
r^2	0.1420	0.1420	0.1420	1.0000	0.0146	0.1055	0.1452
Allele frequency	0.5010	0.5010	0.5010	0.1248	0.9070	0.5748	0.4954
Lowest p-value	4275	0	0	-	370	964	4391



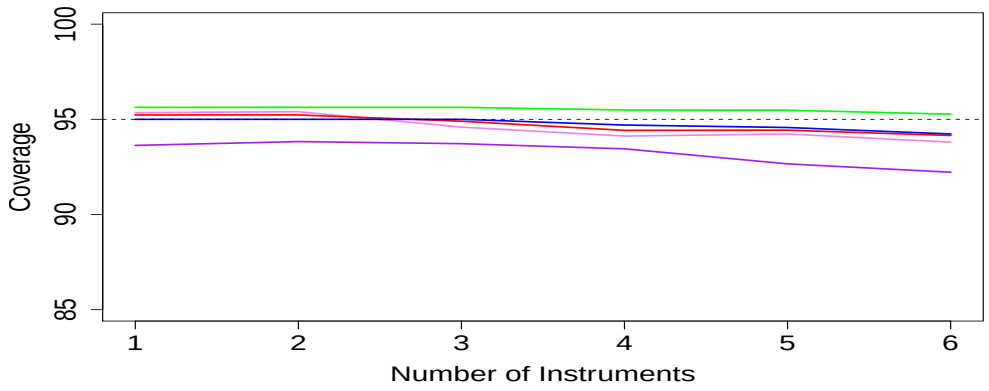
(a) Bias



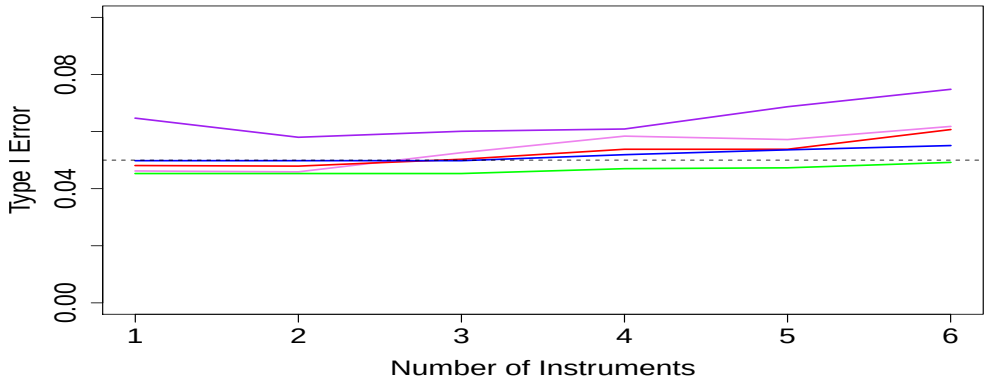
(b) Median Bias



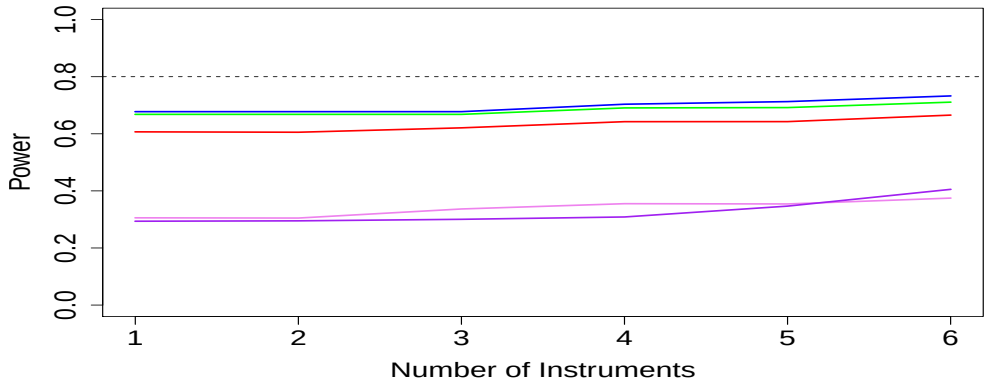
(c) RMSE



(d) Coverage



(e) Type I Error



(f) Power

Figure 6.3: Evaluation criteria of 2SLS with GENOME simulated SNPs, based on the selection of 1,2,...6 lowest p-valued SNPs and applying them jointly. Colours represents 5 different DNAs, see legend. For bias, the dotted and solid lines represent 10% and 0 bias respectively. The dotted line in coverage, type I error and power is nominal level of 95%, 5% and 0.8 respectively.

6.4 Experiment 3: Best Policy

6.4.1 Aims

So far, there is not enough evidence of a required number of instruments for an ideal estimate from 2SLS. Thus, this experiment will focus on the "best" policy in selecting instruments to estimate efficiently from 2SLS the causal relationship between outcome and risk factor of interest, within one sample. The reason for best being in inverted commas is that best is dependent on the investigators' view on which evaluation criterion is the most important, i.e. bias or RMSE.

The following policies for selecting instruments will be considered in this experiment;

- The SNP with the lowest p-value from the regression of X on each SNP.
- All of the SNPs available in the data.
- Any SNPs with p-values < 0.05 from the regression of X on each SNP.
- Any SNPs with F-statistics > 10 from the regression of X on each SNP.

As this experiment aims to identify the best selecting policy for the derivation of an MR estimate under the circumstance where only one dataset is available, the regression of X on each SNP and 2SLS will be performed on the same dataset. The first two policies act as the extreme ends of the spectrum in instrument selection. In MR studies, the instruments come from Genome-wide association studies (GWAS) where they consider the significance threshold at 5×10^{-8} [20]. However this is for hundreds of thousands to millions of tested SNPs and hence, with 6 SNPs, the third policy will use a p-value < 0.05, the commonly used statistical significance. The last policy is the guideline of avoiding the weak instrument problem for Mendelian randomisation studies suggest by Lawlor et al. [182].

The aim of this section is to examine the change in the "best" policy with different sample sizes and/or number of instruments.

6.4.2 Design

In the first scenario, 7 SNPs from the same gene are generated using GENOME as defined in Section 4.2. The middle SNP will be assumed to be causal which will explain 2% of the variation in X. X will explain 6% of the variation in Y. X and Y

will be normally distributed and with sample sizes of 5,000. The causal SNP will be discarded and the 2SLS will be based on selected non-causal SNPs. The simulation will be repeated 10,000 times.

In each simulation, X will be regressed on each SNP in turn and the SNPs will be included into 2SLS based on the criterion of each policy described above. The bias, median bias, root mean square error (RMSE), coverage and power will be derived. Type I error will be estimated by a separate simulation where X is not directly affecting Y. The method described will be repeated again for sample sizes of 10,000 to 50,000, in the steps of 10,000.

The design for the second scenario will be exactly the same apart from the first step; GENOME will simulate 21 SNPs from the same gene.

6.4.3 Results: Sample Size

Figure 6.4a to 6.4f give the six evaluation criteria for all four polices with different sample sizes. As expected the performance of 2SLS benefits with increasing sample size, for the four policies.

For all the policies, the mean and median bias declines towards zero (the solid line) as sample size increases, Figure 6.4a and 6.4b. Notice the median bias for policy of F-statistics > 10 and p-values < 0.05 is the same as the policy of all SNPs as instruments after a sample size of 20,000, but not for mean bias. This gives evidence that with increasing sample size, eventually all the SNPs have sufficient F-statistics and p-values. As sample size increases, the RMSE (Figure 6.4c), combination of bias and variation of the MR estimate, decreases for all 4 polices.

Figure 6.4d shows all the policies had approximately 95% nominal coverage (the dotted line) for sample sizes greater than 20,000. Figure 6.4e indicates that all the policies had approximately 5% significance level (the dotted line) for sample sizes over 10,000. All the policies had more than 0.8 power after sample size of 15,000, the dotted line on Figure 6.4f.

See Table appendix C.3 for the table representations of these figures.

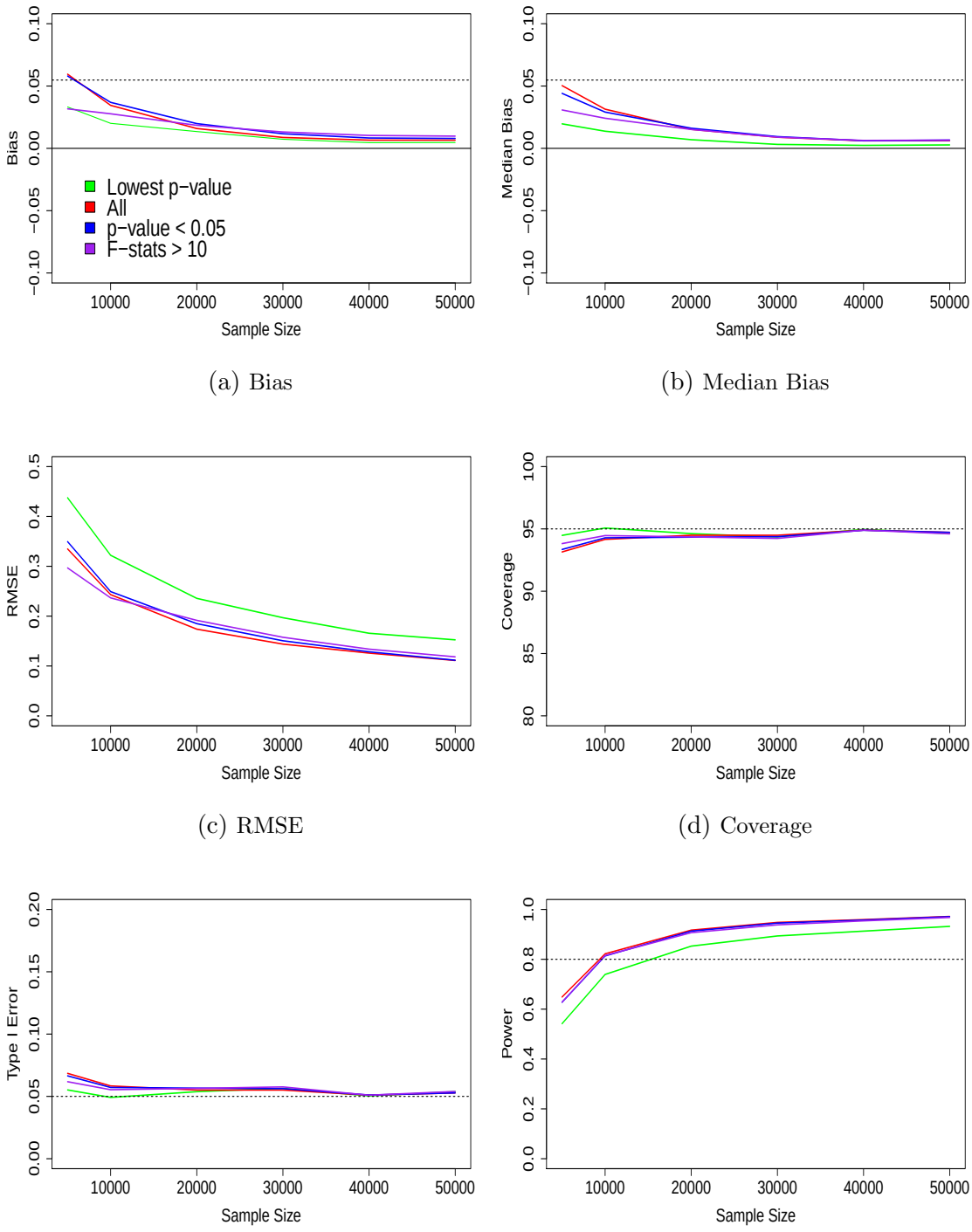
6.4.4 Results: 20 SNPs

Figure 6.5a and 6.5f are exactly the same as the above scenario, but with 20 potential instruments. For all four policies, the 6 evaluation criteria show an improvement, with increasing sample size. For the table representation of these figures, see Table

appendix C.4.

Both the median and the mean bias have deceased as the sample size increased. However unlike the bias from 6 SNPs, to achieve 10% bias (the dotted line) for all four polices, the sample size needs to be greater than 10,000. Moreover, 6 SNPs required 20,000 individuals for the polices of all the SNPs as instruments, p-values < 0.05 and F-statistics > 10 to concur, while 20 SNPs required approximately 35,000 individuals, Figure 6.5a and 6.5b. As the sample size increases, the RMSE has decreased for all policies, Figure 6.5c. For sample sizes of below 20,000, coverage is less for 20 SNPs than for 6 SNPs, Figure 6.5d. The instrument with the least p-value had approximately 95% nominal coverage for all the sample sizes. Figure 6.5e shows that not all the policies had 5% significance level. For sample size about 10,000, all the polices had power greater than 0.8 (the dotted line in Figure 6.5f), but this is misleading as type I error hasn't been controlled.

Selecting instruments by the lowest p-value indicated a decrease in weak instrument bias, as sample size increases. Eventually, all the bias is caused by selection. The difference in bias and RMSE between 6 and 20 SNPs, is due to the number of SNPs available; when there are more SNPs to choose from, the more likely to find a SNP that is highly correlated with the causal SNP and in turn more accurate estimation of the causal effect. The weak instrument bias from all the SNPs as instruments, will eventually disappear with adequate sample size. One of the causes for weak instrument bias is that not enough of the variation is explained to justify the number of instruments. This is supported by Figures 6.5a and 6.5b, where 20 SNPs have greater bias than 6 SNPs for all the sample sizes in Figures 6.4a and 6.4b. For the policies of selecting instruments by p-values < 0.05 and F-statistics > 10, the source of bias is a little more complicated. When sample size is small, these policies are affected by both selection and weak instrument bias. The selection bias disappears with adequate sample size, as all the SNPs fulfilled the various criteria, shown by the average number of instruments in 2SLS in Table 6.3 and the fact that their evaluation criteria becomes the same as the policy with all SNPs as instruments. The increase in sample size, meant all the SNPs would eventually have sufficient F-statistics, but if the number of available SNPs increases, the sample size will need to be even greater for the three policies to become equivalent.



(a) Bias

(b) Median Bias

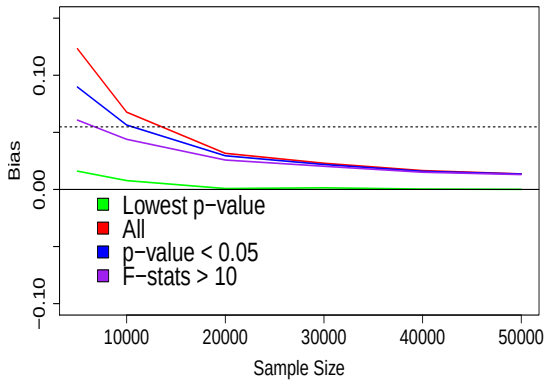
(c) RMSE

(d) Coverage

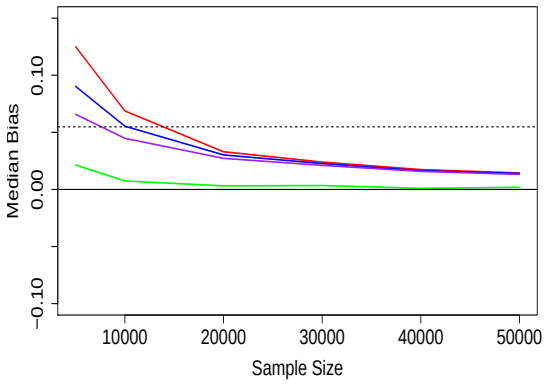
(e) Type I Error

(f) Power

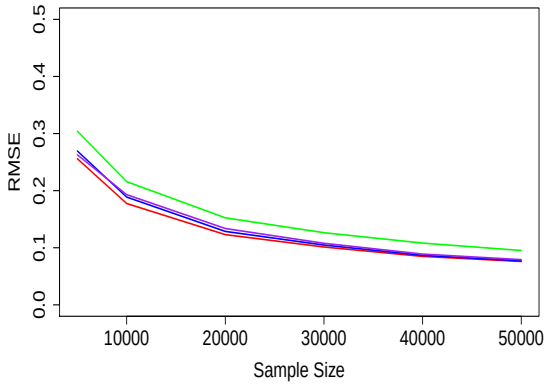
Figure 6.4: Evaluation criteria of best policy for 6 non-causal SNPs simulated from GENOME with different sample sizes. The instrument selection policies are represented by the colours of the lines, see legend. The black solid line in bias is zero bias. The dotted lines in bias, coverage and TIE represent 10% bias, 95% coverage and 5% significance level respectively.



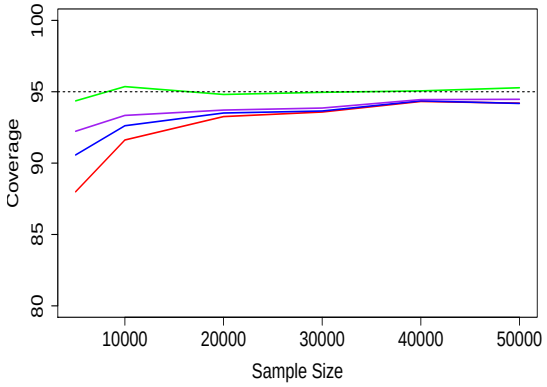
(a) Bias



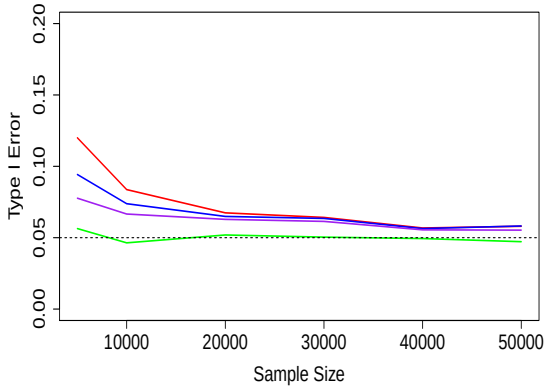
(b) Median Bias



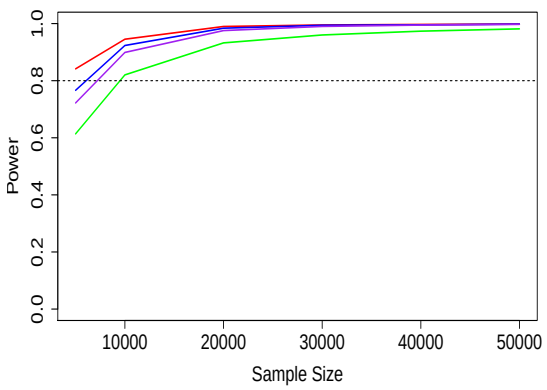
(c) RMSE



(d) Coverage



(e) Type I Error



(f) Power

Figure 6.5: Evaluation criteria of best policy for 20 non-causal SNPs simulated from GENOME with different sample sizes. The instrument selection policies are represented by the colours of the lines, see legend. The black solid line in bias is zero bias. The dotted line in bias, coverage and TIE is 10% bias, 95% coverage and 5% significance level respectively.

Table 6.3: Number of genes fulfilled the policies and average number of instruments included

	6 SNPs		20 SNPs	
	Number of genes analysed	Average number of instruments	Number of genes analysed	Average number of instruments
		n=5,000		
Lowest p-value	10000	1	10000	1
All	10000	6	10000	20
p-value < 0.05	9980	5	9980	14
F-statistic > 10	9796	3	9796	10
		n=10,000		
Lowest p-value	10000	1	10000	1
All	10000	6	10000	20
p-value < 0.05	9998	5	9998	17
F-statistic > 10	9957	4	9957	13
		n=20,000		
Lowest p-value	10000	1	10000	1
All	10000	6	10000	20
p-value < 0.05	10000	6	10000	19
F-statistic > 10	9994	5	9994	16
		n=30,000		
Lowest p-value	10000	1	10000	1
All	10000	6	10000	20
p-value < 0.05	10000	6	10000	19
F-statistic > 10	9997	6	9997	17
		n=40,000		
Lowest p-value	10000	1	10000	1
All	10000	6	10000	20
p-value < 0.05	10000	6	10000	20
F-statistic > 10	9999	6	9999	18

	n=50,000			
Lowest p-value	10000	1	10000	1
All	10000	6	10000	20
p-value < 0.05	10000	6	10000	20
F-statistic > 10	10000	6	10000	19

Conclusion

From examining the results from the 6 evaluation criteria, the "best" policy for the increasing sample size and increasing number of SNPs is the policy of lowest p-valued SNP, from the unadjusted association with X. As this policy produced MR estimates with the least bias and median bias, approximately 95% coverage and 5% significance level giving the correct power. However, there is a greater variation between the MR estimates for this particular policy, hence there is a risk of procuring a very inaccurate estimate from a sample.

The decision on "best" policy is dependent on the investigators' definition for the "best" and what is best for their sample. For example, if I had 6 SNPs from 5,000 individuals, I would apply the policy where my instruments are SNPs with F-statistics > 10, as it will give a slightly biased MR estimate. However if I had 20 SNPs, I would give a second thought on using the same policy, as it gave on average more than 10% bias.

6.5 Discussion

The performance of 2SLS with non-causal SNPs as instruments is dependent on the sample size and the instruments' correlation with the causal SNP. The inclusion of multiple instruments would decrease the variation in the MR estimate. However, in the presence of weak instrument 2SLS is biased towards the OLS [29] and many weak instrument bias will occur if the included instruments does not explain more of the variation in risk factor than the existing set of instruments [220, 227]. See Section 2.5.3 for more detailed discussion of statistical limitation caused by weak instruments. Therefore, proposing a restriction on the selection of instruments could potentially solve this problem but it comes at the cost of selection bias.

Adopting policies will result in a combination of weak instrument and selection bias. The instrument with the lowest p-value would most likely give the least bias, especially when the sample size is large or there are plenty of SNPs to choose from. However this is not always the case for any one sample. All the SNPs as instruments is only sufficient when the number of instruments corresponds to the sample size, i.e. more instruments included requires greater sample size. So far, the evaluation criteria from F-statistics > 10 (or slightly smaller F-statistics; p-values < 0.05) seems to have the best of both bias and variation of the estimate. Despite this balancing act, restricting SNPs with F-statistics > 10 can run the risk of not having any SNPs that fulfil the criterion. In addition, this particular guideline doesn't always guarantee an MR estimate with less than 10% bias in any one sample.

If time wasn't a limitation, there are several other parameter settings that could be incorporated into the experiment. For most of the experiments, the causal SNP only explains 2% of the variation in risk factor of interest. However this is fairly uncommon. GWAS have a history of discovering genetic variants that explain less than 1% of the variation in the risk factor of interest. Throughout the experiments in this chapter, I have restricted the simulations to only generate SNPs with allele frequencies to be greater than 0.05; the uncertainty in the MR estimate from 2SLS may be very large when including SNPs with allele frequency less than 0.05. The instruments used in MR studies were generally established from Genome-wide association studies (GWAS) to avoid overfitting [48]. Therefore to follow the common practice, I could have included another experiment to examine whether the bias from the MR estimate would reduce when the best instruments are found in a separate dataset, i.e. a discovery study.

To conclude, the definition of the "best" procedure is dependent on the investigator, as every policy has its advantages and disadvantages. Another equally, if not more important aspect is the characteristic of the investigator's data, i.e. the sample size and the number of SNPs available. Davies et al. [85] have recently compared different estimators with many weak instruments; in the analysis, all the SNPs were selected as instruments without a selection policy. As it is difficult to distinguish "best" procedure for any one data, the next chapter will investigate the efficiency of other estimators with many dependent SNPs.

Chapter 7

A Comparison of Estimators

7.1 Introduction

The previous chapter concluded that for many dependent instruments, the causal effect estimate from two-stage least squares (2SLS) was severely biased if all of the genetic variants were included individually without instrument selection. Chapter 3 found many econometricians have recommended alternatives to 2SLS, some of which aim to reduce the many weak instruments bias. These include limited information maximum likelihood (LIML) [61], continuously updating estimator (CUE) [207] and two-step generalised method of moments (GMM) [126]. Davies et al. [85] have recently published a comparison of these estimators in the framework of Mendelian randomisation (MR) with many *independent* weak instruments and found that LIML is the most efficient estimator in the case of homoskedasticity. This chapter aims to find a better estimator of the causal effect for many *dependent* instruments by comparing 2SLS, GMM, CUE and LIML.

The genetic information in Section 2.4 suggests an average gene should approximately have 200 SNPs and 90 of them have MAF greater than 0.1 [258]. Hence, this chapter will consider scenarios where genotypes of 10 to 90 SNPs typical of an average length gene are available, with MAF greater than 0.1. SNPs from the same gene are more likely to be inherited together and therefore are correlated within the population. None of these SNPs has a direct effect on but is associated with the risk factor of interest through their correlation with the functional SNP, which will be referred to as the causal SNP. 2SLS, GMM, CUE and LIML will be assessed by including the non-causal SNPs as instruments, without selection. The previous chapter found using a selection criteria introduces bias in 2SLS and selecting

instruments with F-statistic > 10 does not always guarantee an accurate estimate of the causal effect, perhaps one of these alternatives will be able to give unbiased estimates of the causal effect without instrument selection.

The summary of each experiment is as follows; the first experiment will investigate the effect of MAF on the four estimators. The SNPs will have patterned LD, i.e. their correlations will depend on their distances from the causal SNP. In view of the variety in the structure of SNPs (see for example the regional plots in Schizophrenia Psychiatric Genome-Wide Association Study Consortium et al. [247]), the second experiment will examine whether there is a difference in performance with four distinct genetic patterns. Each pattern will differ in correlation structure with the causal SNP. The simulations so far have assumed that the correlation between SNPs depends on their physical proximity. However it is also dependent on other factors such as recombination rate and genealogy tree [258]. Hence, the third experiment will use the GENOME simulator [185], where it integrates these correlation factors into the simulation of SNP haplotypes. Since GENOME simulates realistic genetic patterns, it does not allow users to control the genetic properties (i.e. MAF and LD). In order to distinguish the effect of each genetic factor on the estimators, I will use my own code for SNP simulation for the first and second experiment.

Each experiment will have its own design, results and conclusion section. The simulation method will follow Chapter 4, with minor changes reported in the design sections. The results will be presented graphically with corresponding tables in the Appendix. The chapter ends with a discussion of the conclusions from each experiment, limitations and implications for the next chapter.

7.2 Experiment 1: Minor Allele Frequency

7.2.1 Aims

The aim of this simulation experiment is to determine the most efficient estimator for handling many dependent instruments. MAF could potentially affect their performance. Hence, the secondary aim is to investigate the effect of MAF on the four estimators.

7.2.2 Design

The genotypes of 10 to 90 SNPs from 2,000 individuals will be available as instruments. 2,000 was chosen, as GRAPHIC Study [274] has approximately 2,000 individuals and the dataset will be used in Section 10.5. The SNPs will have patterned LD, where their correlation with each other will be between 0 and 1, depending on their physical proximity, i.e. as the distance between one SNP to the other tends to 0, the correlation between them tends to 1, as described in Section 4.2. MAF for all the non-causal SNPs is 0.45 and for the causal SNP is 0.5. The middle SNP will be assumed as causal and explains 2% of the variation in the risk factor (X). X will explain 6% of the variation in the outcome (Y). X and Y will be normally distributed with sample size of 2,000. The causal SNP will be discarded and the comparison of estimators will be based on all of the non-causal SNPs as instruments. The simulation is repeated 10,000 times.

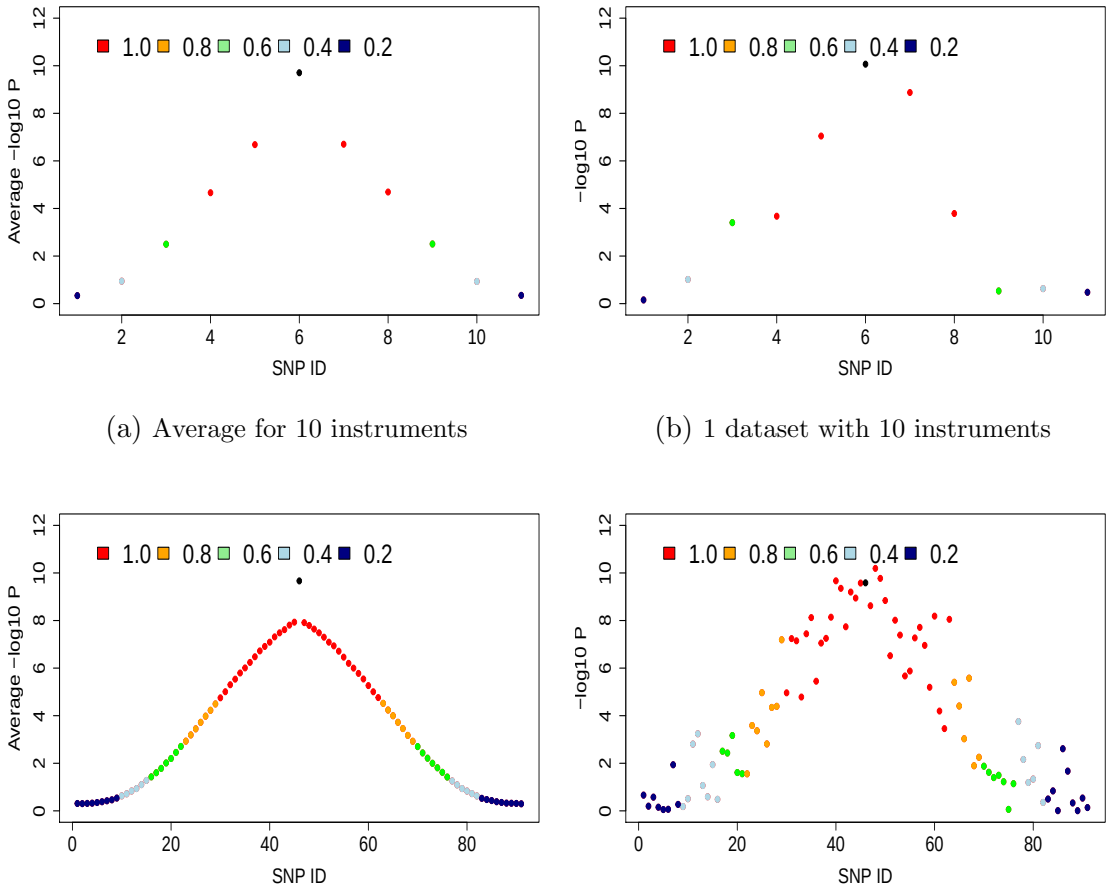
The Winsorised bias and root mean squared error (RMSE), percentage of outliers, coverage and power will be calculated from the simulations. A separate simulation on which X is not directly related to Y will be used to estimate the type I error (TIE). See Chapter 4 for the definition of these evaluation criteria.

Two further simulations will identify whether the MAF of the instruments will influence the performance of the four estimators; the next scenario will change the MAF for non-causal and causal SNPs to 0.1 and 0.05 respectively. The third scenario will investigate SNPs with MAF that varies between 0.1 to 0.5 as instruments, and the causal SNP will have MAF of 0.5.

7.2.3 Results

Common allele

Figure 7.1 gives the regional plot where SNP ID minics the SNP positions in a gene region and $-log_{10}P$ is the $-log_{10}$ of the p-value from the association for X on each SNP. The colours demonstrate their correlation with the causal SNP (positioned in the middle); the further the SNP is from the middle, the weaker the correlation. Figures 7.1a and 7.1c present the averages of p-values and correlations from 10,000 datasets for 10 and 90 SNPs respectively. Figures 7.1b and 7.1d demonstrate the variety of patterns within the simulation, showing p-values and correlations for 10 and 90 SNPs, from a random sample.



(a) Average for 10 instruments

(b) 1 dataset with 10 instruments

(c) Average for 90 instruments

(d) 1 dataset with 90 instruments

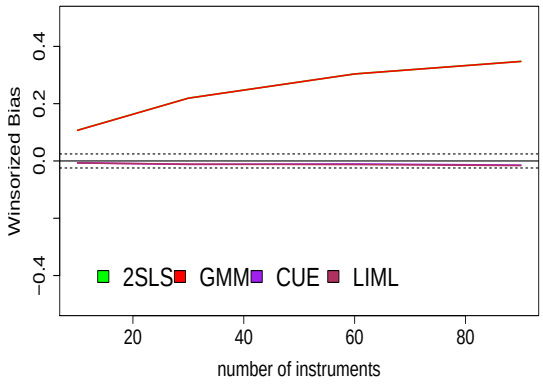
Figure 7.1: The regional association plots for 10 and 90 SNPs with MAF of 0.45, from the average of 10,000 datasets and also a random dataset. The p-value (P) is the mean p-value from the regression of each SNP on X. On the x-axis is SNP ID mimicking chromosome position, and on the y-axis for 10,000 datasets is average -$\log_{10} P$ and a single dataset is -$\log_{10} P$. Colour coding (from red to navy) denotes strong to weak correlation with the causal SNP, see the legend; 1.0 are correlations less than 1 and greater 0.8, and 0.8 are correlations less than 0.8 and greater than 0.6, so on and so forth.

Figures 7.2a to 7.2e give the evaluation criteria for the four estimators with 10 to 90 instruments. As expected, under the condition of homoskedasticity from simulated data, the performance from 2SLS and GMM are equivalent [116]. The results seen here are in agreement with the previous chapters; 2SLS becomes biased when not enough of the variation is explained by the number of instruments. Winsorised bias from LIML and CUE overlap (Figures 7.2a), where they both have within 10%

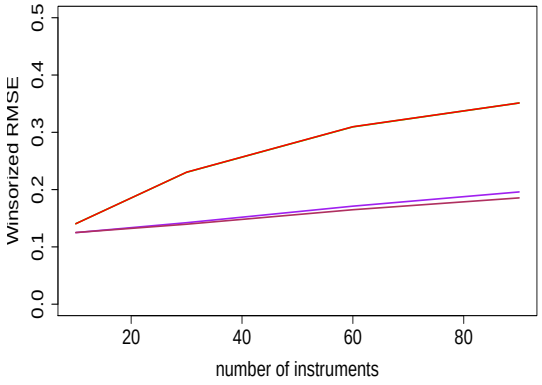
bias, as LIML and CUE ignore nuisance instruments when there are enough strong instruments. Winsorisation is discussed in Section 4.4. The Winsorised standard error of the bias (Appendix D.1), demonstrates the increased bias in 2SLS and GMM over LIML and CUE is not due to sampling error, as their 95% confidence intervals do not overlap. Unsurprisingly, 2SLS and GMM become more biased as the number of instruments increases, and biases are similar for CUE and LIML, except for 10 instruments. The increase in bias from 10 to 30, 60 and 90 instruments suggests CUE and LIML suffer from many weak instruments bias to a certain extent.

The Winsorised RMSE becomes worse as number of instruments increases. Note the accuracy of RMSE was approximately $\mp 0.1\%$ for 10,000 datasets, see Section 4.5. CUE and LIML have the lowest RMSE, Figure 7.2b, but have higher percentage of outliers (Figure 7.2c); this show with many weak instruments CUE and LIML can give extreme estimates. The outliers are caused by the approximately zero SNP association with X (β_{ZX}), consider causal effect estimation from one instrument would be very large ($\approx \frac{\beta_{ZY}}{\beta_{ZX}}$), if the numerator is almost zero. Figure 7.2c shows the percentage of outliers for 2SLS and GMM is decreasing with increasing instruments. This is due to the reduction in the variation of causal effect estimate between datasets. Note that while the percentage of outliers have decreased with increasing instruments, the causal effect estimates are more biased.

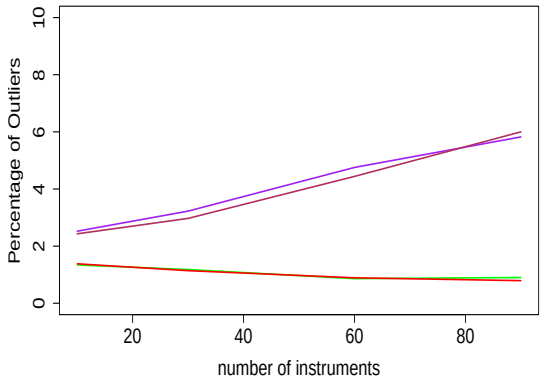
All of the estimators drop from 95% coverage (the dotted line) as the number of instruments increases, LIML stays close to the nominal coverage for all numbers of instruments. Coverage has $\mp 0.4\%$ accuracy from 10,000 datasets, hence the differences between estimators are not from sampling error. Figure 7.2e gives the type I error (TIE); all of the estimators increase from 5% significance level as the number of instruments increases. Therefore the power in Figure 7.2f is misleading; power of over 0.8 from 2SLS and GMM will need be adjusted for TIE of over 0.4.



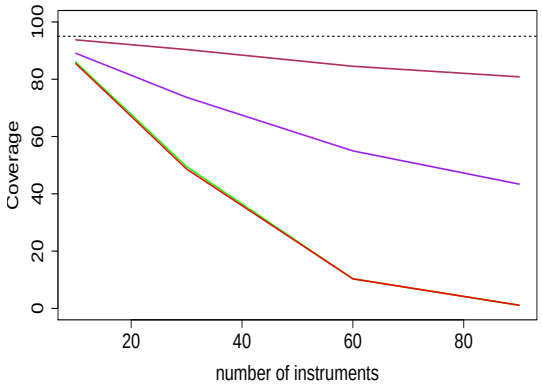
(a) Winsorized Bias



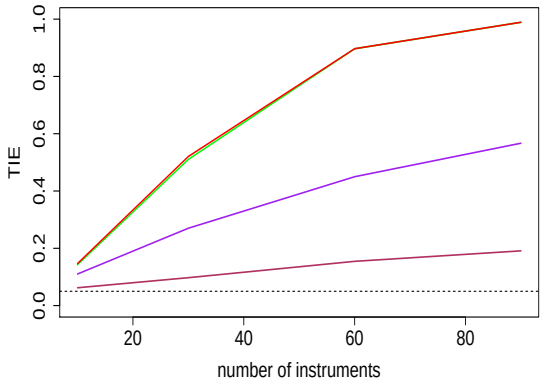
(b) Winsorized RMSE



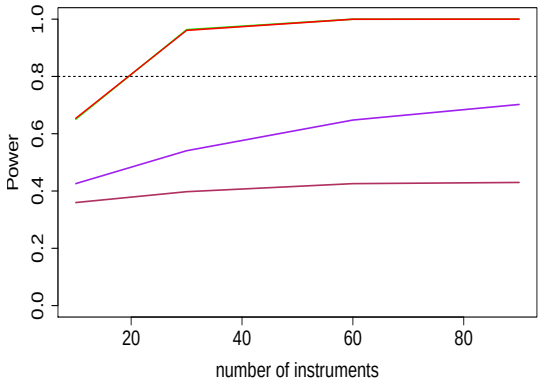
(c) Outlier



(d) Coverage



(e) Type I Error



(f) Power

Figure 7.2: Evaluation criteria for including SNPs with the same common MAF in 2SLS, two-step GMM, CUE and LIML. The estimators are coloured as green, red, purple and maroon respectively. The black solid line in Winsorised bias is zero bias. The dotted line in Winsorised bias, coverage and TIE is 10% bias, 95% coverage and 5% significance level respectively. Note 2SLS and two-step GMM are identical under homoscedasticity.

Low minor allele frequency

Figure 7.3 show the regional plot of the average -$\log_{10}$ p-value from the association with X and correlation with the causal SNP for 90 instruments. As shown by the regional plot, the further the SNP is from the causal SNP (coloured black) the weaker the correlation and lower the -$\log_{10}$ p-value. The variation explained by the non-causal SNPs is the same as the previous section, but instead of having MAF of 0.5, they have MAFs of 0.1 and the causal SNP has MAF of 0.05.

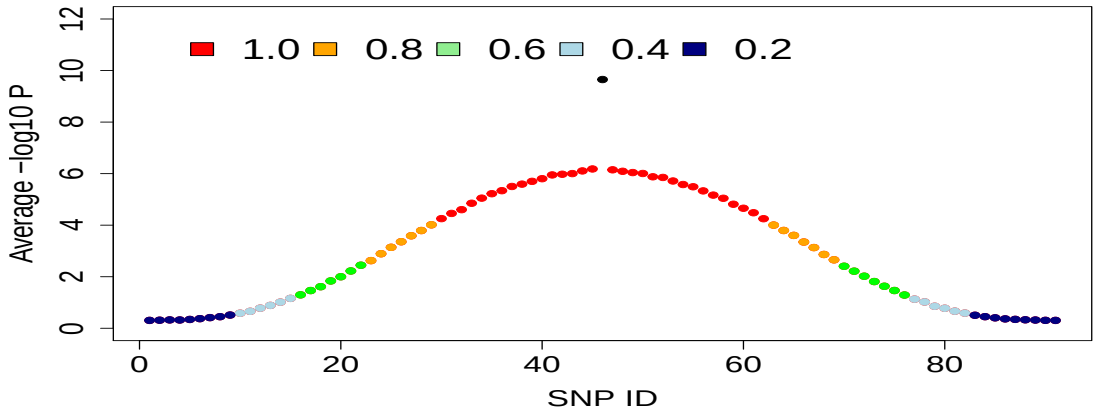


Figure 7.3: The regional association plots for 90 SNPs with MAF of 0.05, on simulated X in 10,000 datasets. The p-value (P) is the mean p-value from the regression of each SNP on X. On the x-axis is SNP ID mimicking chromosome position, and on the y-axis is -$\log_{10} P$. Colour coding (from red to navy) denotes strong to weak correlation with the causal SNP, see the legend; 1.0 are correlations less than 1 and greater 0.8, and 0.8 are correlations less than 0.8 and greater than 0.6, so on and so forth.

Table 7.1 is the comparison of performance when including SNPs with common and low MAF in the four estimators. As for the previous case with common MAF, CUE and LIML are more efficient in terms of Winsorised bias and RMSE when compared to 2SLS and GMM. Note that the change from negative to positive Winsorised bias in CUE does not mean there is an optimum number of instruments for zero bias, this only demonstrates that there are too many extremely positive estimates in CUE for Winsorisation to exclude. This is evident from the greater Winsorised RMSE (0.06% accuracy from 10,000 datasets), percentage of outliers, and the increase in CUE's non-convergence with increasing instruments (Appendix D.3).

All of the evaluation criteria of each estimator deteriorated from common to low

MAF, except for Winsorised bias which remains similar, Table 7.1. Power has been left in Appendix D.3 as it is misleading without the adjustment for the increasing TIE. CUE struggled more to minimize its function in the low MAF case. For 90 instruments 259 datasets could not converge compared with 56 for the common MAF case, see Appendix D.1 and D.3 .

Table 7.1: The evaluation criteria when including SNPs with the same common (Com.) and low MAF, in 2SLS, CUE and LIML. Inst., and S.E. are Instruments and Standard Error of winsorised average respectively.

	Inst.	2SLS	GMM	CUE	LIML
Winsorised Bias(S.E.)					
Com.	10	0.1074 (0.0009)	0.1069 (0.0009)	-0.0064 (0.0012)	-0.0071 (0.0012)
	30	0.2193 (0.0007)	0.2191 (0.0007)	-0.0120 (0.0014)	-0.0113 (0.0014)
	60	0.3039 (0.0006)	0.3037 (0.0006)	-0.0105 (0.0017)	-0.0124 (0.0016)
	90	0.3477 (0.0005)	0.3474 (0.0005)	-0.0146 (0.0020)	-0.0159 (0.0018)
Low	10	0.1355 (0.0010)	0.1362 (0.0010)	-0.0101 (0.0015)	-0.0125 (0.0015)
	30	0.2616 (0.0008)	0.2622 (0.0008)	-0.0036 (0.0018)	-0.0088 (0.0017)
	60	0.3367 (0.0006)	0.3378 (0.0006)	0.0028 (0.0023)	-0.0157 (0.0021)
	90	0.3733 (0.0005)	0.3740 (0.0005)	0.0324 (0.0027)	-0.0099 (0.0024)
Winsorised RMSE					
Com.	10	0.1405	0.1404	0.1249	0.1250
	30	0.2302	0.2304	0.1422	0.1396
	60	0.3095	0.3097	0.1711	0.1649
	90	0.3512	0.3512	0.1959	0.1855
Low	10	0.1698	0.1712	0.1544	0.1538
	30	0.2722	0.2735	0.1812	0.1747
	60	0.3421	0.3435	0.2309	0.2140
	90	0.3769	0.3780	0.2635	0.2375
Percentage of Outlier					
Com.	10	1.34	1.38	2.52	2.43
	30	1.18	1.14	3.22	2.97
	60	0.86	0.89	4.75	4.44

Continued on next page

Table 7.1 – *Continued from previous page*

		Inst.	2SLS	GMM	CUE	LIML
		90	0.90	0.79	5.82	6.00
	Low	10	1.50	1.47	3.58	3.42
		30	1.30	1.29	5.86	5.25
		60	1.02	1.07	7.29	6.34
		90	1.01	1.03	8.87	8.65
Coverage	**Com.**	10	85.72	85.33	89.01	93.76
		30	48.91	47.94	73.01	90.24
		60	10.08	9.85	54.02	84.37
		90	1.05	1.05	42.09	80.33
	Low	10	83.55	82.78	86.03	93.53
		30	41.07	39.65	63.58	88.66
		60	6.78	6.43	42.93	82.73
		90	0.70	0.76	32.72	77.47
Type I Error	**Com.**	10	0.1427	0.1467	0.1102	0.0622
		30	0.5104	0.5206	0.2701	0.0974
		60	0.8966	0.8968	0.4499	0.1545
		90	0.9890	0.9891	0.5669	0.1911
	Low	10	0.1645	0.1722	0.1420	0.0645
		30	0.5891	0.6035	0.3686	0.1130
		60	0.9321	0.9357	0.5727	0.1726
		90	0.9930	0.9924	0.6778	0.2250

Varying minor allele frequency

Figure 7.4 is the regional plot for 90 SNPs, which includes their association with X (p-value) and their correlation with the causal SNP. The variation explained by the

non-casual SNPs are the same as the common allele section, i.e. variation explained is dependent on their correlation with the causal SNP. The difference is that the MAF of the non-causal SNPs, instead of all having MAF of 0.5, this section allows each SNP to vary between 0.1 and 0.5.

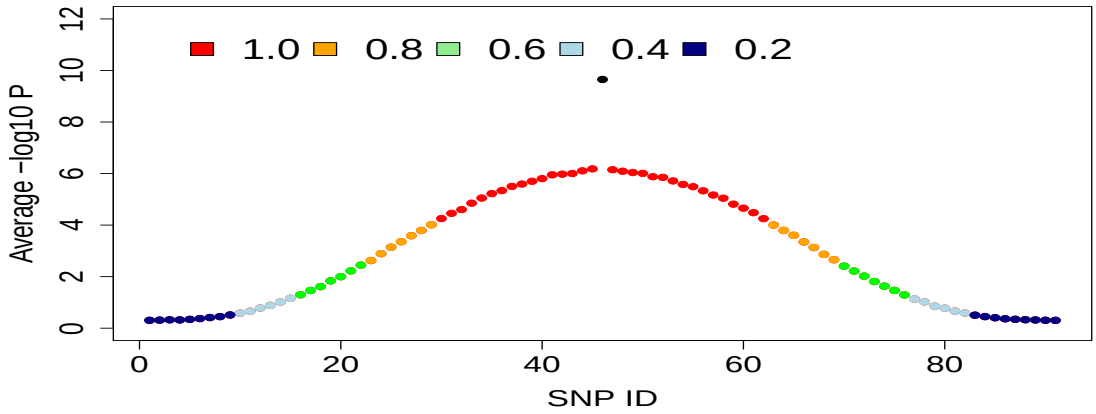


Figure 7.4: The regional association plots for 90 SNPs with the variable MAF between 0.1 and 0.5, on simulated X in 10,000 datasets. The p-value (P) is the mean p-value from the regression of each SNP on X. On the x-axis is SNP ID mimicking chromosome position, and on the y-axis is -$\log_{10} P$. Colour coding (from red to navy) denotes strong to weak correlation with the causal SNP, see the legend; 1.0 are correlations less than 1 and greater 0.8, and 0.8 are correlations less than 0.8 and greater than 0.6, so on and so forth.

Table 7.2 show similar conclusions for SNPs with common and low MAF; 2SLS and GMM have higher Winsorised bias and RMSE in comparison to CUE and LIML for 10 to 90 instruments. However, the percentage of outliers from 2SLS and GMM is less than LIML and CUE. LIML was closest to 95% nominal coverage and 5% significance level for all the different numbers of instruments.

Unsurprisingly the difference in evaluation criteria between including SNPs with common and variable MAF are minimal. Since there is a mixture of SNPs with common and low MAF, LIML and CUE are more certain to obtain estimates from the strong instruments and ignore the nuisance instruments that suffers from finite sample bias. 10,000 datasets give approximately $\mp 0.1\%$, $\mp 0.5\%$ and $\mp 1\%$ accuracy to RMSE, Coverage and type I error respectively.

Table 7.2: The evaluation criteria when including SNPs with the same common (Com.) and variable (Var.) MAF, in 2SLS, CUE and LIML. Inst., and S.E. are Instruments and Standard Error of winsorised average respectively.

		Inst.	2SLS	GMM	CUE	LIML
Win. Bias (S.E.)	**Com.**	10	0.1074 (0.0009)	0.1069 (0.0009)	-0.0064 (0.0012)	-0.0071 (0.0012)
		30	0.2193 (0.0007)	0.2191 (0.0007)	-0.0120 (0.0014)	-0.0113 (0.0014)
		60	0.3039 (0.0006)	0.3037 (0.0006)	-0.0105 (0.0017)	-0.0124 (0.0016)
		90	0.3477 (0.0005)	0.3474 (0.0005)	-0.0146 (0.0020)	-0.0159 (0.0018)
	Var.	10	0.1130 (0.0010)	0.1133 (0.0010)	-0.0127 (0.0014)	-0.0128 (0.0014)
		30	0.2287 (0.0007)	0.2291 (0.0007)	-0.0083 (0.0015)	-0.0081 (0.0015)
		60	0.3069 (0.0006)	0.3062 (0.0006)	-0.0172 (0.0018)	-0.0183 (0.0017)
		90	0.3488 (0.0005)	0.3489 (0.0005)	-0.0071 (0.0020)	-0.0097 (0.0019)
Win. RMSE	**Com.**	10	0.1405	0.1404	0.1249	0.1250
		30	0.2302	0.2304	0.1422	0.1396
		60	0.3095	0.3097	0.1711	0.1649
		90	0.3512	0.3512	0.1959	0.1855
	Variable	10	0.1491	0.1495	0.1364	0.1359
		30	0.2400	0.2407	0.1490	0.1463
		60	0.3125	0.3122	0.1795	0.1724
		90	0.3524	0.3527	0.1988	0.1859
Percentage of Outlier	**Com.**	10	1.34	1.38	2.52	2.43
		30	1.18	1.14	3.22	2.97
		60	0.86	0.89	4.75	4.44
		90	0.90	0.79	5.82	6.00
	Variable	10	1.35	1.37	2.62	2.72
		30	0.91	0.89	3.68	3.69
		60	0.84	0.85	4.78	4.57

Continued on next page

Table 7.2 – *Continued from previous page*

		Inst.	2SLS	GMM	CUE	LIML
		90	0.74	0.77	5.95	6.15
Coverage	**Com.**	10	85.72	85.33	89.01	93.76
		30	48.91	47.94	73.01	90.24
		60	10.08	9.85	54.02	84.37
		90	1.05	1.05	42.09	80.33
	Variable	10	85.10	84.88	88.09	93.52
		30	46.69	45.48	70.95	89.50
		60	9.59	9.36	52.43	83.95
		90	1.00	1.04	41.97	80.03
Type I Error	**Com.**	10	0.1427	0.1467	0.1102	0.0622
		30	0.5104	0.5206	0.2701	0.0974
		60	0.8966	0.8968	0.4499	0.1545
		90	0.9890	0.9891	0.5669	0.1911
	Variable	10	0.1486	0.1512	0.1192	0.0648
		30	0.5330	0.5452	0.2906	0.1049
		60	0.9041	0.9064	0.4760	0.1601
		90	0.9900	0.9896	0.5808	0.1994

7.2.4 Conclusions

In the comparison of estimators, LIML had within 10% bias, lowest RMSE, nominal coverage, and closest to the 5% significance level. Even though LIML gives the least bias for many instruments, it does suffer from finite sample bias in SNP association with X, evident from the increase in standard error of the bias from common to low MAF. Hence, LIML is more likely to estimate an extreme value with rare SNPs. CUE also demonstrated the same property.

7.3 Experiment 2: Patterns

7.3.1 Aims

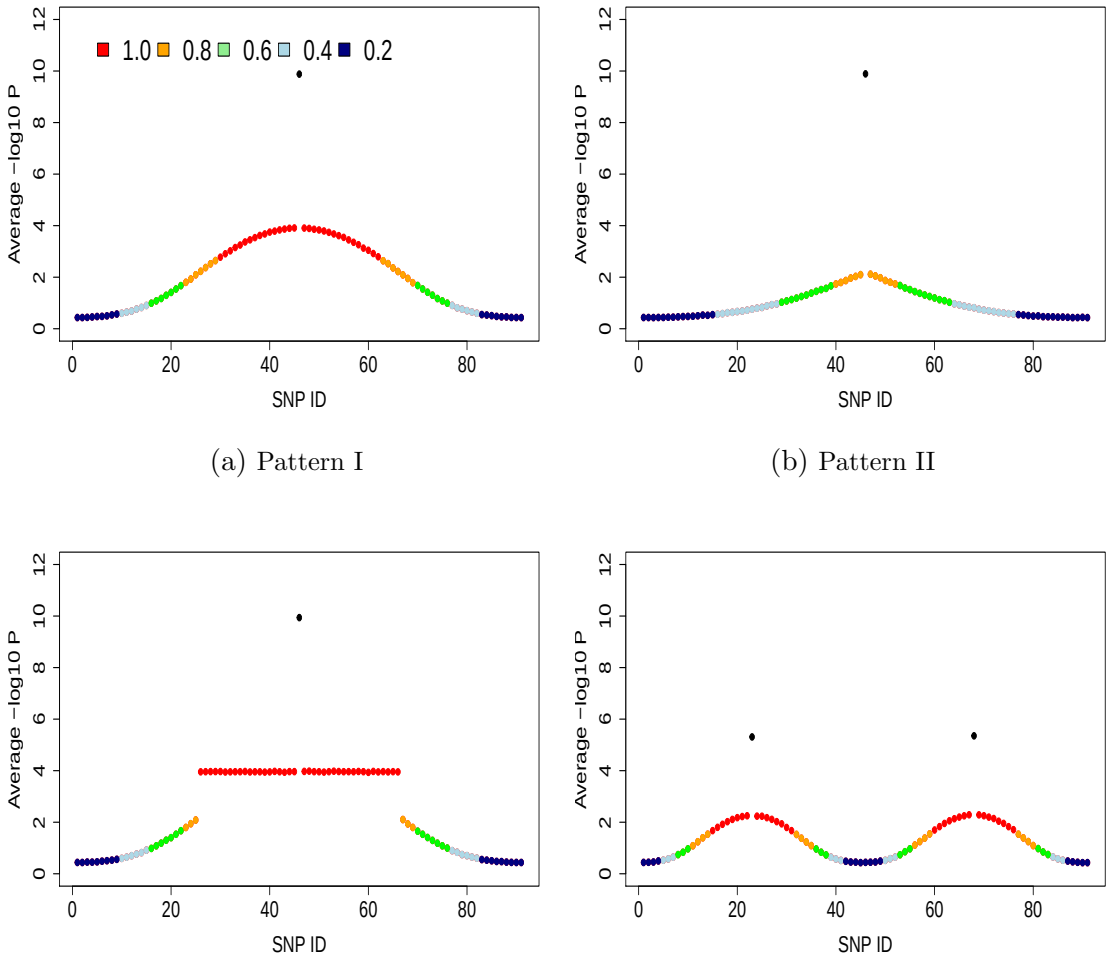
The regional plots in Section 7.2 are not the only genetic patterns seen in GWAS. The correlation between SNPs do not always distribute themselves evenly according to their genetic position. This section will examine the performance of 2SLS, GMM, CUE and LIML for other genetic patterns. The performance of each estimator will be based on instruments that are non-causal SNPs. The aim of this experiment will be to identify the difference in performance between the four estimators in four distinct genetic patterns, as illustrated in Figure 7.5d.

7.3.2 Design

In all four patterns, the minor allele frequency (MAF) for the non-causal SNPs will be variable between 0.1 and 0.5, and will be 0.5 for the causal SNP. For Patterns I to III, the middle SNP will be assumed to be causal and will explain 2% of the variation in the risk factor (X). Pattern IV has two causal SNPs, each explaining 1% of the variation in X and thus the p-values are lower in comparison to the other patterns. X will explain 6% of the variation in the outcome (Y). X and Y will be normally distributed with sample size of 2,000. The causal SNP will be discarded and the instruments included in 2SLS, GMM, CUE and LIML will be the non-causal SNPs. The simulation will be repeated 10,000 times.

7.3.3 Results

Figure 7.5 shows the regional plot of each pattern; the SNP correlations with the causal SNP (coloured black) decline as the distance increases in Pattern I, and hence the p-value from the association of X decreases. In Pattern II, the SNP correlations with the causal SNP are considerably weaker than Pattern I. Pattern III has a plateau effect where SNPs closest to causal SNP have similar correlations with a sharp drop at recombination points. Pattern IV has two causal SNPs, each explained 1% of the variation in X and thus the p-values are lower in comparison to the other patterns.



(a) Pattern I

(b) Pattern II

(c) Pattern III

(d) Pattern IV

Figure 7.5: The regional association plots for the four patterns with 90 SNPs of 10,000 simulations. The p-value is the mean p-value from the regression of each SNP on X. On the x-axis is SNP ID mimicking chromosome position, and on the y-axis is -$\log_{10} P$. The black dot is the causal SNP. Colour coding (from red to navy) denotes strong to weak correlation with the causal SNP, see the legend; 1.0 are correlations less than 1 and greater 0.8, and 0.8 are correlations less than 0.8 and greater than 0.6, so on and so forth.

Winsorised Bias

Figure 7.6 shows the Winsorised bias for Patterns I to IV. 2SLS and GMM estimates are more biased than CUE and LIML, and bias increases with number of instruments included in the algorithm. Estimates from CUE for Patterns I, III and IV have stayed within 10% bias for all instruments, however for Pattern II its bias is no

longer within 10%, caused by the greater number of weak instruments. LIML is a slightly more efficient estimator as its bias remains the same as the number of instruments increases, even for Pattern II, shown by the Winsorised S.E. for bias in Appendix D.4 to D.7. The change from negative to positive bias in CUE for all the patterns is due to the increase in number and magnitude of extreme estimates with number of instruments (as shown by the increasing Winsorised S.E., RMSE and percentage of outliers). As there are too many extreme estimates Winsorisation can not trim enough of them to not influence the mean causal effect estimate. The same explanation applies to the change of negative to positive bias in LIML for Pattern II.

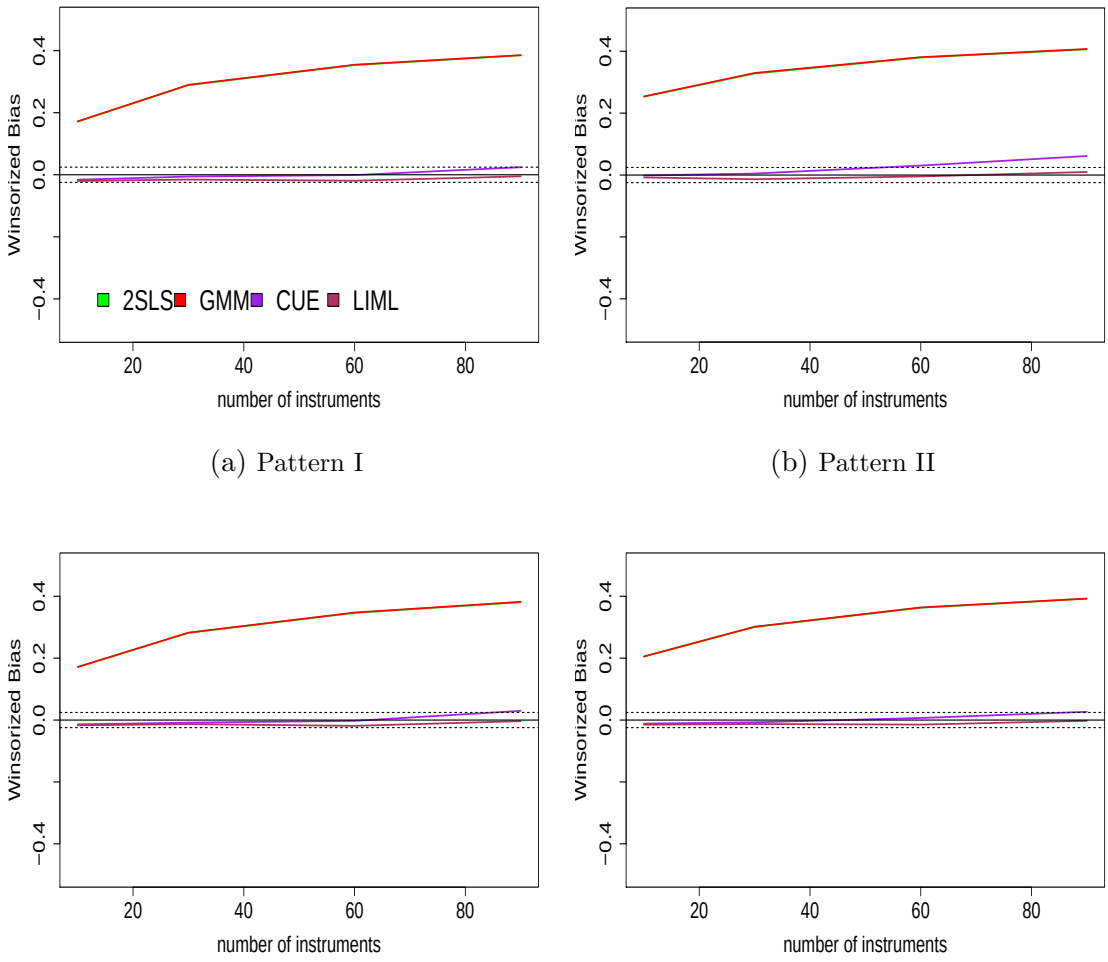
Winsorised RMSE

The Winsorised RMSE from 2SLS, GMM, CUE and LIML for all four patterns are shown in Figure 7.7. Patterns I and III have slightly better RMSE from the 4 estimators. This is because of the simulation method; these patterns are more likely to generate strong instruments. This is contrast with the higher RMSE for Pattern II in Figure 7.7b where there are weaker instruments. The RMSE increases with number of instruments for all four algorithm, this is because the proportion of weak instruments has increased where the estimators becomes less certain.

These plots show that for each pattern the CUE and LIML have lower RMSE in comparison to 2SLS and GMM, except for Pattern II with 10 instruments; CUE and LIML are efficient when there are enough strong instruments for the estimators to ignore the random noise caused by the weak instruments [253] which follows the decrease in RMSE with 30 instruments, the variation of causal effect estimates between datasets reduces, as the chance of having a strong instrument out of 30 is higher than 10 for every datasets. However, RMSE increases again after 30, as there are more weak instruments.

When comparing RMSE from LIML and CUE, LIML seem to be in a slightly better position; this is caused by finite sample bias in the first stage regression. Even though LIML has a similar weighting algorithm to CUE, the LIML minimising function is equivalent to 2SLS divided by variance of the OLS estimates. Since 2SLS is biased towards OLS with weak instruments, LIML avoids the ratio of 1 which reduces the bias to a certain extent. On the other hand CUE cannot distinguish SNPs with genuine correlation with the causal SNP, resulting from the finite sample bias in the first-stage regression. Then uncorrelated SNPs that have zero association

with X will be given the same weighting as correlated SNPs, and approximating extreme estimate as a result.



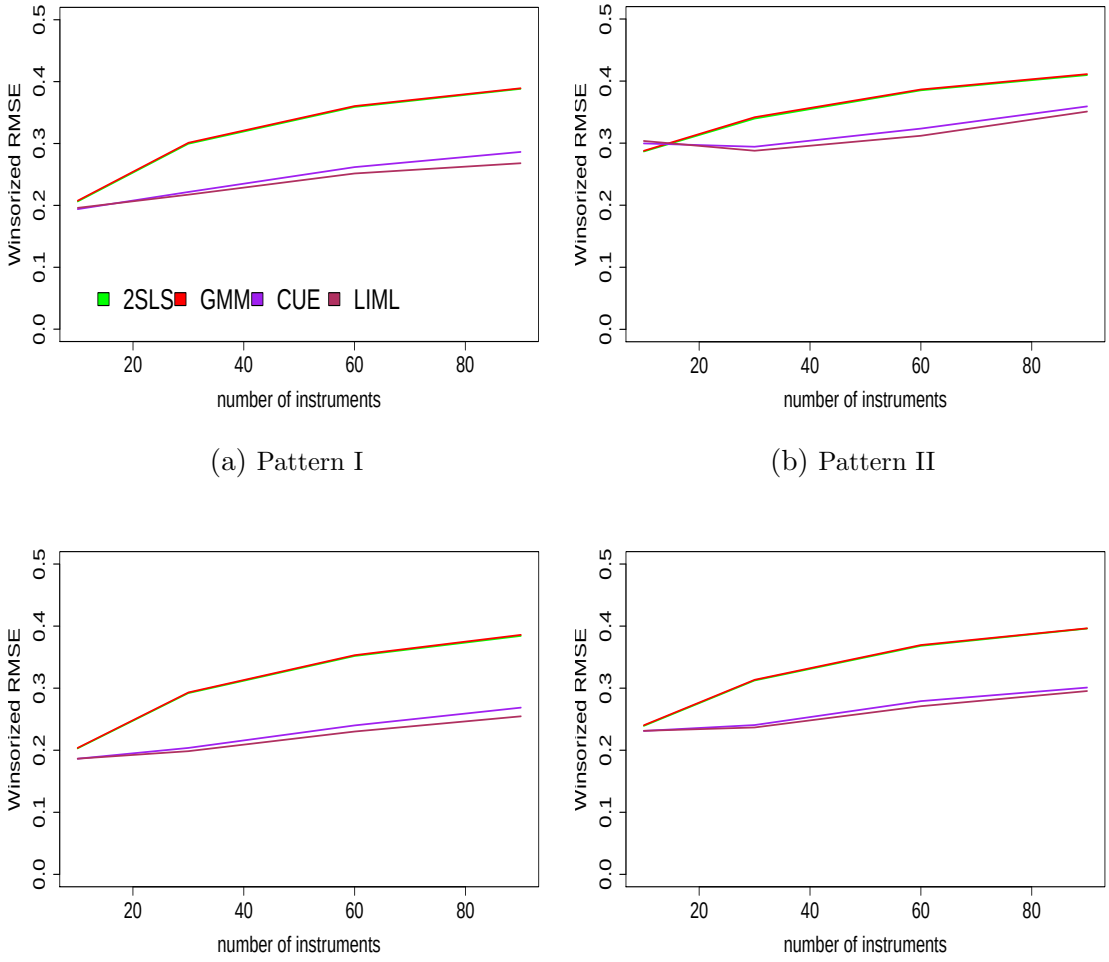
(a) Pattern I

(b) Pattern II

(c) Pattern III

(d) Pattern IV

Figure 7.6: Winsorised bias from 2SLS, two-step GMM, CUE and LIML for all four Patterns. The estimators are coloured as green, red, purple and maroon respectively. The estimation from 2SLS and two-step GMM are similar, therefore the green and the red are superimposed. The black solid line in each plot is zero bias and the dotted line is 10% bias.



(a) Pattern I

(b) Pattern II

(c) Pattern III

(d) Pattern IV

Figure 7.7: Winsorised RMSE from 2SLS, two-step GMM, CUE and LIML for all four Patterns. The estimators are coloured as green, red, purple and maroon respectively. The estimation from 2SLS and two-step GMM are similar, therefore the green and the red are superimposed.

Percentage of Outliers

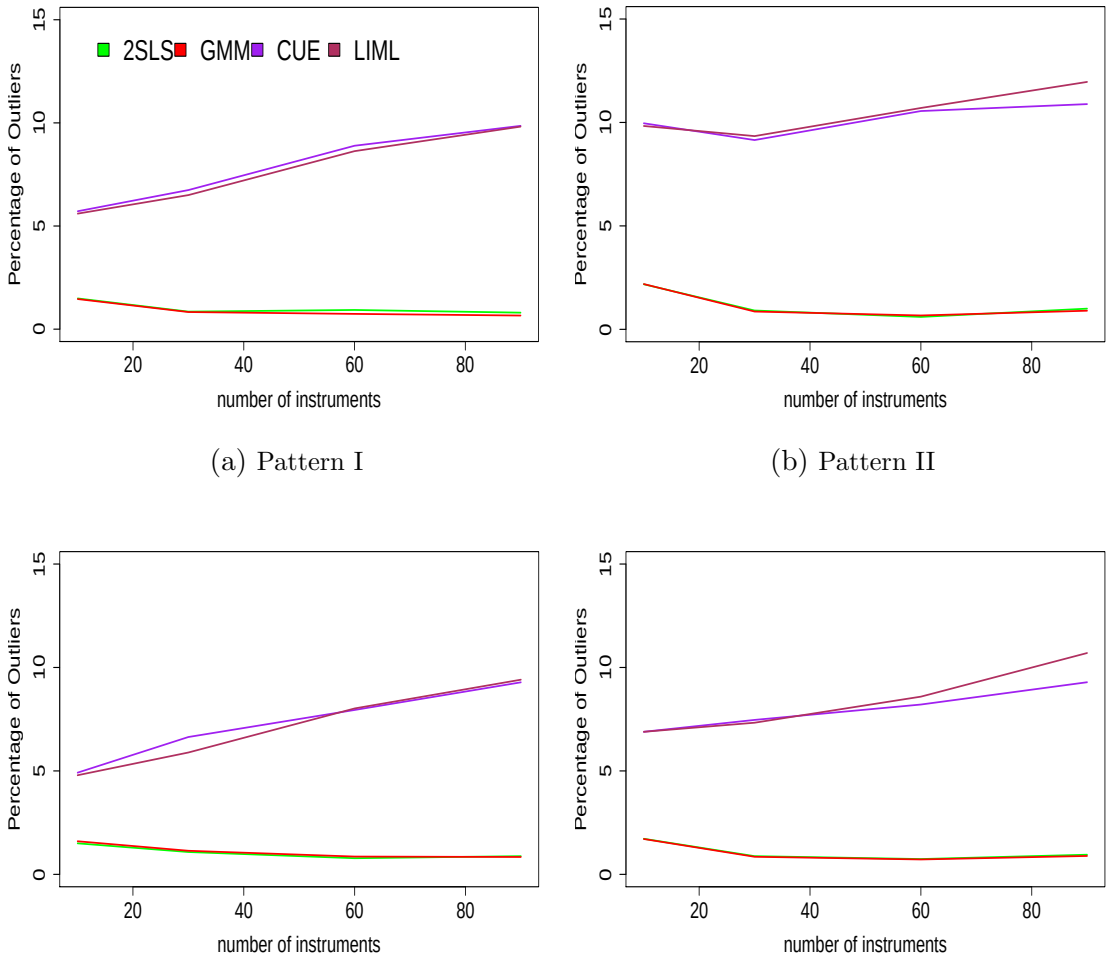
Here we start to see the disadvantages to CUE and LIML, in their percentage of outliers (Figure 7.8). When the proportion of weak instruments is too great, CUE and LIML become unstable, their range of estimates becomes significantly wider, as demonstrated by their higher percentage of outliers compared to 2SLS and GMM, shown in Appendix D.4 to D.7. Unlike CUE and LIML, the percentage of outliers for 2SLS and GMM are declining as number of instruments increases. This is because

the variation between causal effect estimate of each dataset is decreasing with more instruments included in the algorithm, as seen in Chapter 6, which reduces the interquartile range. Nonetheless, the 2SLS and GMM estimates are severely biased shown by Figure 7.6.

The percentage of outliers for CUE and LIML increases with number of instruments for all four patterns. Pattern I shows minor difference between them, where CUE has a higher percentage. By contrast Pattern II suggests that LIML has more outliers with 90 instruments, but since about 3% of the data could not converge for CUE (Appendix D.5), the difference in performance between LIML and CUE becomes less convincing. Similar to Winsorised RMSE, outliers have dropped by increasing from 10 to 30 instruments, where the same explanation applies. At 30 instruments, LIML and CUE have demonstrated a contrast in Pattern III; this is due to CUE placing too much weight on SNPs that gives near zero association with X which overestimate the causal effect. Again, Pattern IV displays a higher percentage of outliers for LIML than CUE, but again this is misleading as there were approximately 1.5% and 2% of data with non-converging CUE estimates for 60 and 90 instruments respectively, shown in (Appendix D.7).

Coverage

Figure 7.9 demonstrate the coverage from all the estimators for Pattern I to IV. Pattern I to IV are in agreement with each other; coverage declines as number of instruments increases and LIML estimates are closer to the 95% nominal coverage. As expected Pattern II has slightly lower coverage in comparison to the other patterns and a crossing between 2SLS and CUE at 10 instruments which demonstrates the issue with the CUE weighting algorithm when there are not enough strong instruments, as mentioned previously in the discussion on Winsorised RMSE. The difference in coverage between the 4 estimators, except between 2SLS and GMM is not due to sampling error, as 10,000 datasets gives approximately 0.45% accuracy to the 95% coverage.



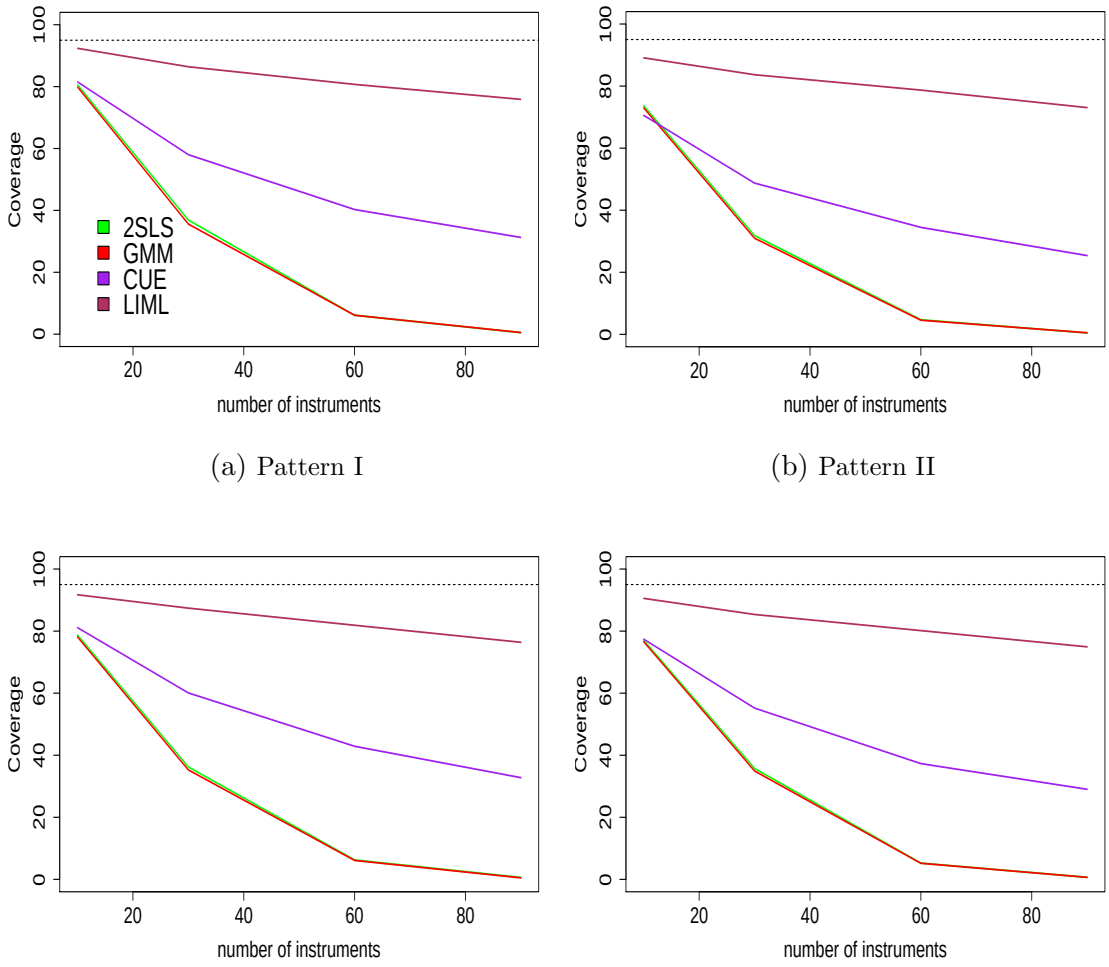
(a) Pattern I

(b) Pattern II

(c) Pattern III

(d) Pattern IV

Figure 7.8: Percentage of Outliers from 2SLS, two-step GMM, CUE and LIML for all four Patterns. The estimators are coloured as green, red, purple and maroon respectively. The estimation from 2SLS and two-step GMM are similar, therefore the green and the red are superimposed.



(a) Pattern I

(b) Pattern II

(c) Pattern III

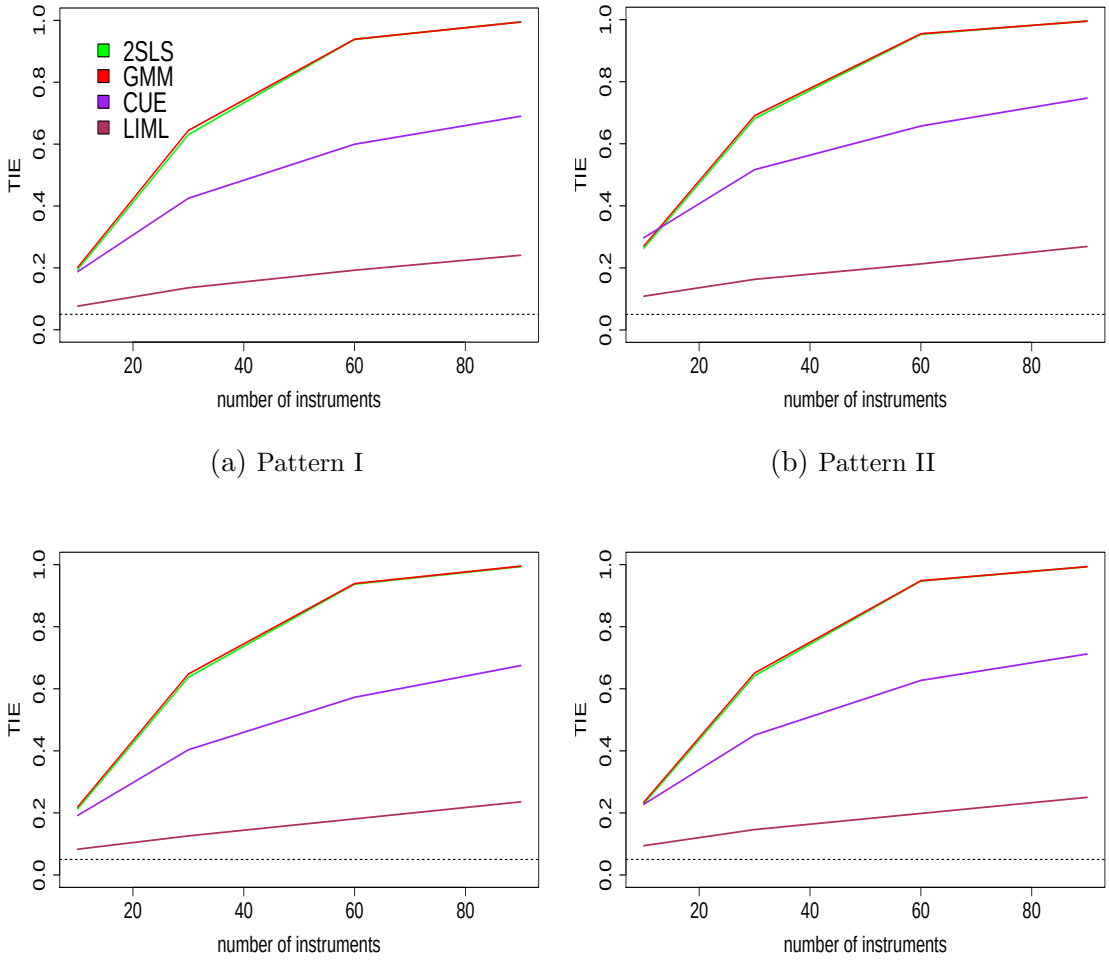
(d) Pattern IV

Figure 7.9: Coverage from simulation of the comparison of 2SLS, two-step GMM, CUE and LIML for all four Patterns. The estimators are coloured as green, red, purple and maroon respectively. The estimation from 2SLS and two-step GMM are similar, therefore the green and the red are superimposed. The dotted line in each plot is the 95% coverage.

Type I Error and Power

Type I error (TIE) and power from 2SLS, GMM, CUE and LIML are shown in Figures 7.10 and 7.11 respectively. TIE increases from 5% significance level (the dotted line) as the number of instruments increases and therefore the power curve in Figure 7.11 becomes deceiving for all 4 estimators. LIML estimates are closer to 5% significance level for 10 to 90 instruments. 10,000 datasets give approximately

$\mp 1\%$ accuracy to 5% significant level, hence the cross-over between 2SLS and CUE with 10 instruments is not due to sampling error, but more likely due to the lack of strong instruments from Pattern II causing CUE fails to detect no effect.



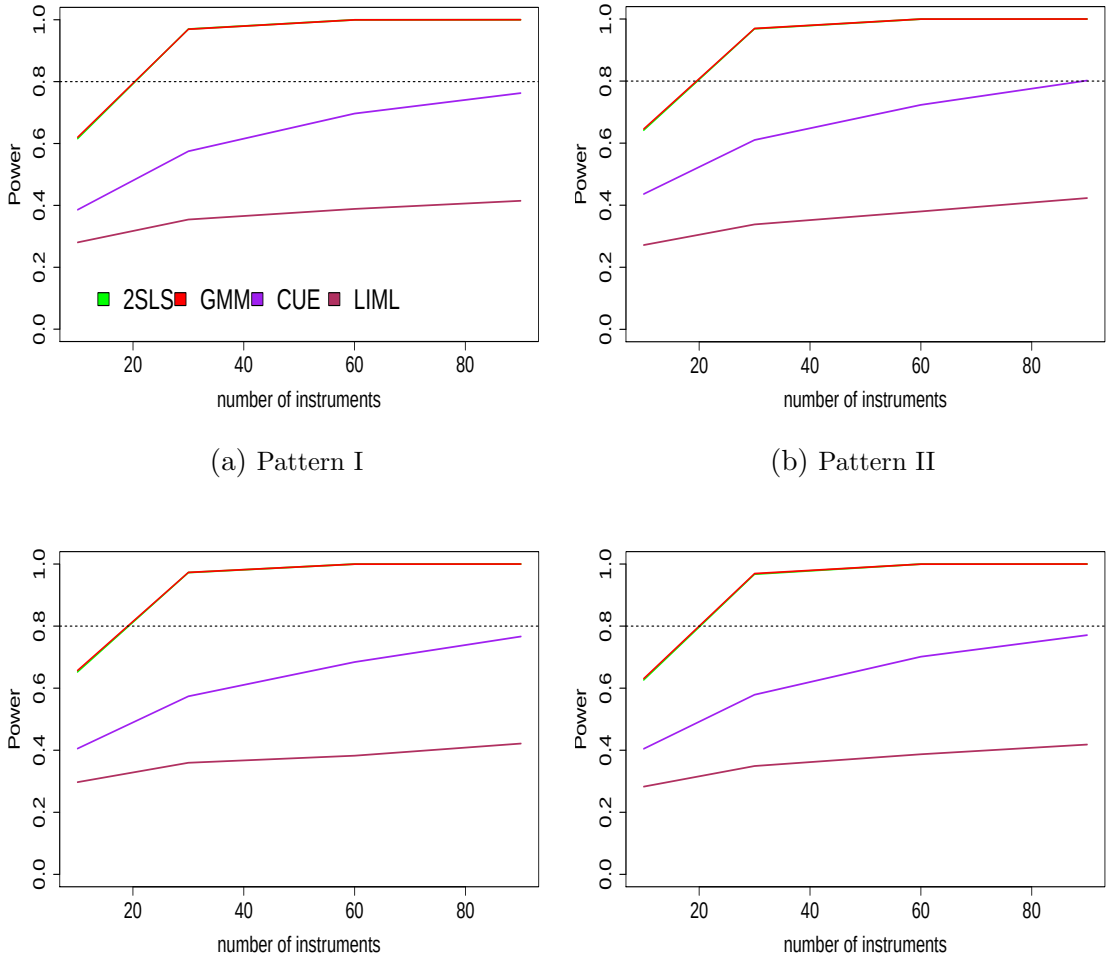
(a) Pattern I

(b) Pattern II

(c) Pattern III

(d) Pattern IV

Figure 7.10: TIE from 2SLS, two-step GMM, CUE and LIML for all four Patterns. The estimators are coloured as green, red, purple and maroon respectively. The dotted line in each plot is the 5% significance level.



(a) Pattern I

(b) Pattern II

(c) Pattern III

(d) Pattern IV

Figure 7.11: Power from 2SLS, two-step GMM, CUE and LIML for all four Patterns. The estimators are coloured as green, red, purple and maroon respectively. The dotted line in each plot is the 0.8 power.

7.3.4 Conclusions

The differences in genetic pattern did not affect LIML's efficiency in comparison to 2SLS, GMM and CUE; this was similar to the previous section where LIML provided the best evaluation criteria, except for percentage of outliers which increases with number of instruments within the algorithm. 60% of the time LIML and CUE provide unbiased estimates. However the assurance of accurate estimate of the causal effect heavily relies on the accuracy of the SNP associations with X; the finite sample bias in the first-stage regression could be detrimental.

So far I have been able to control the MAF and the correlation of each SNP, to identify the potential factors affecting the four estimators. In order to create an ever more realistic genetic structure, a GENOME simulator will be used in the next experiment.

7.4 Experiment 3: GENOME

7.4.1 Aims

This experiment will apply SNPs from an average length gene as instruments, where they will be simulated by GENOME [185]. GENOME generates realistic genetic correlations; the correlation between SNPs does not always depend on their physical proximity and high correlation usually occurs in relatively small random blocks, known as haplotype blocks [258]. GENOME does not allow the user to control the number of SNPs with specified MAF and the correlation between SNPs. Consequently it is more difficult to identify the factors affecting the performance of the four estimators, and hence for the previous experiments I have designed my own SNPs simulation method.

The experiments in this Chapter are considering the case where the causal SNP is unmeasured and therefore the likely explanation would be that the functional allele is rare as the rare alleles are harder to detect. Then the aim of this experiment will be to monitor the performance of 2SLS, GMM, CUE and LIML with common (MAF > 0.1) and non-causal SNPs as instruments, to see whether the evaluation criteria of each estimator with realistic genetic patterns differs from the previous experiments with artificial genetic patterns.

7.4.2 Design

There are approximately 200 SNPs in an average sized gene [112], hence GENOME will be set to generate genotypes of 200 SNPs for 2,000 individuals as described in Section 4.2. The potential instruments are SNPs with common MAF (> 0.1) and the causal SNP will be randomly chosen among SNPs with low MAF (< 0.1). The causal SNP will explain 2% of the variation in X and 6% of the variation in Y will be explained by X. X and Y will be normally distributed with sample size of 2,000. All of the rare SNPs will be discarded, therefore the performance of the estimators

will only be based on common non-causal SNPs as instruments. The simulation will be repeated 10,000 times and evaluated as in Section 4.4.

7.4.3 Results

Table 7.3 gives the summary of the number of SNPs with common MAF from 10,000 datasets. Most of the datasets have approximately 30 instruments, but the number of SNPs is variable between 12 to 50. There is one dataset that had only SNPs with rare MAF (< 0.05), under "NA" in Table 7.3.

As expected, due to the nature of the simulation, GENOME does not have the options to control the MAF of the SNPs and their correlation with another, and consequently produced some datasets with many weak instruments. Hence, the range of causal effect estimates from CUE and LIML is very wide. Even though 2SLS and GMM causal effect estimates are less varied (standard error) they are biased, whereas CUE and LIML have more accurate estimates on average, shown by Winsorised bias in Table 7.4. LIML and CUE are minimizing the same function in the case of homoskedasticity [84] and hence the similarity in their bias and RMSE. CUE has the lower Winsorised RMSE and proportion of outliers; nonetheless the difference with LIML's estimates seems minimal as there were 89 datasets for which CUE did not converge.

The coverage for LIML is nearest to the 95% nominal coverage in comparison to the other estimators (Table 7.4). LIML's type I error is closer to the 5% significance level and hence its power is more believable whereas the other estimators has seemingly greater power, but it is exaggerated as adjust for the greater type I error is required.

The extreme estimate values from CUE with standard deviation of the mean is 98488, is due to CUE's weights relying on the causal effect estimate of each instrument. Therefore the SNPs that have high correlations with the causal SNP will not be given more weight, as their estimates are affected by finite sample bias. Figure 7.12 is an example that demonstrates this mechanism, where CUE estimated the causal effect to be -1.6316×10^6 and LIML gave 0.6986 (the true effect is 0.2449). Even though they both gave extreme estimates, CUE was worse. The first thing to note is that there were 40 instruments and most of them have estimates of zero association with X, β_{ZX}, were all below 0.1 (Figure 7.12b). The next thing is that the SNPs are highly correlated with the causal SNP and have MAF < 0.2, which will cause finite sample bias in estimates of their association with X from the

Table 7.3: Summary of GENOME simulated genetic instruments

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
Number of Instruments	12	28	31	31	34	50	1

first regression. Consequently CUE will give similar weight to estimated β_{ZX} on each SNP, which as a single instrument overestimates the causal effect (equivalently $\frac{\beta_{ZY}}{\beta_{ZX}}$), where the algorithm divides β_{ZY} by approximately zero.

Chao and Swanson [61] found LIML has Cauchy-type tails in its finite sample distribution, which means it is a median unbiased estimator but badly behaved at the ends, hence the reason for Winsorisation. In the presence of weak instruments, weak could be due to finite sample bias and/or low variation explained. Then the minimising function for LIML, Q_{LIML}, minimizes by increasing its denominator (equivalent to the OLS estimate) to avoid a ratio of 1, as explained in Experiment 2 of this Chapter. Figure 7.13 is an example of LIML estimating -1603 for the causal effect and CUE gave -5.599; there are 33 instruments and only two instruments have correlation > 0.3 with the causal SNP (Figure 7.13c) and both have MAF < 0.3 i.e. low variation explained and lack of data respectively. The inaccuracy in CUE is not as large, as there is smaller number of SNPs with near zero β_{ZX}.

Figure 7.14 is an example where CUE and LIML had the most accurate causal effect estimate; there are 24 instruments and most of them have a strong correlation with causal SNP, high $-log_{10}$(p-value) and β_{ZX} above 0.1. Even though there are quite a few SNPs with near zero β_{ZX}, but there are enough strong instruments to outweigh them.

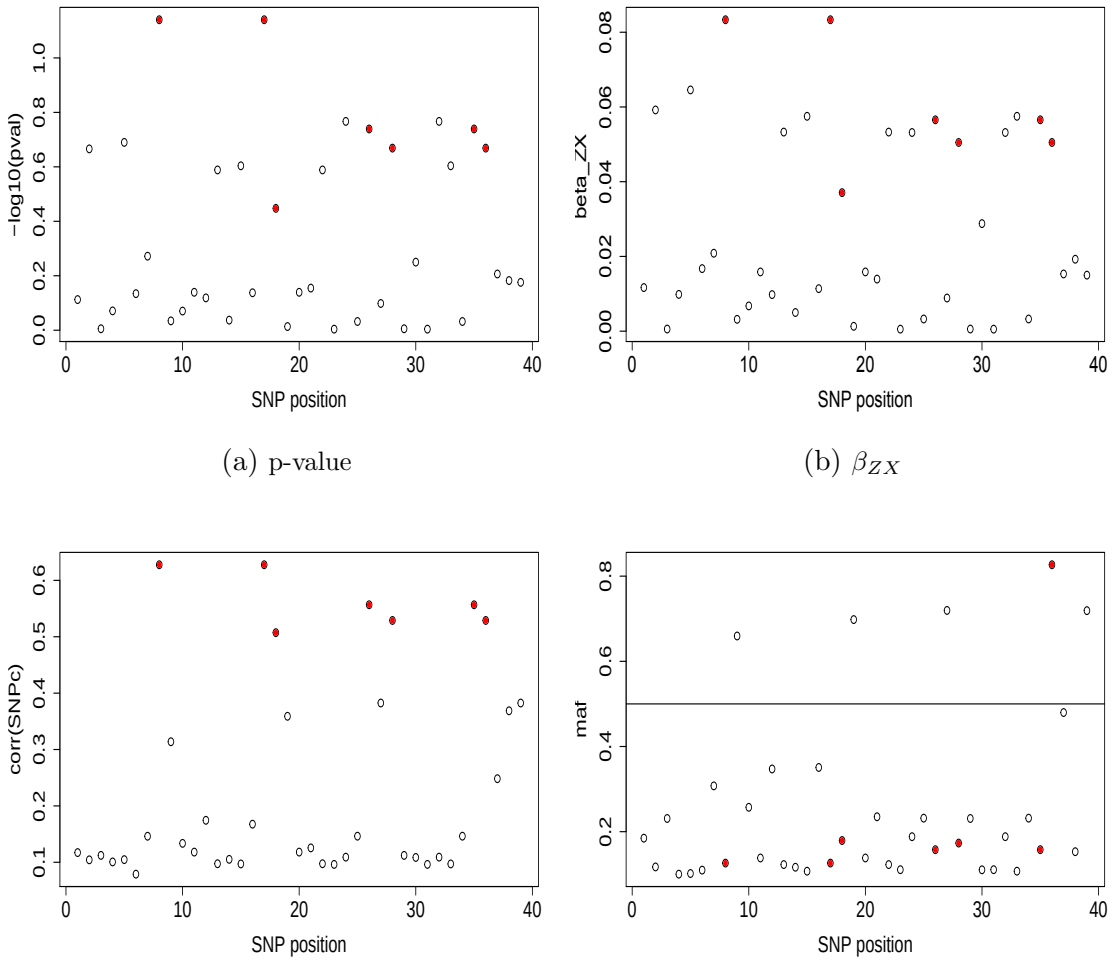
7.4.4 Conclusions

The GENOME simulated SNPs showed that for most of the datasets CUE and LIML gave less biased causal effect estimates than 2SLS and GMM. LIML estimates gave approximately 95% nominal coverage and 5% significance level, whereas the other estimators did not. Hence, we have arrived at the same conclusion as in the previous experiment.

The outliers in CUE and LIML were caused by many weak instruments. Perhaps excluding weak instruments will be a solution to reduce the extreme causal effect estimates from CUE and LIML, or not analyse a particular dataset if it contains too many weak instruments. This theory will be investigated in the next experiment.

	Mean $\hat{\beta_{XY}}$ (S.D.)	Winsor. S.E.	Winsor. Bias	Winsor. RMSE	Prop. of Outliers	Coverage	TIE	Power	Non-Convergence
2SLS	0.4307 (0.1398)	0.0008	0.1844	0.2027	0.0186	66.1466	0.3381	0.8692	0
GMM	0.4307 (0.1413)	0.0009	0.1847	0.2035	0.0193	65.2365	0.3476	0.8723	0
CUE	811.0794 (98488)	0.0014	-0.0087	0.1425	0.0477	77.9213	0.2251	0.4991	89
LIML	0.0780 (16.47)	0.0014	-0.0116	0.1431	0.0490	91.0591	0.0893	0.3712	0

Table 7.4: Evaluation Criteria from 2SLS, Two-step GMM, CUE and LIML with GENOME simulated genetic instruments. $\hat{\beta_{XY}}$ is the causal effect estimates and standard deviation of its mean (S.D.) are not Winsorised. The true β_{XY} is 0.2449.



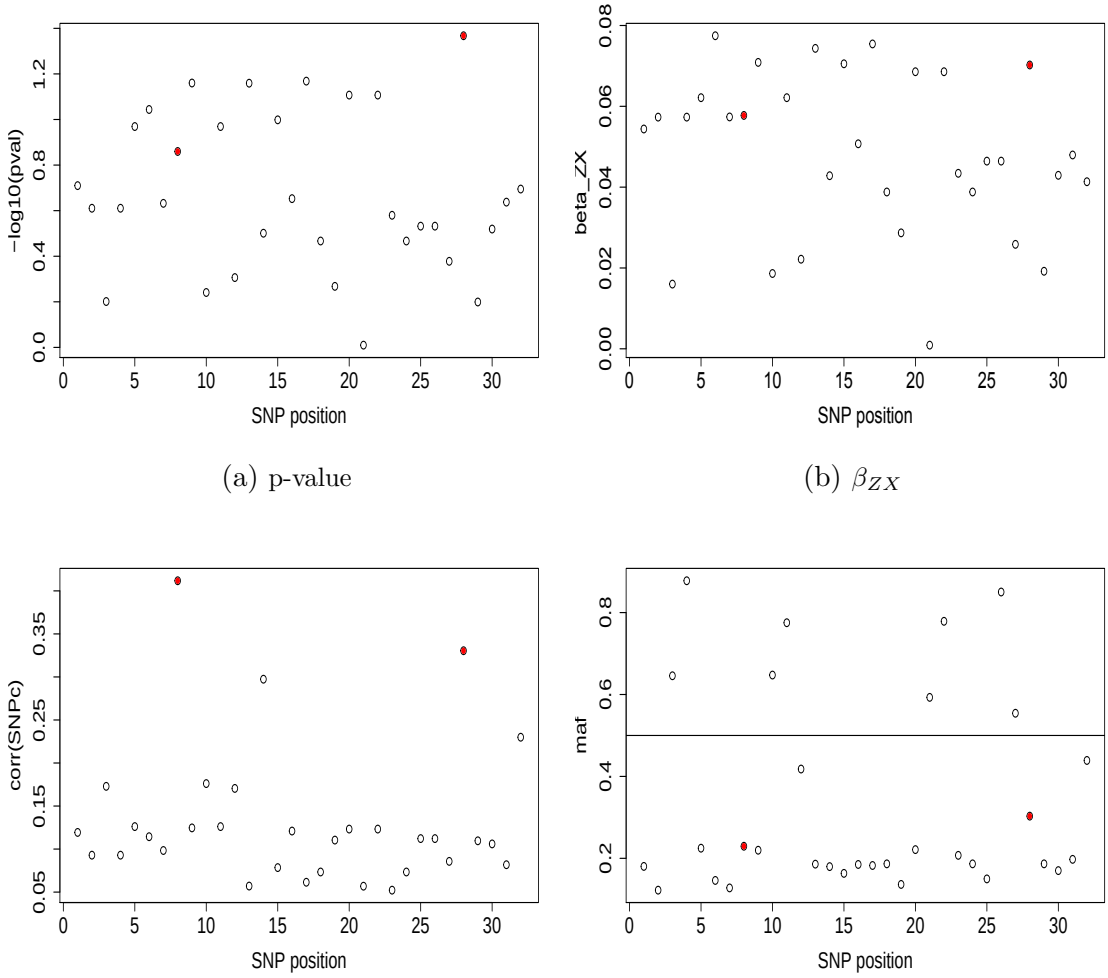
(a) p-value

(b) β_{ZX}

(c) Correlation with SNP_c

(d) MAF

Figure 7.12: An example of a dataset yielding an inaccurate causal effect estimate for CUE. The four plots give each SNPs' p-value, β_{ZX}, correlation with SNPc and MAF. P-value and β_{ZX} is $-log_{10}$(p-value) and coefficient estimate from the association with X. The SNP coloured in red has correlation > 0.5 with the causal SNP.



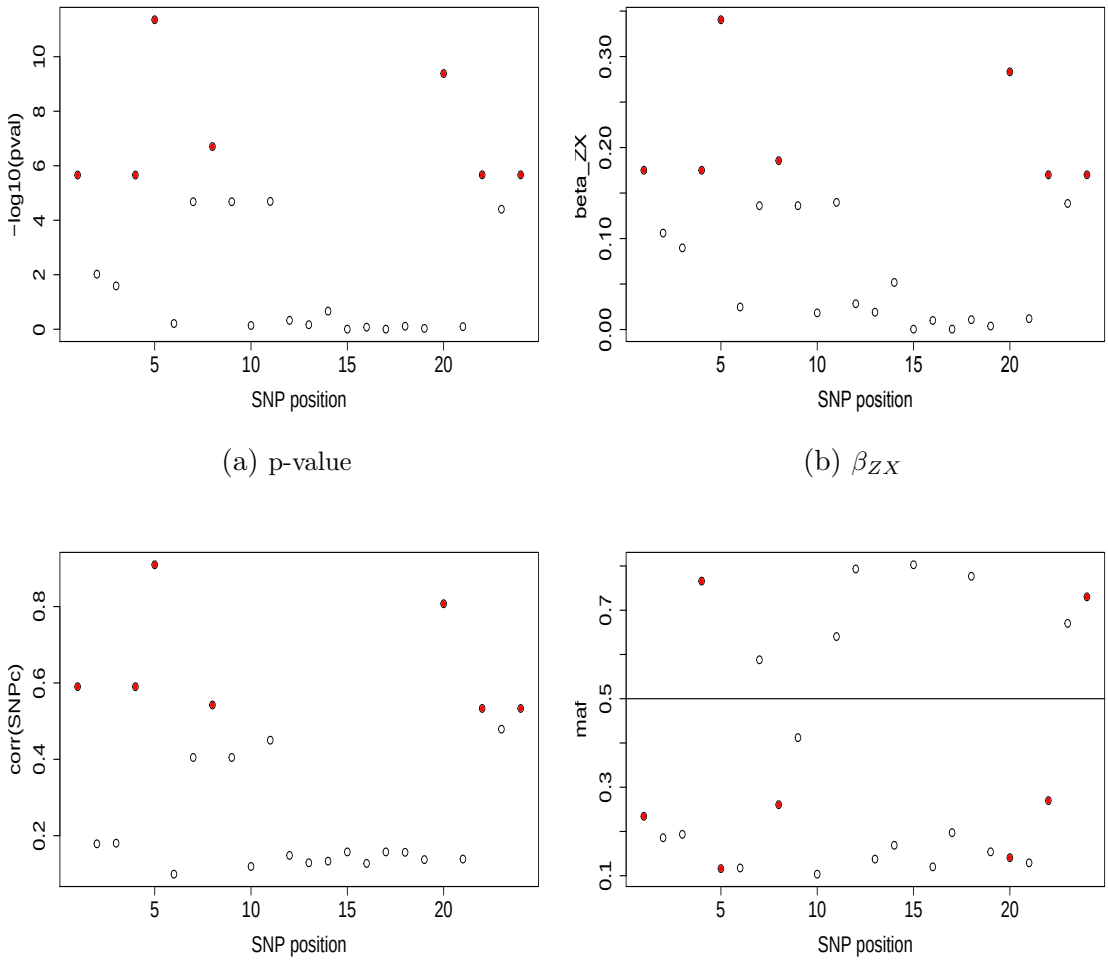
(a) p-value

(b) β_{ZX}

(c) Correlation with SNP_c

(d) MAF

Figure 7.13: An example of a dataset yielding an inaccurate causal effect estimate for LIML. The four plots give each SNPs' p-value, β_{ZX}, correlation with SNPc and MAF. P-value and β_{ZX} is $-log_{10}$(p-value) and coefficient estimate from the association with X. The SNP coloured in red has correlation > 0.3 with the causal SNP.



(a) p-value

(b) β_{ZX}

(c) Correlation with SNP_c

(d) MAF

Figure 7.14: The dataset that gives the most accurate causal estimate for both LIML and CUE. The four plots give each SNP's p-value, β_{ZX}, correlation with SNPc and MAF. P-value and β_{ZX} is $-log_{10}$(p-value) and coefficient estimate from the association with X. The SNP coloured in red has correlation > 0.5 with the causal SNP.

7.5 Experiment 4: Selection of Data and Instruments in GENOME

7.5.1 Aims

As seen from the previous experiment, GENOME simulated datasets had too many weak instruments which caused the 2SLS, two-step GMM, CUE and LIML to give extreme causal effect estimates. The next experiment will examine the most popular criterion for the selection of instruments, an F-statistic greater than 10, to see whether this is able to reduce the number of outliers. In order to demonstrate selection bias, the choice of instruments will be from the same dataset as used in the MR analysis.

7.5.2 Design

The design of this experiment will be identical to the previous experiment, except this experiment includes two procedures to exclude weak instruments from the 10,000 datasets, involving the F-statistic of the individual SNP association with X;

1. All the SNPs will be used as individual instruments, if datasets in the analysis possess SNPs with any F-statistics > 10.

2. With datasets from the previous criterion, only include SNPs with F-statistic of > 10 as individual instruments in the analysis.

Henceforth the first and second criteria will be referred to as filtering and selection criteria respectively.

The F-statistic of 10 is equivalent to a p-value of 0.001 for sample size of 2,000. The performance of 2SLS, two-step GMM, CUE and LIML will be monitored by mean and median bias, root mean squared error (RMSE), coverage and power from datasets that satisfies the criteria. Winsorisation will not be used in this experiment as the criteria should not give the extreme estimates seen previously. Type I error (TIE) will be calculated from a separate simulation where X does not directly influence Y.

7.5.3 Results

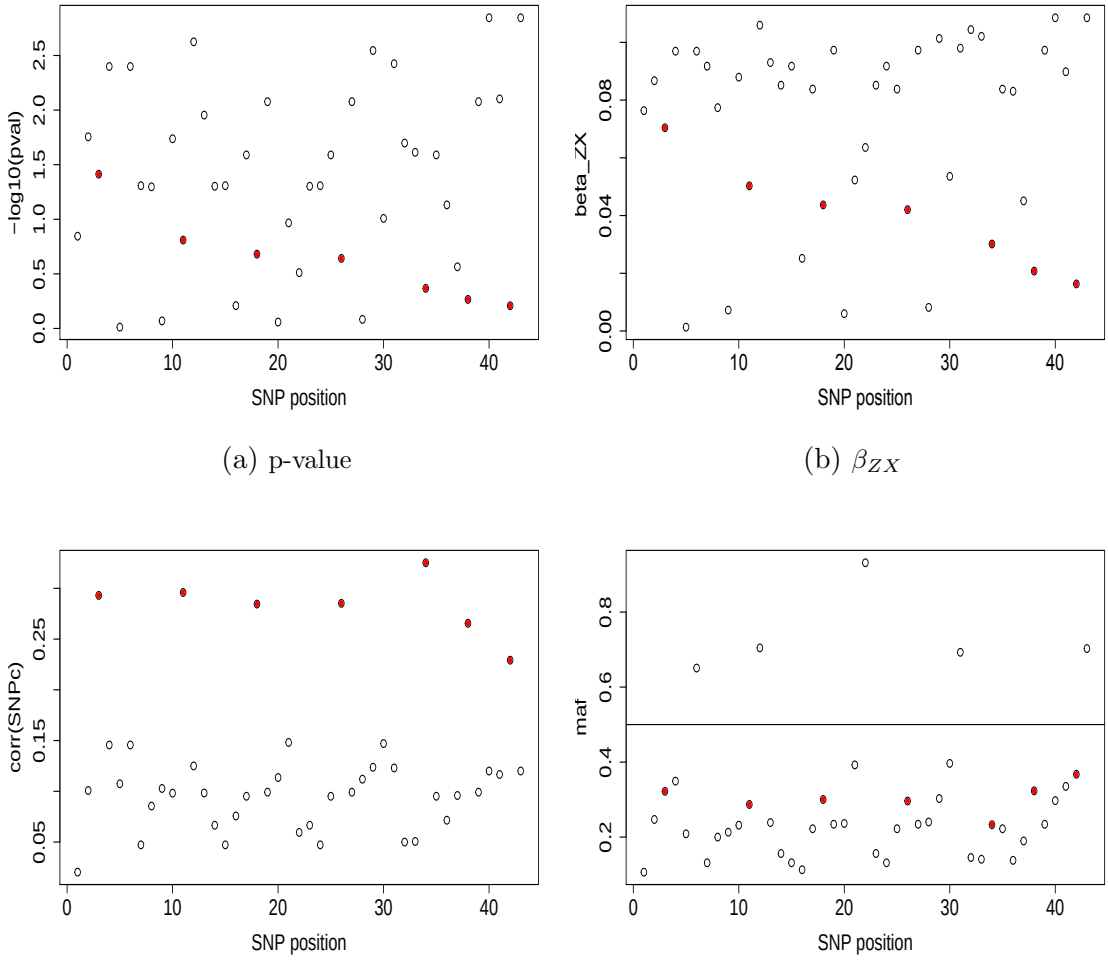
Table 7.5 show that less than half of the 10,000 datasets satisfy both of the conditions and the mean causal effect estimate for both CUE and LIML show that they no longer give extreme estimates, except for one dataset (Figure 7.15) which caused the former to produce an outlier; even though this dataset had instruments with F-statistics > 10 but their correlations with the causal SNP were below 0.2. In addition, this dataset had too many instruments with approximately zero association with X, 43 SNPs in total. Note that LIML also estimated an inaccurate causal effect estimate from this dataset; -0.1553. This show that an F-statistic > 10 does not always manage to exclude the weak instruments, in line with Palmer et al. [220]'s findings. The standard error for CUE and LIML have reduced after both criteria.

As expected the 2SLS and GMM have the least bias under selection criteria, as the amount of variation explained is sufficient for the number of instruments. LIML prefers filtering as there are enough strong instruments within the dataset and it is therefore able to ignore the nuisance instruments. CUE should have performed better under filtering criterion too, but the one dataset mentioned above is pulling the mean bias away from zero, as demonstrated by the relatively smaller median bias. The difference between mean and median bias from the four estimator is minor when compared to no selection.

Even though the median bias from CUE and LIML for datasets without selection is less than filtering and selection, due to datasets with many weak instruments, the RMSE is much worse without selection. Interestingly, the least bias for 2SLS and GMM gave the highest RMSE, this is because by chance the instruments chosen within some datasets are not actually correlated with the causal SNP and therefore inaccurately estimating the causal effect; the difference in mean and median bias is because of the small number of datasets with extreme estimates.

LIML still had near to nominal coverage in comparison to the other estimators, even with selection and filtering. Out of all four estimators, LIML is closest to the 5% significance level. Thus, the 0.8 power from 2SLS and GMM is misleading, since they do not have 5% TIE nominal level for any criteria.

CUE struggles to converge in some datasets even in the absence of weak instruments (F-statistic < 10).



(a) p-value

(b) β_{ZX}

(c) Correlation with SNP_c

(d) MAF

Figure 7.15: The dataset that give the most inaccurate causal estimate for CUE. The four plots give each SNPs' p-value, β_{ZX}, correlation with SNPc and MAF. P-value and β_{ZX} is $-log_{10}$(p-value) and coefficient estimate from the association with X. The SNP coloured in red has correlation > 0.2 with the causal SNP.

7.5.4 Conclusions

In conclusion, even after applying the selection and filtering criteria, LIML is still the most efficient estimator when examining all of the evaluation criteria.

In terms of preventing outliers in CUE and LIML, the selection criterion have reduced the standard error and RMSE in comparison to no selection, however this introduces selection bias, since the filtering criterion has a lower bias. In comparison the filtering criterion prevented outliers from LIML but not from CUE, as seen from

the relatively large difference between mean and median bias.

7.6 Discussion

The experiments in this chapter have concluded limited information maximum likelihood (LIML) to be the preferred estimator for many dependent instruments, in comparison to the two-stage least squares (2SLS), two-step generalised method of moments (GMM) and continuously updating estimator (CUE). The errors in the simulated datasets were under the condition of homoskedasticity, see Section 3.2.2. LIML gave the least Winsorised bias and RMSE, approximately 95% coverage and 5% significance level for different numbers of instruments. However, LIML does also suffers from weak instrument bias as was evident from the selection and filtering experiment, where the number of outliers have been diminished.

2SLS and GMM estimators have the same algorithm under the condition of homoskedasticity [116], hence the similarity in their performance throughout the experiments. The results from this chapter and Chao and Swanson [61] demonstrate the superiority of LIML for many weak instruments; LIML's minimising function is essentially 2SLS divided by OLS estimates, and 2SLS is biased towards the OLS in the presence of weak instruments [253], hence the minimising function would avoid a ratio of 1. However, if the association between X and the instrument is weak, either caused by finite sample bias and/or the correlation with the causal SNP is weak, leading to an inaccurate first-stage regression then LIML would avoid a ratio of 1 by finding a value that calculates a very large OLS estimate to compensate for the large 2SLS [61]. CUE uses the same weighting algorithm as GMM, but instead of incorporating 2SLS estimates to approximate weights for each instrument, it goes through different possible values for the causal effect and updates the overall causal effect estimate, to minimise its function. However, if the sample size is small the instrument association with X will approximate to zero which overestimates the causal effect as a result.

The conclusion from this chapter is consistent with the results in Davies et al. [85], where they have also compared 2SLS, CUE and LIML and found LIML to be the least biased estimator for many weak independent instruments in the case of homoskedasticity. They used the median and IQR of the causal effect estimates and so did not present the extreme estimates, stating that "LIML estimator and CUE have occasionally very large outliers". The outliers were also shown in the

	Mean $\hat{\beta}_{XY}$ (S.D.)	Standard Error	Mean Bias	Median Bias	RMSE	Coverage	TIE	Power	Non-Convergence
No Selection (N = 10,000)									
2SLS	0.4307 (0.1398)	0.0014	0.1857	0.1843	0.2324	66.1466	0.3381	0.8692	0
GMM	0.4307 (0.1413)	0.0014	0.1857	0.1850	0.2334	65.2365	0.3476	0.8723	0
CUE	811.0794 (98488)	989.35	810.83	0.0009	98487	77.9213	0.2251	0.4991	89
LIML	0.0780 (16.47)	0.1647	-0.1670	-0.0019	16.4703	91.0591	0.0893	0.3712	0
Including instruments with F-statistics > 10 (N = 4,347)									
2SLS	0.3794 (0.2382)	0.0036	0.1344	0.1288	0.2735	85.1392	0.1484	0.4810	0
GMM	0.3794 (0.2383)	0.0036	0.1345	0.1285	0.2736	84.9781	0.1502	0.4842	0
CUE	0.3566 (0.2532)	0.0038	0.1117	0.1015	0.2767	86.2903	0.1376	0.4412	7
LIML	0.3561 (0.2514)	0.0038	0.1111	0.0998	0.2749	87.6006	0.1238	0.4145	0
Datasets with instruments with F-statistics > 10 (N = 4,347)									
2SLS	0.4605 (0.1474)	0.0022	0.2156	0.2127	0.2612	58.3621	0.4159	0.9068	0
GMM	0.4604 (0.1497)	0.0023	0.2155	0.2117	0.2624	57.3269	0.4267	0.9080	0
CUE	-637.13 (41860)	637.4111	-637.38	0.0396	41860	74.7971	0.2568	0.5662	34
LIML	0.2615 (0.9497)	0.0144	0.0165	0.0370	0.9497	87.1176	0.1286	0.4339	0

Table 7.5: Evaluation Criteria from 2SLS, Two-step GMM, CUE and LIML with GENOME simulated genetic instruments. The true β_{XY} is 0.2449.

experiments of this Chapter. However, there were no further discussions of these outliers in their paper.

Staiger and Stock [253] have proposed selecting instruments with F-statistics > 10 from their association with X to reduce the weak instrument bias from 2SLS, which have been widely implemented ever since. The last experiment also applied the same selection criterion for including instruments in LIML and CUE to prevent the extreme outliers. Note that an F-statistics > 10 does not always manage to exclude weak instruments, shown by Experiment 4 and Palmer et al. [220]. Nevertheless, Econometricians have designed other selection criteria specifically for LIML and CUE, such as minimising mean squared error to choose an instrument set for LIML [92], and Andrews [10] demonstrated that a Bayesian information criterion (BIC) with upward and downward testing yields consistent estimates from GMM (CUE is a form of GMM).

The mean causal effect estimate of 10,000 datasets with many instruments from CUE and LIML was not accurate, as their standard error without Winsorisation were relatively large, therefore more simulations should have been performed. However, the algorithm of CUE and LIML with many instruments is computationally intensive; to run 10,000 datasets with 90 instruments took approximately 100 hours. Pattern IV is imitating two haplotype blocks within a region, where the two causal SNPs in separate blocks have independent effect on exposure and are not in LD, shown in Figure 7.5d. Further simulation is required for patterns where these two causal SNPs have varying correlation between them, but still have independent effect on the exposure of interest. This could give similar conclusion to the scenario in Section 5.5.1, where the two non-causal SNPs both explain the same amount of variation due to their high correlation with the causal SNP, nevertheless when they are in high LD with each other, including both as instruments did not explain any more variation.

This chapter has concluded that from the four estimators, LIML was the preferred estimator for many weak instruments but it can still be affected by sample size and variation explained. In addition, the conclusion was made from 60% of the datasets, hence LIML is only a median unbiased estimator and can still behaved badly in some samples. Intuitively, the next step would be to compare the classical to the Bayesian approach, to see whether Bayesian methods could provide consistent estimates of the causal effect for many weak instruments or be able to select relevant instruments internally without introducing selection bias.

Chapter 8

Bayesian Approaches to Mendelian Randomisation

8.1 Introduction

From the previous chapter we have seen that limited information maximum likelihood (LIML) is only median unbiased, there are still cases where it can be severely biased when there are many weak instruments. Using the F-statistic and p-value have reduced the weak instrument bias, however it comes with the cost of selection bias if instrument selection was performed on the data under analysis, and some datasets may not have SNPs that have fulfilled the selection criteria. The selection of instruments using the p-value or F-statistic cannot distinguish between situations in which (1) there is not sufficient data to detect an effect and (2) the data provides evidence for the null hypothesis [139]. By computing the posterior effect probability, the Bayesian approach is able to differentiate these two scenarios. To demonstrate, consider an example from Hoeting et al. [139], where for treatment A $P(\beta \neq 0|Data) = 22\%$ and treatment B $P(\beta \neq 0|Data) = 2\%$, which indicates there is uncertainty in a conclusion that treatment A had no effect whereas for treatment B there is evidence of no effect from the data. For more detailed debate of the limitations of p-value and the advantage to the Bayesian posterior probability, see Marden [195].

Bayesian approaches offer a systematic and structured way of incorporating external biological knowledge into the statistical analysis [23]. Many publications have discussed the plausibility of the genetic effect from their analysis using external biological information [114]. Therefore why not incorporate this information

into statistical analysis? The inclusion of prior knowledge is developing in genetic association studies; Rannala and Reeve [233] was the first to use the human genome sequence as an informative prior for Bayesian gene mapping. Cantor et al. [57] have reviewed different prior knowledge for SNP selection for genetic association studies from meta-analysis, gene interaction and pathway analysis. The application of bioinformatics techniques have been advocated to prioritise the most biologically relevant genes for further investigation [201]. Minelli et al. [200] have incorporated experts' opinion with the findings in meta-analysis for gene prioritisation and its companion paper, Thompson et al. [267] used these findings as weightings in Bayesian analysis. In the context of Mendelian randomisation, instead of selecting instruments in an arbitrary fashion, instruments that demonstrate biological relevance to the exposure of interest can be given more weight or instruments that have shown association with the exposure of interest in a meta-analysis.

Biological mechanisms are often complex, but as the Bayesian paradigm is based on probability, it provides a straightforward and consistent way of modelling biological complexity [290]. Consider an example where there are three possible models; in the classical setting, comparisons of these three models are dealt with indirectly by hypothesis testing with the null model, whereas the Bayesian method estimates the posterior probability for each model to give evidence of their plausibility [249]. Thanks to the advancement of Markov chain Monte Carlo (MCMC), a sampling algorithm in Bayesian computation [192], it is computationally efficient to fit realistic models to complex datasets. Datasets with measurement errors, missing observations, multilevel correlation structures and latent variables can be incorporated into the Bayesian model [96]. Conversely, due to the high dimensional integration, the optimisation approaches for fitting such models can be difficult when the assumptions of normality and non-linearity are violated.

The advantages of the Bayesian approach as an alternative to the classical have been well recognised and advocated in genetic association [23, 256] and observational epidemiological studies [96, 119]. Therefore, contemplating the benefits of the Bayesian approach and its recognition in genetic association and observational epidemiological studies, I will compare the Bayesian to the classical methods in Mendelian randomisation.

The purpose of this chapter is to describe a Bayesian algorithm that gives good estimates for many weak instruments. Bayesian methods are not widely used in Mendelian randomisation studies but two literature searches will be performed;

(1) a review of the current methodological literature for Bayesian approaches to Mendelian randomisation and (2) econometric literature of the Bayesian approaches to instrumental variable analysis.

8.2 A review of Bayesian approaches to Mendelian randomisation

The earliest mention of Bayesian approach to Mendelian Randomisation (MR) was by Thompson et al. [266], where their main aim was to obtain the causal effect through applying summary data from multiple genotype-phenotype (ZX) and genotype-outcome (ZY) studies without the need to include the within-study correlation between ZX and ZY. For binary outcomes, McKeigue et al. [198] have designed a Bayesian approach to Mendelian randomisation, which also allows for missing data. Burgess et al. [47] introduced a solution to the problem of weak instruments by utilising data from meta-analysis via the Bayesian method, which takes account of different SNPs measured in different studies and heterogeneity between studies. Their next paper proposed 4 Bayesian methods for imputing the missing data in MR studies; multiple imputations, SNP imputation, latent variables and haplotype imputation [43].

Jones et al. [153] published a detailed paper on the impact of the choice of priors in estimating the causal parameter, where they advised against vague priors for datasets with weak instruments or small sample sizes. However, an informative prior has to be chosen carefully as there is large variation in parameter estimates with different priors. They have also examined the performance of three models with different parametrisations, naming them as the "full", correlated errors and independent errors model. Consider the equation for Mendelian Randomisation;

$$X = \alpha_0 + \alpha_1 Z + \alpha_2 U + \varepsilon_x \tag{8.1}$$

$$Y = \beta_0 + \beta_1 X + \beta_2 U + \varepsilon_y \tag{8.2}$$

where Z is the genetic instrument, X is the risk factor, Y is the outcome, U is the confounder, $\varepsilon_x \sim N(0, \tau_x)$ and $\varepsilon_y \sim N(0, \tau_y)$. The "full" model makes assumptions on conditional distribution of U, X and Y;

$$U \sim N(0,1) \tag{8.3}$$

$$X|(Z,U) \sim N(\mu_x, \tau_x^2) \tag{8.4}$$

$$Y|(X,U) \sim N(\mu_y, \tau_y^2) \tag{8.5}$$

where $\mu_x = \alpha_0 + \alpha_1 Z + \alpha_2 U$ and $\mu_y = \beta_0 + \beta_1 X + \beta_2 U$ and therefore have eight independent priors on eight parameters; $\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2, \tau_x, \tau_y$. Instead of modelling the confounder independently, it can be described as a correlated error between X and Y, and the equation for Y is rewritten as the reduced form where the X in Equation 8.2 has been substituted in terms of Z. Hence the correlated errors model has the structure;

$$[X, Y|Z] \sim MVN\left(\begin{bmatrix} \alpha_0 + \alpha_1 Z \\ \beta_0 + \beta_1\alpha_0 + \beta_1\alpha_1 Z \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \lambda \\ \lambda & \sigma_y^2 \end{bmatrix}\right) \tag{8.6}$$

where $\sigma_x^2 = Var[X|Z]$, $\sigma_y^2 = Var[Y|Z]$, $\lambda = Cov(X, Y|Z) = \rho\sigma_x\sigma_y$ and ρ is the correlation between X and Y. Finally, the independent errors model is defined by the following form;

$$X|Z \sim N(\mu_x, \sigma_x^2) \tag{8.7}$$

$$Y|Z \sim N(\mu_y, \sigma_y^2) \tag{8.8}$$

where $\mu_x = \alpha_0 + \alpha_1 Z$ and $\mu_y = \beta_0 + \beta_1\alpha_0 + \beta_1\alpha_1 Z$. This model is equivalent to 2SLS where you predict X from Z and substitute the predicted X into Y. The independent errors model structure is taken from Burgess et al. [47]. Jones et al. [153] found that the "full" model gives the least biased causal effect estimate when the sample size is small and the strength of the instrument is relatively weak. With informative priors the correlated errors model becomes more efficient. The independent model only performs well when the confounding between X and Y is weak, as the model assumes that the error terms in 8.6 are independent, which does not follow the 3 core assumptions for instrumental variable (Chapter 1). Thompson et al. [266] have also applied the correlated error model to their problem, but they were meta-analysing multiple MR studies, and modelling the correlation between X and Y within a study.

Another model has been suggested by Burgess [40], where they compared the

classical and the Bayesian approach to Mendelian randomisation. The classical approaches were 2SLS, the Wald estimator and LIML. The Bayesian approaches were the Bayesian method and adjusted Bayesian, the former is equivalent to the independent errors model described in Jones et al. [153] and the latter was previously adopted by Burgess and Thompson [43] shown below;

$$X_i \sim N(\mu_x, \sigma_x^2) \tag{8.9}$$

$$Y_i|X_i = x_i \sim (\mu_y + \frac{\sigma_y}{\sigma_x}\rho(x_i - \mu_x), (1-\rho^2)\sigma_y^2) \tag{8.10}$$

They claimed that their simulation of Mendelian randomisation with continuous outcome demonstrated that the causal effect estimate from the adjusted Bayesian model had closest to 95% nominal coverage and had less bias in comparison to the classical approaches. However the superiority of the Bayesian approach is less apparent for the cases where weak instruments are present. Unsurprisingly, for the case of weak instruments, they found that the unadjusted Bayesian method was as biased as 2SLS. In their simulation, including the weak instrument case, LIML had the least bias but coverage was lower than the adjusted Bayesian model.

8.3 A review of Bayesian approaches to instrumental variable analysis

In comparison to the Mendelian randomisation literature there is much more research on Bayesian approaches in econometrics. Due to its vast methodological development, I will give a brief summary of the Bayesian approaches to instrumental variable problem, where different approaches cover a wide range of scenarios and assumptions by designing topic specific priors and parameterisation. Lopes and Polson [190] have also reviewed these approaches.

Kleibergen and Zivot [169] is perhaps the most cited paper, as they derived the Bayesian equivalents to the classical approaches of 2SLS and LIML. They examined the structural, restricted reduced (RRF) and unrestricted reduced (URF) forms of the IV model with vague and Jeffreys priors. The URF model is the same as the adjusted Bayesian model in Burgess [40], seen in Section 8.2. The URF with Jeffreys priors on the parameters are equivalent to LIML, thus Hoogerheide et al. [142] have proposed the "natural conjugate priors" that is more informative than the Jeffreys

prior.

The shape of the posterior distribution for the causal effect is bimodal when the instruments are weak and the effect of confounding is strong [141, 305]. Hoogerheide et al. [141] have developed a two-step procedure to capture this multi-modal shape of the posterior distribution, which the Gibbs sampler could not handle. The stages were (1) approximate the target posterior density via the construction of a neural network, (2) embed the posterior in a Metropolis-Hastings algorithm. An alternative sampling method for weak instruments was introduced by Zellner et al. [305], using an extension to Direct Monte Carlo (DMC), named Acceptance-Rejection within Direct Monte Carlo (ARDMC). They have shown through simulation that ARDMC is not only able to obtain the bimodal shape but is also more computationally efficient and gives more accurate numerical standard errors than the conventional Gibbs sampler.

Chib and Greenberg [64] described a semi-parametric Bayesian approach to IV models, where they do not assume the exogenous covariates (both instruments and measured confounders) to be in a parametric form for either the first or second regression. Their algorithm compares the different functional forms using Bayes factors. Their simulation studies suggest that their method is able to capture the non-linear form of the covariates for continuous outcome. However for the case of binary outcome, a large sample size is required.

There have been debates on the distribution of errors from X and Y; most econometricians use a bivariate normal distribution to model the correlated errors between X and Y. However some argue that the assumption of normality is not realistic. Conley et al. [74] have developed a semi-parametric Bayesian approach to this specific problem, assigning a Dirichlet process prior to allow for flexibility in the prior distribution of the errors between X and Y. Rossi [240] found this Bayesian semi-parametric procedure with Dirichlet process prior to have lower median bias and coverage in comparison to the classical approaches in the presence of many weak instruments. A nonparametric approach to Bayesian instrumental variables has been examined by Liao and Jiang [186] and references therein. Kato [163] has also implemented quasi-Bayesian analysis to the nonparametric instrumental variables model. Wiesenfarth et al. [288] have modified the algorithm from Conley et al. [74] to allow for even more flexible choice on prior distribution for the error terms. Their method does not assume linearity for the covariates effects and its credible interval is not affected by the distributional assumptions. Salois and Balcombe [244] have

addressed heteroskedasticity in either or both of the first and second regressions of the IV model by assuming Student t-distribution for the error terms and placing Gamma priors on the variance of the errors.

The exclusion restriction, assumption that an instrument does not have a direct effect on the outcome and is not associated with the error, is an untestable assumption for instrumental variable analysis. In order to allow for the uncertainty of this assumption, Kraay [174] designed priors for the correlation between the instruments and the errors from Y.

Bayesian model averaging (BMA) approaches are becoming more popular among econometricians for large number of exogenous variables, as a way to avoid overfitting the regression model. Koop et al. [172] introduced BMA into the framework of instrumental variable analysis. They argued that investigators may have uncertainty about whether their variables belong to the group of endogenous, exogenous or instruments, and BMA would be a way of incorporating this uncertainty. However, it is computationally intensive, if there is a large number of variables (m), as BMA will be comparing 2^m models. Hence they used Reversible Jump Markov Chain Monte Carlo (RJMCMC) to jump through model space and improved the speed of the algorithm by incorporating different priors, depending on which variable is under consideration. Alternatively, for smaller model spaces, Lenkoski et al. [184] have designed Two-stage Bayesian Model Averaging (2BMA) which is similar to 2SLS, and unlike Koop et al. [172], the algorithm approximates the marginal likelihoods. Through the simulations they showed that 2BMA does not suffer from many instrument bias unlike its classical counterparts. They used a Unit Information Prior (UIP) on the regression coefficients, which is a multivariate normal distribution with the mean taken from the coefficients estimated by the maximum likelihood estimator (MLE) and the variance is the inverse of the average information contained in one observation. UIP was motived by Raftery [232], where they argue the coefficients of the full model (i.e. without BMA) from a Bayesian approach are the same as obtained from MLE and this prior distribution has the same amount of information as a single observation.

8.4 Conclusion

The comparison of classical and Bayesian approaches has demonstrated the advantages of the use of Bayesian methods, except in the case of weak instruments [40].

The advice for reducing weak instrument bias is only to include instruments with a significant p-value from their association with risk factor. However selecting instruments by p-value cannot distinguish between situations in which (1) there are not enough data to detect an effect and (2) the data provide evidence for the null hypothesis [139].

Many econometric literature have addressed the weak instrument problem, more specifically Lenkoski et al. [184] have provided a solution to *many* weak instruments problem, through the use of Bayesian Model Averaging (BMA). BMA reduces the weak instrument bias by averaging the estimated causal effect from models with different sets of instruments. The selection of instruments is conditional on the likelihood of the data and the given prior. BMA also gives the posterior probability for each instrument and evidence to support the null hypothesis of "no causal effect". As we rarely know the location of the causal variant, BMA offers a way of comparing multiple models with different instruments without selection by p-value. As described above, BMA have been shown to be less affected by many weak instruments than the classical algorithms in Econometrics. Therefore the next chapter will examine the properties of BMA approach to Mendelian randomisation.

Chapter 9

Bayesian Model Averaging

9.1 Introduction

Suppose that there are m dependent SNPs, then by selecting which to use, there will be 2^m possible Mendelian randomisation models. The literature and our simulations both suggest that Mendelian randomisation will be more accurate if weak or redundant SNPs are excluded, but there is a risk that the process of selection would itself introduce bias. SNP selection may introduce bias, as the SNP may only be significant in one sample under analysis or the estimator may overestimate the true genetic effect, a phenomenon more commonly known as the 'winners curse'. Bayesian Model Averaging (BMA) considers the 2^m possible models and averages over them based on their posterior probabilities.

This chapter begins with a tutorial of BMA and its implementation in `OpenBUGS` and R, with examples. Next, the chapter compares `OpenBUGS` and R for BMA. Section 4 discusses the convergence of Instrumental Variable Bayesian Model Averaging (IVBMA), reaching convergence means having a stationary posterior distribution that is independent of the initial values [192]. For example in the BMA case, the MCMC sampler would have examined all the possible models at the beginning and eventually settles to a handful of models with the highest likelihoods. This section will aim to deduce the number of iterations required to reach convergence in different scenarios. The chapter will end with an investigation of the mechanism by which a SNP is chosen as an instrument in IVBMA.

9.2 Bayesian Model Averaging

Considering the fact that we are not entirely sure about which instruments will be best, BMA selects instruments and includes model uncertainty. Neglecting all the possible models has been described as a "quiet scandal" [33]. As we have seen in the previous chapters the number of outliers from LIML declined with instrument selection, however when selection is applied to the same data as is used in the analysis this will cause selection bias. BMA acts in the way of a sensitivity analysis that considers all the possible estimates when based on different instruments. This section will give the theory of BMA and a practical example to aid the understanding of BMA. Hoeting et al. [139] have written a comprehensive tutorial on BMA.

If there are K possible models, then the posterior distribution of the quantity of interest Λ given data D is;

$$pr(\Lambda|D) = \sum_{k=1}^{K} pr(\Lambda|M_k, D)pr(M_k|D). \tag{9.1}$$

This is an average of the posterior distributions under each of the models (M_k), weighted by their posterior model probability. In (9.1), $M_1, \ldots, M_K$ are the models considered. The posterior probability for model M_k is given by

$$pr(M_k|D) = \frac{pr(D|M_k)pr(M_k)}{\sum_{l=1}^{K} pr(D|M_l)pr(M_l)}. \tag{9.2}$$

where

$$pr(D|M_k) = \int pr(D|\theta_k, M_k)pr(\theta_k|M_k)d\theta_k \tag{9.3}$$

is the integrated likelihood of model M_k, θ_k is the vector of parameters of model M_k, $pr(\theta_k|M_k)$ is the prior density of θ_k under model M_k, $pr(D|\theta_k, M_k)$ is the likelihood and $pr(M_k)$ is the prior probability that M_k is the true model.

The posterior mean and variance of Λ are;

$$E[\Lambda|D] = \sum_{k=0}^{K} \hat{\Lambda}_k pr(M_k|D),$$
$$Var[\Lambda|D] = \sum_{k=0}^{K} (Var[\Lambda|D, M_k] + \hat{\Lambda}_k^2)pr(M_k|D) - E[\Lambda|D]^2,$$

where $\hat{\Lambda}_k = E[\Lambda|D, M_k]$.

The main advantage of applying BMA to Mendelian randomisation is to reduce the many weak instrument bias by allowing flexibility in instrument inclusion without introducing selection bias.

9.2.1 OpenBUGS

To actually evaluate Equation 9.1 to give the posterior distribution is extremely difficult, therefore simulations are used to approximate the distribution, the method more commonly known as Markov Chain Monte Carlo (MCMC). MCMC is a sampling method that constructs a markov chain so that the target distribution is the stationary distribution of the chain, and the chain converges to a distribution that is the stationary distribution [178]. Generally, a markov chain can be constructed by Metropolis-Hastings algorithm, where it samples values from a proposal distribution and accepts or rejects proposed value using an acceptance probability. If the proposed value is accepted then markov chain moves to the proposed value, otherwise the chain stays at the current value. See Robert and Casella [237] for more detail on Metropolis-Hastings algorithm.

Bayesian models are commonly fitted in `OpenBUGS` [263], where the name stands for "**B**ayesian inference **U**sing **G**ibbs **S**ampling". Gibbs sampler is a special case of Metropolis-Hastings algorithm and approximates distributions for more than one parameter [192] by generating a multi-dimensional markov chain. This section will give the model code for Mendelian randomisation (MR) with and without BMA, henceforth "full" and "BMA" model respectively, in `OpenBUGS`. An example will show how to obtain results and will explain their implications.

Consider the following MR model with three potential genetic instruments, SNP_1, SNP_2 and SNP_3;

$$[X, Y] \sim MVN\left(\begin{bmatrix} \alpha_0 + \alpha_1 SNP_1 + \alpha_2 SNP_2 + \alpha_3 SNP_3 \\ \beta_0 + \beta_1 X \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \lambda \\ \lambda & \sigma_y^2 \end{bmatrix}\right) \tag{9.4}$$

The error of X and Y is monitored by bivariate normal distribution. σ_x^2 and σ_y^2 is the variance of X and Y respectively and $\lambda = Cov(X, Y) = \rho\sigma_x\sigma_y$ and ρ is the correlation between X and Y, i.e. the amount of confounding between them. Then the code for the "full" model is as follows;

```
model {
```

```
  for( i in 1 : N ) {
    mu[i,1]<-alpha[1] + alpha[2] * SNP1[i]
    + alpha[3] * SNP2[i] + alpha[4]* SNP3[i]
    mu[i,2] <- beta[1] +  beta[2]*XY[i,1]
    XY[i,1:2] ~ dmnorm(mu[i,1:2],Sigma.inv[1:2,1:2])
  }
  #priors for first-stage regression
   for (j in 1 : 4) {
    alpha[j] ~  dnorm(0,1);}
 #priors for second-stage regression
  for (j in 1 : 2) {
    beta[j] ~  dnorm(0,1); }

  #priors for correlated errors
  Sigma.inv[1:2,1:2] ~ dwish(R[1:2,1:2],3)
  Sigma[1:2,1:2]<-inverse(Sigma.inv[1:2,1:2])
}
```

I have given the coefficients for first and second stage regression a normal prior, $\sim N(0,1)$, which is fairly informative [153]. A vague prior is given to the covariance matrix of X and Y in the form of an Inverse-Wishart, $\mathcal{W}^{-1}(\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}, 3)$ [192]. These priors were not applied for a specific reason, merely as a demonstration. These priors can be changed to different distributions to accommodate different problems, see Lunn et al. [192] for a detailed tutorial on priors in Bayesian analysis.

To integrate BMA into an MR model, add variable indicators to each of the first stage regression coefficients. For my case, these are `g[1]` to `g[4]` where `g[]` is coded 0 and 1 to indicate the absence and the presence of a variable in the current model respectively. The model code has thus been altered to the following;

```
       model {
  for( i in 1 : N ) {
    mu[i,1]<-g[1]*alpha[1] + g[2]*alpha[2] * SNP1[i]
    + g[3]*alpha[3] * SNP2[i] + g[4]*alpha[4]* SNP3[i]
    mu[i,2] <- beta[1] +  beta[2]*XY[i,1]
    XY[i,1:2] ~ dmnorm(mu[i,1:2],Sigma.inv[1:2,1:2])
```

```
  }
   #priors for first-stage regression
   for (j in 1 : 4) {
    alpha[j] ~  dnorm(0,1);
  }
 #priors for second-stage regression
  for (j in 1 : 2) {
    beta[j] ~  dnorm(0,1);
  }
 #priors for variable indicators
  for (j in 1 : 4) {
    g[j] ~ dbern(0.5); }

  #priors for correlated errors
  Sigma.inv[1:2,1:2] ~ dwish(R[1:2,1:2],3)
  Sigma[1:2,1:2]<-inverse(Sigma.inv[1:2,1:2])

  #Defining Model Code
  mdl <- g[1]*1+g[2]*2+g[3]*4+g[4]*8

  #Defining vector with model indicators
  for (j in 1 : models) {
    pmdl[j]<-equals(mdl , j); }
}
```

The prior for variable indicators is a Bernoulli distribution with parameter 0.5; this prior implies each variable is equally likely to be chosen for inclusion in the model. As there are more than two variables in the first-regression, I have defined model indicators in the code, so that `OpenBUGS` will give the output of the probability of each model being chosen in the MCMC iterations;

```
  #Defining Model Code
  mdl <- g[1]*1+g[2]*2+g[3]*4+g[4]*8

  #Defining vector with model indicators
```

```
for (j in 1 : models) {
  pmdl[j]<-equals(mdl , j); }
```

`pmdl` is the model indicator and `mdl` shows which variables are within the particular model. Shown by `mdl`, presence of intercept (`g[1]`) is given the value of 1, presence of SNP_1 (`g[2]`) is given the value of 2, etc. Hence `pmdl[1]` is the model that consists of the intercept (`g[1]`$\times 1$), `pmdl[2]` has SNP_1 (`g[2]`$\times 2$), `pmdl[3]` has intercept and SNP_1 (`g[1]`$\times 1$+`g[2]`$\times 2$), and so on and so forth.

9.2.2 Example

This example aims to demonstrate the difference between the full and BMA models. Consider a dataset with information on a risk factor (X), disease outcome of interest (Y) and genotypes of 3 SNPs for 2,000 individuals. Out of the 3 potential instruments, only one SNP is causal and explains 2% of the variation in X, the other SNPs are only associated with X through their correlation with the causal SNP (correlation is approximately 0.9). X explains 6% of the variation in Y. X and Y are normally distributed.

The model code is the same as above, the full and the BMA model will be run with 50,000 iterations on the same dataset, and the first 10,000 iterations will be dismissed as a burn-in. Burn-in is the initial, non-stationary portion of the chain which are usually dismissed, as they will influence the overall mean of the posterior distribution. All initial values for each parameter are randomly generated. The R package *rbugs* provides a "click-free" way of operating `OpenBUGS`; *rbugs* sends the code to `OpenBUGS`, which returns the results in coda format. The R package *coda* then reads the coda to give the mean, standard deviation, naive standard error, time-series standard error and quantiles for each variable. The time-series standard error is adjusted for autocorrelation (i.e. the correlation between each iteration).

The output for the full model, including all the available variables from the data is;

```
        Iterations = 1:40000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 40000
```

```
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:
                Mean      SD  Naive SE Time-series SE
Sigma[1,1]  9.572e-01 0.03011 0.0001506      0.0001524
Sigma[1,2]  3.538e-01 0.13072 0.0006536      0.0064359
Sigma[2,1]  3.538e-01 0.13072 0.0006536      0.0064359
Sigma[2,2]  8.729e-01 0.10783 0.0005391      0.0053724
alpha[1]   -1.413e-02 0.03943 0.0001972      0.0005278
alpha[2]   -2.772e-02 0.04387 0.0002193      0.0002896
alpha[3]    2.264e-01 0.03370 0.0001685      0.0006931
alpha[4]   -3.220e-02 0.03046 0.0001523      0.0002930
beta[1]    -3.843e-02 0.03185 0.0001593      0.0012257
beta[2]     3.869e-01 0.13461 0.0006731      0.0067916

2. Quantiles for each variable:
                2.5%        25%        50%        75%     97.5%
Sigma[1,1]  9.001e-01  9.367e-01  9.565e-01  9.770e-01 1.018e+00
Sigma[1,2]  1.110e-01  2.652e-01  3.481e-01  4.350e-01 6.264e-01
Sigma[2,1]  1.110e-01  2.652e-01  3.481e-01  4.350e-01 6.264e-01
Sigma[2,2]  7.283e-01  7.973e-01  8.523e-01  9.240e-01 1.135e+00
alpha[1]   -9.200e-02 -4.077e-02 -1.395e-02  1.273e-02 6.278e-02
alpha[2]   -1.145e-01 -5.720e-02 -2.726e-02  1.797e-03 5.763e-02
alpha[3]    1.609e-01  2.038e-01  2.264e-01  2.489e-01 2.926e-01
alpha[4]   -9.277e-02 -5.267e-02 -3.178e-02 -1.146e-02 2.699e-02
beta[1]    -9.775e-02 -6.025e-02 -3.957e-02 -1.824e-02 2.809e-02
beta[2]     1.065e-01  3.035e-01  3.920e-01  4.783e-01 6.374e-01
```

The true causal effect is 0.2449. The full model estimated 0.3869 (`beta[1]`) with 95% credible interval of 0.1065 to 0.6374, and the true effect is still within the credible interval. The full model estimated a positive confounding effect (`Sigma[1,2]` and `Sigma[2,1]`), where the correlation between X and Y is estimated at 0.3538 (the true is 0.5).

The output for the BMA model for the same dataset is;

```
        Iterations = 1:40000
Thinning interval = 1
```

```
Number of chains = 1
Sample size per chain = 40000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:
                 Mean      SD  Naive SE Time-series SE
Sigma[1,1]  9.570e-01 0.03016 0.0001508      1.508e-04
Sigma[1,2]  4.070e-01 0.14183 0.0007092      7.890e-03
Sigma[2,1]  4.070e-01 0.14183 0.0007092      7.890e-03
Sigma[2,2]  9.182e-01 0.12778 0.0006389      6.899e-03
alpha[1]   -3.542e-03 0.97131 0.0048566      4.815e-03
alpha[2]   -5.053e-03 0.97442 0.0048721      4.872e-03
alpha[3]    1.923e-01 0.02053 0.0001027      3.321e-04
alpha[4]    1.481e-03 0.97354 0.0048677      4.868e-03
beta[1]    -2.511e-02 0.03247 0.0001624      1.401e-03
beta[2]     3.315e-01 0.14577 0.0007289      7.889e-03
g[1]        5.245e-02 0.22294 0.0011147      3.043e-03
g[2]        5.275e-02 0.22354 0.0011177      2.242e-03
g[3]        1.000e+00 0.00000 0.0000000      0.000e+00
g[4]        5.727e-02 0.23237 0.0011619      3.689e-03
pmdl[1]     0.000e+00 0.00000 0.0000000      0.000e+00
pmdl[2]     0.000e+00 0.00000 0.0000000      0.000e+00
pmdl[3]     0.000e+00 0.00000 0.0000000      0.000e+00
pmdl[4]     8.459e-01 0.36105 0.0018052      4.671e-03
pmdl[5]     4.735e-02 0.21239 0.0010619      2.852e-03
pmdl[6]     4.675e-02 0.21111 0.0010555      2.083e-03
pmdl[7]     2.725e-03 0.05213 0.0002607      4.064e-04
pmdl[8]     0.000e+00 0.00000 0.0000000      0.000e+00
pmdl[9]     0.000e+00 0.00000 0.0000000      0.000e+00
pmdl[10]    0.000e+00 0.00000 0.0000000      0.000e+00
pmdl[11]    0.000e+00 0.00000 0.0000000      0.000e+00
pmdl[12]    5.170e-02 0.22142 0.0011071      3.338e-03
pmdl[13]    2.300e-03 0.04790 0.0002395      4.654e-04
pmdl[14]    3.200e-03 0.05648 0.0002824      5.993e-04
pmdl[15]    7.500e-05 0.00866 0.0000433      5.561e-05
```

```
pmdl[16]    0.000e+00 0.00000 0.0000000      0.000e+00

2. Quantiles for each variable:
                2.5%        25%        50%        75%     97.5%
Sigma[1,1]  8.994e-01  9.363e-01  9.564e-01  9.773e-01 1.018e+00
Sigma[1,2]  1.360e-01  3.100e-01  4.083e-01  5.015e-01 6.856e-01
Sigma[2,1]  1.360e-01  3.100e-01  4.083e-01  5.015e-01 6.856e-01
Sigma[2,2]  7.364e-01  8.240e-01  8.995e-01  9.895e-01 1.216e+00
alpha[1]   -1.934e+00 -6.346e-01 -1.733e-02  6.248e-01 1.937e+00
alpha[2]   -1.940e+00 -6.333e-01 -1.779e-02  6.246e-01 1.940e+00
alpha[3]    1.546e-01  1.787e-01  1.915e-01  2.048e-01 2.362e-01
alpha[4]   -1.933e+00 -6.251e-01 -2.198e-02  6.261e-01 1.948e+00
beta[1]    -8.942e-02 -4.675e-02 -2.470e-02 -3.138e-03 3.775e-02
beta[2]     4.719e-02  2.341e-01  3.293e-01  4.315e-01 6.117e-01
g[1]        0.000e+00  0.000e+00  0.000e+00  0.000e+00 1.000e+00
g[2]        0.000e+00  0.000e+00  0.000e+00  0.000e+00 1.000e+00
g[3]        1.000e+00  1.000e+00  1.000e+00  1.000e+00 1.000e+00
g[4]        0.000e+00  0.000e+00  0.000e+00  0.000e+00 1.000e+00
pmdl[1]     0.000e+00  0.000e+00  0.000e+00  0.000e+00 0.000e+00
pmdl[2]     0.000e+00  0.000e+00  0.000e+00  0.000e+00 0.000e+00
pmdl[3]     0.000e+00  0.000e+00  0.000e+00  0.000e+00 0.000e+00
pmdl[4]     0.000e+00  1.000e+00  1.000e+00  1.000e+00 1.000e+00
pmdl[5]     0.000e+00  0.000e+00  0.000e+00  0.000e+00 1.000e+00
pmdl[6]     0.000e+00  0.000e+00  0.000e+00  0.000e+00 1.000e+00
pmdl[7]     0.000e+00  0.000e+00  0.000e+00  0.000e+00 0.000e+00
pmdl[8]     0.000e+00  0.000e+00  0.000e+00  0.000e+00 0.000e+00
pmdl[9]     0.000e+00  0.000e+00  0.000e+00  0.000e+00 0.000e+00
pmdl[10]    0.000e+00  0.000e+00  0.000e+00  0.000e+00 0.000e+00
pmdl[11]    0.000e+00  0.000e+00  0.000e+00  0.000e+00 0.000e+00
pmdl[12]    0.000e+00  0.000e+00  0.000e+00  0.000e+00 1.000e+00
pmdl[13]    0.000e+00  0.000e+00  0.000e+00  0.000e+00 0.000e+00
pmdl[14]    0.000e+00  0.000e+00  0.000e+00  0.000e+00 0.000e+00
pmdl[15]    0.000e+00  0.000e+00  0.000e+00  0.000e+00 0.000e+00
pmdl[16]    0.000e+00  0.000e+00  0.000e+00  0.000e+00 0.000e+00
```

The BMA model provides the probability that each variable is chosen in the first regression (`g[1]` to `g[4]`) and the probability that each model is the current model (`pmdl`), where there are 16 possible combination of models from 4 variables ($2^4 = 16$). The BMA model estimated the causal effect as 0.3315 and 95% credible interval of 0.0472 to 0.6117. Unlike the full model, BMA shows which instrument and model was preferred; the causal SNP had the highest probability, as shown by mean of `g[3]` in comparison to other variable. `pmdl[4]`, model 4 was chosen the most which included only the causal SNP as a variable (pmdl= `g[4]` $\times 4$). Note that some models were never chosen.

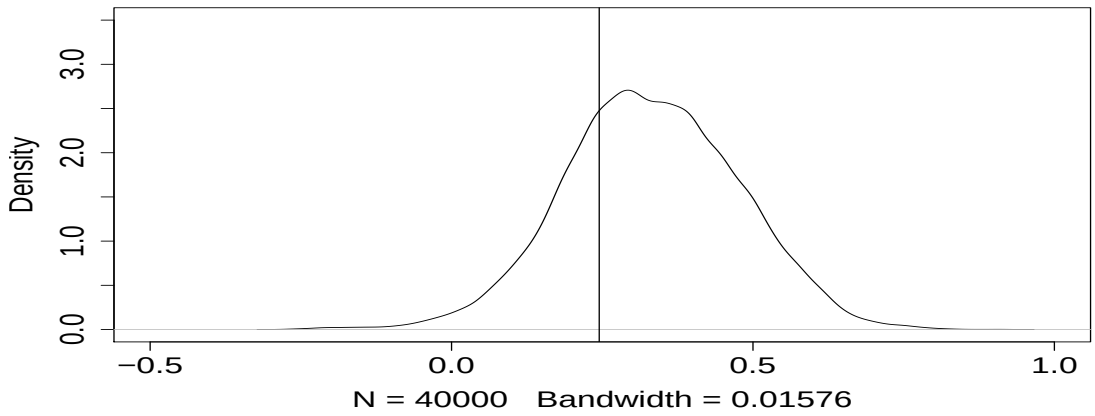


Figure 9.1: The overall posterior distribution of all the models. The vertical line is the true causal effect.

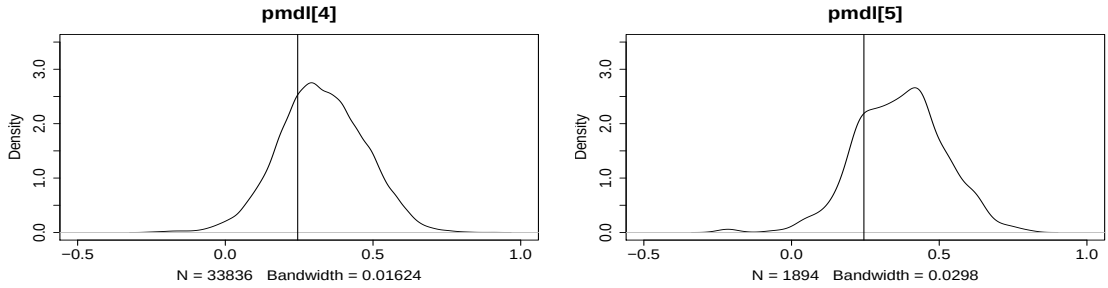
BMA approximates its posterior distribution by averaging the posterior distribution from each model which is weighted by the model probability; Figure 9.1 is the overall posterior distribution for all the models. The overall posterior model is the average of the posterior distributions for all the models in Figure 9.2, each weighted by their model probability. The posterior from `pmdl[4]` have the highest weighting, as its probability is 0.8459 and `pmdl[15]` has the lowest weight with probability of 0.000075. Hence, that is why Figure 9.1 looks identical to Figure 9.2a.

9.2.3 Conclusion

`OpenBUGS` is a very general MCMC program and it can be inefficient for specific cases [192]; Example 9.2.2 took 30 minutes for 50,000 iterations with 4 potential covariates (2^4 =16 potential models as a result). *ivbma* is an R package that incorporates

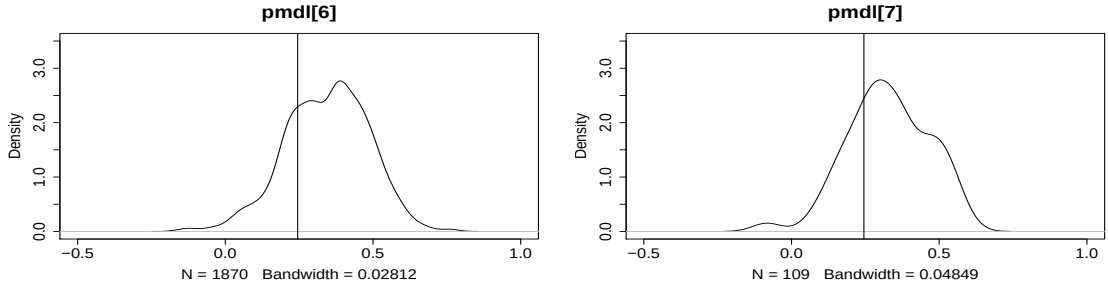
algorithms to optimise the instrumental variable Bayesian model averaging approach and reduce the computation time.

The number of iterations was chosen for illustration, this will be examined in detail at Section 9.4.



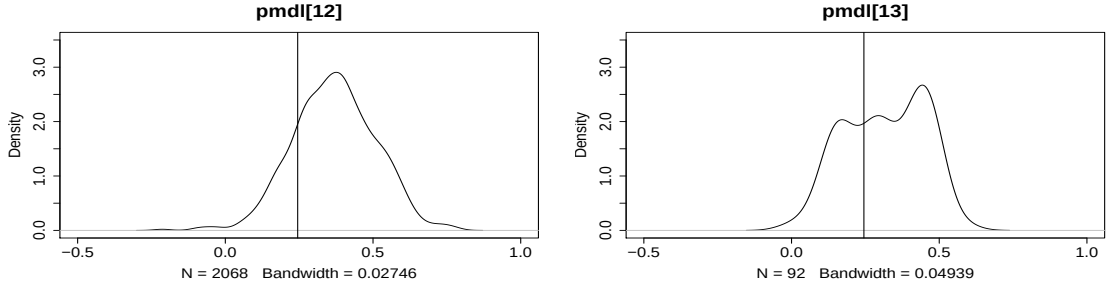
(a) SNP_c

(b) $Intercept + SNP_c$



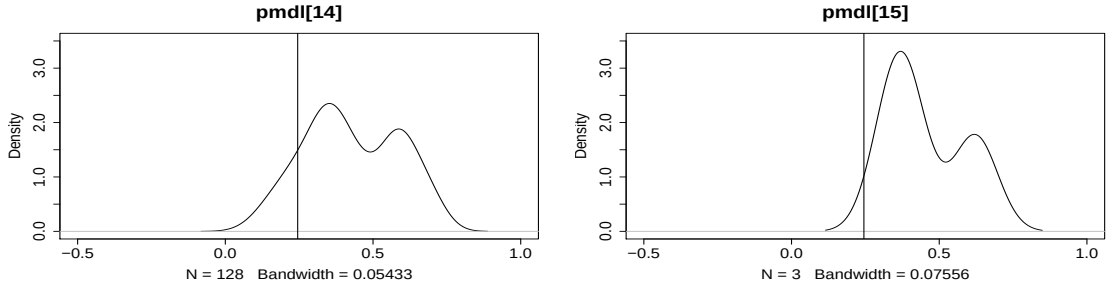
(c) $SNP_1 + SNP_c$

(d) $Intercept + SNP_1 + SNP_c$



(e) $SNP_c + SNP_2$

(f) $Intercept + SNP_c + SNP_2$



(g) $SNP_1 + SNP_c + SNP_2$

(h) $Intercept + SNP_1 + SNP_c + SNP_2$

Figure 9.2: The posterior distribution for each model and N is the number of times the specific model was chosen. The captions are variables that is within the model. The vertical line is the true causal effect.

9.3 R Package: IVBMA

Karl and Lenkoski [160] have designed algorithms specifically for instrumental variable Bayesian model averaging (IVBMA) and later written an R package, *ivbma*, to implement this approach. IVBMA is more efficient when compared to 2BMA [184] and Koop et al. [172]'s algorithm, see Section 8.3. The former is less computationally efficient. For the same dataset 2BMA took 15 hours whereas IVBMA took 10 minutes. In addition, 2BMA had mixing difficulties and required further tuning to improve the mixing between models, mixing in BMA mostly means how easy it is for a sampler to go from one model to another, see the next section for more detail.

IVBMA relies on Markov Chain Monte Carlo Model Composition (MC3) within a Gibbs sampler, which is just a special case of a Metropolis-within-Gibbs algorithm. MC3 can be considered as a Metropolis-Hastings step in the space of the models; MC3 moves through model space, accepting or rejecting a model via Conditional Bayes Factor. MC3-within-Gibbs sampler is more efficient in regards to mixing than `OpenBUGS` when there are many models in the model space [210]. The IVBMA model is;

$$\boldsymbol{Y} = \boldsymbol{X}\beta_{XY} + \boldsymbol{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{9.5}$$

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\delta} + \boldsymbol{W}\boldsymbol{\tau} + \boldsymbol{\eta} \tag{9.6}$$

where $\binom{\varepsilon_i}{\eta_i} \sim N_2(0, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$. $\boldsymbol{Y}$, the response variable and the endogenous explanatory factor $\boldsymbol{X}$ are both $n \times 1$. $\boldsymbol{W}$ denotes an $n \times p$ matrix of further explanatory variables and $\boldsymbol{Z}$ contains the instrumental variables with an $n \times q$ matrix.

The MC3-within-Gibbs sampler creates a sequence $\theta^{(1)} \dots \theta^{(S)}$ where;

$$\theta^{(s)} = \left\{ \boldsymbol{\rho}^{(s)}, \mathcal{M}_{sec}^{(s)}, \boldsymbol{\lambda}^{(s)}, \mathcal{M}_{fst}^{(s)}, \Sigma^{(s)} \right\},$$

$\boldsymbol{\rho}^{(s)} = [\beta_{XY}, \boldsymbol{\gamma}]$, $\boldsymbol{\lambda}^{(s)} = [\boldsymbol{\delta}, \boldsymbol{\tau}]$, $\mathcal{M}_{sec}$ and $\mathcal{M}_{fst}$ is the model space for Equation 9.5 and 9.6 respectively. Given the current state $\rho^{(s)}$ and data $\mathcal{D}$, the simplified *ivbma* algorithm starts;

1. Sample $\mathcal{M}'_{sec}$ from the neighbourhood of $\mathcal{M}_{sec}^{(s)}$, i.e. models that differ from

$\mathcal{M}_{sec}^{(s)}$ by one variable. Then calculate

$$\alpha = \frac{pr(\mathcal{D}|\mathcal{M}'_{sec}, \boldsymbol{\lambda}^{(s)}, \Sigma^{(s)})}{pr(\mathcal{D}|\mathcal{M}_{sec}^{(s)}, \boldsymbol{\lambda}^{(s)}, \Sigma^{(s)})} 1\left\{\mathcal{M}'_{sec}, \mathcal{M}_{fst}^{(s)} \in \mathcal{A}\right\}$$

with probability $min\{\alpha, 1\}$ set $\mathcal{M}_{sec}^{(s+1)} = \mathcal{M}'_{sec}$, otherwise $\mathcal{M}_{sec}^{(s+1)} = \mathcal{M}_{sec}^{(s)}$

2. Sample $\boldsymbol{\rho}^{(s+1)}$ from the conditional posterior distribution of $\boldsymbol{\rho}_{\mathcal{M}_{sec}^{(s+1)}}$, i.e. posterior distribution for coefficients of the new model in the second stage regression.

3. Sample $\mathcal{M}'_{fst}$ from the neighbourhood of $\mathcal{M}_{fst}^{(s)}$. Then calculate

$$\alpha = \frac{pr(\mathcal{D}|\mathcal{M}'_{fst}, \boldsymbol{\rho}^{(s+1)}, \Sigma^{(s)})}{pr(\mathcal{D}|\mathcal{M}_{fst}^{(s)}, \boldsymbol{\rho}^{(s+1)}, \Sigma^{(s)})} 1\left\{\mathcal{M}_{sec}^{(s+1)}, \mathcal{M}'_{fst} \in \mathcal{A}\right\}$$

with probability $min\{\alpha, 1\}$ set $\mathcal{M}_{fst}^{(s+1)} = \mathcal{M}'_{fst}$, otherwise $\mathcal{M}_{fst}^{(s+1)} = \mathcal{M}_{fst}^{(s)}$

4. Sample $\boldsymbol{\lambda}^{(s+1)}$ from the conditional posterior distribution of $\boldsymbol{\lambda}_{\mathcal{M}_{fst}^{(s+1)}}$, i.e. posterior distribution for coefficients of the new model in the first stage regression.

5. Use $\boldsymbol{\lambda}^{(s+1)}$ and $\boldsymbol{\rho}^{(s+1)}$ to calculate $\boldsymbol{\varepsilon}^{s+1}$ and $\boldsymbol{\eta}^{(s+1)}$ and sample $\boldsymbol{\Sigma}^{(s+1)}$ from the conditional posterior distribution of $\boldsymbol{\Sigma}$.

See Karl and Lenkoski [160] for the derivation of Bayes Factor in Step (1) and (3), and the full equation of the conditional posterior distributions for each parameter.

ivbma imposes the following prior for each parameter:

$$\begin{aligned}
[\beta_{XY}, \boldsymbol{\gamma}] &\sim N(0, 1), \\
[\boldsymbol{\delta}, \boldsymbol{\tau}] &\sim N(0, 1), \\
[\mathcal{M}_{fst}, \mathcal{M}_{sec}] &\sim Bern(0.5), \\
\boldsymbol{\Sigma} &\sim \mathcal{W}^{-1}(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, 3).
\end{aligned}$$

where N, $Bern$ and $\mathcal{W}^{-1}$ stands for the Normal, Bernoulli and Inverse-Wishart distribution respectively. *ivbma* does not provide an option to alter these priors. Lenkoski et al. [184] provide the full instructions and options available for the R package.

9.3.1 Comparison of OpenBUGS and R

The aim of this subsection is to demonstrate that the parameters of interest estimated by `OpenBUGS` and R for the same dataset are similar.

The single dataset will consist of risk factor (X), outcome of interest (Y) and genotype of 3 SNPs for 2,000 individuals; only one SNP, SNP_c, explains 10% of the variation in X. The two non-causal SNPs will have correlation of 0.9 to SNP_c, thus they will be associated with X. X will explain 6% of the variation in Y. X and Y will have a normal distribution. To reduce computation time stronger instruments have been simulated here, as the algorithm in `OpenBUGS` is inefficient and needs a large number of iterations to converge for weak instruments.

In order to give evidence of convergence to the posterior distribution of the causal effect estimate and the model space, as per advised by Karl and Lenkoski [160], the experiment will run 50,000 iterations with a 10,000 burnin. Five chains will be run, each with different initial values (randomly chosen) to ensure all five chains reach the same posterior [192]. The choice of priors in `OpenBUGS` will be specified to be equivalent to those in *ivbma*. As *ivbma* allows for variable uncertainty in both the first and second regressions, the model code for `OpenBUGS` in Example 9.2.2 will include variable indicators for the second regression, see Appendix E.1 for the addition to the model code.

Table 9.1 shows that the probability of X included in the second regression, mean and 95% credible interval of the causal effect estimate (β_{XY}) are the same to at least 2 significant places between *ivbma* and `OpenBUGS`. Note that the β_{XY} estimate from `OpenBUGS` was corrected to 0 when X is not in the model, the reason for this alteration will be explained below.

Table 9.1: The summary causal effect estimates from *ivbma* and `OpenBUGS` for 5 chains. Prob. X is the probability of X being included in the second regression.

	IVBMA		BUGS	
Chain	Prob. X	Mean (95% Credible Interval)	Prob.X	Mean (95% Credible Interval)
1	0.9980	0.2037 (0.11 - 0.29)	1.0000	0.2046 (0.11 - 0.29)
2	0.9991	0.2059 (0.12 - 0.29)	0.9873	0.2031 (0.10 - 0.29)
3	0.9991	0.2039 (0.11 - 0.29)	0.9941	0.2049 (0.11 - 0.29)
4	0.9991	0.2048 (0.12 - 0.29)	0.9952	0.2052 (0.11 - 0.29)
5	0.9994	0.2053 (0.12 - 0.29)	0.9994	0.2059 (0.11 - 0.29)

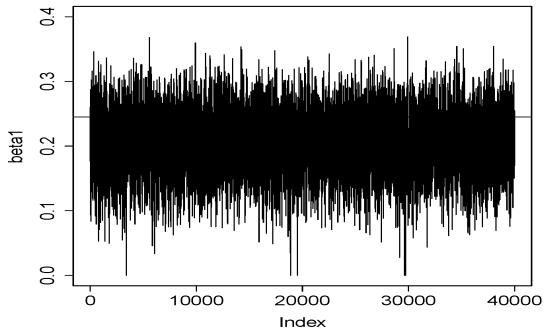
50,000 iterations and a 10,000 burn-in have given evidence of the convergence of the posterior distribution of β_{XY} for *ivbma*, as shown by the "caterpillar-shape" appearance from all 5 chains [192], in Figure 9.3, and the similar results from all 5 chains (Table 9.1). The difference between *ivbma* and `OpenBUGS` other than the computation time, is their output; when X is not in the model *ivbma* automatically sets the output of β_{XY} to 0, whereas `OpenBUGS` does not; hence the difference between Figure 9.3 and Figure 9.4. When X is included in the model the posterior distribution is the combination of likelihood of the data and prior distribution, when X is not in the model, the posterior estimate is from the prior only. Our aim is to find the unconfounded effect of X and Y using instrumental variable analysis, thus we are only interested in β_{XY} when X is in the model. Manually setting the `OpenBUGS` output of β_{XY} to 0 when indicator variable for X is 0 (Figure 9.5), Figure 9.3 and Figure 9.5 are similar.

For the BMA problem, it is essential to check the convergence in model space, as different models will give different coefficient estimates. The model choice in the first regression will influence the causal effect estimate through its association with X. Model choice for the first regression is given by Figure 9.6 and 9.7. These two trace plots show that *ivbma* has approximately the same mixture of models as in `OpenBUGS`. The similarity in model choice from each chain indicates convergence in model space from *ivbma* and `OpenBUGS`.

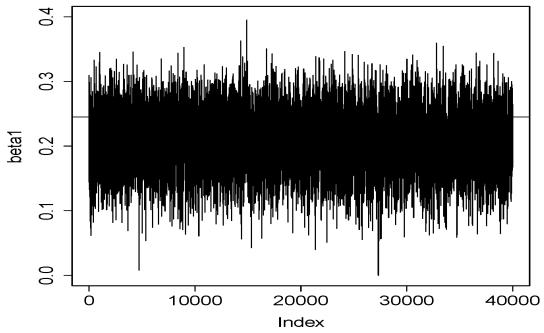
9.3.2 Conclusion

This section have given a description of IVBMA and illustrated the similarity between *ivbma* and `OpenBUGS`. Differences are mainly their computation time; for the same dataset, number of iterations and chains *ivbma* took 15 minutes, whereas `OpenBUGS` took 2 hours and 30 minutes to run.

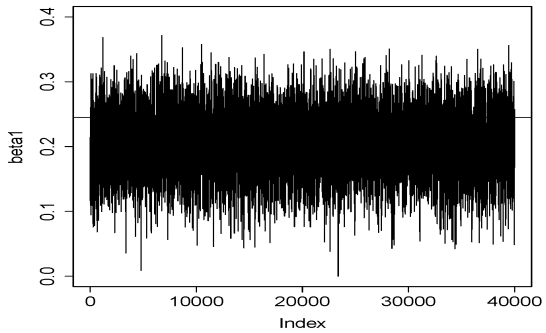
The convergence of the posterior distribution is an importance aspect to Bayesian analysis and Karl and Lenkoski [160] suggested 50,000 iterations to reach convergence for their R package. However the scenarios considered in their paper were applicable to econometrics. Therefore the next section will examine the convergence of *ivbma* for different problems that are realistic in a genetic setting.



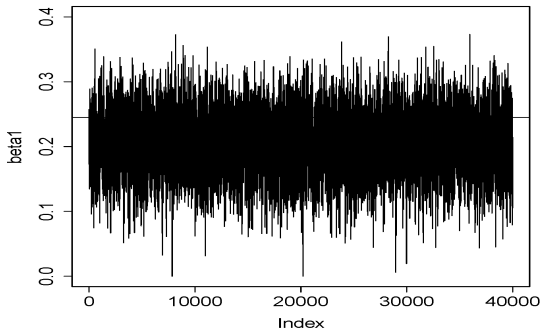
(a) Chain 1



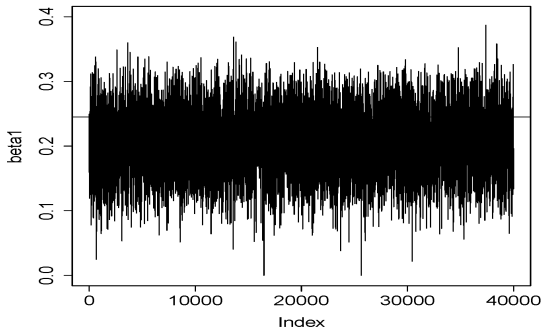
(b) Chain 2



(c) Chain 3

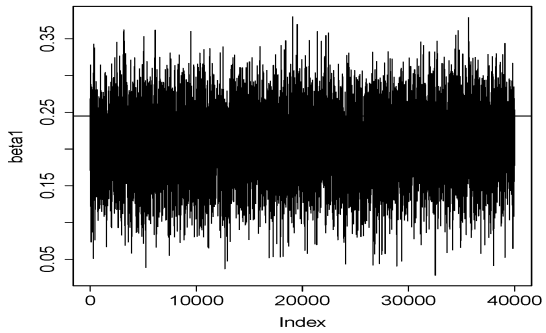


(d) Chain 4

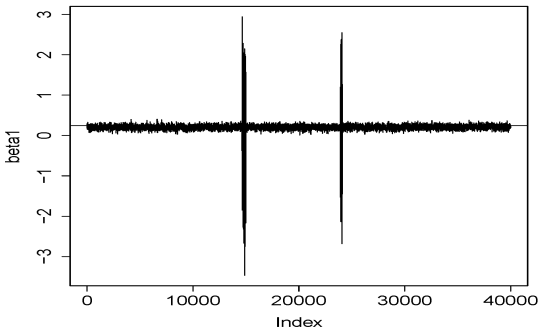


(e) Chain 5

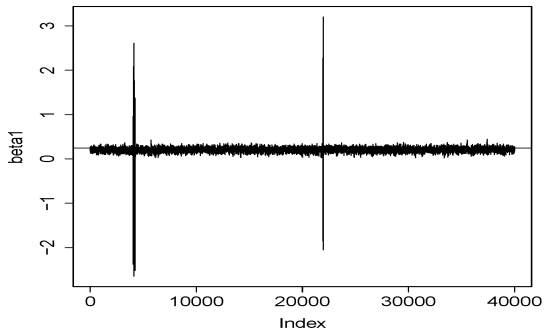
Figure 9.3: Trace plot of $\hat{\beta}_{XY}$ from *ivbma* for 5 chains. The horizontal line is the true causal effect.



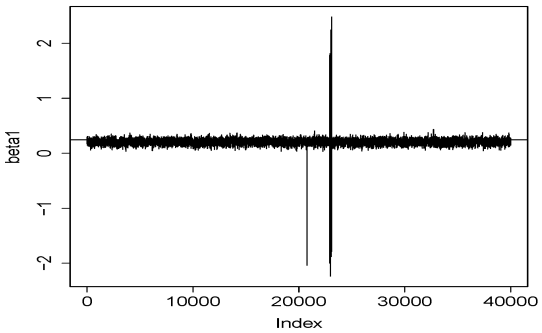
(a) Chain 1



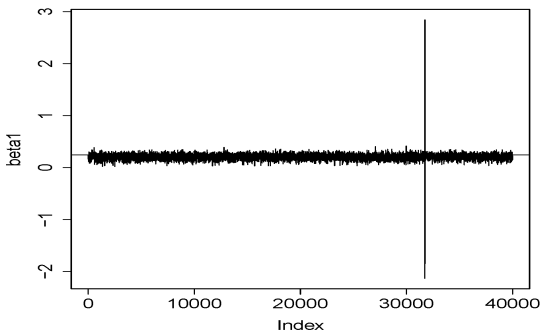
(b) Chain 2



(c) Chain 3

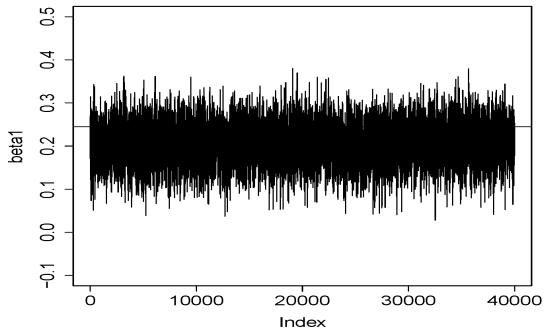


(d) Chain 4

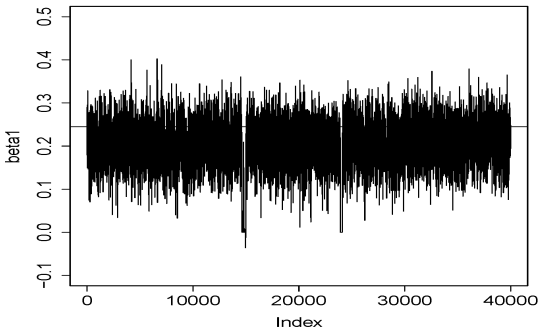


(e) Chain 5

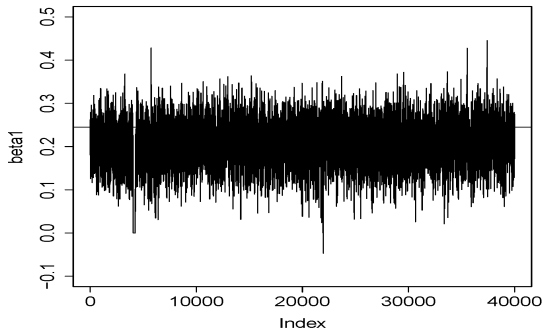
Figure 9.4: Trace plot of $\hat{\beta}_{XY}$ from `OpenBUGS` for 5 chains. The horizontal line is the true causal effect (0.2449).



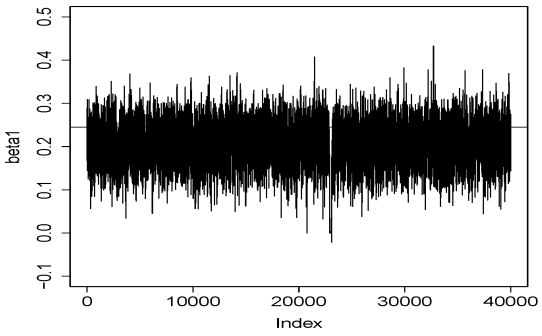
(a) Chain 1



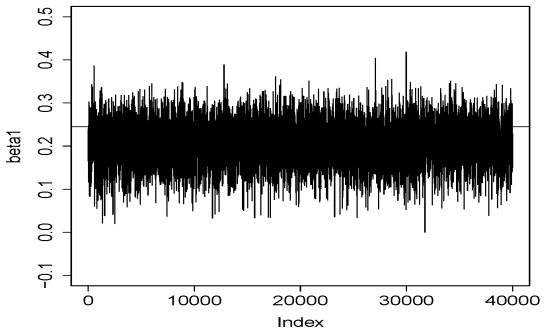
(b) Chain 2



(c) Chain 3

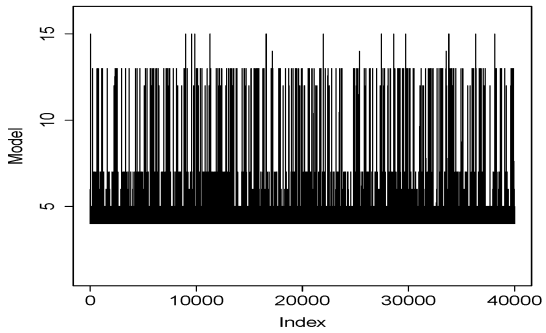


(d) Chain 4

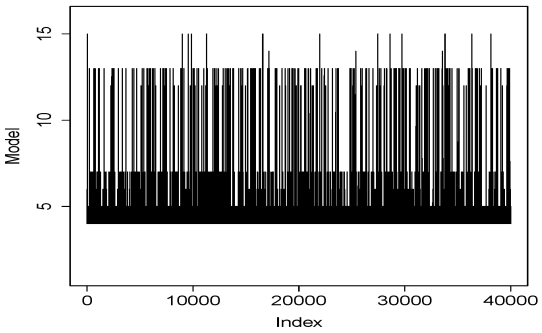


(e) Chain 5

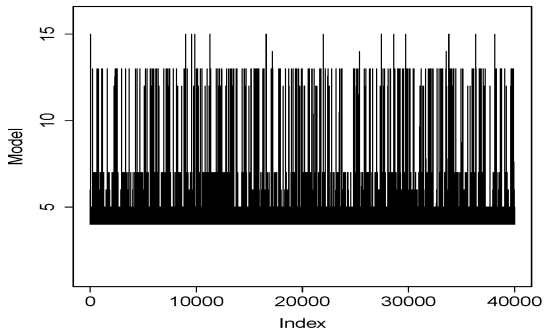
Figure 9.5: Trace plot of $\hat{\beta}_{XY}$ from `OpenBUGS` for 5 chains, note that $\hat{\beta}_{XY}$ have been set to 0 when X is not included in the second regression. The horizontal line is the true causal effect (0.2449).



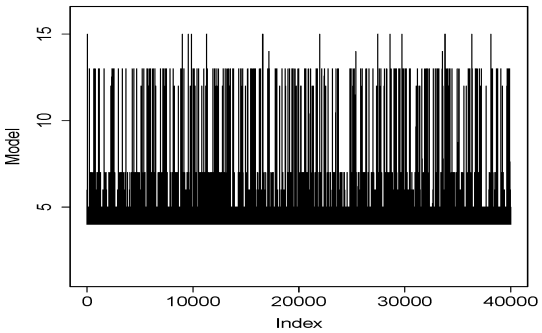
(a) Chain 1



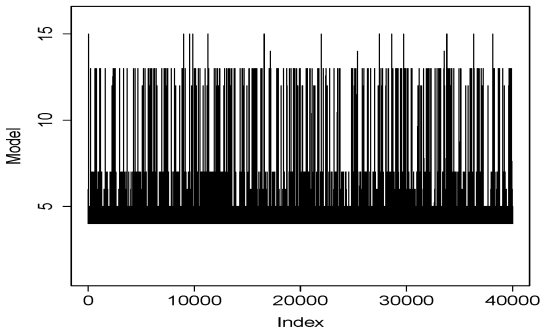
(b) Chain 2



(c) Chain 3

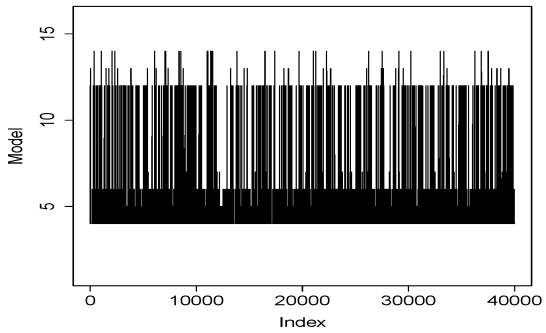


(d) Chain 4

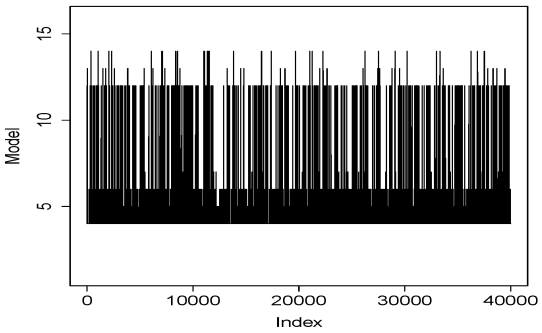


(e) Chain 5

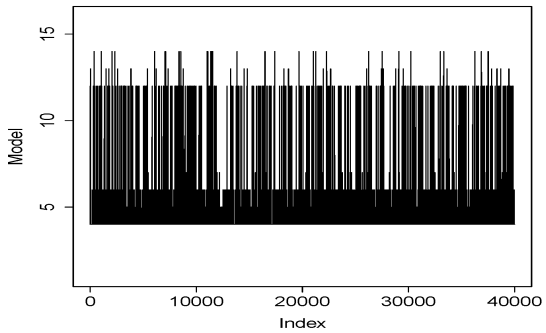
Figure 9.6: Trace plot of model choice from *ivbma* for 5 chains.



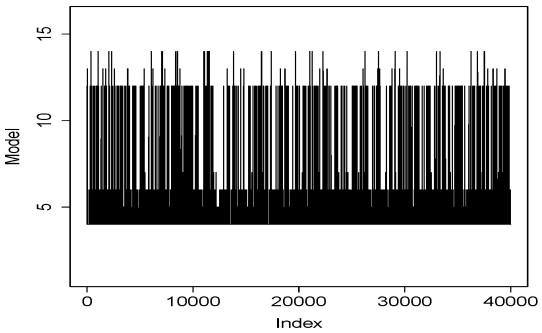
(a) Chain 1



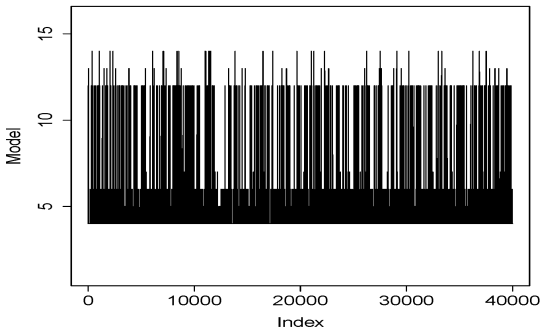
(b) Chain 2



(c) Chain 3



(d) Chain 4



(e) Chain 5

Figure 9.7: Trace plot of model choice from `OpenBUGS` for 5 chains.

9.4 Convergence and mixing in IVBMA

Convergence of the posterior distribution of each parameter in BMA is slower than for full model without variable selection. This long computation time is partly due to the prior imposed on each variable; when a variable is not in the model under consideration then it obtains its coefficient from the prior. The variable could have an extreme coefficient if the prior was too vague (large standard deviation). Then the likelihood of the model with this particular variable would be lower than the model without this variable and it would be rejected. Mixing is another important aspect to BMA to ensure all the possible models were considered and the flexibility of the sampler to jump from one model to another for the same reasoning as convergence, mixing is also affected by vague priors. Hence, BMA will require more iterations in order for a variable to have a chance of obtaining a reasonable coefficient from its prior and so be included in the model. The discussion of different priors for increasing the speed of convergence and mixing in model space is discussed by O'Hara et al. [213]. This section will focus on the role of the prior imposed by the R package, *ivbma*.

Karl and Lenkoski [160] have suggested 50,000 iterations and 10,000 burn-ins to be sufficient to reach convergence for their method. However, they have only looked at scenarios realistic in the econometric literature. This section will investigate whether 50,000 iterations is enough to reach convergence in a Mendelian randomisation problem. The performance of multiple chains with different initial values is a form of convergence diagnostic [192] and inspecting the trace plot will also check for sufficient mixing. Trace plots is a continuous line that shows the values a parameter have against the iteration number. 5 chains will be run with randomly generated initial values. To further confirm the convergence of these 5 short chains, a longer chain will be run, which will have 500,000 iterations and due to the limited computer memory, the burn-in will be 250,000.

The key factors in Mendelian randomisation are the number of instruments, the minor allele frequency (MAF) of the SNPs and amount of confounding between the risk factor (X) and the outcome of interest (Y). All the experiments will monitor the causal effect estimate by its mean and 95% credible interval, the probability that X is included in the second regression model and the total visited probability. The total visited probability [111] is defined as the probability of the 10 models for the first regression being visited, where the 10 models are defined as the top 10 most visited models from the long chain.

9.4.1 Experiment 1: Number of instruments

Aim

The greater the number of instruments the larger the model space, so in order for IVBMA to consider every model, more iterations will be required. Thus, this experiment aims to check convergence with increasing number of instruments.

Datasets

The dataset will consist of the genotypes of 10 SNPs, a risk factor (X) and an outcome of interest (Y) from 2,000 individuals. There is only one causal SNP which will explains 2% of the variation in X. The rest of the SNPs will be associated with X through their correlation to the causal SNP; they will have patterned LD, where their correlation will vary between 0.1 and 0.9 depending on their distance from the causal SNP, as described in Section 4.2. The MAF for the non-causal SNPs will be generated randomly between 0.1 and 0.5 and the causal MAF will be 0.05. X will explain 6% of the variation in Y. X and Y will be normally distributed.

As we are comparing the convergence for different numbers of instruments, datasets with 30, 60 and 90 potential instruments will also be simulated.

Results

Table 9.2 shows the results extracted from *ivbma* for 4 datasets where they differ in numbers of potential instruments. The more instruments there are the more possible models for the first regression to choose from. For example, 60 instruments will have 2^{60} possible models. This creates uncertainty in causal effect estimates and there are more models to consider as demonstrated by the contrast in 95% credible intervals between long and short chains when there are more than 30 instruments. Also the reduction in total visited probability, declined from 0.6 to 0.01 for 10 to 90 instruments respectively. The time-series standard error in the long chain have also increased with number of instruments which suggests the accuracy of the posterior mean of β_{XY} are decreasing, as there are more models to consider with greater number of instruments.

As there are more instruments, there are more parameters and more models for the algorithm to explore. The time taken for 90 instruments with 500,000 iterations and 250,000 burn-in was under an hour, Table 9.3.

Figure 9.8 shows the trace plot of the posterior distribution of the estimated causal effect, from the first 15,000 iterations for each number of instruments. Figure 9.8 only gives the trace plot of one short chain, as all 5 chains were very similar, the rest of the short chains are in the Appendix E.2 to E.5. The mixing between long and short chains are similar, jumping from one coefficient to another is the reflection of the jumping between models in the first regression (i.e. different and/or more instruments included). Hence, the advise of the "a fat caterpillar" shaped in trace plot (see Figure 9.3) by Lunn et al. [192] is not really applicable for BMA.

Table 9.2: Convergence Diagnostic by comparing 5 short chains with 1 long chain. The short chain had 50,000 iterations with 10,000 burn in. The long chain had 500,000 iterations and 250,000 burn in. The true β_{XY} is 0.2449. SE is standard error. Prob. X is the probability of X included in the second regression. Total Visit. Prob. is the visited probability of the set of models chosen in the first regression; the set consist of the top 10 models from the longer chain.

Ins.	Chain	Mean $\hat{\beta}_{XY}$	SE	Time-Series SE	95% Credible Int.		Prob. X	Total Visit. Prob.
10	1	0.0856	0.0008	0.0154	-0.1613	0.5020	0.4448	0.6930
	2	0.0923	0.0008	0.0155	-0.0900	0.4972	0.4351	0.6684
	3	0.1048	0.0009	0.0188	-0.1454	0.5496	0.4651	0.6943
	4	0.1008	0.0008	0.0152	-0.0679	0.5143	0.4494	0.6829
	5	0.1366	0.0009	0.0180	-0.0779	0.5544	0.5313	0.6600
	Long	0.0995	0.0004	0.0068	-0.1330	0.5259	0.4584	0.6786
30	1	0.2839	0.0011	0.0220	0.0000	0.6645	0.7621	0.1061
	2	0.2766	0.0011	0.0237	0.0000	0.6801	0.7274	0.1051
	3	0.2954	0.0010	0.0183	0.0000	0.6603	0.8106	0.1015
	4	0.2536	0.0011	0.0218	0.0000	0.6540	0.7174	0.1068
	5	0.2819	0.0010	0.0181	0.0000	0.6529	0.7877	0.1056
	Long	0.2906	0.0004	0.0079	0.0000	0.6597	0.7881	0.1034
	1	0.2017	0.0011	0.0222	-0.0142	0.6165	0.6219	0.0285

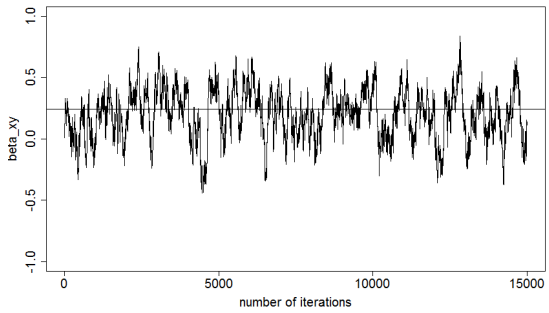
Continued on next page

Table 9.2 – *Continued from previous page*

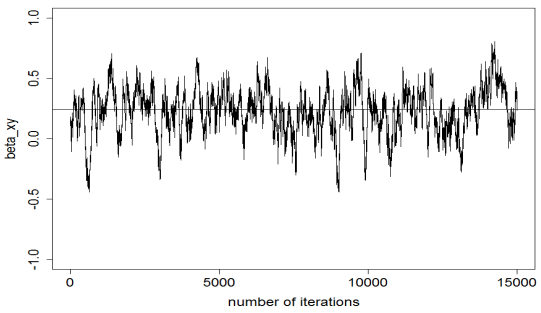
Ins.	Chain	Mean $\hat{\beta}_{XY}$	SE	Time-Series SE	95% Credible Int.		Prob. X	Total Visit. Prob.
	2	0.2025	0.0011	0.0243	-0.0614	0.6215	0.6291	0.0346
	3	0.2275	0.0011	0.0198	-0.0162	0.6333	0.6947	0.0287
	4	0.2103	0.0012	0.0277	-0.1478	0.6661	0.6597	0.0358
	5	0.1744	0.0011	0.0245	-0.0912	0.6489	0.5621	0.0263
	Long	0.1801	0.0004	0.0093	-0.0724	0.6077	0.5939	0.0315
	1	0.2068	0.0011	0.0226	-0.0157	0.6452	0.6356	0.0106
	2	0.2043	0.0010	0.0202	-0.0107	0.5802	0.6602	0.0101
90	3	0.2228	0.0010	0.0207	0.0000	0.6164	0.6775	0.0121
	4	0.1659	0.0010	0.0206	-0.0512	0.5933	0.5789	0.0137
	5	0.2154	0.0011	0.0252	-0.0306	0.6510	0.6496	0.0128
	Long	0.1962	0.0004	0.0096	-0.0377	0.6236	0.6210	0.0137

Table 9.3: Time taken (seconds) for *ivbma* to run these scenarios.

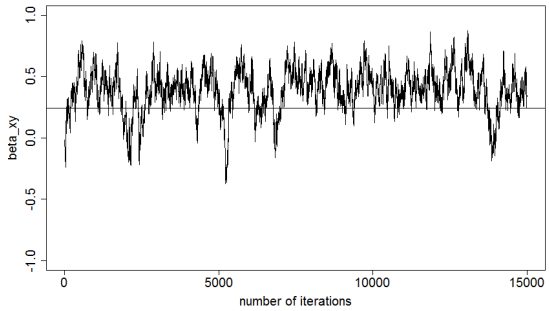
Number of Instruments	5 Chains	Long Chain
10	668.87	1320.02
30	1002.68	2008.77
60	1177.39	2273.98
90	1632.43	3163.83



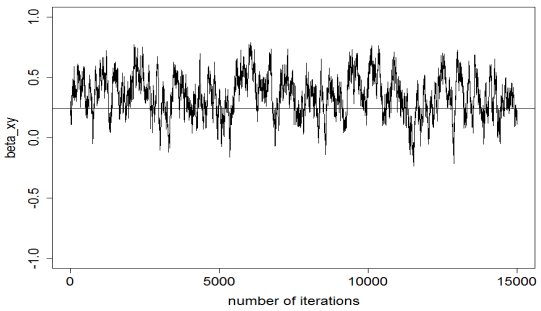
(a) 10 Instruments & Short Chain



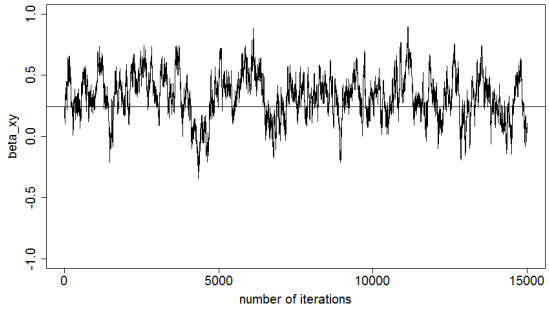
(b) 10 Instruments & Long Chain



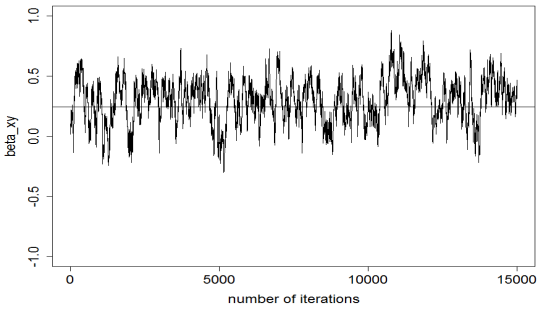
(c) 30 Instruments & Short Chain



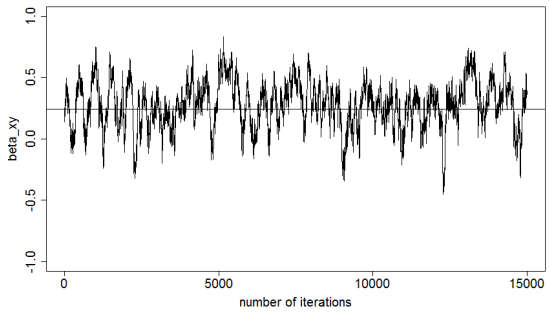
(d) 30 Instruments & Long Chain



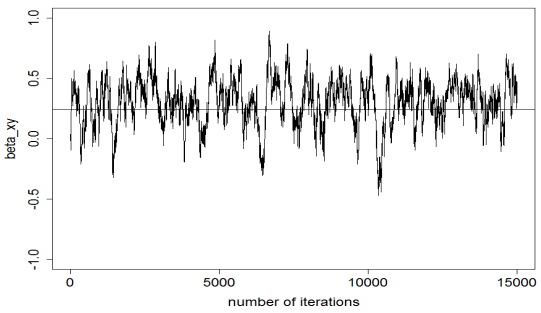
(e) 60 Instruments & Short Chain



(f) 60 Instruments & Long Chain



(g) 90 Instruments & Short Chain



(h) 90 Instruments & Long Chain

Figure 9.8: Trace plot of the causal effect estimate ($\hat{\beta}_{XY}$) from 10,30,60 and 90 instruments with short and long chain. Short and long chain consist of 50,000 and 500,000 iterations with 10,000 and 250,000 burn-in respectively. The horizontal line is the true β_{XY} (0.2449).

9.4.2 Experiment 2: Minor allele frequency

Aim

This experiment investigates whether more iterations are required when the SNPs are less common.

Datasets

This section only considers dataset with 10 SNPs, the same as in Section 9.4.1. Another similar dataset will be simulated but non-causal SNPs have MAF of 0.1.

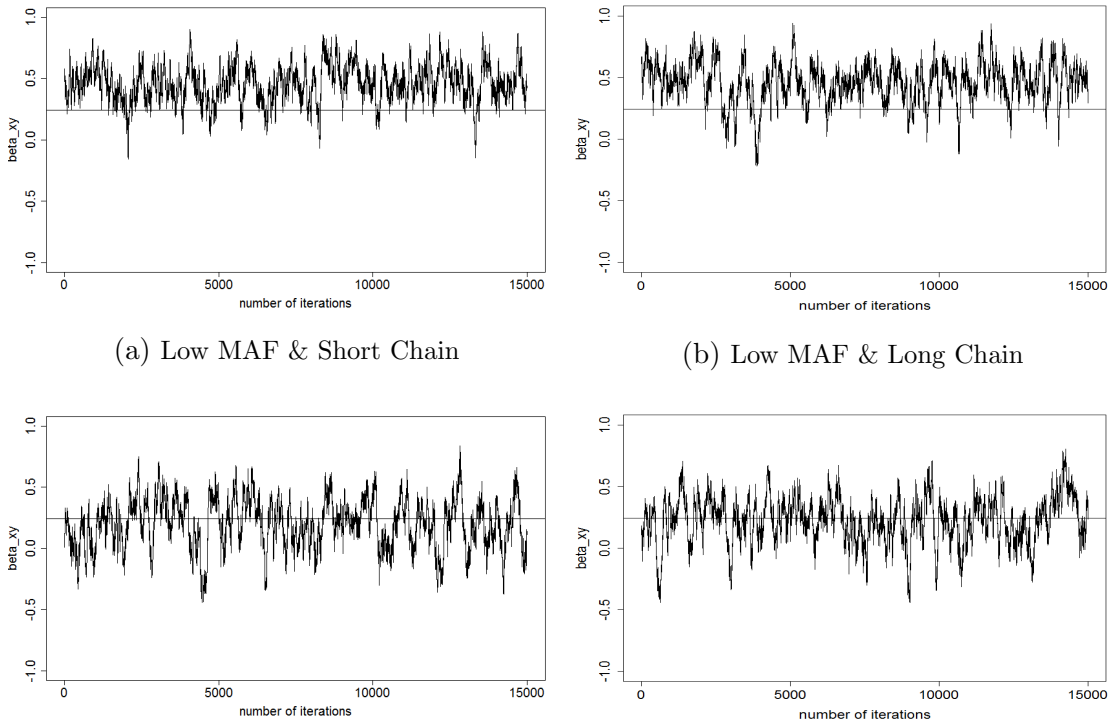
Results

Table 9.4 give the mean estimated β_{XY} and 95% credible interval, probability of including X and total model visited probability for datasets with 10 instruments, the results for datasets with 30, 60 and 90 instruments are in Appendix Table E.1. In the low frequency MAF case, the short and long chains have estimated similar β_{XY} and 95% credible interval, more similarly than the chains with variable MAF. This is because for the dataset with variable MAF, SNPs differ in their correlation with the causal SNP because of their MAF. Whereas with a low frequency MAF, each SNP will have the same correlation. The probability of including X and total model visited probability are also similar between long and short chains for both low and variable MAF case. The time-series standard error is similar between low and variable MAF.

Figure 9.9 shows that mixing is the same between low and variable (Var.) MAF; the sampler is moving from one model to another very quickly. The trace plots for more potential instruments included, see Appendix Figure E.6 to E.9 for 30 to 90 instruments respectively.

Table 9.4: Convergence diagnostic of 10 instruments with low and variable (Var.) MAF by comparing 5 short chains with 1 long chain. The short chain had 50,000 iterations with 10,000 burn in. The long chain had 500,000 iterations and 250,000 burn in. The true β_{XY} is 0.2449. SE is standard error. Prob. X is the probability of X included in the second regression. Total Visit. Prob. is the visited probability of the set of models chosen in the first regression; the set consist of the top 10 models from the longer chain.

MAF	Chain	Mean $\hat{\beta}_{XY}$	SE	Time-Series SE	95% Credible Int.		Prob. X	Total Visit. Prob.
Low	1	0.4145	0.0010	0.0161	0.0000	0.7490	0.9104	0.5053
	2	0.4226	0.0010	0.0157	0.0000	0.7543	0.9148	0.4861
	3	0.3855	0.0010	0.0178	0.0000	0.7225	0.8771	0.4956
	4	0.4338	0.0009	0.0140	0.0000	0.7367	0.9240	0.5132
	5	0.4167	0.0010	0.0148	0.0000	0.7340	0.9192	0.5056
	Long	0.4272	0.0004	0.0065	0.0000	0.7473	0.9252	0.5069
Var.	1	0.0856	0.0008	0.0154	-0.1613	0.5020	0.4448	0.6930
	2	0.0923	0.0008	0.0155	-0.0900	0.4972	0.4351	0.6684
	3	0.1048	0.0009	0.0188	-0.1454	0.5496	0.4651	0.6943
	4	0.1008	0.0008	0.0152	-0.0679	0.5143	0.4494	0.6829
	5	0.1366	0.0009	0.0180	-0.0779	0.5544	0.5313	0.6600
	Long	0.0995	0.0004	0.0068	-0.1330	0.5259	0.4584	0.6786



(a) Low MAF & Short Chain

(b) Low MAF & Long Chain

(c) Var. MAF & Short Chain

(d) Var. MAF & Long Chain

Figure 9.9: Trace plot of the causal effect estimate ($\hat{\beta}_{XY}$) from 10 instruments with low MAF. Short and long chain consist of 50,000 and 500,000 iterations with 10,000 and 250,000 burn-in respectively. The horizontal line is the true β_{XY} (0.2449).

9.4.3 Experiment 3: Confounding Effect

Aims

The aim of this section is to see whether the confounding effect between X and Y affects the number of iterations needed to have evidence of convergence.

Datasets

One dataset has the same simulation method as Section 9.4.1, the other two datasets will differ by (1) the direction and (2) the amount of confounding. Section 4.3 shows that U, the confounder, has a positive effect on both X and Y, and the causal SNP explains 2% of the variation in X. Then 98% of the variation is shared equally between confounding(49%) and random error(49%). In the next dataset, U will have a negative effect on Y, a change in direction of confounding from the

original simulation. Finally, the amount of confounding will be increased, so that the confounding effect explains 73.5% and random error explains 24.5% of the variation in X. The datasets will be referred to as positive, negative and strong confounding.

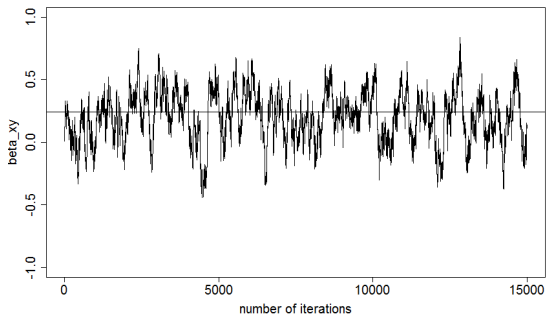
Results

Table 9.5 show that the causal effect estimates are similar between short and long chains for all effects of confounding. For strong confounding, the upper and lower limits of the 95% credible interval in the short chain are not as accurate as the long chain. There is very little difference between long and short chains for the probability of X and the total visited probability, in all of the datasets. The contrast in probability of X between negative confounding and the rest of the confounding effect is because they each have their own datasets which differ by instrument strength.

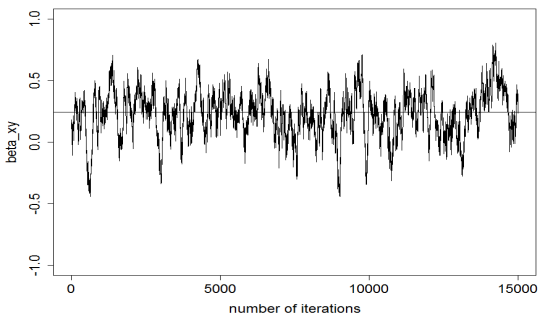
The trace plots for long and short chain have similar mixing properties, Figure 9.10. The short chains for strong confounding have a more "snake-like" appearance in its trace plot than the other confounding effects. This indicates worse mixing for strong confounding, due to the prior for the covariance matrix assuming no confounding between X and Y; as a consequence *ivbma* struggled to obtain the covariance matrix. The trace plots of negative and strong confounding for all 5 short chains are in Appendix Figure E.10 and E.11 respectively.

Table 9.5: Convergence Diagnostic of 10 instruments with different confounding effect by comparing 5 short chains with 1 long chain. The short chain had 50,000 iterations with 10,000 burn in. The long chain had 500,000 iterations and 250,000 burn in. The true β_{XY} is 0.2449. SE is standard error. Prob. X is the probability of X included in the second regression. Total Visit. Prob. is the visited probability of the set of models chosen in the first regression; the set consist of the top 10 models from the longer chain.

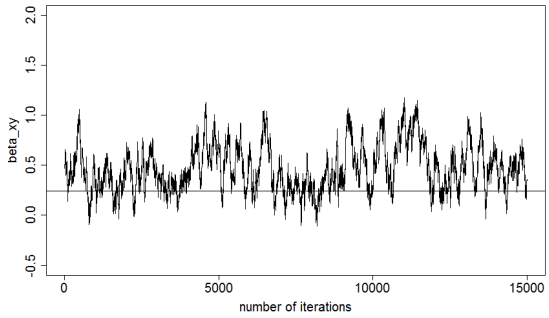
Conf.	Chain	Mean $\hat{\beta}_{XY}$	SE	Time-Series SE	95% Credible Int.		Prob. X	Total Visit. Prob.
Positive	1	0.0856	0.0008	0.0154	-0.1613	0.5020	0.4448	0.6930
	2	0.0923	0.0008	0.0155	-0.0900	0.4972	0.4351	0.6684
	3	0.1048	0.0009	0.0188	-0.1454	0.5496	0.4651	0.6943
	4	0.1008	0.0008	0.0152	-0.0679	0.5143	0.4494	0.6829
	5	0.1366	0.0009	0.0180	-0.0779	0.5544	0.5313	0.6600
	Long	0.0995	0.0004	0.0068	-0.1330	0.5259	0.4584	0.6786
Negative	1	0.4312	0.0014	0.0297	0.0000	1.0558	0.8951	0.7213
	2	0.4064	0.0013	0.0270	0.0000	0.9535	0.8576	0.7181
	3	0.4292	0.0013	0.0240	0.0000	0.9532	0.8977	0.7400
	4	0.3777	0.0015	0.0349	0.0000	0.9754	0.7804	0.7271
	5	0.4680	0.0014	0.0304	0.0000	1.0914	0.9116	0.7287
	Long	0.4379	0.0005	0.0115	0.0000	1.0038	0.8842	0.7249
Strong	1	0.0849	0.0009	0.0235	-0.2470	0.4869	0.4556	0.7671
	2	0.0530	0.0009	0.0264	-0.3916	0.4592	0.4309	0.7782
	3	0.0727	0.0008	0.0171	-0.1352	0.4800	0.4050	0.7859
	4	0.0923	0.0010	0.0277	-0.3282	0.5217	0.4802	0.7617
	5	0.0908	0.0009	0.0242	-0.3049	0.4971	0.4953	0.7667
	Long	0.0797	0.0003	0.0079	-0.1404	0.4704	0.4312	0.7774



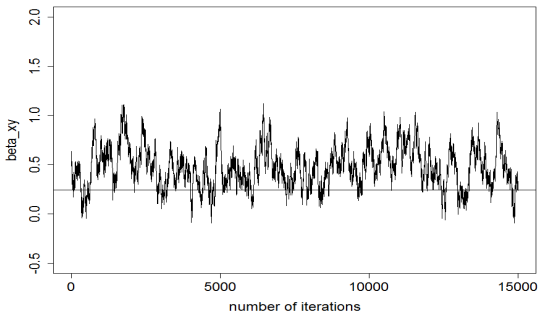
(a) Positive Confounding & Short Chain



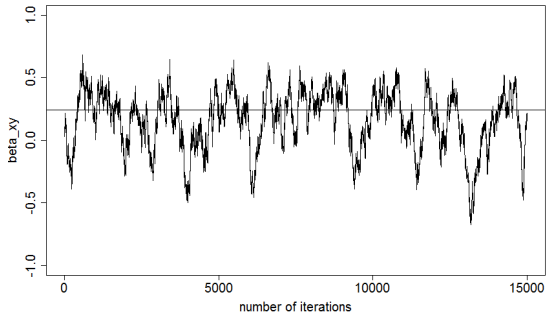
(b) Positive Confounding & Long Chain



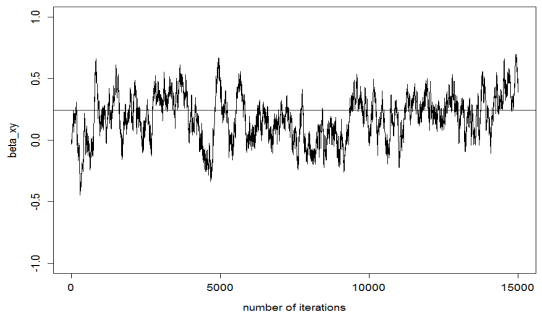
(c) Negative Confounding & Short Chain



(d) Negative Confounding & Long Chain



(e) Strong Confounding & Short Chain



(f) Strong Confounding & Long Chain

Figure 9.10: Trace plot of the causal effect estimate ($\hat{\beta}_{XY}$) from 10 instruments with low MAF. Short and long chain consist of 50,000 and 500,000 iterations with 10,000 and 250,000 burn-in respectively. The horizontal line is the true β_{XY} (0.2449).

9.4.4 Conclusion

If there are many instruments then the choice of models will be large, therefore *ivbma* will require more iterations to go though the different models. However there was evidence of convergence for mean $\hat{\beta}_{XY}$ in the short chain and its 95% credible interval.

The inclusion of SNPs with lower MAF does not necessarily require more iterations. However, if the investigator is interested in the credible interval for variable MAF case, then more iterations will be needed to acquire convergence in the credible interval. Variation in MAF and correlation with the causal SNP creates uncertainty in causal effect estimates.

The short chain does show sufficient mixing and evidence of convergence for causal effect estimate, probability of X and total visited probability in each form of confounding. Nevertheless when the confounding effect is strong, more iterations are required to demonstrate evidence of convergence in the 95% credible interval.

From these results, we can tell that the number of iterations to have evidence of convergence is dependent on the aims of the analysis. If an accurate 95% credible interval is the aim then running more iterations will be needed. However, if the investigators are only interested in the the point estimates, then simulating a short chain will suffice.

9.5 Selection of instruments in IVBMA

9.5.1 Aim

The aim of this section is to understand whether IVBMA is able to select instruments that are highly correlated with the causal SNP, and the effect of MAF on SNP selection.

9.5.2 Design

The genotype of 10 SNPs from 2,000 individuals will be available as instruments. The SNPs will have patterned LD, where their correlation with each other will be between 0 and 1, depending on their physical proximity, as described in Section 4.2. MAF for all the non-causal SNPs is 0.45 and for the causal SNP is 0.5. The middle SNP will be assumed as causal and explains 2% of the variation in the risk factor (X). X will explain 6% of the variation in the outcome (Y). X and Y will be normally distributed with sample size of 2,000. The causal SNP will be discarded and the potential instruments will be in the form of non-causal SNPs. The simulation is repeated 200 times.

To understand the effect of MAF, two further scenarios will be simulated; in the low frequency MAF case the non-causal and the causal SNP have MAF of 0.1

and 0.05 respectively. For the variable MAF case, the non-causal MAF will vary between 0.1 to 0.5 and the causal SNP will have 0.05.

The average of several parameters from the 200 datasets will be monitored; causal effect estimate with its standard error (S.E.), probability of X being included in the second regression with its interquartile range (IQR), correlation of each SNP with the causal SNP (SNP_c), their estimated association with X ($\hat{\beta}_{ZX}$) and their probability of being included as an instrument.

9.5.3 Results

For all MAF, on average over the 200 datasets, IVBMA (the method) usually selects one of the SNPs that is highly correlated with the causal SNP; SNP_5 and SNP_6 have the highest correlation with SNP_c for all cases of MAF, they have very similar probability of being included in the model as an instrument and their probability adds up close to 1, which indicates that one or the other is under model consideration in each MCMC iteration. For random MAF case, the probability of being an instrument for SNP_5 or SNP_6 is not quite 1. This is because in some datasets SNP_5 and SNP_6 have low MAF, and IVBMA selects the next highly correlated SNP with common MAF in order to obtain more information.

Table 9.6: Comparing the average parameters from 200 datasets between different MAF cases; $\hat{\beta}_{XY}$ is the causal effect estimate (the true is 0.2449) and the standard error of its mean (S.E.). Prob. X is the probability of X being included in the second regression and its interquartile range (IQR). Corr. with SNP_c is each SNP's correlation with the causal SNP. $\hat{\beta}_{ZX}$ is their estimated association with X. Prob. SNP is their probability of being included as an instrument.

MAF	Mean $\hat{\beta}_{XY}$ (S.E.)	Prob. X (IQR)	SNP	Corr.SNP_c	Mean $\hat{\beta}_{ZX}$	Prob. SNP
Random	0.1705 (0.0221)	0.5567 (0.2990)	1	0.0894	-0.0012	0.0630
			2	0.3095	0.0004	0.0640
			3	0.5680	0.0032	0.0866
			4	0.8013	0.0164	0.1871
			5	0.9413	0.0469	0.3806
			6	0.9413	0.0459	0.3776

Continued on next page

Table 9.6 – *Continued from previous page*

MAF	Mean $\hat{\beta}_{XY}$ (S.E.)	Prob. X (IQR)	SNP	Corr.SNP_c	Mean $\hat{\beta}_{ZX}$	Prob. SNP
			7	0.8007	0.0168	0.1829
			8	0.5689	0.0022	0.0770
			9	0.3081	-0.0009	0.0727
			10	0.0895	-0.0025	0.0722
			1	0.0887	0.0002	0.0444
			2	0.3084	0.0007	0.0450
			3	0.5683	0.0030	0.0736
			4	0.8005	0.0115	0.1480
Common	0.1902 (0.0180)	0.6931 (0.5169)	5	0.9413	0.0763	0.5411
			6	0.9414	0.0670	0.4898
			7	0.8013	0.0111	0.1458
			8	0.5689	0.0028	0.0819
			9	0.3102	0.0018	0.0593
			10	0.0878	0.0001	0.0381
			1	0.0889	-0.0002	0.0717
			2	0.3067	0.0004	0.0801
			3	0.5698	0.0062	0.1015
			4	0.8014	0.0265	0.2164
Lower	0.1672 (0.0223)	0.5612 (0.3191)	5	0.9416	0.0780	0.4463
			6	0.9416	0.1032	0.5555
			7	0.8013	0.0214	0.1952
			8	0.5701	0.0079	0.1213
			9	0.3082	0.0004	0.0696
			10	0.0892	-0.0013	0.0885

Continued on next page

Table 9.6 – *Continued from previous page*

MAF	Mean $\hat{\beta}_{XY}$ (S.E.)	Prob. X (IQR)	SNP	Corr.SNP_c	Mean $\hat{\beta}_{ZX}$	Prob. SNP

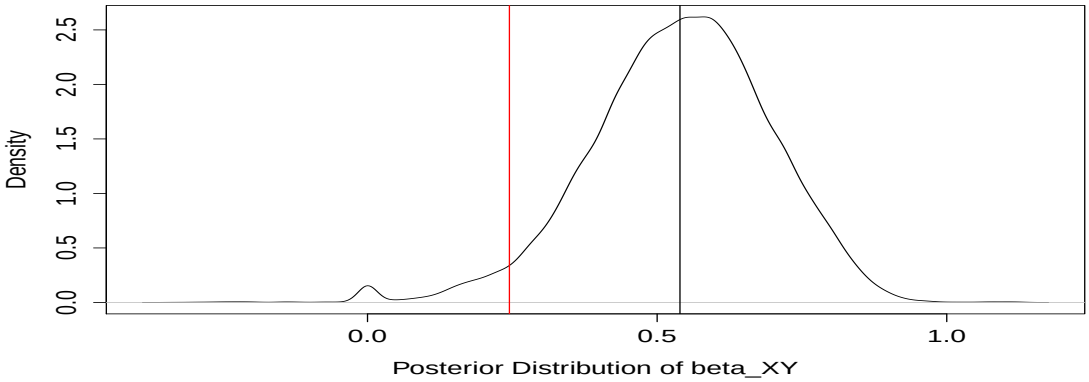


Figure 9.11: Posterior distribution of causal effect estimates ($\hat{\beta}_{XY}$) for dataset A, a dataset that had the highest probability of X being included in the second regression (0.9942). The red and black vertical line is the true causal effect (0.2449) and mean of $\hat{\beta}_{XY}$ (0.5394).

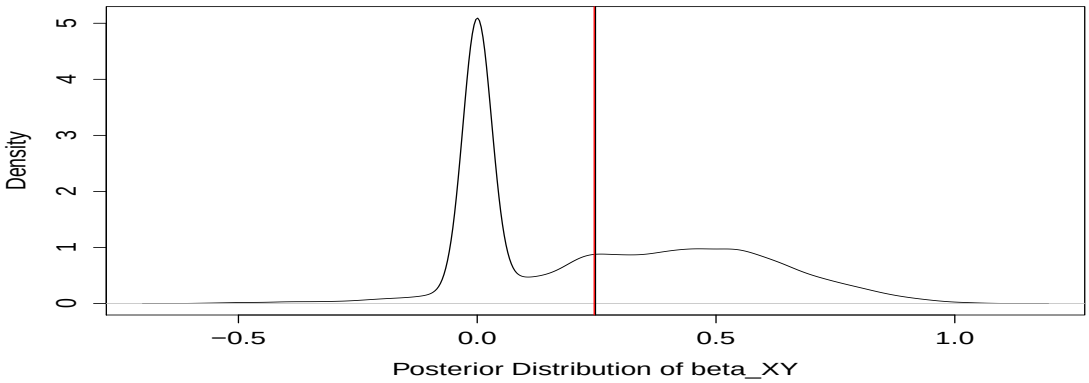


Figure 9.12: Posterior distribution of causal effect estimates ($\hat{\beta}_{XY}$) for dataset B, a dataset that had mean estimate approximately to the true causal effect and the probability of X is 0.6387. The red and black vertical line is the true causal effect (0.2449) and mean of $\hat{\beta}_{XY}$ (0.2476).

The common SNP analysis had the least bias, low frequency had the most and random MAF had bias in between them (True β_{XY} =0.2449). There is not enough information from low frequency SNPs to support X in the second regression, as is evident from the large S.E. of β_{XY} in both variable and low frequency MAF in comparison to common MAF.

ivbma sets the coefficient to zero if its variable is not included in the current model, thus mean estimates are pulled towards zero. Consider the results for datasets A and B, selected from the 200 datasets in the simulation. They both consist of SNPs with variable MAF. Dataset A gave the highest probability of X being included in the second regression and causal effect estimate for dataset B was closest to the true. Figure 9.12 show even though dataset B estimated approximately the true causal effect, IVBMA was uncertain of X, so there were large number of models that did not include X, the mean causal effect estimate is pulled towards zero as a result, which just happens to be towards the true value. Whereas dataset A is more certain of X and has hence less models with zero estimates.

9.5.4 Conclusion

Most of the time IVBMA chooses SNPs that are highly correlated with the causal SNP. However the low MAF of the SNPs does creates uncertainty for the causal relationship between X and Y. Thus, the researcher should examine the posterior distribution of the causal effect from IVBMA for their dataset to assess the reliability of its mean estimate.

9.6 Discussion

This chapter have given a description of Bayesian Model Averaging (BMA), BMA in instrumental variable analysis (IVBMA) and their implementation in `OpenBUGS`. However, `OpenBUGS` is inefficient for IVBMA as `OpenBUGS` was designed for a wide range of Bayesian applications. For exactly the same dataset, R package *ivbma* took a tenth of the time of `OpenBUGS` and its causal effect estimates were similar including the 95% credible interval. After deciding on the most efficient software, the convergence and mixing were the next issues to be considered. The accuracy of the posterior credible interval decreased as the number of instruments increased, because the number of possible models increases. Instruments with low MAF do not

change the rate of convergence. If the confounding between exposure and outcome of interest is strong then the rate of convergence is slow, because the prior from *ivbma* implies there is no confounding effect. Overall, the results of these scenarios suggest that if an accurate 95% credible interval is necessary then more iterations will be required. IVBMA selects SNPs that are highly correlated with the causal SNP most of the time, except when these SNPs have low MAF and then the next highly correlated SNPs with greater MAF will be chosen.

The designers of *ivbma*, Karl and Lenkoski [160] demonstrated their algorithm to be more computationally efficient than 2BMA [184] by performing analyses on the same dataset. Furthermore, they have shown IVBMA's quick convergence and lack of mixing difficulties in comparison to Koop et al. [172]'s method. Karl and Lenkoski [160] have also found the convergence of the 95% credible interval from *ivbma* deteriorates as number of potential covariates in the regression increases.

The point estimate of the causal effect is the mean of its posterior distribution which includes the case when X is not in the model and if IVBMA is uncertain of X then the mean will be pulled toward 0, since *ivbma* assigns a 0 coefficient for the causal effect in the absence of X. Thus, it is important for the researcher to plot the posterior distribution of the causal effect estimates. Along the same line of thought, if `OpenBUGS` is used, despite its inefficiency, then the user must ensure that, when X is not included, the output of causal effect estimates should be changed to 0, as those estimates are taken from the prior distribution. As a rule, in the context of BMA, the posterior inclusion probability of an explanatory variable, Kass and Raftery [161] have suggested for <50% as evidence for no effect of the explanatory variable on the outcome, 50-75% as weak evidence for an effect, 75-95% as evidence and >95% as strong evidence of explanatory variable having an effect on the outcome.

The rate of convergence in IVBMA also depends on the priors for parameters of interest. The priors specified by *ivbma* are perhaps a little informative but with reason. A vague prior can cause slow convergence and mixing. When a variable is not in the model, its coefficients is extracted from its vague prior. If by chance an extreme value is obtained then it is much harder for the model with this variable to be under consideration, as the extremity will give a low likelihood and thus rejected. A review of appropriate priors for BMA have been discussed by O'Hara et al. [213].

Chapter 10

A Comparison of Bayesian and Classical Approaches

10.1 Introduction

The most commonly implemented estimator for the average causal effect in Mendelian randomisation is two-stage least squares (2SLS) [84]. However as seen in Chapter 7 it is severely biased with weak instruments. Limited information maximum likelihood (LIML) is known to reduce the weak instrument bias seen in 2SLS [61]. Chapter 7 concluded that LIML is the most efficient estimator for the causal effect with many weak instruments. Nonetheless LIML occasionally gives extreme estimates. The number of these outliers were lower when weak instruments were excluded, but selection bias was introduced when the same dataset was used for instrument selection and the analysis. Instrumental variable Bayesian model averaging (IVBMA) offers an alternative to the classical approach, where the instrument selection is within the analysis of causal effect. IVBMA removes instruments that do not explain any additional variation and incorporates model uncertainty into the causal effect estimate. The objective of this chapter is investigate whether the mechanism of IVBMA have an advantage over the classical approaches, 2SLS and LIML.

This chapter describes 3 experiments. Experiment 1 compares the performance of the 3 estimators when potential instruments have common, low and variable MAF. Experiment 2 examines whether the change in LD pattern alters the performance of the 3 estimators. The simulation method in experiment 1 and 2 is designed to control the SNPs' MAF and LD. However it does not integrate evolutionary factors such as population structure, natural selection and recombination into the

algorithm. Experiment 3 uses GENOME to simulate realistic genetic patterns for the genotypes of the SNPs and aims to discover the possible differences in performance with realistic genetic patterns.

10.2 Experiment 1: Minor Allele Frequency

10.2.1 Aim

As we have seen from Section 9.4.3, IVBMA's performance can be affected by minor allele frequency (MAF); if the allele frequency is low then there is not enough information for IVBMA to be certain of the association between X and Y. This experiment aims to compare the efficiency of IVBMA to the classical approaches with different MAF.

10.2.2 Design

There will be 200 datasets, each dataset will contain from 10 to 90 SNPs, the risk factor (X) and the outcome of interest (Y) for 2,000 individuals. The SNPs will have patterned LD, their correlation will be dependent on their physical distance to the causal SNP, as described in Section 4.2. The position of the causal SNP will be in the middle and will explain 2% of the variability in X. 6% of the variation in Y will be explained by X. X and Y will be normally distributed. The causal SNP will be discarded before the comparison of methods. Hence the potential instruments are the non-causal SNPs.

This experiment will consider three cases of MAF; variable, common and low. For the variable MAF, the causal SNP will have MAF of 0.5 and non-causal MAFs will be randomly generated between 0.1 and 0.5. In the common case, the causal and non-causal SNPs will have MAF of 0.45 and 0.5 respectively. The causal SNP will have MAF of 0.05 and non-causal with 0.1 in the low MAF case.

For IVBMA, each dataset will be run with 50,000 MCMC iterations and burn-in of 10,000. Section 9.4 showed that even with many potential instruments, and thus more models to consider, the point estimates remains similar between long and short chains. Therefore I will keep the number of iterations the same for all numbers of instruments and different cases of MAF.

The evaluation criteria of the 3 estimators will be monitored from 200 datasets; Winsorised bias and root mean square error (RMSE), percentage of outliers and

coverage.

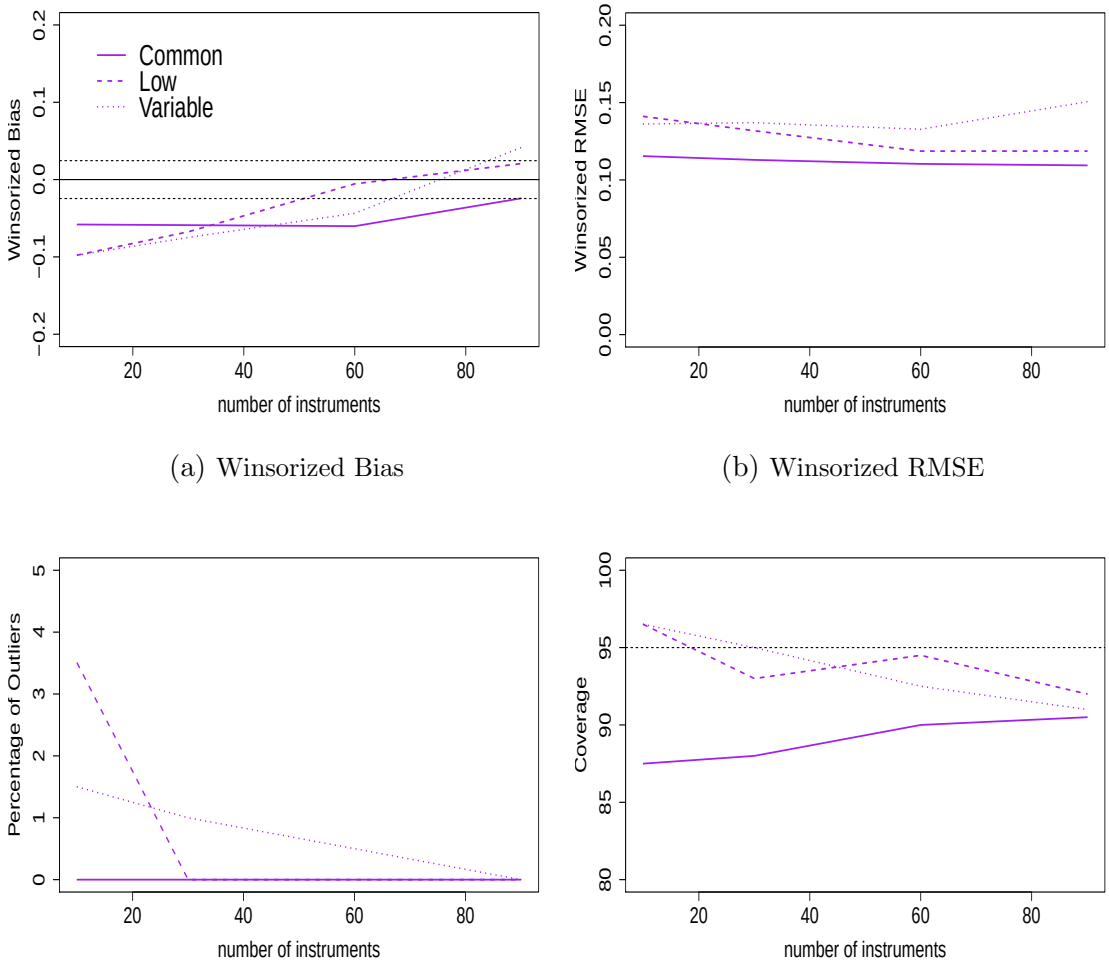
10.2.3 Results

Instrumental variable Bayesian model averaging

For common and variable MAF, the Winsorised bias from IVBMA for all instruments are similar (Figure 10.1), shown by the overlapping confidence interval, calculated from Winsorised S.E. (Appendix Table F.1). Except for the decrease in bias from 60 to 90 instruments, this is because there are more SNPs to choose from, and hence the analyses are more likely to select the SNPs with strong correlations with the causal SNP. For low MAF, the cause of the bias is slightly more complicated, with small number of instruments and with low MAF, IVBMA is less certain of the causality of X, hence the causal effect estimates are pulled towards zero, due to the normal prior of $N(0, 1)$. When there are more instruments, the IVBMA has more potential instruments to choose from, but low MAF causes finite sample bias in the first regression. As shown by the change of negative to positive bias from 60 to 90 instruments; the combination of estimates is pulled towards zero and the estimates are affected by finite sample bias giving an illusion of zero bias. Then the positive bias in 90 instruments demonstrates that there are more instruments with strong correlation with the causal SNP but they are affected by finite sample bias.

10 instruments with common MAF had the least bias, as there is more information to confirm the presence of X in the second regression (Figure 10.2). Then all of the MAF gave the same bias for 30 instruments, since variable and low MAF have gained more information with an increased number of SNPs. For 60 instruments, Figure 10.1a shows a difference in bias between variable and common MAF, but the large standard error (see Appendix Table F.1) means that their bias is the similar. The low MAF gave approximately zero bias at 60 instruments. This is because some of the datasets in the simulation are affected by finite sample bias giving positive estimates which balances the negative estimates from datasets that are still unsure about X being included in the second regression. Thus, with 90 instruments the low MAF case becomes more certain of X but is affected by finite sample bias, consequently increasing the bias. The variable MAF also shows positive bias, this is because of the range of estimates from each datasets is wide, shown by Winsorised RMSE (Figure 10.1b). Some datasets will have SNPs strongly correlated with the causal SNP but have low MAF, and other datasets will have SNPs with common

MAF and strong correlation.



(a) Winsorized Bias

(b) Winsorized RMSE

(c) Outlier

(d) Coverage

Figure 10.1: Comparing evaluation criteria of IVBMA from instruments with different MAF. The purple coloured straight, dashed and dotted lines are common, low and variable MAF respectively. The black solid line in Winsorised bias is zero bias. The dotted black line in (a) and (d) is 10% bias and 95% nominal coverage respectively.

Both Winsorised RMSE from common and low MAF decreases with additional instruments, Figure 10.1b. In the variable MAF case, IVBMA estimates show differences in RMSE between 30 to 90 instruments; The reduction of RMSE from 30 to 60 instruments, is caused by the certainty of X. There is an increase in RMSE from 60 to 90 instruments, since the overall bias is reduced, the variation in estimates between datasets is greater, as explained before. In the variable MAF case the in-

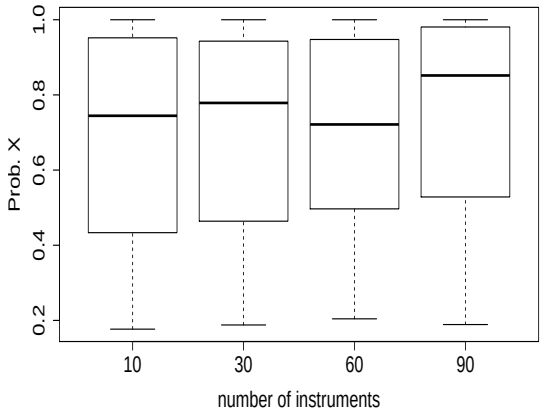
struments in each datasets will vary by their MAFs and their correlations with the causal SNP. Due to the certainty of X included in the second regression and the first regression is not affected by finite sample bias, the common MAF have the lowest RMSE.

There are some outliers in IVBMA estimates when the number of instruments is small, and in the case where MAF is low and variable, Figure 10.1c. This is not surprising considering low MAF will have finite sample bias and the variation in MAF will cause each dataset to have a different selection of potential instruments. This is evident from the decreasing number of outliers when there are more instruments to choose from. However the percentage of outliers in IVBMA is never too large, as the prior restricts the range of the causal effect estimate can have.

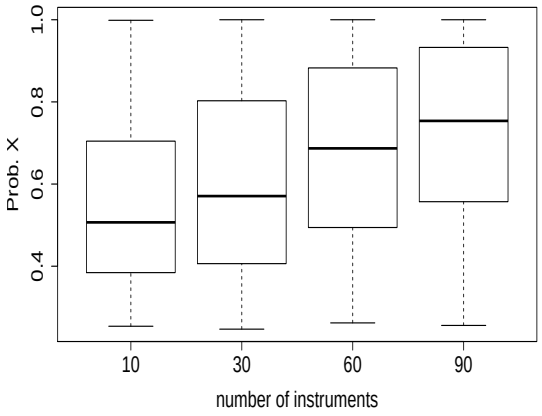
For all three cases of MAF, the coverage remains the same with increasing instruments, Figure 10.1d. The coverage from low and variable MAF is closer to the nominal level than that from common MAF. This is because of the uncertainty of X in low and variable MAF, which resulted in wider credible intervals for the causal effect estimates.

Comparison of classic and Bayesian approaches

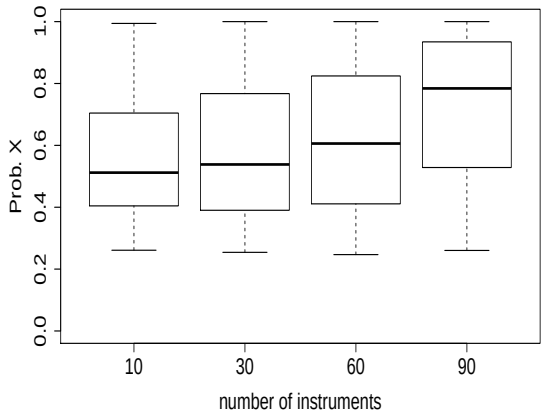
There are no differences in Winsorised bias between LIML and IVBMA for 10 and 90 instruments, due to their large Winsorized standard error (Appendix Table F.1). IVBMA has more inaccurate causal effect estimates than LIML, due to its uncertainty for X, shown by the low median probability of X in Figure 10.2a. Thus its causal effect estimates are pulled towards zero (the prior of the causal effect estimate has a mean of zero). When the number of instruments reaches 90, then there is more evidence for IVBMA to include X and LIML is the more biased. Note that Figure 10.3a does not imply zero bias from LIML with 60 instruments. The estimator had too many extreme values even for Winsorisation as is evident from the large percentage of outliers in Figure 10.1c. For the low MAF case, the Winsorised bias is similar between LIML and IVBMA, as their Winsorised standard error is large (Appendix Table F.1). An exception is for 90 instruments where LIML is affected by the many weak instruments problem and IVBMA has more confidence in X (Figure 10.2b). For the case of variable MAF, IVBMA reaches similar bias as LIML with increase in number of instruments, since IVBMA's certainty of X grows with additional instruments.



(a) Common MAF



(b) Low MAF

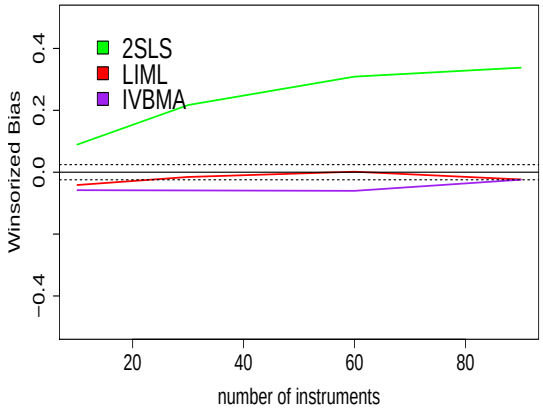


(c) Variable MAF

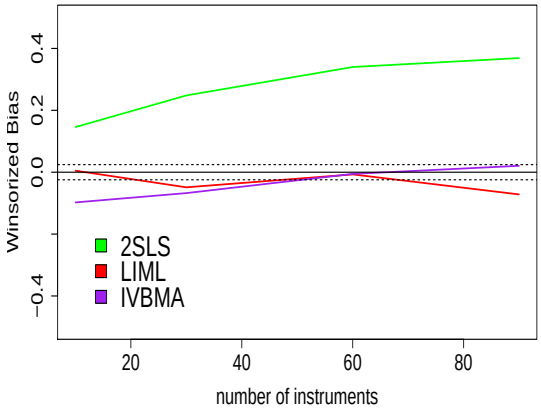
Figure 10.2: Probability of X for common, variable and low MAF. Probability of X, is the probability of X included in the second regression of IVBMA.

IVBMA has the lowest Winsorisation RMSE for common MAF (Figure 10.4a), which is due to the effect of the informative prior. The difference between estimators is not due to sampling error. In the low MAF case, the Winsorised RMSE for the two classical approaches increases with number of instruments, while IVBMA decreases, Figure 10.4b). As mentioned before, the decrease is due to IVBMA's increasing confidence in X. Note for 10 instruments, Figure 10.4b shows that the RMSE from IVBMA is greater than LIML, the difference is caused by sampling error. RMSE has 0.4% accuracy from 200 datasets. Similar conclusions can be drawn with variable MAF, except for 90 instruments, as there is more variation between SNP correlations

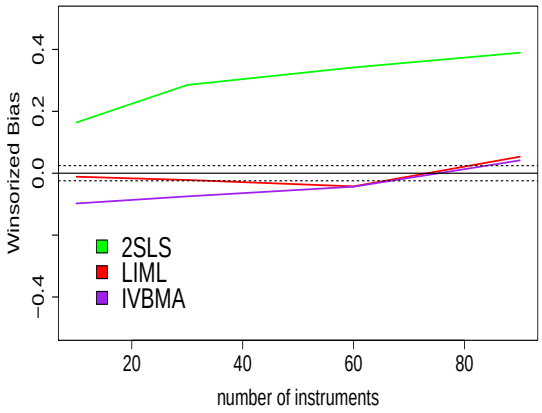
to the causal SNP and their MAF.



(a) Common MAF



(b) Rare MAF



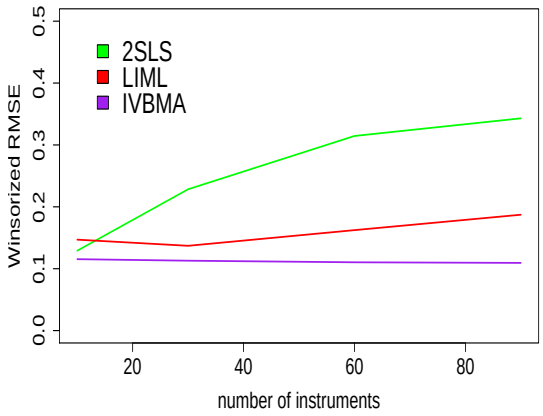
(c) Variable MAF

Figure 10.3: Winsorised bias from 2SLS, LIML and IVBMA for common, variable and low MAF. The estimators are coloured as green, red and purple respectively. The black solid line in each plot is zero bias and the black dotted lines are 10% bias.

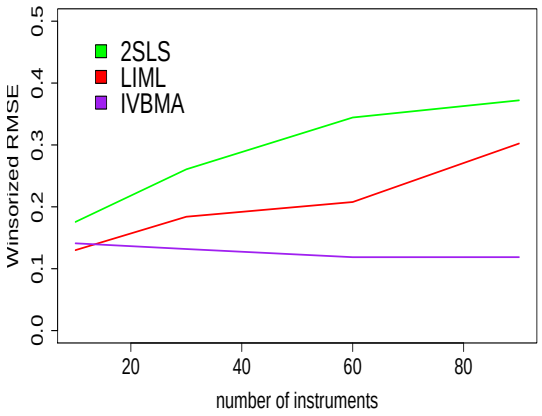
IVBMA has the advantage of a prior to restrict the range that a causal effect estimate can have. Hence the percentage of outliers is approximately zero for all cases of MAF, Figure 10.5. However there are a few exceptions; for 10 instruments with low MAF, and 10 and 30 instruments with variable MAF, IVBMA has more outliers than 2SLS, which because of its uncertainty about X being included in the second regression.

The coverage for IVBMA remains close to the 95% nominal level for all cases of

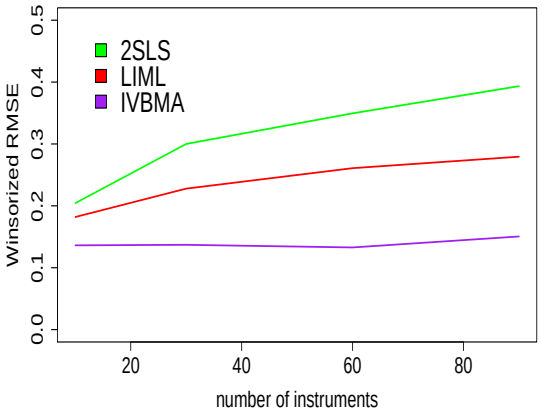
MAF, whereas the classical approaches decrease away from the nominal level with increasing number of instruments. At 10 instruments LIML seems to have better coverage than IVBMA this is due to sampling error. 95% coverage has 3% accuracy from 200 dataset, using the formula in Section 4.5.



(a) Common MAF

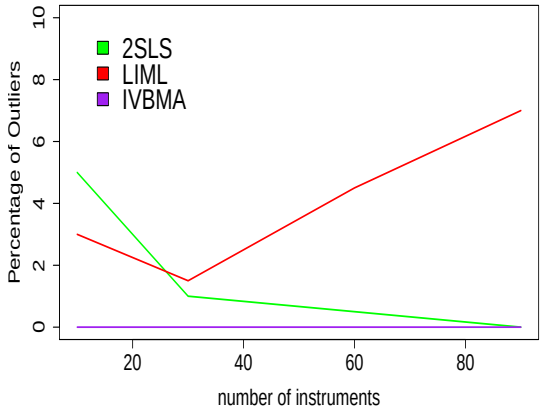


(b) Low MAF

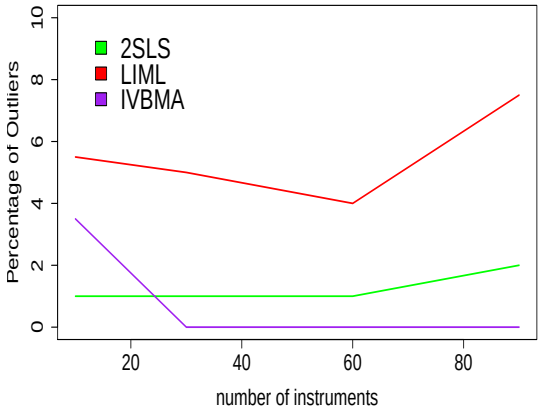


(c) Variable MAF

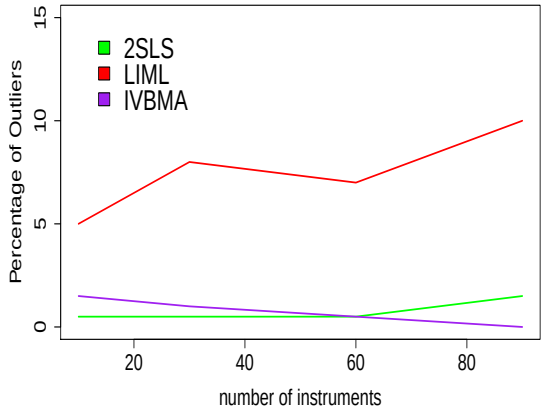
Figure 10.4: Winsorised RMSE from 2SLS, LIML and IVBMA for common, variable and low MAF. The estimators are coloured as green, red and purple respectively.



(a) Common MAF

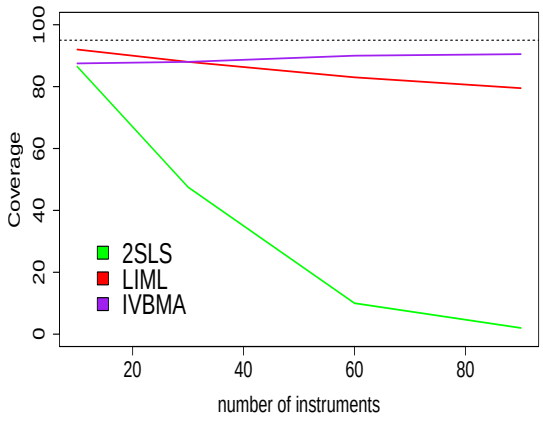


(b) Low MAF

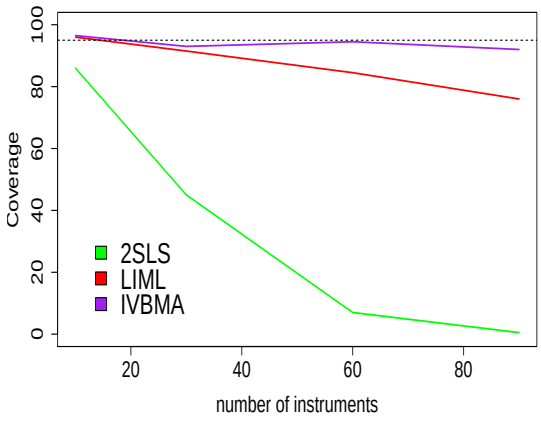


(c) Variable MAF

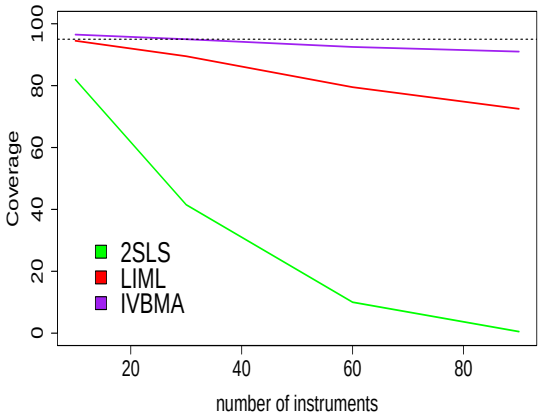
Figure 10.5: Outlier from 2SLS, LIML and IVBMA for common, variable and low MAF. The estimators are coloured as green, red and purple respectively.



(a) Common MAF



(b) Low MAF



(c) Variable MAF

Figure 10.6: Coverage from 2SLS, LIML and IVBMA for common, variable and low MAF. The estimators are coloured as green, red and purple respectively. The black dotted line in each plot is 95% nominal coverage.

10.2.4 Conclusion

For many instruments IVBMA is generally the better option, as its certainty of X being included in the second regression increases. These results also demonstrate the sensitivity of priors on causal effect estimates, as there are hardly any outliers, unlike LIML. The performance of IVBMA depends on which potential instruments are included in the datasets. If the highly correlated SNPs have rare MAF, then it will have finite sample bias even if it has lots of instruments to choose from.

10.3 Experiment 2: Patterns

10.3.1 Aim

In the previous experiment, the simulation distributed the genetic correlation evenly according to their physical distance to the causal SNP. This experiment examines the effect of changing the simulation method on the performance of 2SLS, LIML and IVBMA.

10.3.2 Design

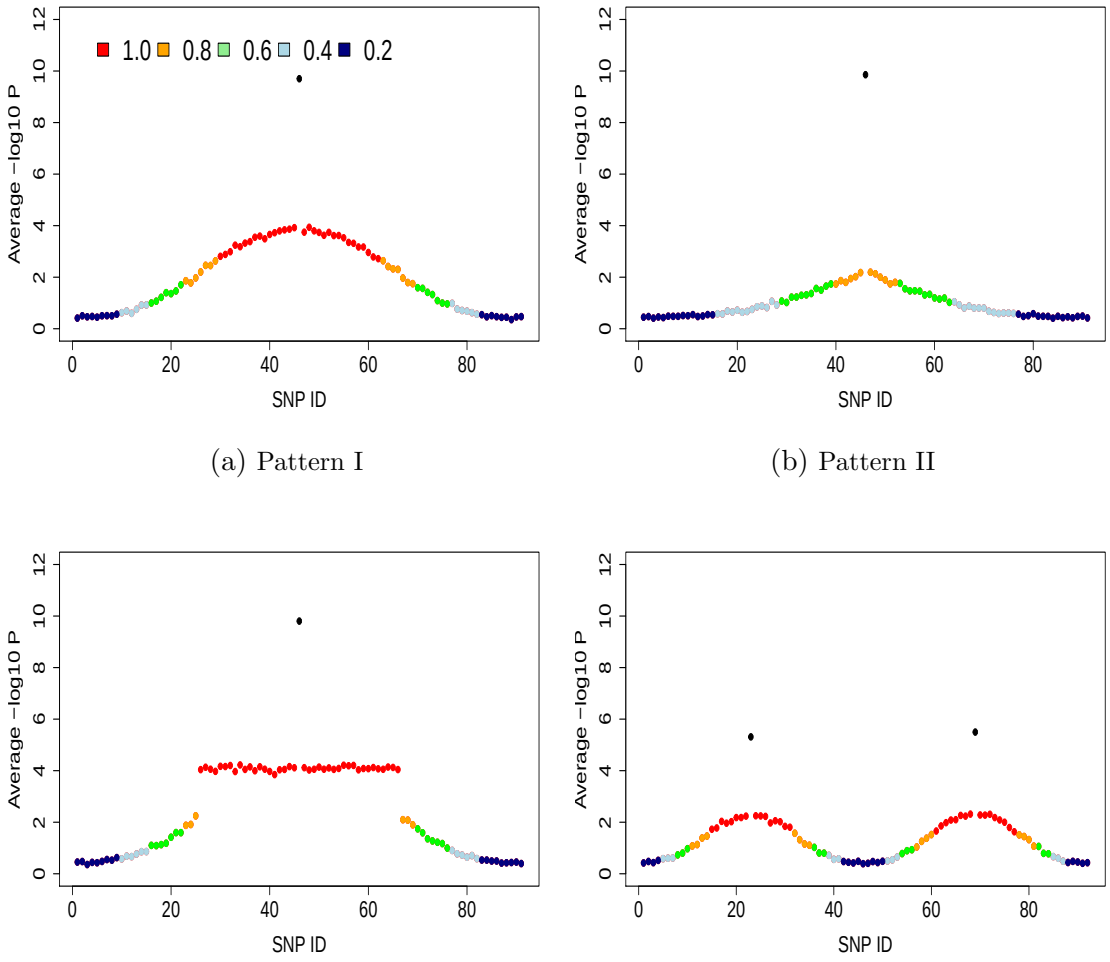
The simulation of the genotype of 10 to 90 SNPs, risk factor (X) and outcome of interest (Y) from 2,000 individuals are the same as experiment 1. The MAF of causal SNP will be 0.5 and non-causal will vary between 0.1 and 0.5. The only difference from experiment 1 is the genetic correlation pattern; Pattern I will have patterned LD, same as experiment 1, Pattern II will have the same shape as Pattern I but the SNPs have weaker correlation to the causal SNP. Pattern III will resemble the haplotype blocks; strong correlation with the causal SNP within a block and at recombination (outside the block) the correlation drops drastically. Pattern IV will have two functional variants, where each explains 1% of the variation in X and each will have the same correlation pattern as Pattern I. The detail description of the simulation method is in Section 4.2.

For IVBMA, each dataset will be run with 50,000 MCMC iterations and burn-in of 10,000, in all the different genetic patterns and numbers of instruments.

The simulation will be repeated 200 times and for each simulation Winsorised bias and RMSE, percentage of outliers and coverage will be monitored.

10.3.3 Results

Figure 10.7 show the regional plots for Patterns I,II,III and IV. As described in the design section; the correlations decreases evenly with the increase in distance to the causal SNP (black dot) for Pattern I. Pattern II have the correlation distribution but with weaker start point of the correlation. The correlations in Pattern III have a plateau effect of a haplotype block, then drop sharply at the recombination points. Pattern IV has two causal SNPs, each with the same correlation pattern as Pattern I and each explains 1% of the variation on X. Thus the p-value for the association with X is smaller when compared to Pattern I.



(a) Pattern I

(b) Pattern II

(c) Pattern III

(d) Pattern IV

Figure 10.7: The regional association plots for the four patterns with 90 SNPs of 200 simulations. The p-value is the mean p-value from the regression of each SNP on X. On the x-axis is SNP ID mimicking chromosome position, and on the y-axis is -$\log_{10} P$. Colour coding (from red to navy) denotes strong to weak correlation with the causal SNP; see also the legend within the plot. The black dot is the causal SNP

Instrumental variable Bayesian model averaging

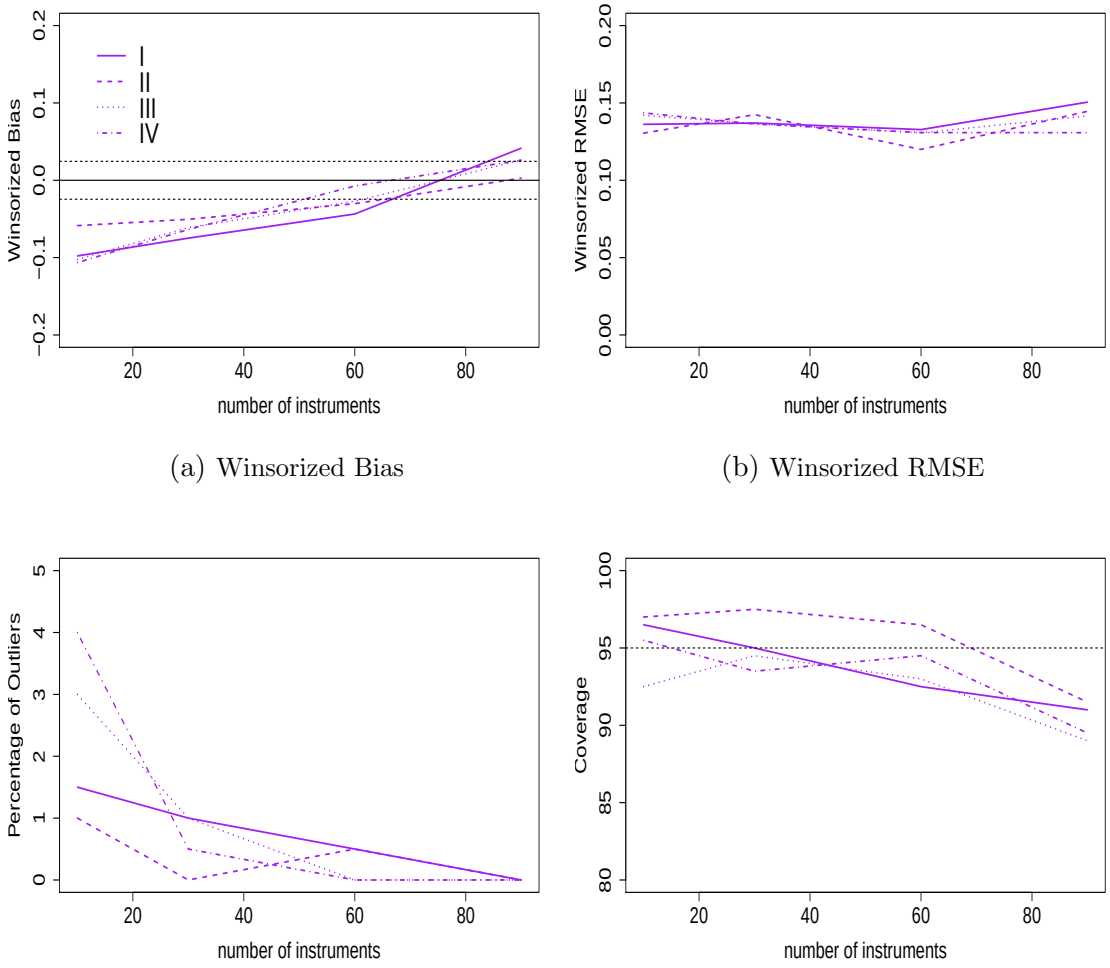
Figure 10.8 give the evaluation criteria of IVBMA for all patterns. Winsorised bias is decreasing with increasing number of instruments for all patterns, Figure 10.8a, since IVBMA is more certain of X (as shown by Figure 10.9) and has less estimates of zero for the causal effect. There is no difference in Winsorised bias between the patterns for all numbers of instruments, except for 10 instruments; Pattern II have

more weakly correlated SNPs in comparison to the other patterns, hence its mean MCMC causal effect estimate is pulled towards zero, as zero estimates are from the uncertainty of X being included in the model. The difference in Winsorised RMSE is caused by the variation in causal effect estimates of 200 datasets, as Winsorised bias showed no differences between the patterns. This is because each pattern have varying qualities of instrument; their correlation with the causal SNP and MAF. IVBMA becomes more confident of X with the number of instruments and hence the number of outliers decreases for all patterns. The coverage from all the patterns is similar, Figure 10.8d, and the coverage does not change with more instruments.

Comparison of classic and Bayesian approaches

When the number of instruments is small, IVBMA is more biased than LIML for all the patterns, as IVBMA is uncertain of X and estimates zero for causal effect as a result, Figure 10.10. At 90 instruments for pattern III and IV, IVBMA has become positively biased rather than negatively, as it is more certain of X and is not estimating a zero causal effect. LIML has too many outliers for Winsorisation to remove. For Pattern II there are hardly any differences in bias for 30, 60 and 90 instruments. IVBMA does not have any advantage with the choice of instruments as all the instruments are weakly correlated with the causal SNP. However the Winsorised standard errors from LIML are large in comparison to IVBMA, Appendix Table F.2.

IVBMA having the lowest Winsorised RMSE for all patterns in comparison to the classical approaches is not due to sampling error (Figure 10.11), as RMSE was calculated to have 0.4% accuracy from 200 datasets. IVBMA has priors to restrict the range a causal effect estimate can have, and its RMSE and percentage of outliers is smaller than LIML. For all patterns, as IVBMA is uncertain of X being included in the second regression with small numbers of instruments, 2SLS has less outliers by comparison, Figure 10.12. At 10 instruments LIML has similar coverage to IVBMA (accuracy of 3% for 95% coverage), but decreases from the 95% nominal level with an increase in the number of instruments, whereas IVBMA remains at approximately 95% for all patterns.



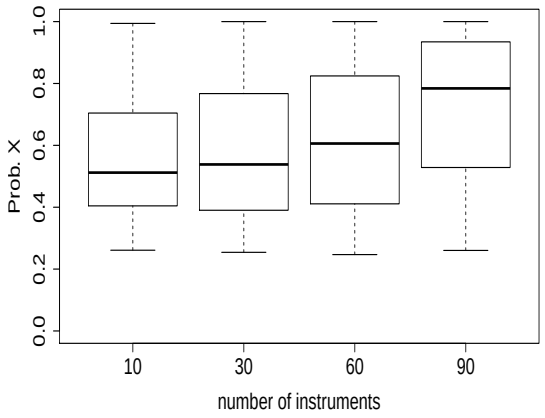
(a) Winsorized Bias

(b) Winsorized RMSE

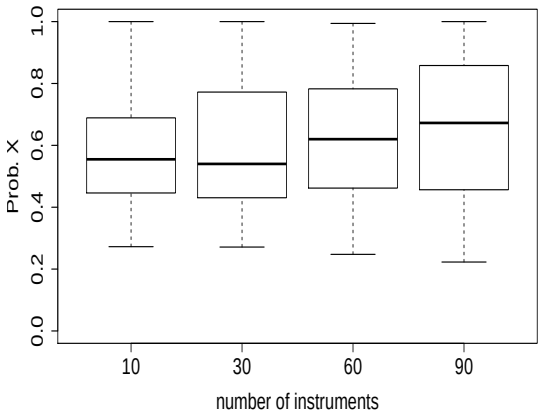
(c) Outlier

(d) Coverage

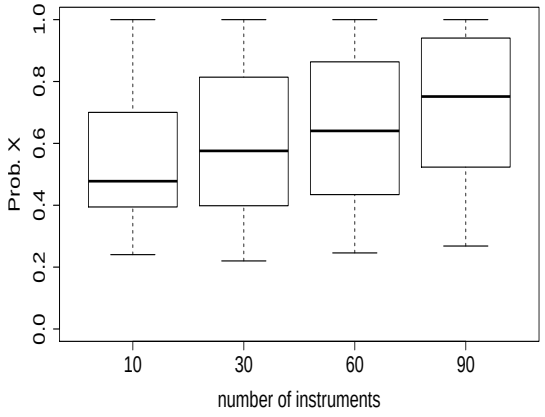
Figure 10.8: Comparing evaluation criteria of IVBMA from four different patterns. The purple coloured straight, dashed and dotted lines are common, low and variable MAF respectively. The black solid line in Winsorised bias is zero bias. The dotted line in Winsorised bias and coverage is 10% bias and 95% coverage respectively.



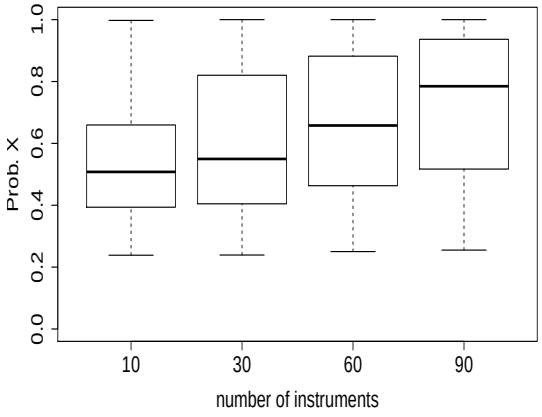
(a) Pattern I



(b) Pattern II

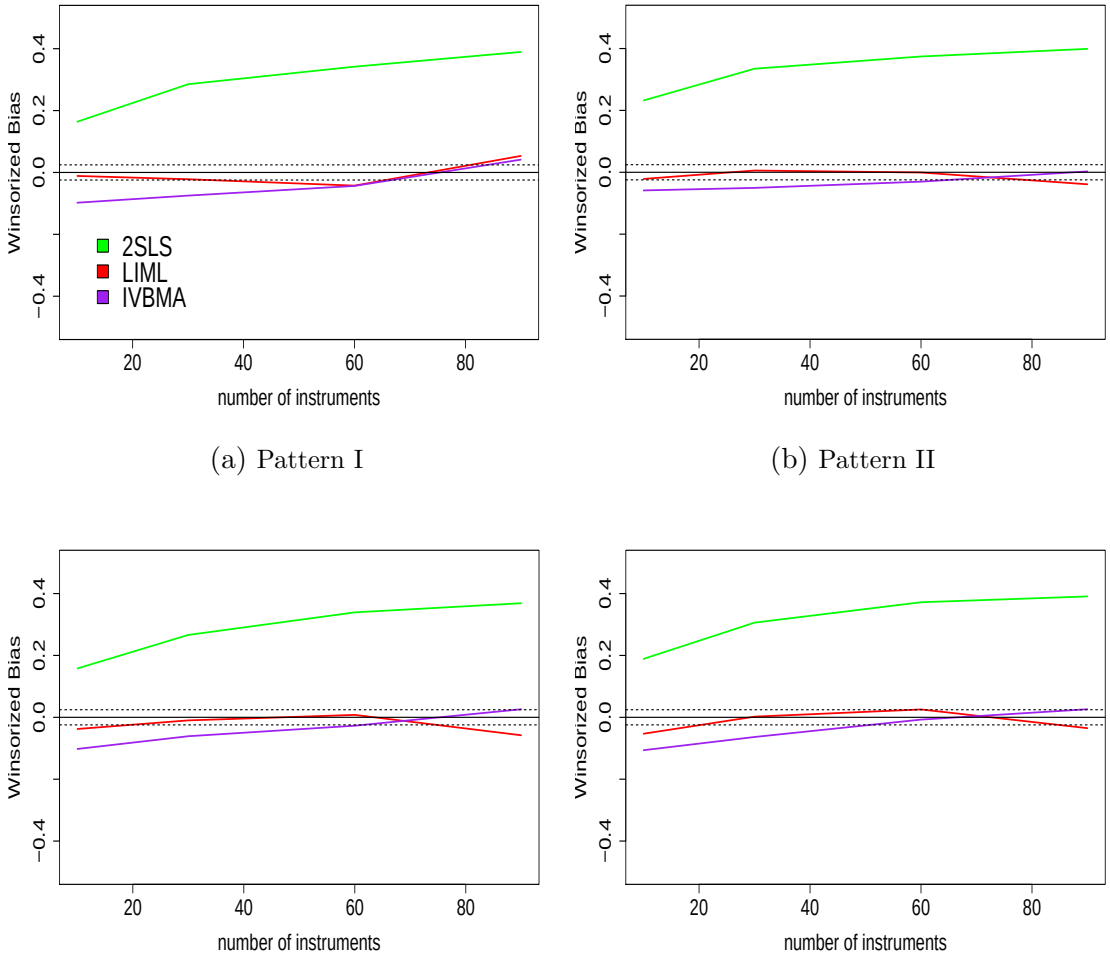


(c) Pattern III



(d) Pattern IV

Figure 10.9: Probability of X for all four patterns. Probability of X, is the probability of X included in the second regression of IVBMA.



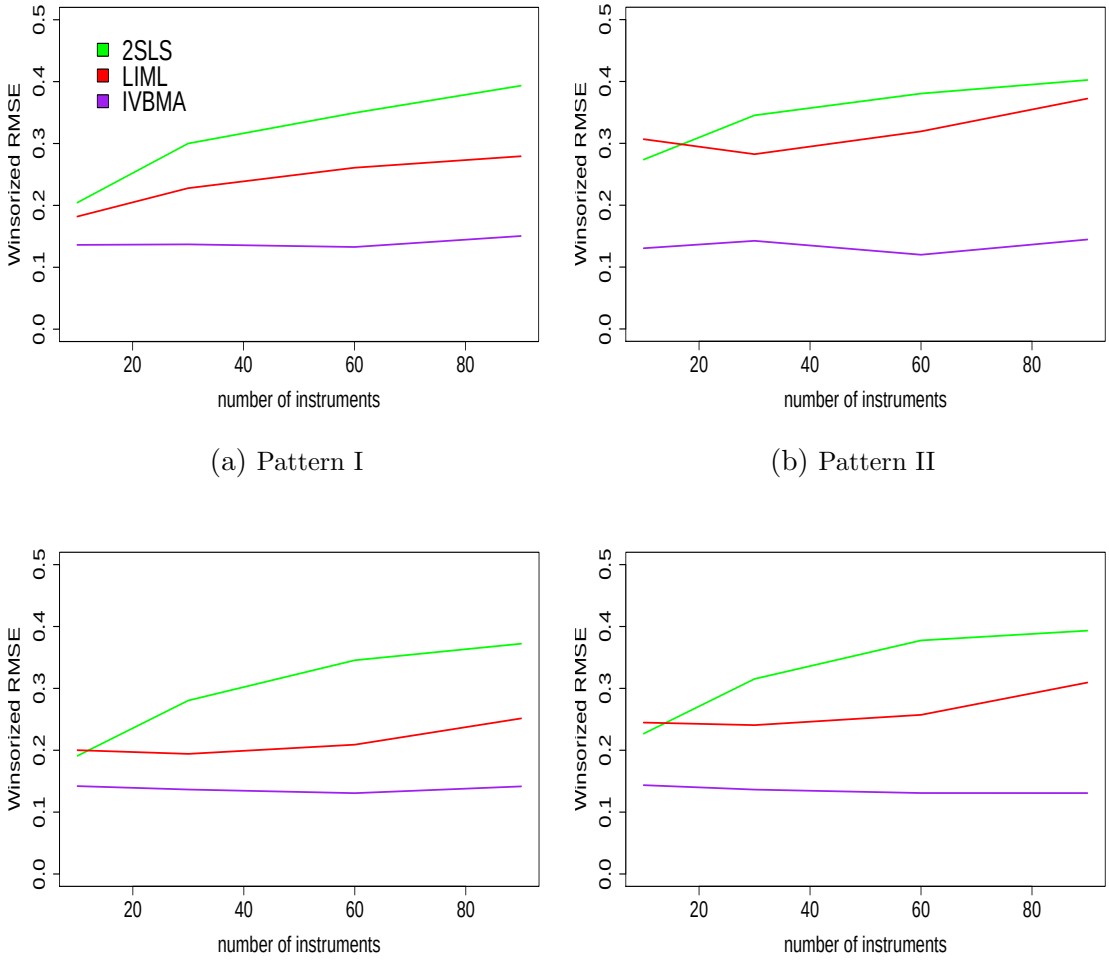
(a) Pattern I

(b) Pattern II

(c) Pattern III

(d) Pattern IV

Figure 10.10: Winsorised bias from 2SLS, LIML and IVBMA for all four Patterns. The estimators are coloured as green, red and purple respectively. The black solid line in each plot is zero bias and the dotted line is 10% bias.



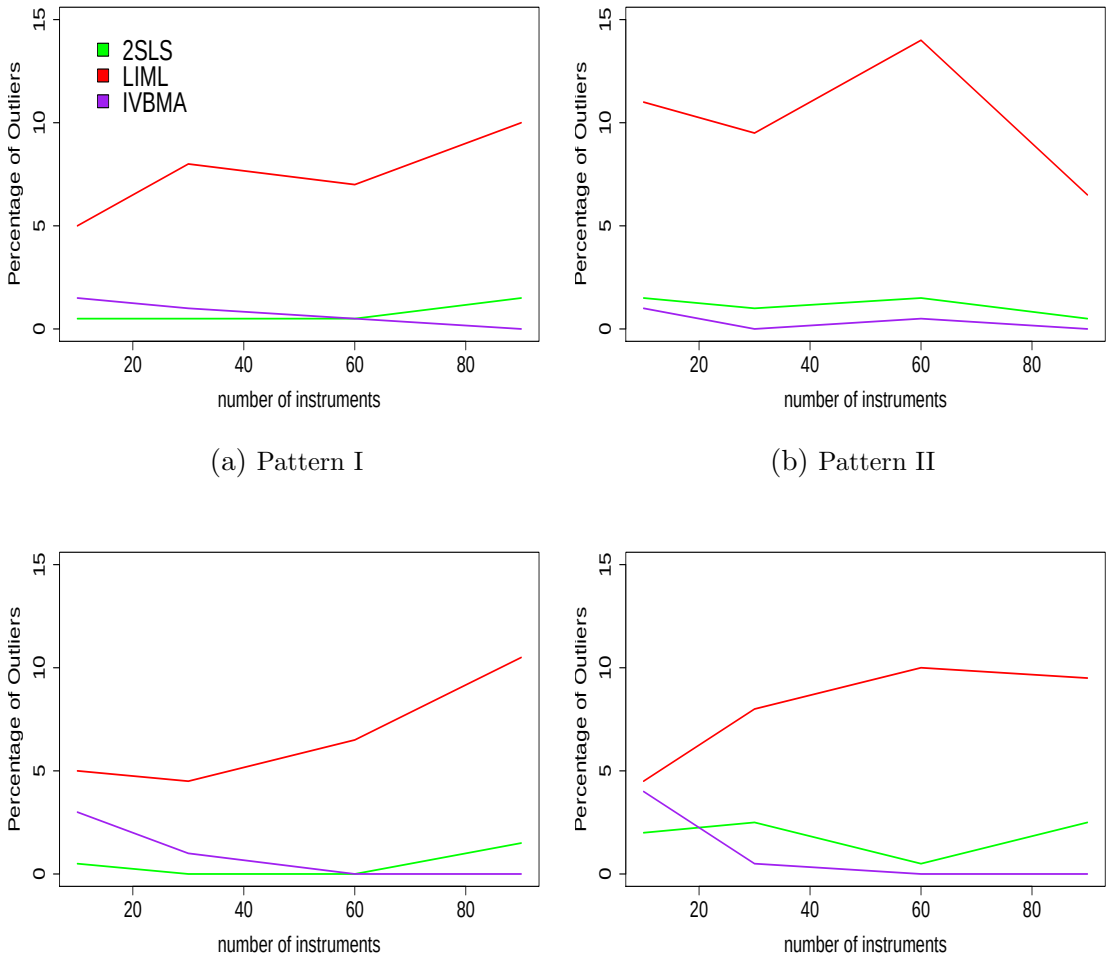
(a) Pattern I

(b) Pattern II

(c) Pattern III

(d) Pattern IV

Figure 10.11: Winsorised RMSE from 2SLS, LIML and IVBMA for all four Patterns. The estimators are coloured as green, red and purple respectively.



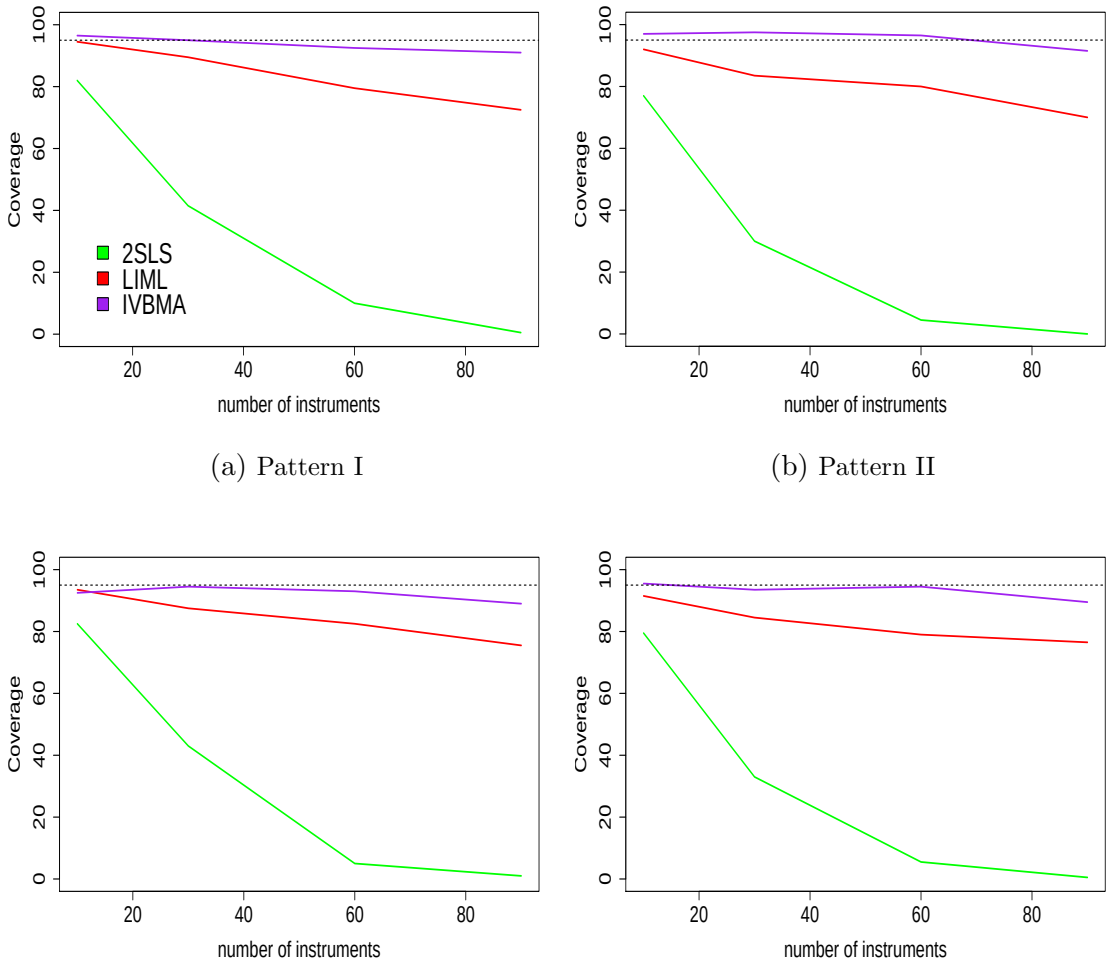
(a) Pattern I

(b) Pattern II

(c) Pattern III

(d) Pattern IV

Figure 10.12: Winsorised RMSE from 2SLS, LIML and IVBMA for all four Patterns. The estimators are coloured as green, red and purple respectively.



(a) Pattern I

(b) Pattern II

(c) Pattern III

(d) Pattern IV

Figure 10.13: Coverage from 2SLS, LIML and IVBMA for all four Patterns. The estimators are coloured as green, red and purple respectively. The black dotted line in each plot is 95% nominal coverage.

10.3.4 Conclusion

The different genetic patterns do not change the conclusion that IVBMA is more efficient with many instruments. Even though IVBMA has similar bias to LIML, it have the least RMSE and percentage of outliers, and has closer to the nominal coverage. IVBMA's performance does depend on how much information it has gained from the instruments. Lack of information will create uncertainty for the inclusion of X and thus zero causal effect estimates could arise.

10.4 GENOME

10.4.1 Aim

To understand the effect of MAF and correlation patterns on the performance of 2SLS, LIML and IVBMA, I have designed my own simulation methods for the generation of SNPs. However, the procedure are not realistic, thus GENOME simulator [185] will be used in this experiment. In the previous experiment, GENOME was not implemented as it does not allow for any control over the distribution of MAF or correlation between SNPs. This experiment aims to see whether there is a difference in conclusion from the previous experiment with realistic genetic patterns.

10.4.2 Design

The options in GENOME will be specified to simulate genotype of 200 SNPs for 2,000 individuals, as there are approximately 200 SNPs in an averaged sized gene[112]. The causal SNP will be randomly chosen among SNPs with rare MAF (< 0.1) and potential instruments will be SNPs with MAF> 0.1. The chosen causal SNP will explain 2% of the variation in X. 6% of the variation in Y will be explained by X. X and Y will be normally distributed with sample size of 2,000. All of the rare SNPs will be discarded and the comparison of estimators will be based on non-causal SNPs with MAF> 0.1 as potential instruments. The simulation will be repeated 200 times.

For IVBMA, 50,000 MCMC iterations will be run with a 10,000 burn-in for all GENOME simulated datasets. The performance of 2SLS, LIML and IVBMA will be measured by Winsorised bias and root mean squared error (RMSE), percentage of outliers and coverage.

Table 10.1: Summary of GENOME simulated genetic instruments.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
Number of Instruments	3	25	41	41	58	87	1

10.4.3 Results

Table 10.1 give the number of potential instruments. Most of the datasets have approximately 40 SNPs available and only one dataset had all SNPs with rare MAF

which were excluded from the analysis.

Table 10.2: Summary of IVBMA's Probability of X being included in the second regression for 200 GENOME simulated datasets.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
Prob. X	0.21	0.40	0.59	0.61	0.83	1.00	1.00

IVBMA has lower Winsorised bias in comparison to the classic methods and it is not due to the sampling error, Table 10.3. IVBMA also have the lower RMSE, proportion of outliers and has closest to the 95% nominal coverage. These results are plausible as IVBMA is allowed to make its own choice of instrument, whereas LIML and 2SLS do not and thus suffer from many weak instrument bias. Table 10.2 shows that for most of the 200 datasets, IVBMA included X in the second regression 60% of the time.

Table 10.3: Evaluation Criteria from 2SLS, LIML and IVBMA with GENOME simulated genetic instruments. $\hat{\beta}_{XY}$ is the causal effect estimates. True β_{XY} is 0.2449

	Mean $\hat{\beta}_{XY}$ (S.D.)	Winsor. S.E.	Winsor. Bias	Winsor. RMSE	Prop. of Outliers	Coverage
2SLS	0.4172 (0.1572)	0.0071	0.1635	0.1915	0.0101	73.87
LIML	0.7550 (9.0494)	0.0609	-0.2815	0.9017	0.0804	10.55
IVBMA	0.1957 (0.1664)	0.0080	-0.0691	0.1317	0.0000	94.47

10.4.4 Conclusions

The comparison of Bayesian and Classic approaches with GENOME simulated genetic data also demonstrates IVBMA's efficiency in estimating the causal effect by having the least Winsorised bias and RMSE, proportion of outliers and nominal coverage.

10.5 GRAPHIC Study: *FTO* gene, body mass index and blood pressure

There are many epidemiological studies demonstrating the association of obesity and blood pressure. However the exact mechanism of this relationship is unknown. Hence it is difficult to identify all of the potential confounding [94, 215, 254]. We hope to estimate the causal effect of body mass index (BMI) and mean 24 hour systolic blood pressure (SBP) using instruments from the *FTO* gene.

10.5.1 Data

The real data are from the Genetic Regulation of Arterial Pressure of Humans in the Community (GRAPHIC) study [274]. This population-based cohort has recruited 2037 white European participants within 520 nuclear families from Leicestershire, UK. To not further complicate the MR estimate with family effects, only parents' data were considered. Individuals were included if they had complete data for body mass index (BMI) and mean 24-hour systolic blood pressure (SBP).

10.5.2 Instruments

The *FTO* gene has been previously identified as BMI-related in GWAS [104]. The genotypes of 207 SNPs in the *FTO* gene were available. A SNP was included in the analyses if it (1) had less than 1% of missing data, (2) MAF was greater than 0.1 and (3) was in Hardy Weinberg Disequilibrium. After quality control there were 173 BMI-related SNPs. The genotypes of the SNPs were coded 0, 1, or 2 representing the BMI-increasing alleles within an individual.

10.5.3 Results

There were 2037 participants in the GRAPHIC study. After quality control, there were 1026 unrelated-individuals with complete record of BMI and mean 24-hr SBP. Table 10.4 gives the summary statistics of characteristics from the selected individuals.

Table 10.4: GRAPHIC study unrelated-individuals characteristics, N=1028

	N	Mean	Standard deviation
Gender (male)	1028(514)	-	-
Age (years)	1028	52.71	4.63
BMI (kg/m^2)	1028	27.44	4.28
Mean 24-hr SBP (mm Hg)	1026	120.61	12.02

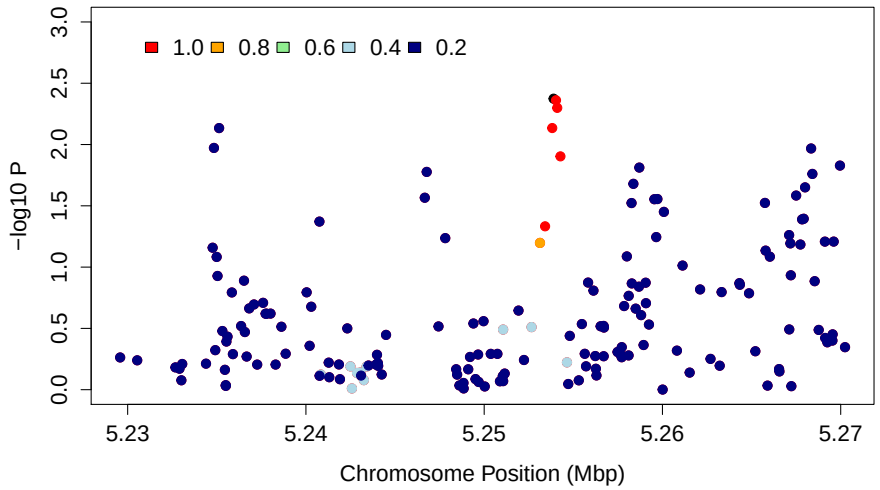


Figure 10.14: Regional association plots for BMI-related *FTO* variants. Regional P value plots where the p-value is from the regression of each SNP on BMI. On the x axis is SNP ID in the ascending order of chromosome position, and on the y axis is -$\log_{10} P$. Colour coding (from red to blue) denotes LD information; see also the legend within the plot.

All the SNPs from GRAPHIC would be considered as weak instruments, as their individual F-statistics from the association with BMI were less than 10 [253]. Even for the lead SNP, the F-statistic is 8 and explained approximately 1% of the variation in BMI (Appendix Table F.3). Figure 10.14 and Appendix Table F.3 show that out of 173 SNPs there are only 6 in strong correlation with the lead SNP. There is another SNP that was significant and had a similar effect size for BMI as the lead SNP and they were independent of each other. This resembles Pattern IV in the simulation study but with weaker instruments.

Table 10.5 gives the estimated coefficient, 95% confidence interval, standard

error and p-value for the effect of BMI on blood pressure with 1026 individuals. OLS regression estimated 0.90 (95% CI: 0.74,1.07) increase in blood pressure (mm Hg) with 1-unit increase in BMI. The estimates from 2SLS with all the SNPs as instruments approximates the OLS results. LIML estimated the effect of BMI in the other direction; LIML derived 5.72 (95% CI: -14.94,3.50) drop in blood pressure from 1-unit increase in BMI. However the standard error is relatively high in comparison to other methods and includes the null effect within its confidence interval. IVBMA with 250,000 iterations and 50,000 burnin, estimated a positive causal effect of BMI on systolic blood pressure, 4.3901 (95% CI: 4.35, 4.43) unit increase for 1-unit of increase of BMI.

Table 10.5: The effect of BMI on systolic blood pressure (mm Hg), where N=1026 and 173 instruments. SE is standard error. Poster. Prob. is inclusion posterior probability of BMI.

Method	Coefficient (95% Confidence Interval)	SE	p-value
OLS	0.9013 (0.7369, 1.0658)	0.0838	<0.001
2SLS	0.8613 (0.4682, 1.2544)	0.2006	<0.001
LIML	-5.7190 (-14.9426, 3.5046)	4.7060	0.2246
	Coefficient (95% Credible Interval)	Time-series SE	Poster. Prob.
IVBMA	4.3901 (4.3485, 4.4321)	0.0001	1.0000

10.5.4 Conclusion

2SLS are biased towards the OLS estimate with weak instruments [253]. Chapter 7 have shown LIML to give a median unbiased estimate but it can have extreme outliers if the dataset only has weak instruments. This is the case for GRAPHIC study; the highest F-statistic for all the SNPs was 8. Therefore the results from LIML are less believable. IVBMA estimated a positive relationship between SBP and BMI and was certain of their causal relationship, evident the inclusion posterior probability of 1. Based from the favourable simulations results for IVBMA, I am more inclined to believe there is a positive causal relationship between SBP and BMI from the GRAPHIC study.

However, recent Mendelian Randomisation studies have not found such a large magnitude of effect between BMI and SBP. In 2009, Timpson et al. [272] showed in

the Copenhagen General Population study of 36,851 participants that SBP of 3.85 (95% CI:1.88, 5.83) unit change per 10% increase in BMI, using *FTO* and *MC4R* as the two instruments. A meta-analysis of 30 studies on the effect of *FTO* genotype against SBP, resulted in 0.89 (95% CI: 0.48, 1.31) unit increase in SBP for 1-unit of increase in BMI [100]. Holmes et al. [140] have demonstrated from individual-level data of 6 studies (N=30,136) that per unit increase of BMI, SBP increased by 0.70 (95% CI: 0.24, 1.16), where the genetic instrument is in the form of genetic score from 14 SNPs, weighted by the coefficients from a discovery study [16].

It is surprising their estimated causal effects are similar to the 2SLS with 173 instruments from my analysis, even though the simulations from previous chapters have given evidence that 2SLS is severely biased with many instruments. However, Timpson et al. [272] have shown the estimate from the observational analysis are within the confidence interval from the instrumental variable analysis, and no significant difference between the two estimates. Perhaps the confounding effect on BMI and SBP are bidirectional and have canceled out, whereas IVBMA have suffered here, as all instruments have equal prior probability, consequently SNP selection is driven by data; some of the SNPs with strong association to BMI that also have a positive effect on confounding are selected more frequently in IVBMA than the SNPs that are weakly associated with BMI and have a negative effect on confounding. Further investigation of IVBMA will be required in scenarios where the genetic instruments are associated with confounding.

10.6 Discussion

From all the experiments, IVBMA has shown advantages over the classical approaches, as its causal effect estimates have given the least Winsorised bias and RMSE, percentage of outliers and are closest to 95% coverage in comparison to 2SLS and LIML. The genotypes of the SNPs are coded 0, 1 and 2. Thus, a SNP with low MAF will lack information, as it has more 0s. Then, the finite sample bias is introduced to the first regression when the instruments have low MAF, which then biases the causal effect estimate from IVBMA. The variation in MAF between the SNPs gave greater RMSE from IVBMA in comparison to SNPs all having either common or low MAF. This is because some datasets consist of SNPs with common MAF and strong correlation with the causal SNP, and some dataset have SNPs with the same strong correlation but have low MAF. Therefore IVBMA is more likely to

select the next strongest correlated SNP that possess higher MAF, as shown in Section 9.5. Datasets with SNPs that are weakly correlated with the causal SNP will increase IVBMA's uncertainty about the presence of X in the second regression. This uncertainty creates less confidence in the mean causal effect estimates from IVBMA. If IVBMA is uncertain of X in a dataset then the posterior distribution will have a bimodal shape with one peak at 0 for models without X and the other peak at mean estimates from models with X. As a result the mean of the causal effect estimate is pulled between the two peaks and by chance the mean is at the true causal effect, as seen in the previous chapter.

As explained above when IVBMA is uncertain of the presence of X in the second regression, it is possible by chance that the mean causal effect estimate is centred at the true value. Therefore to check whether this situation occurred, the posterior distribution of the causal effect estimate should have been plotted, but storing the results of 200 datasets, each with 40,000 iterations will require a large computer memory. The long computation time of the Bayesian approach has induced a couple of limitations to this simulation study; (1) simulations were only repeated 200 times, which did not give a high accuracy level for the classical approaches in comparison to 10,000 datasets which have 3%, 4% and 0.4% for 95% coverage, 0.0227 bias and 0.228 RMSE respectively, derived from Section 4.5. However the low accuracy level was taken into account when comparing to the Bayesian approach. (2) The number of iterations for IVBMA was low, 50,000, Chapter 9 found that 250,000 iterations is required to show evidence of convergence in 95% credible interval with 90 instruments.

In the IVBMA literature, Karl and Lenkoski [160] did not compare *ivbma* to a classical approach, but they have shown through an empirical example that IVBMA have similar estimates as 2BMA (see Section 8.3), except IVBMA favours more parsimonious models. Lenkoski et al. [184] did compare 2BMA to 2SLS. Therefore I will compare my simulation results to Lenkoski et al. [184]. They have found from 500 datasets, each with 10 instruments and 100 individuals that 2BMA gave 45% less bias and 46% less mean squared error in comparison to 2SLS, which is similar to my results where there are 10 SNPs as potential instruments; IVBMA have lower Winsorised bias and RMSE than 2SLS. Note that Koop et al. [172] did not provide a simulation study to compare the classical approach to BMA in instrumental regression model.

The main reason for IVBMA having a smaller RMSE and percentage of outliers

in comparison to the classical approaches was due to its prior, as it restricts the range that the causal effect estimate can have. In the Bayesian literature, most of the priors were designed to improve the convergence and the mixing in BMA [213]. There is very little information on how to design priors based on biological knowledge, which is an important area for further research, and therefore the next and final chapter, which concludes the thesis, will include a discussion of the biological priors for a Mendelian randomisation analysis.

Chapter 11

Discussion

The motivation for investigating many dependent instruments in Mendelian randomisation is in the scenario where the causal variant(s) was unknown or not genotyped. If the independent instrument is not the causal or highly correlated with the causal variant(s) and therefore it does not fully explain the true variation, consequently causing weak instrument bias. Genome-wide association studies have also utilised the linkage disequilibrium between SNPs to ensure the coverage of ungenotyped causal variant [19, 69, 89]. In addition, a variant showing significant association with the exposure in one population may not be the case in another [239, 260], as a consequence this genetic variant as an instrument may overestimate the true causal effect. The same motivation have been expressed in the context of Mendelian randomisation by Burgess et al. [52] and Wang et al. [283], however they have not focused on many dependent SNPs in individual-level data.

The primary aim of this thesis was to investigate the use of many dependent SNPs from a single region for Mendelian randomisation. The three main questions were;

1. Are there any gains from using many dependent SNPs from the same gene?
2. What is the most efficient estimator for many dependent SNPs?
3. In comparison to the classical approaches, are the Bayesian approaches more efficient with many dependent SNPs?

11.1 Summary of Findings

11.1.1 Q1: Potential gains

If the causal SNP is unknown or unmeasured, then the answer to question 1 is yes. Chapter 5 showed that the variation explained by a single non-causal SNP is dependent on the linkage disequilibrium (LD) that it has with the causal SNP and the true variation explained by the causal SNP. RMSE is the potential gain, but at the cost of bias, coverage and type I error (the gain in power is therefore misleading); Figure 6.1c and Figure 6.1a showed there is a 22% decrease in RMSE but a 6% increase in bias between a single weak dependent SNP and multiple dependent SNPs. However, the increase in bias from many instruments for two-stage least squares (2SLS) is also seen in independent SNPs [219, 227]. 2SLS is biased if the increasing number of instruments does not increase the variation explained [227]. An instrument selection policy can ensure that a SNP is not included in the instrument set if it does not increase the variation explained by the existing set. Section 6.4 demonstrated that bias can be introduced if instrument selection is performed in the dataset under analysis and it is difficult to determine the best instrument selection policy for any one sample.

Unless the sample size is large or there is external information for instrument selection, many dependent SNPs should not be used in 2SLS. This conclusion is not unique to dependent instruments but also applies to independent instruments. Both Burgess et al. [52] and Wang et al. [283] have found the use of multiple SNPs from the same gene can increase power and precision of the estimation of a causal effect compared to the use of a proxy SNP. However, they did not assess many dependent SNPs as individual instruments in 2SLS. The implication is that another estimator should be implemented, an unbiased estimator for many dependent instruments that does not require an instrument selection policy to reduce weak instrument bias.

11.1.2 Q2: Classical estimators

The answer to Question 2 is that Limited Information Maximum Likelihood (LIML) is the most efficient estimator, as it has the least bias and variation, it has the closest to nominal coverage and significance level for many weak instruments, as shown in Chapter 7. For example, Table 7.4, for GENOME simulated SNPs, LIML had 94% less bias and 29% lower RMSE than 2SLS. A similar conclusion was reached

by Davies et al. [85] with many independent instruments when the condition of homoskedasticity is satisfied. However, Davies et al. [85] also found that LIML has extreme outliers, as shown in Table 7.4 2SLS has 62% fewer outliers than LIML. Table 7.5 demonstrates that the proportion of extreme estimates was reduced if datasets with weak instruments are excluded, or if only SNPs with F-statistics > 10 were included. As LIML is only median unbiased, the interpretation of the causal effect must be done with caution when the analysed dataset that only contains weak instruments. The estimation of a causal effect by classical approaches to Mendelian randomisation is still subject to many weak instruments bias and implementing instrument selection policies on a dataset under analysis does introduce selection bias, which motivates the investigation of Bayesian approaches as a better alternative.

11.1.3 Q3: Classical and Bayesian estimators

I found that Bayesian analysis is more efficient than the classical approach for many dependent instruments. The simulations in Chapter 10 has shown that Instrumental variable Bayesian model averaging (IVBMA) gave the least bias, variation, outliers, and the best nominal coverage in comparison to the classical approaches. For GENOME simulated SNPs, Table 10.3, IVBMA gave 75% and 85% less bias and RMSE than LIML, respectively, and IVBMA did not estimate any extreme estimates, whereas LIML had 8%. IVBMA provides a solution to *many* weak instruments bias. IVBMA usually selects the SNP that has the highest LD with the causal SNP unless it has a low minor allele frequency (MAF), in which case it selects the next highest LD where the SNP has a greater MAF, as shown in Section 9.5. The reason why IVBMA does not have any extreme estimates like LIML is because the priors restrict the range of the estimates, which is another important advantage of BMA. The inclusion of biological priors into the estimation of the causal effect will be discussed later in this chapter.

I have focussed on the literature on the comparison of results with two-stage Bayesian Model Averaging (2BMA)[160], as Lenkoski et al. [184] did not compare their IVBMA with the classical approaches and Karl and Lenkoski [160] have shown through an empirical example that IVBMA gives similar estimates as 2BMA, except IVBMA favours more parsimonious models. They have found from 500 datasets, each with 10 instruments and 100 individuals, that 2BMA gave lower bias and mean squared error than the results from 2SLS, which is similar to my results where there are 10 SNPs as seen in Chapter 10.

My simulations suggest that IVBMA is the most suitable for many dependent SNPs, as it has the ability to select SNPs that have high correlations with the causal SNP, and do not have large outliers like LIML, due to the priors that strengthen the causal effect estimation.

11.2 Challenges and Limitations

Apart from the other violations to the IV assumption discussed in Chapter 2, pleiotropy is the most likely violation for many dependent instruments, as highlighted by the difference in causal effect estimates from previous findings and applied example in Section 10.5. If a instrument is in linkage disequilibrium with another genetic variant that provides a different pathway to the disease outcome, then the IV assumptions are violated. However, the violation from pleiotropy is not unique to many dependent instruments [220]. Through well-established genetic functions the bias from pleiotropy can minimised. In addition, Egger regression [31] can be implemented as a sensitivity analysis to detect the bias from the pleiotropic effect, method described in Section 2.5.1.

Some of the aspects of the simulation settings may not be representative of real world genetic data. Throughout this thesis, most of the simulations generated a causal SNP to explain 2% variation in X, but this is uncommon as most of GWAS significant SNPs explained less than 1% of the variation. The potential instruments were simulated to have MAF ranging from 0.1 to 0.5, and Gibbs et al. [112] estimated 10 million SNPs in the human genome and only 5 million have MAF$>$ 0.1. However, in GWAS, rare SNPs are usually excluded from the analysis, as they lack power [6]. These two genetic properties will create weaker instruments than the instruments from my simulation method for Mendelian randomisation and therefore the magnitude of bias will be greater with finite samples and have lower power.

The simulations in this thesis have assumed several simplifying assumptions including using an additive genetic model, no gene-gene interactions and a linear effect for X on Y. If the genetic modelling assumptions are not valid then it is possible to incorporate this knowledge, discussed in detail elsewhere [182, 227]. X and Y were simulated as continuous variables, however epidemiological studies often express causal effects as a risk ratio or odds ratio, in which case alternative algorithms must be considered as binary outcomes raise different and somewhat more complicated statistical challenges [90]. The tried and tested approximation methods for

binary outcomes with multiple instruments in MR studies includes two-stage logistic models [218], structural mean models (SMM) [30] and SMM using GMM (which allow instruments to be modelled independently, unlike SMM itself which requires instruments to be combined as one) [66]. It is also unclear which algorithm would be most appropriate with multiple instruments for a binary outcome.

An additional assumption of the simulations was that the condition of homoskedasticity was satisfied. The results in Chapter 7 showed that 2SLS and CUE are identical under homoskedasticity, consistent with the findings in the literature [116, 129]. If this assumption is violated then CUE should outperform LIML [85, 207], as LIML is known to be inconsistent under the case of heteroskedasticity and many weak instruments [62]. In the econometrics literature, apart from CUE there are other alternatives and modifications of LIML, namely, a jackknife version [132] and asymptotic optimal modification [177] of LIML, and even a jackknife version of GMM (which claims to be consistent with unknown heteroskedasticity) [25]. However, these alternatives do come with further assumptions and have not yet been tested in a Mendelian randomisation setting.

My simulations only considered individual-level data of SNP genotypes, exposure and outcome of interest. Due to data protection, individual level data are not always accessible. Hence Burgess et al. [52] have developed meta-analysis based algorithms to include many correlated SNPs from summarised data as instruments in Mendelian randomisation. Although their simulation also examined correlated SNPs, there are differences in comparison to my simulation method. Their strongest pairwise correlation was 0.5 and they did not assume a single functional variant SNP in the genetic region but all the SNPs in the region have their own direct effect on the exposure and the correlation did not have a direct effect on the variation explained. In theory, the contrast in simulation methods should not affect the estimation of the causal effect as both SNP simulation algorithms gave each SNP an association with the exposure, but the simulated SNPs in this thesis are arguably weaker instruments and more realistic as their variation explained is dependent on the functional variant.

LIML, CUE and IVBMA are computationally intensive algorithms; CUE and LIML took approximately 100 hours for 10,000 datasets, each with 90 instruments. As time was limited, Winsorisation had to be applied to the bias and RMSE as LIML and CUE had extreme outliers. Even with 10,000 replicates, their mean causal effect estimates were not accurate, and the standard errors were still relatively large. Davies et al. [85] also simulated 10,000 datasets, where they used the median and

inter-quartile range (IQR) of causal effect estimates. The time taken is even longer for the Bayesian approach as it is simulation within simulation. For example, 200 datasets, each with 90 instruments, took approximately 18 hours. Karl and Lenkoski [160] also ran 200 datasets for their comparison of BMA with 2SLS.

Allele score is one of the popular methods in Mendelian randomisation and was also compared in Davies et al. [85], which this thesis did not evaluate. The main reason was that deriving a weighted allele score requires external data. However this is not always easy for multiple SNPs in a single gene region as there is a limited number of reference populations [52]. As discussed in Chapter 1, the same SNPs do not always appear in all of the populations and the effect sizes for the genetic effect may vary between populations. For example, Imamura et al. [149] have found two loci with the type II diabetes risk allele specific to the Japanese population. Allele scores can be unweighted, but Palmer et al. [220] warned combining multiple SNPs into a single instrument by unweighted allele score has lower power compared to including them as individual instruments.

11.3 Ongoing and Further Work

There are two main papers I aim to write following the work of this thesis. The first has a working title of "A Comparison of Estimators with Many Dependent Instruments in Mendelian Randomisation", based on the work from Chapter 7. Davies et al. [85] have compared the classical approaches with many weak instruments, however they did not discuss the possible outliers that exists in LIML and CUE. Therefore I believe this paper could bring to the attention of investigators, that LIML and CUE are not necessarily always unbiased for any one sample. In addition this paper demonstrates that multiple *dependent* SNPs from a single genetic region can be used as instruments with estimators that are commonly applied with *independent* instruments. The second paper I aim to publish has the working title of "Instrumental Variable Bayesian Model Averaging to Mendelian Randomisation", based on the work of Chapters 9 and 10. The concept of the IVBMA approach is a novel approach in Mendelian randomisation, and with evidence that IVBMA outperforms the classical approaches, I believe this will have substantial impact in the field of instrument selection for Mendelian randomisation.

Prior sensitivity is an important aspect of Bayesian analysis. The R package for IVBMA, *ivbma* does not provide options to change the priors imposed by the

package hence, for the flexibility of prior specification, `OpenBUGS` will need to be used instead which is less computationally efficient than the R package, as shown in Section 9.3.1. *ivbma* uses fairly informative priors, perhaps necessary as vague priors will slow down the Bayesian algorithm and create mixing difficulties. Therefore, given more time, the *ivbma* R package could be extended to allow for prior specification, and to investigate the magnitude of the effect of the change in priors on the causal effect estimation. Alternatively, the computation time in `OpenBUGS` could be shortened by designing efficient priors. Designing priors for Bayesian variable selection with correlated variables is slightly more complex than uncorrelated variables [70]. Simulations of priors mentioned by O'Hara et al. [213] should be performed to find the most appropriate statistical prior for Mendelian randomisation with many correlated SNPs.

Wang et al. [283]'s paper was published in December 2015, where they compared instruments from a SNP selection method via stepwise regression to haplotypes as instruments and found haplotype-IV has greater power gain than the SNP-IV method (see Section 1.2). However, previous work by Clayton et al. [69] found that with more "tag" SNPs, the prediction for a genetic association study from SNPs and haplotypes are similar. Since Wang et al. [283]'s work was published in the later stage of this thesis, an extension to this thesis would be to compare their haplotype-IV method to the IVBMA approach for the estimation of the causal effect with many dependent SNPs.

A further possible extension to this thesis would be to consider a family effect in Mendelian randomisation. Section 10.5 introduced the GRAPHIC study [274], where the study had information from each family but only parent's data were used in the analysis, as the family data would affect the variation of the causal effect estimate. There are two possible methods to include the family effect into causal effect estimation; Morris et al. [204] have derived an algorithm similar to G-estimation, which tests whether the residual in Y from the second regression of 2SLS is independent of the genetic instrument with a range of causal effect estimates that allows for within family correlations. However the efficiency of this test could not be determined as they have only reported the results from real data. In 2010, Morris et al. [205] proposed the use of structural equation modelling (SEM) as a technique to incorporate measurement error into causal inference for general pedigree data. SEM consists of measurement and structural models, where the measurement model evaluates the latent variables from observed variables and the structural model estimates

the relationships between latent variables and observed variables. Due to the mechanism of the two models, SEM is able to allow for genetic variance components, such as environmental and polygenic effects. They have also developed an R package *strum* [252], in which Mendelian randomisation is considered as a special case of SEM. However, they have not provided any form of evidence for the consistency of their algorithm in estimating the causal effect.

11.4 Biological Priors

The assumptions of Mendelian randomisation should be validated by biological knowledge [42, 219] and Bayesian approaches are a way of incorporating this prior knowledge. However, most of the Bayesian statistical literature focuses on priors that improves the mixing and speed of convergence; O'Hara et al. [213] provides an insightful review of priors and samplers for Bayesian variable selection. Therefore, this section aims to discuss the possible biological knowledge that can be quantified into Bayesian priors.

The implementation of Bayesian variable selection in genetic association has been purposed by Fridley [106], where he reviewed some of the algorithms in O'Hara et al. [213] with SNPs as potential predictors, but there was very little discussion on how to quantify priors into the model space. Due to the small genetic effect, GWAS uses SNP prioritisation to increase statistical power [107, 200, 267]. This approach can be applied as weights in the model space of the gene and exposure regression. One can put weight on a single GWAS significant SNP, which can be found in GWAS Central at *www.gwascentral.org*. Alternatively, we can check whether any of the SNPs are proxies for the GWAS significant SNP in SNAP (SNP Annotation and Proxy Search) at *www.broadinstitute.org*. Similar weighting can be given to models that include SNPs that are from the same haplotype blocks, as these SNPs are more likely to have a similar exposure association from their correlation with the functional variant(s). SNPs with protein coding function should be given more weight, as a SNP in a functional protein domain may increase the probability of true association [200]. Information on the function of SNPs can be found using Ensembl at *www.ensembl.org* [3]. The prior for regression coefficients in the first regression can be extracted from GWAS Central at *www.gwascentral.org*, where information on effect sizes and standard errors are available for GWAS significant SNPs. Preferably, information from meta-analysis of multiple populations should

be used, as the genetic effect can be different between populations.

A cautionary note for correlated SNPs; including two identical SNPs would cause the "dilution" problem (Discussion by George in Clyde [70]), as this is essentially giving these SNPs twice the weight than the rest of the SNPs. Incorporating prior information for correlated variables can be complicated [146]. For example, consider an OLS regression model with two variables that are correlated with each other. If one variable is removed from the model then the coefficient for the variable in the model will change. Hence for BMA with correlated variables, assigning priors to each variable should be conditional on which variable is already in the model. However to implement this theory for many variables will be complicated and computationally intensive, as one has to set priors for every possible model for each individual variable.

Decisions on model space for the second regression is a debatable issue; both Koop et al. [172] and Karl and Lenkoski [160] allows X to come in and out of the second regression model and hence the causal effect can be zero. This theory is in keeping to the true Bayesian philosophy, as it is not assuming a structure for the model, that X is affecting Y. If the investigator is sure of the causal relationship between X and Y then the model space for the second regression is not necessary. However, if the investigator is not confident then priors for model space should be imposed. An advantage of including priors for the model space is that BMA will give the posterior probability of X being included in the model and that enables us to quantify the uncertainty of the presence of a causal effect. The prior for the causal effect, β, could be obtained from the literature in the form of meta-analysis of the relationship between outcome and risk factor of interest, and use the overall mean and standard deviation as the prior.

Jones et al. [153] have found that the structure of an instrumental variable model is robust to the prior misspecification of covariance matrix, as the model does not directly estimate the causal effect from the covariance matrix. Nevertheless, the covariance matrix does effect the precision of the causal effect estimate. In addition, BMA accounts for the model uncertainty in the causal effect estimate. Thus, without a more informative prior on the covariance matrix, the precision of the causal effect could potentially be very wide. The amount of confounding between Y and X can be indicated by clinical trials and epidemiological studies; if available, examine the change in coefficient of X on Y regression before and after being adjusted for measured confounders. If the coefficient decreases this implies presence of positive

confounding and if it increases this suggests negative confounding [48].

Minelli et al. [200] have published an extensive research on prioritising SNPs in GWAS findings based on prior knowledge from both experts' opinion and empirical evidence. With the same motivation as GWAS, biological priors can increase the statistical power in the estimation of a causal effect and support the validity of Mendelian randomisation assumptions. As more biological information is becoming available with time, it is possible to design biological priors for Bayesian analysis. Therefore, further research in quantification of biological priors in Bayesian approaches to Mendelian randomisation should be conducted jointly between statistician and biologist.

11.5 Practical Implications

This thesis fills the gap in methodological work on the use of many dependent SNPs from a gene as instruments in Mendelian randomisation, in situations where the causal SNP is unknown or unmeasured. My simulations suggest that Bayesian model averaging is preferable to the classical approaches, with or without SNP selection. Furthermore, as time passes and more biological knowledge is accumulated, and we will be able to incorporate this knowledge into the Bayesian approach of Mendelian randomisation.

Appendix A

Simulating Mendelian Randomisation

A.1 Genotype of multiple SNPs

The simulation aims to create a variable, henceforth X, that is correlated with variable Z, and have a normal distribution $\sim (0, 1)$, hence the linear regression equation is;

$$X = \beta_1 Z + \beta_2 \varepsilon \tag{A.1}$$

where Z and ε have normal distribution of $\sim (0, 1)$. From a fundamental statistics textbook and the desired correlation of ρ_{XZ} between X and Z, equation above can be rewritten to,

$$X = \rho_{XZ}\sqrt{\frac{Var(X)}{Var(Z)}}Z + (1 - \rho_{XZ})\sqrt{\frac{Var(X)}{Var(\varepsilon)}}\varepsilon \tag{A.2}$$

where $Var()$ is the variance of its variables. To simulate X to have normal distribution $\sim (0, 1)$, hence $Var(X)$ must satisfy,

$$\left(\rho_{XZ}\sqrt{\frac{Var(X)}{Var(Z)}}\right)^2 Var(Z) + \left((1 - \rho_{XZ})\sqrt{\frac{Var(X)}{Var(\varepsilon)}}\right)^2 Var(\varepsilon) = 1 \tag{A.3}$$

since fundamental of statistics states in a linear regression $Var(X) = \beta_1^2 Var(Z) + \beta_2^2 Var(\varepsilon)$

As variance of X and ε is 1, then

$$(\rho_{XZ}\sqrt{Var(X)})^2 + ((1-\rho_{XZ})\sqrt{Var(X)})^2 = 1 \tag{A.4}$$

After some rearrangements,

$$Var(X) = \frac{1}{\rho_{XZ}^2 + (1-\rho_{XZ})^2} \tag{A.5}$$

Now replace $Var(X)$, Equation A.2 becomes,

$$X = \frac{\rho_{XZ}Z + (1-\rho_{XZ})\varepsilon}{\sqrt{\rho_{XZ}^2 + (1-\rho_{XZ})^2}} \tag{A.6}$$

Appendix B

One and Two Instruments in 2SLS

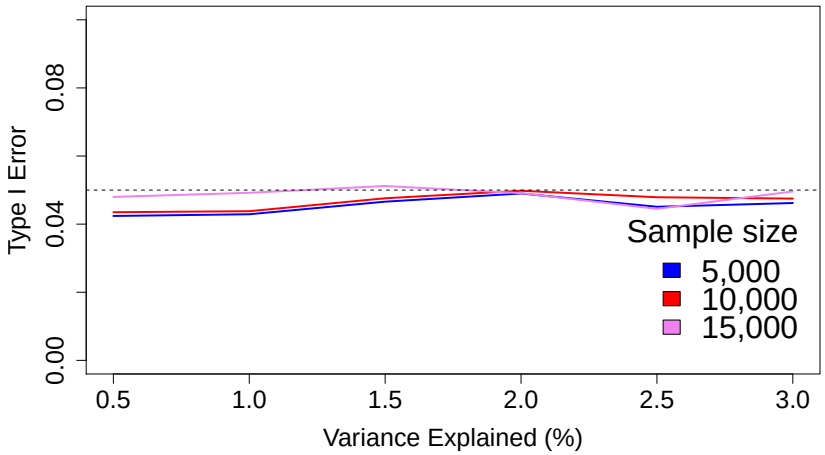


Figure B.1: Type I error of 2SLS against variance explained by SNP_c with different sample size. The blue, red and violet coloured lines are sample sizes of 5,000, 10,000 and 15,000 respectively. The black dotted line is the 5% significance level.

Table B.1: Evaluation criteria of 2SLS against variance explained by SNP_c with different sample size. $\hat{\beta}_{XY}$ is the estimated β_{XY}. The true β_{XY} is 0.5480. S.D. is the standard deviation of mean of $\hat{\beta}_{XY}$. S.E. is the standard error from the standard deviation.

Var. Explained	Mean $\hat{\beta}_{XY}$	S.D.	S.E.	Bias	RMSE	Coverage	Power	Type I Error	F-statistics
Sample size 5,000									
0.50	0.5009	0.4844	0.0048	-0.0472	0.4866	95.76	0.30	0.04	25.12
1.00	0.5201	0.3217	0.0032	-0.0279	0.3229	95.70	0.44	0.04	50.48
1.50	0.5322	0.2609	0.0026	-0.0158	0.2614	95.34	0.57	0.05	76.11
2.00	0.5370	0.2233	0.0022	-0.0110	0.2235	95.09	0.67	0.05	102.00
2.50	0.5324	0.1976	0.0020	-0.0156	0.1983	95.49	0.75	0.05	128.15
3.00	0.5389	0.1812	0.0018	-0.0091	0.1814	95.38	0.82	0.05	154.58
Sample size 10,000									
0.50	0.5300	0.3222	0.0032	-0.0180	0.3227	95.7	0.45	0.04	50.24
1.00	0.5381	0.2206	0.0022	-0.0099	0.2208	95.6	0.68	0.04	100.99
1.50	0.5413	0.1801	0.0018	-0.0068	0.1802	95.2	0.82	0.05	152.25
2.00	0.5432	0.1547	0.0015	-0.0048	0.1548	95.0	0.91	0.05	204.04
2.50	0.5445	0.1381	0.0014	-0.0036	0.1381	95.2	0.96	0.05	256.36
3.00	0.5421	0.1251	0.0013	-0.0059	0.1252	95.2	0.98	0.05	309.22
Sample size 15,000									
0.50	0.5376	0.2582	0.0026	-0.0104	0.2584	95.2	0.57	0.05	75.37
1.00	0.5387	0.1781	0.0018	-0.0093	0.1783	95.1	0.82	0.05	151.50
1.50	0.5401	0.1485	0.0015	-0.0079	0.1487	94.9	0.93	0.05	228.40
2.00	0.5443	0.1259	0.0013	-0.0038	0.1260	95.1	0.98	0.05	306.08
2.50	0.5452	0.1115	0.0011	-0.0028	0.1115	95.5	0.99	0.04	384.56
3.00	0.5454	0.1029	0.0010	-0.0026	0.1029	95.0	1.00	0.05	463.86

Table B.2: Evaluation criteria of 2SLS with a non-causal SNP, SNP_1, as instrument. r^2 is the correlation between SNP_1 and SNP_c. Var is the percentage of variance in X that is explained by SNP_c. $\hat{\beta}_{XY}$ is the estimated β_{XY}. The true β_{XY} is 0.5480. S.D. is the standard deviation of mean of $\hat{\beta}_{XY}$. S.E. is the standard error from the standard deviation.

r^2	Mean $\hat{\beta}_{XY}$	S.D.	S.E.	Bias	RMSE	Coverage	Power	Type I Error
Var=1%								
0.1	-0.1694	44.3582	0.4436	-0.7174	44.3617	95.7	0.15	0.04
0.2	0.3078	6.7134	0.0671	-0.2402	6.7174	95.5	0.20	0.04
0.3	0.4477	0.8507	0.0085	-0.1003	0.8566	95.7	0.23	0.04
0.4	0.4855	0.5624	0.0056	-0.0625	0.5658	95.8	0.27	0.04
0.5	0.5004	0.4862	0.0049	-0.0476	0.4885	95.7	0.30	0.04
0.6	0.5107	0.4296	0.0043	-0.0373	0.4312	95.8	0.33	0.04
0.7	0.5171	0.3913	0.0039	-0.0310	0.3925	95.7	0.36	0.04
0.8	0.5213	0.3616	0.0036	-0.0267	0.3626	95.5	0.38	0.05
0.9	0.5241	0.3398	0.0034	-0.0240	0.3406	95.5	0.42	0.04
1.0	0.5281	0.3196	0.0032	-0.0199	0.3202	95.3	0.44	0.05
Var=2%								
0.1	0.3471	9.0533	0.0905	-0.2009	9.0551	95.7	0.19	0.04
0.2	0.4641	0.6512	0.0065	-0.0839	0.6565	96.1	0.26	0.04
0.3	0.4941	0.4445	0.0044	-0.0539	0.4477	95.8	0.32	0.04
0.4	0.5111	0.3693	0.0037	-0.0370	0.3712	95.8	0.38	0.04
0.5	0.5228	0.3230	0.0032	-0.0253	0.3240	95.7	0.44	0.04
0.6	0.5289	0.2935	0.0029	-0.0192	0.2941	95.6	0.50	0.04
0.7	0.5317	0.2692	0.0027	-0.0163	0.2696	95.3	0.54	0.05
0.8	0.5348	0.2485	0.0025	-0.0132	0.2488	95.3	0.59	0.05
0.9	0.5364	0.2341	0.0023	-0.0116	0.2344	95.3	0.64	0.05
1.0	0.5369	0.2215	0.0022	-0.0111	0.2218	95.2	0.67	0.05
Var=3%								
0.1	0.4478	1.1706	0.0117	-0.1002	1.1748	96.0	0.23	0.04

0.2	0.5007	0.4405	0.0044	-0.0473	0.4430	96.0	0.33	0.04
0.3	0.5200	0.3428	0.0034	-0.0281	0.3439	95.7	0.41	0.04
0.4	0.5273	0.2908	0.0029	-0.0207	0.2915	95.7	0.49	0.04
0.5	0.5316	0.2598	0.0026	-0.0165	0.2603	95.3	0.57	0.05
0.6	0.5339	0.2382	0.0024	-0.0141	0.2386	95.4	0.63	0.05
0.7	0.5368	0.2188	0.0022	-0.0112	0.2191	95.2	0.69	0.05
0.8	0.5374	0.2034	0.0020	-0.0106	0.2037	95.3	0.74	0.05
0.9	0.5404	0.1907	0.0019	-0.0076	0.1909	95.1	0.78	0.05
1.0	0.5417	0.1806	0.0018	-0.0063	0.1807	95.1	0.82	0.05

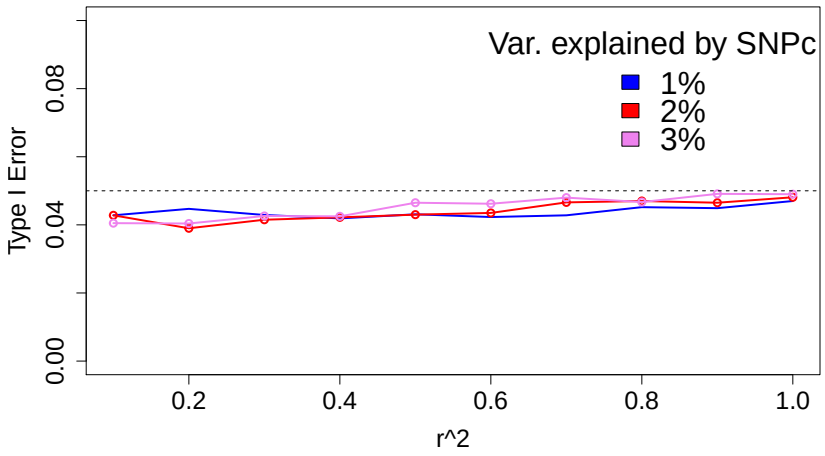
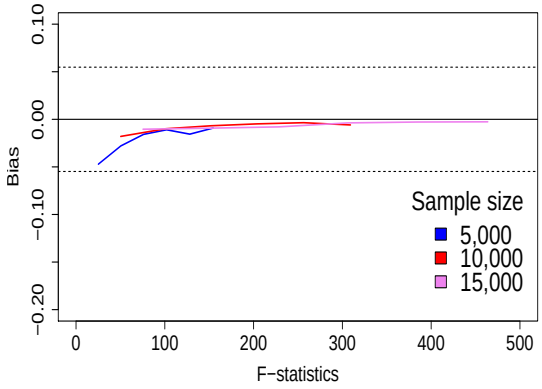
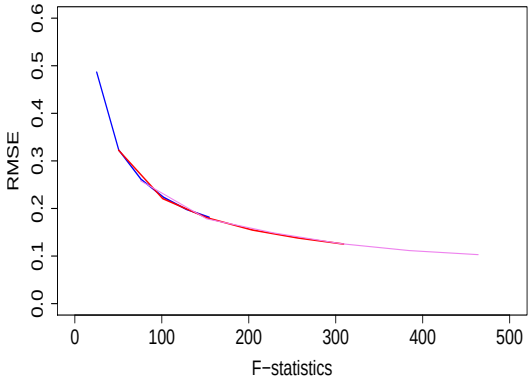


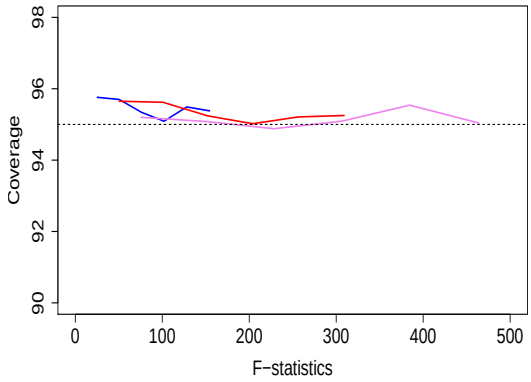
Figure B.3: Type I error of 2SLS with SNP_1 as instrument against correlation (r^2) between SNP_c and SNP_1 for when SNP_c explains 1%, 2% and 3% variation in X, shown by the blue, red and violet coloured lines accordingly. The dotted line is the 5% significance level.



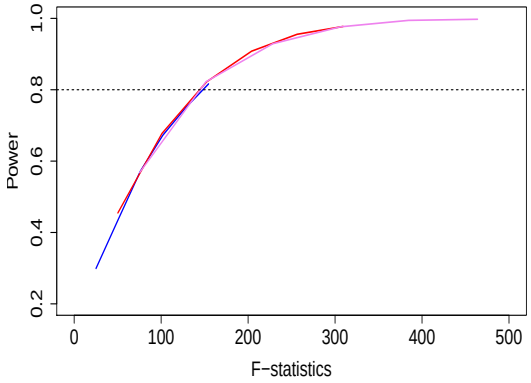
(a) Bias



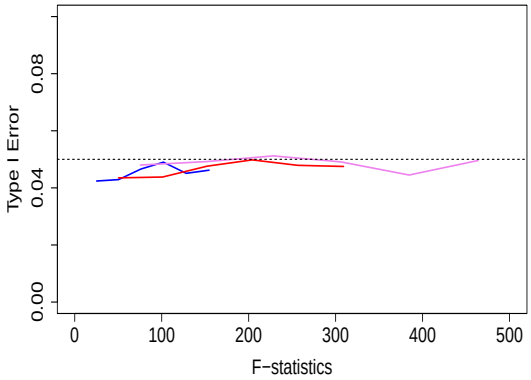
(b) RMSE



(c) Coverage



(d) Power



(e) TIE

Figure B.2: Evaluation criteria of 2SLS with SNP_c as instrument against the expected F-statistics with different sample size. The blue, red and violet coloured lines are sample sizes of 5,000, 10,000 and 15,000 respectively. The black solid line in Bias is labelling zero bias. The black dotted lines in Bias, Coverage, Power and TIE is the 10% bias, 95% nominal coverage, 0.8 power and 5% significance level respectively.

Table B.3: Evaluation criteria of 2SLS with SNP_c and SNP_1 as instruments. r^2 is the correlation between SNP_1 and SNP_c. $\hat{\beta}_{XY}$ is the estimated β_{XY}. The true β_{XY} is 0.5480. S.D. is the standard deviation of mean of $\hat{\beta}_{XY}$. S.E. is the standard error from the standard deviation.

r^2	Mean $\hat{\beta}_{XY}$	S.D.	S.E.	Bias	RMSE	Coverage	Power	Type I Error
SNP_c only	0.5390	0.2206	0.0022	-0.0091	0.2208	95.0	0.68	0.05
0.9	0.5500	0.2187	0.0022	0.0020	0.2187	94.8	0.70	0.05
0.8	0.5498	0.2189	0.0022	0.0018	0.2189	94.7	0.70	0.05
0.7	0.5498	0.2189	0.0022	0.0018	0.2189	94.8	0.70	0.05
0.6	0.5500	0.2189	0.0022	0.0019	0.2189	94.8	0.70	0.05
0.5	0.5500	0.2189	0.0022	0.0020	0.2189	94.8	0.70	0.05
0.4	0.5502	0.2189	0.0022	0.0022	0.2189	94.7	0.70	0.05
0.3	0.5502	0.2189	0.0022	0.0022	0.2189	94.8	0.70	0.05
0.2	0.5504	0.2190	0.0022	0.0024	0.2190	94.9	0.70	0.05
0.1	0.5509	0.2184	0.0022	0.0029	0.2184	94.7	0.70	0.05

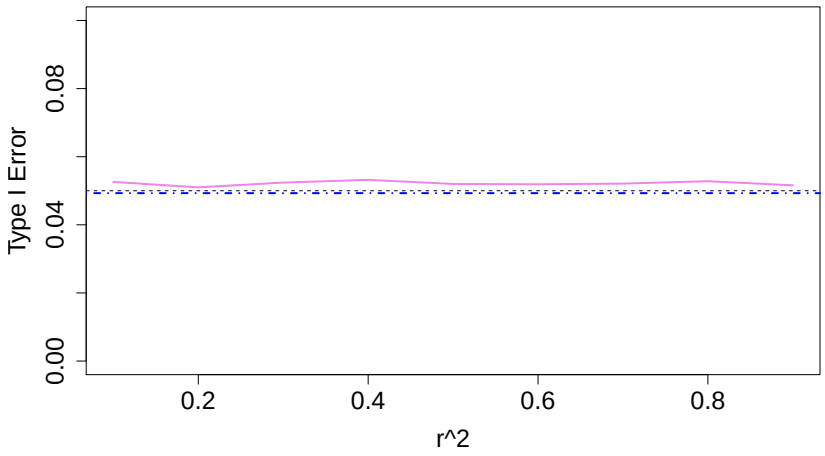


Figure B.4: Evaluation criteria of 2SLS with instruments of SNP_c and SNP_1, and 2SLS with only SNP_c, against the correlation (r^2) between SNP_c and SNP_1. 2SLS with only SNP_c is represented by the blue dashed and 2SLS with SNP_c and SNP_1 is the violet line. The dotted line is the 5% significance level.

Appendix C

Multiple Dependent Instruments in 2SLS

Table C.1: Evaluation criteria to measure the performance of 2SLS for different strengths of LD, maximum ρ, based on selecting the best 1,2,...6 SNPs and using them jointly in a 2SLS MR. $\hat{\beta}_{XY}$ is the estimated β_{XY}. The true β_{XY} is 0.5480. S.D. is the standard deviation of mean of $\hat{\beta}_{XY}$. S.E. is the standard error from the standard deviation.

	Mean $\hat{\beta}_{XY}$	S.D.	S.E.	Median Bias	Bias	RMSE	Coverage	Power	Type I error
$\rho = 0.1$									
1	1.1304	1.1355	0.0114	0.5824	0.5787	1.2761	91.45	0.25	0.09
2	1.1430	0.9070	0.0091	0.5950	0.6101	1.0846	86.23	0.36	0.14
3	1.1547	0.8241	0.0082	0.6067	0.6289	1.0233	83.83	0.41	0.17
4	1.1675	0.7831	0.0078	0.6195	0.6355	0.9985	82.50	0.44	0.18
5	1.1731	0.7665	0.0077	0.6251	0.6425	0.9890	81.73	0.45	0.19
6	1.1787	0.7582	0.0076	0.6307	0.6460	0.9862	81.53	0.46	0.19
$\rho = 0.2$									
1	0.8379	0.8585	0.0086	0.2899	0.3192	0.9061	91.34	0.27	0.08
2	0.8247	0.6977	0.0070	0.2767	0.3114	0.7505	89.67	0.34	0.10
3	0.8563	0.6401	0.0064	0.3083	0.3408	0.7104	88.49	0.38	0.12
4	0.8812	0.6123	0.0061	0.3332	0.3658	0.6971	87.34	0.41	0.13

5	0.9037	0.5975	0.0060	0.3557	0.3859	0.6953	86.45	0.43	0.14
6	0.9199	0.5884	0.0059	0.3719	0.4004	0.6961	85.68	0.44	0.15
$\rho = 0.3$									
1	0.7026	0.6458	0.0065	0.1546	0.1921	0.6640	92.37	0.30	0.07
2	0.6619	0.5447	0.0054	0.1139	0.1503	0.5565	92.56	0.34	0.08
3	0.6915	0.5062	0.0051	0.1434	0.1796	0.5261	91.85	0.38	0.09
4	0.7239	0.4850	0.0049	0.1759	0.2063	0.5159	90.90	0.41	0.10
5	0.7521	0.4738	0.0047	0.2041	0.2327	0.5159	89.94	0.44	0.11
6	0.7772	0.4669	0.0047	0.2292	0.2580	0.5201	88.88	0.46	0.11
$\rho = 0.4$									
1	0.6399	0.4852	0.0049	0.0919	0.1229	0.4938	93.36	0.35	0.07
2	0.5912	0.4278	0.0043	0.0432	0.0735	0.4300	94.06	0.38	0.06
3	0.6156	0.4090	0.0041	0.0675	0.0969	0.4145	93.32	0.41	0.06
4	0.6389	0.4004	0.0040	0.0909	0.1215	0.4106	92.68	0.44	0.07
5	0.6654	0.3910	0.0039	0.1174	0.1433	0.4082	91.67	0.46	0.08
6	0.6952	0.3823	0.0038	0.1472	0.1709	0.4096	90.48	0.50	0.09
$\rho = 0.5$									
1	0.6368	0.4136	0.0041	0.0888	0.1165	0.4230	93.02	0.41	0.06
2	0.5925	0.3721	0.0037	0.0445	0.0687	0.3747	93.85	0.44	0.06
3	0.6012	0.3579	0.0036	0.0532	0.0795	0.3618	93.57	0.47	0.06
4	0.6168	0.3518	0.0035	0.0688	0.0937	0.3584	92.92	0.49	0.06
5	0.6356	0.3465	0.0035	0.0875	0.1109	0.3574	92.29	0.51	0.07
6	0.6639	0.3394	0.0034	0.1159	0.1398	0.3586	91.07	0.55	0.09
$\rho = 0.6$									
1	0.6118	0.3539	0.0035	0.0638	0.0825	0.3596	93.85	0.45	0.06
2	0.5781	0.3213	0.0032	0.0301	0.0492	0.3227	94.69	0.49	0.06
3	0.5860	0.3117	0.0031	0.0379	0.0595	0.3139	94.46	0.51	0.06
4	0.5994	0.3064	0.0031	0.0514	0.0720	0.3107	94.14	0.53	0.06
5	0.6147	0.3029	0.0030	0.0667	0.0859	0.3102	93.51	0.56	0.07
6	0.6348	0.2984	0.0030	0.0868	0.1055	0.3107	92.86	0.59	0.08
$\rho = 0.7$									

1	0.5983	0.3107	0.0031	0.0502	0.0710	0.3147	94.13	0.52	0.05
2	0.5730	0.2888	0.0029	0.0249	0.0459	0.2899	94.74	0.55	0.05
3	0.5789	0.2825	0.0028	0.0309	0.0516	0.2842	94.45	0.57	0.05
4	0.5889	0.2795	0.0028	0.0409	0.0598	0.2824	94.25	0.58	0.06
5	0.6010	0.2766	0.0028	0.0529	0.0702	0.2816	93.80	0.60	0.06
6	0.6185	0.2722	0.0027	0.0705	0.0894	0.2812	93.15	0.63	0.07
$\rho = 0.8$									
1	0.5916	0.2766	0.0028	0.0435	0.0566	0.2800	94.01	0.58	0.06
2	0.5774	0.2623	0.0026	0.0294	0.0413	0.2640	94.40	0.60	0.06
3	0.5766	0.2577	0.0026	0.0286	0.0398	0.2593	94.38	0.62	0.05
4	0.5825	0.2558	0.0026	0.0345	0.0457	0.2581	94.07	0.63	0.06
5	0.5921	0.2542	0.0025	0.0441	0.0563	0.2580	93.74	0.65	0.06
6	0.6062	0.2513	0.0025	0.0582	0.0695	0.2579	93.10	0.67	0.07
$\rho = 0.9$									
1	0.5801	0.2325	0.0023	0.0321	0.0434	0.2347	94.55	0.68	0.06
2	0.5739	0.2298	0.0023	0.0259	0.0370	0.2312	94.68	0.69	0.06
3	0.5722	0.2291	0.0023	0.0242	0.0355	0.2304	94.58	0.69	0.06
4	0.5741	0.2286	0.0023	0.0261	0.0372	0.2300	94.53	0.69	0.06
5	0.5816	0.2273	0.0023	0.0335	0.0443	0.2297	94.19	0.70	0.06
6	0.5958	0.2257	0.0023	0.0478	0.0573	0.2307	93.59	0.72	0.06

Table C.2: Evaluation criteria of 2SLS with GENOME simulated SNPs, based on the selection of 1,2,...6 lowest p-valued SNPs and applying them jointly. $\hat{\beta}_{XY}$ is the estimated β_{XY}. The true β_{XY} is 0.5480. S.D. is the standard deviation of mean of $\hat{\beta}_{XY}$. S.E. is the standard error from the standard deviation.

	Mean $\hat{\beta}_{XY}$	S.D.	S.E.	Bias	Median Bias	RMSE	Coverage	Power	Type I Error
Gene 1									
1	0.5182	0.4886	0.0049	-0.0298	0.0156	0.4895	95.35	0.31	0.05
2	0.5167	0.4881	0.0049	-0.0313	0.0150	0.4891	95.40	0.30	0.05

3	0.5590	0.4671	0.0047	0.0109	0.0563	0.4672	94.59	0.34	0.05
4	0.5811	0.4563	0.0046	0.0331	0.0773	0.4574	94.12	0.36	0.06
5	0.5780	0.4558	0.0046	0.0300	0.0758	0.4568	94.23	0.35	0.06
6	0.6003	0.4482	0.0045	0.0523	0.0957	0.4512	93.80	0.37	0.06
Gene 2									
1	0.5406	0.2492	0.0025	-0.0074	0.0076	0.2493	95.23	0.61	0.05
2	0.5392	0.2491	0.0025	-0.0088	0.0058	0.2493	95.23	0.61	0.05
3	0.5493	0.2470	0.0025	0.0013	0.0149	0.2470	94.90	0.62	0.05
4	0.5624	0.2440	0.0024	0.0143	0.0281	0.2444	94.42	0.64	0.05
5	0.5627	0.2438	0.0024	0.0147	0.0282	0.2443	94.43	0.64	0.05
6	0.5759	0.2406	0.0024	0.0279	0.0391	0.2422	94.15	0.67	0.06
Gene 3									
1	0.5366	0.2186	0.0022	-0.0114	-0.0021	0.2189	95.63	0.67	0.05
2	0.5366	0.2186	0.0022	-0.0114	-0.0021	0.2189	95.63	0.67	0.05
3	0.5366	0.2186	0.0022	-0.0114	-0.0021	0.2189	95.63	0.67	0.05
4	0.5478	0.2164	0.0022	-0.0002	0.0087	0.2164	95.49	0.69	0.05
5	0.5486	0.2162	0.0022	0.0006	0.0095	0.2162	95.48	0.69	0.05
6	0.5586	0.2142	0.0021	0.0105	0.0198	0.2145	95.27	0.71	0.05
Gene 4									
1	0.5386	0.2240	0.0022	-0.0094	0.0013	0.2242	95.00	0.68	0.05
2	0.5386	0.2240	0.0022	-0.0094	0.0013	0.2242	95.00	0.68	0.05
3	0.5386	0.2240	0.0022	-0.0094	0.0013	0.2242	95.00	0.68	0.05
4	0.5521	0.2212	0.0022	0.0041	0.0139	0.2213	94.70	0.70	0.05
5	0.5584	0.2198	0.0022	0.0103	0.0197	0.2200	94.57	0.71	0.05
6	0.5712	0.2173	0.0022	0.0232	0.0320	0.2185	94.23	0.73	0.06
Gene 5									
1	0.6056	0.5818	0.0058	0.0576	0.1026	0.5846	93.63	0.29	0.06
2	0.5881	0.5572	0.0056	0.0401	0.0894	0.5586	93.83	0.29	0.06
3	0.5956	0.5527	0.0055	0.0476	0.0989	0.5547	93.72	0.30	0.06
4	0.6047	0.5451	0.0055	0.0566	0.1054	0.5480	93.45	0.31	0.06
5	0.6508	0.5185	0.0052	0.1027	0.1509	0.5286	92.66	0.35	0.07

6	0.6477	0.4503	0.0045	0.0997	0.1373	0.4611	92.22	0.41	0.07

Table C.3: Evaluation criteria of best policy for 6 non-causal SNPs simulated from GENOME with different sample sizes. $\hat{\beta}_{XY}$ is the estimated β_{XY}. The true β_{XY} is 0.5480. S.D. is the standard deviation of mean of $\hat{\beta}_{XY}$. S.E. is the standard error from the standard deviation.

Policy	Mean $\hat{\beta}_{XY}$ (S.D.)	S.E.	Bias	Median bias	RMSE	Coverage	Power	Type I error
n=5,000								
Lowest p-value	0.5813 (0.44)	0.0044	0.0332	0.0197	0.4375	94.48	0.54	0.06
All	0.6076 (0.33)	0.0033	0.0595	0.0504	0.3349	93.15	0.65	0.07
p-value<0.05	0.6061 (0.34)	0.0035	0.0580	0.0441	0.3495	93.35	0.63	0.07
F-statistic>10	0.5797 (0.30)	0.0031	0.0317	0.0308	0.2968	93.82	0.63	0.06
n=10,000								
Lowest p-value	0.5681 (0.32)	0.0032	0.0201	0.0137	0.3221	95.07	0.74	0.05
All	0.5825 (0.24)	0.0024	0.0345	0.0314	0.2432	94.16	0.82	0.06
p-value<0.05	0.5849 (0.25)	0.0025	0.0369	0.0291	0.2492	94.28	0.81	0.06
F-statistic>10	0.5757 (0.24)	0.0024	0.0277	0.0241	0.2366	94.46	0.82	0.06
n=20,000								
Lowest p-value	0.5615 (0.24)	0.0024	0.0135	0.0069	0.2357	94.62	0.85	0.05
All	0.5639 (0.17)	0.0017	0.0159	0.0152	0.1739	94.49	0.92	0.06

p-value<0.05	0.5679 (0.18)	0.0018	0.0199	0.0161	0.1849	94.34	0.91	0.06
F-statistic>10	0.5665 (0.19)	0.0019	0.0185	0.0151	0.1915	94.36	0.91	0.06
n=30,000								
Lowest p-value	0.5553 (0.20)	0.0020	0.0073	0.0032	0.1968	94.35	0.89	0.06
All	0.5568 (0.14)	0.0017	0.0014	0.0089	0.1439	94.49	0.95	0.06
p-value<0.05	0.5597 (0.15)	0.0015	0.0117	0.0094	0.1506	94.38	0.95	0.06
F-statistic>10	0.5612 (0.16)	0.0016	0.0131	0.0090	0.1576	94.24	0.94	0.06
n=40,000								
Lowest p-value	0.5526 (0.17)	0.0017	0.0046	0.0024	0.1657	94.96	0.91	0.05
All	0.5546 (0.13)	0.0013	0.0066	0.0061	0.1259	94.90	0.96	0.05
p-value<0.05	0.5563 (0.13)	0.0013	0.0083	0.0061	0.1284	94.89	0.96	0.05
F-statistic>10	0.5584 (0.13)	0.0013	0.0104	0.0063	0.1338	94.90	0.95	0.05
n=50,000								
Lowest p-value	0.5527 (0.15)	0.0015	0.0047	0.0027	0.1525	94.62	0.93	0.05
All	0.5546 (0.11)	0.0011	0.0066	0.0062	0.1111	94.72	0.97	0.05
p-value<0.05	0.5560 (0.11)	0.0011	0.0079	0.0066	0.1116	94.72	0.97	0.05
F-statistic> 10	0.5578 (0.12)	0.0012	0.0098	0.0066	0.1183	94.60	0.97	0.05

Table C.4: Evaluation criteria of best policy for 20 non-causal SNPs simulated from GENOME with different sample sizes. $\hat{\beta}_{XY}$ is the estimated β_{XY}. The true β_{XY} is 0.5480. S.D. is the standard deviation of mean of $\hat{\beta}_{XY}$. S.E. is the standard error from the standard deviation.

Policy	Mean $\hat{\beta}_{XY}$ (S.D.)	S.E.	Bias	Median bias	RMSE	Coverage	Power	Type I error
n=5,000								
Lowest p-value	0.5641 (0.30)	0.0030	0.0161	0.0215	0.3042	94.36	0.61	0.06
All	0.6713 (0.22)	0.0022	0.1233	0.1251	0.2559	87.99	0.84	0.12
p-value<0.05	0.6378 (0.25)	0.0025	0.0897	0.0903	0.2695	90.57	0.77	0.09
F-statistic>10	0.6088 (0.26)	0.0026	0.0608	0.0659	0.2629	92.23	0.72	0.08
n=10,000								
Lowest p-value	0.5558 (0.22)	0.0022	0.0078	0.0074	0.2158	95.36	0.82	0.05
All	0.6157 (0.16)	0.0016	0.0677	0.0687	0.1775	91.62	0.95	0.08
p-value<0.05	0.6044 (0.18)	0.0018	0.0564	0.0554	0.1888	92.62	0.92	0.07
F-statistic>10	0.5919 (0.19)	0.0019	0.0439	0.0448	0.1930	93.34	0.90	0.07
n=20,000								
Lowest p-value	0.5488 (0.15)	0.0015	0.0008	0.0031	0.1524	94.81	0.93	0.05
All	0.5796 (0.12)	0.0012	0.0316	0.0330	0.1230	93.26	0.99	0.07
p-value<0.05	0.5775 (0.13)	0.0013	0.0295	0.0302	0.1286	93.51	0.98	0.06

F-statistic>10	0.5738 (0.13)	0.0013	0.0258	0.0272	0.1337	93.72	0.98	0.06
n=30,000								
Lowest p-value	0.5495 (0.13)	0.0013	0.0014	0.0034	0.1264	94.96	0.96	0.05
All	0.5710 (0.10)	0.0010	0.0229	0.0240	0.1013	93.58	1.00	0.06
p-value<0.05	0.5699 (0.10)	0.0010	0.0219	0.0227	0.1048	93.65	0.99	0.06
F-statistic>10	0.5683 (0.11)	0.0011	0.0203	0.0211	0.1078	93.86	0.99	0.06
n=40,000								
Lowest p-value	0.5483 (0.11)	0.0011	0.0003	0.0010	0.1082	95.06	0.97	0.05
All	0.5645 (0.08)	0.0008	0.0165	0.0174	0.0849	94.32	1.00	0.06
p-value<0.05	0.5639 (0.09)	0.0009	0.0158	0.0163	0.0867	94.36	1.00	0.06
F-statistic>10	0.5631 (0.09)	0.0009	0.0151	0.0158	0.0889	94.45	0.99	0.06
n=50,000								
Lowest p-value	0.5481 (0.10)	0.0010	0.0001	0.0019	0.0952	95.28	0.98	0.05
All	0.5616 (0.07)	0.0007	0.0136	0.0144	0.0761	94.20	1.00	0.06
p-value<0.05	0.5615 (0.08)	0.0008	0.0135	0.0140	0.0771	94.18	1.00	0.06
F-statistic>10	0.5610 (0.08)	0.0008	0.0130	0.0132	0.0792	94.47	1.00	0.06

Appendix D

A Comparison of Estimators

Table D.1: Evaluation criteria for including SNPs with the same common MAF in 2SLS, two-step GMM, CUE and LIML. $\hat{\beta}_{XY}$ is the estimated β_{XY}. The true β_{XY} is 0.2449. S.D. is the standard deviation of mean of $\hat{\beta}_{XY}$. Winsor. is Winsorised. Winsorised S.E. is the standard error from the standard deviation of Winsorised $\hat{\beta}_{XY}$.

	Mean $\hat{\beta}_{XY}$ (S.D.)	Winsor. S.E.	Winsor. Bias	Winsor. RMSE	Prop. of Outliers	Coverage	TIE	Power	Non-Convergence
10 Instruments									
2SLS	0.3475 (0.1474)	0.0009	0.1074	0.1405	0.0134	85.72	0.1427	0.6509	0
GMM	0.3475 (0.1480)	0.0009	0.1069	0.1404	0.0138	85.33	0.1467	0.6545	0
CUE	0.2218 (0.2480)	0.0012	-0.0064	0.1249	0.0252	89.01	0.1102	0.4258	6
LIML	0.2244 (0.2381)	0.0012	-0.0071	0.1250	0.0243	93.76	0.0622	0.3603	0
30 Instruments									
2SLS	0.4628 (0.1121)	0.0007	0.2193	0.2302	0.0118	48.91	0.5104	0.9626	0
GMM	0.4628 (0.1137)	0.0007	0.2191	0.2304	0.0114	47.94	0.5206	0.9620	0
CUE	0.2175 (0.7173)	0.0014	-0.0120	0.1422	0.0322	73.01	0.2701	0.5396	15
LIML	0.2514 (3.6539)	0.0014	-0.0113	0.1396	0.0297	90.24	0.0974	0.3958	0
60 Instruments									
2SLS	0.5483 (0.0921)	0.0006	0.3039	0.3095	0.0086	10.08	0.8966	0.9998	0
GMM	0.5484 (0.0943)	0.0006	0.3037	0.3097	0.0089	9.85	0.8968	0.9998	0
CUE	-377 (69984)	0.0017	-0.0105	0.1711	0.0475	54.02	0.4499	0.6503	30
LIML	0.1996 (1.0284)	0.0016	-0.0124	0.1649	0.0444	84.37	0.1545	0.4241	0
90 Instruments									
2SLS	0.5925 (0.0786)	0.0005	0.3477	0.3512	0.0090	1.05	0.9890	1.0000	0
GMM	0.5923 (0.0811)	0.0005	0.3474	0.3512	0.0079	1.05	0.9891	1.0000	0
CUE	-1099 (140719)	0.0020	-0.0146	0.1959	0.0582	42.09	0.5669	0.7062	56
LIML	0.3009 (9.7116)	0.0018	-0.0159	0.1855	0.0600	80.33	0.1911	0.4263	0

Table D.2: Evaluation criteria for including SNPs with the variable MAF in 2SLS, two-step GMM, CUE and LIML. $\hat{\beta}_{XY}$ is the estimated β_{XY}. The true β_{XY} is 0.2449. S.D. is the standard deviation of mean of $\hat{\beta}_{XY}$. Winsor. is Winsorised. Winsorised S.E. is the standard error from the standard deviation of Winsorised $\hat{\beta}_{XY}$.

	Mean $\hat{\beta}_{XY}$ (S.D.)	Winsor. S.E.	Winsor. Bias	Winsor. RMSE	Prop. of Outliers	Coverage	TIE	Power	Non-Convergence
10 Instruments									
2SLS	0.3536 (0.1562)	0.0010	0.1130	0.1491	0.0135	85.10	0.1486	0.6286	0
GMM	0.3539 (0.1568)	0.0010	0.1133	0.1495	0.0137	84.88	0.1512	0.6334	0
CUE	-51.86 (5205)	0.0014	-0.0127	0.1364	0.0262	88.09	0.1192	0.4021	8
LIML	0.2544 (3.9017)	0.0014	-0.0128	0.1359	0.0272	93.52	0.0648	0.3356	0
30 Instruments									
2SLS	0.4726 (0.1146)	0.0007	0.2287	0.2400	0.0091	46.69	0.5330	0.9633	0
GMM	0.4727 (0.1160)	0.0007	0.2291	0.2407	0.0089	45.48	0.5452	0.9620	0
CUE	1109 (98241)	0.0015	-0.0083	0.1490	0.0368	70.95	0.2906	0.5431	8
LIML	0.2188 (0.5648)	0.0015	-0.0081	0.1463	0.0369	89.50	0.1049	0.3914	0
60 Instruments									
2SLS	0.5517 (0.0926)	0.0006	0.3069	0.3125	0.0084	9.59	0.9041	0.9994	0
GMM	0.5513 (0.0951)	0.0006	0.3062	0.3122	0.0085	9.36	0.9064	0.9991	0
CUE	-189.84 (22501)	0.0018	-0.0172	0.1795	0.0478	52.43	0.4760	0.6415	31
LIML	0.2251 (17.0914)	0.0017	-0.0183	0.1724	0.0457	83.95	0.1601	0.4061	0
90 Instruments									
2SLS	0.5936 (0.0787)	0.0005	0.3488	0.3524	0.0074	1.00	0.9900	1.0000	0
GMM	0.5936 (0.0814)	0.0005	0.3489	0.3527	0.0077	1.04	0.9896	1.0000	0
CUE	-1757 (95100)	0.0020	-0.0071	0.1988	0.0595	41.97	0.5808	0.7190	51
LIML	0.1658 (2.9157)	0.0019	-0.0097	0.1859	0.0615	80.03	0.1994	0.4362	0

Table D.3: Evaluation criteria for including SNPs with the same low MAF in 2SLS, two-step GMM, CUE and LIML. $\hat{\beta}_{XY}$ is the estimated β_{XY}. The true β_{XY} is 0.2449. S.D. is the standard deviation of mean of $\hat{\beta}_{XY}$. Winsor. is Winsorised. Winsorised S.E. is the standard error from the standard deviation of Winsorised $\hat{\beta}_{XY}$.

	Mean $\hat{\beta}$ (S.D.)	Winsor. S.E.	Winsor. Bias	Winsor. RMSE	Prop. of Outliers	Coverage	TIE	Power	Non-Convergence
10 Instruments									
2SLS	0.3754 (0.1681)	0.0010	0.1355	0.1698	0.0150	83.55	0.1645	0.6319	0
GMM	0.3759 (0.1695)	0.0010	0.1362	0.1712	0.0147	82.78	0.1722	0.6369	0
CUE	-334.37 (33310)	0.0015	-0.0101	0.1544	0.0358	86.03	0.1420	0.4038	88
LIML	0.2102 (0.7452)	0.0015	-0.0125	0.1538	0.0342	93.53	0.0645	0.3151	0
30 Instruments									
2SLS	0.5058 (0.1215)	0.0008	0.2616	0.2722	0.0130	41.07	0.5891	0.9681	0
GMM	0.5063 (0.1249)	0.0008	0.2622	0.2735	0.0129	39.65	0.6035	0.9681	0
CUE	-1181 (70019)	0.0018	-0.0036	0.1812	0.0586	63.58	0.3686	0.5662	149
LIML	3.2466 (298)	0.0017	-0.0088	0.1747	0.0525	88.66	0.1130	0.3689	0
60 Instruments									
2SLS	0.5812 (0.0947)	0.0006	0.3367	0.3421	0.0102	6.78	0.9321	0.9994	0
GMM	0.5824 (0.0987)	0.0006	0.3378	0.3435	0.0107	6.43	0.9357	0.9995	0
CUE	2188 (152256)	0.0023	0.0028	0.2309	0.0729	42.93	0.5727	0.6948	174
LIML	0.3084 (14.71)	0.0021	-0.0157	0.2140	0.0634	82.73	0.1726	0.3961	0
90 Instruments									
2SLS	0.6185 (0.0815)	0.0005	0.3733	0.3769	0.0101	0.70	0.9930	1.0000	0
GMM	0.6190 (0.0861)	0.0005	0.3740	0.3780	0.0103	0.76	0.9924	1.0000	0
CUE	-1028 (388354)	0.0027	0.0324	0.2635	0.0887	32.72	0.6778	0.7608	259
LIML	0.3354 (16.17)	0.0024	-0.0099	0.2375	0.0865	77.47	0.2250	0.4172	0

Table D.4: Evaluation criteria of 2SLS, GMM, CUE and LIML for Pattern I. $\hat{\beta}_{XY}$ is the estimated β_{XY}. The true β_{XY} is 0.2449. S.D. is the standard deviation of mean of $\hat{\beta}_{XY}$. Winsor. is Winsorised. Winsorised S.E. is the standard error from the standard deviation of Winsorised $\hat{\beta}_{XY}$.

	Mean $\hat{\beta}_{XY}$ (S.D.)	Winsor. S.E.	Winsor. Bias	Winsor. RMSE	Prop. of Outliers	Coverage	TIE	Power	Non-Convergence
10 Instruments									
2SLS	0.4117 (0.1888)	0.0012	0.1710	0.2063	0.0149	80.55	0.1944	0.6157	0
GMM	0.4123 (0.1901)	0.0012	0.1719	0.2076	0.0146	79.86	0.2014	0.6202	0
CUE	0.1739 (1.4218)	0.0019	-0.0160	0.1939	0.0572	81.50	0.1876	0.3858	97
LIML	0.1799 (5.6557)	0.0019	-0.0198	0.1960	0.0560	92.37	0.0763	0.2804	0
30 Instruments									
2SLS	0.5330 (0.1272)	0.0008	0.2885	0.2997	0.0085	36.88	0.6308	0.9699	0
GMM	0.5342 (0.1296)	0.0008	0.2897	0.3012	0.0083	35.57	0.6443	0.9694	0
CUE	646 (148248)	0.0022	-0.0058	0.2217	0.0674	58.00	0.4251	0.5750	138
LIML	0.1265 (3.7936)	0.0022	-0.0155	0.2173	0.0650	86.42	0.1357	0.3543	0
60 Instruments									
2SLS	0.5983 (0.0993)	0.0006	0.3535	0.3591	0.0093	6.12	0.9387	0.9997	0
GMM	0.5993 (0.1028)	0.0007	0.3546	0.3605	0.0074	6.11	0.9389	0.9996	0
CUE	-1937 (474508)	0.0026	-0.0010	0.2619	0.0889	40.28	0.5996	0.6968	171
LIML	0.2553 (7.3573)	0.0025	-0.0192	0.2516	0.0863	80.72	0.1925	0.3886	0
90 Instruments									
2SLS	0.6300 (0.0820)	0.0005	0.3848	0.3884	0.0080	0.49	0.9950	1.0000	0
GMM	0.6307 (0.0855)	0.0005	0.3855	0.3893	0.0066	0.53	0.9947	1.0000	0
CUE	-10495 (417425)	0.0029	0.0244	0.2864	0.0986	31.27	0.6901	0.7629	230
LIML	0.2815 (13.1534)	0.0027	-0.0051	0.2681	0.0982	75.89	0.2408	0.4148	0

Table D.5: Evaluation criteria of 2SLS, GMM, CUE and LIML for Pattern II.$\bar{\hat{\beta}}_{XY}$ is the estimated β_{XY}. The true β_{XY} is 0.2449. S.D. is the standard deviation of mean of $\hat{\beta}_{XY}$. Winsor. is Winsorised. Winsorised S.E. is the standard error from the standard deviation of Winsorised $\hat{\beta}_{XY}$.

	Mean $\hat{\beta}_{XY}$ (S.D.)	Winsor. S.E.	Winsor. Bias	Winsor. RMSE	Prop. of Outliers	Coverage	TIE	Power	Non-Convergence
10 Instruments									
2SLS	0.4951 (0.2263)	0.0013	0.2533	0.2865	0.0218	73.68	0.2631	0.6415	0
GMM	0.4956 (0.2278)	0.0013	0.2539	0.2876	0.0219	72.98	0.2702	0.6456	0
CUE	-1944 (118005)	0.0030	-0.0014	0.2996	0.0996	70.57	0.2970	0.4360	121
LIML	0.3431 (20.99)	0.0030	-0.0078	0.3035	0.0983	89.10	0.1087	0.2716	0
30 Instruments									
2SLS	0.5723 (0.1384)	0.0009	0.3279	0.3396	0.0091	31.81	0.6812	0.9686	0
GMM	0.5737 (0.1413)	0.0009	0.3298	0.3418	0.0086	30.95	0.6905	0.9695	0
CUE	420.7007 (180201)	0.0030	0.0048	0.2942	0.0915	48.73	0.5165	0.6102	152
LIML	0.1863 (9.7469)	0.0029	-0.0135	0.2878	0.0934	83.66	0.1631	0.3381	0
60 Instruments									
2SLS	0.6243 (0.1001)	0.0006	0.3799	0.3851	0.0060	4.70	0.9528	0.9997	0
GMM	0.6252 (0.1032)	0.0007	0.3808	0.3865	0.0067	4.55	0.9545	0.9998	0
CUE	-3721 (403223)	0.0033	0.0306	0.3234	0.1055	34.47	0.6571	0.7233	191
LIML	-0.1104 (25.52)	0.0031	-0.0045	0.3119	0.1070	78.70	0.2128	0.3799	0
90 Instruments									
2SLS	0.6506 (0.0852)	0.0005	0.4061	0.4097	0.0100	0.45	0.9954	1.0000	0
GMM	0.6523 (0.0887)	0.0006	0.4074	0.4112	0.0090	0.50	0.9950	1.0000	0
CUE	2289.4692 (480992)	0.0036	0.0615	0.3592	0.1089	25.38	0.7471	0.8017	291
LIML	-7.0752 (709.5)	0.0035	0.0099	0.3509	0.1196	73.06	0.2691	0.4230	0

Table D.6: Evaluation criteria of 2SLS, GMM, CUE and LIML for Pattern III. $\hat{\beta}_{XY}$ is the estimated β_{XY}. The true β_{XY} is 0.2449. S.D. is the standard deviation of mean of $\hat{\beta}_{XY}$. Winsor. is Winsorised. Winsorised S.E. is the standard error from the standard deviation of Winsorised $\hat{\beta}_{XY}$.

	Mean $\hat{\beta}_{XY}$ (S.D.)	Winsor. S.E.	Winsor. Bias	Winsor. RMSE	Prop. of Outliers	Coverage	TIE	Power	Non-Convergence
10 Instruments									
2SLS	0.4123 (0.1793)	0.0011	0.1709	0.2029	0.0150	78.75	0.2123	0.6522	0
GMM	0.4127 (0.1806)	0.0011	0.1714	0.2036	0.0160	78.12	0.2188	0.6568	0
CUE	880.9 (115535)	0.0019	-0.0139	0.1863	0.0491	81.12	0.1916	0.4052	70
LIML	0.1818 (0.9200)	0.0019	-0.0171	0.1864	0.0479	91.71	0.0829	0.2971	0
30 Instruments									
2SLS	0.5264 (0.1231)	0.0008	0.2817	0.2921	0.0108	36.22	0.6373	0.9725	0
GMM	0.5272 (0.1254)	0.0008	0.2825	0.2931	0.0114	35.26	0.6475	0.9730	0
CUE	3719 (262269)	0.0020	-0.0084	0.2037	0.0664	60.07	0.4035	0.5738	104
LIML	-0.0468 (14.72)	0.0020	-0.0131	0.1985	0.0589	87.40	0.1259	0.3596	0
60 Instruments									
2SLS	0.5912 (0.0963)	0.0006	0.3464	0.3519	0.0078	6.26	0.9373	0.9998	0
GMM	0.5921 (0.0994)	0.0006	0.3474	0.3531	0.0086	6.09	0.9391	0.9997	0
CUE	-526.6 (207070)	0.0024	-0.0027	0.2400	0.0795	42.88	0.5726	0.6842	184
LIML	0.2146 (3.453)	0.0023	-0.0188	0.2301	0.0802	81.88	0.1808	0.3824	0
90 Instruments									
2SLS	0.6256 (0.0820)	0.0005	0.3807	0.3843	0.0088	0.62	0.9938	1.0000	0
GMM	0.6270 (0.0854)	0.0005	0.3820	0.3859	0.0084	0.48	0.9952	1.0000	0
CUE	-1945 (220215)	0.0027	0.0298	0.2686	0.0928	32.76	0.6750	0.7665	230
LIML	0.1520 (7.1606)	0.0025	-0.0035	0.2547	0.0941	76.39	0.2357	0.4215	0

Table D.7: Evaluation criteria of 2SLS, GMM, CUE and LIML for Pattern IV. $\hat{\beta}_{XY}$ is the mean of estimated β_{XY}. The true β_{XY} is 0.2449. S.D. is the standard deviation of mean of $\hat{\beta}_{XY}$. Winsor. is Winsorised. Winsorised S.E. is the standard error from the standard deviation of Winsorised $\hat{\beta}_{XY}$.

	Mean $\hat{\beta}_{XY}$ (S.D.)	Winsor. S.E.	Winsor. Bias	Winsor. RMSE	Prop. of Outliers	Coverage	TIE	Power	Non-Convergence
10 Instruments									
2SLS	0.4458 (0.2015)	0.0012	0.2050	0.2392	0.0172	77.05	0.2291	0.6263	0
GMM	0.4460 (0.2025)	0.0012	0.2054	0.2402	0.0171	76.59	0.2341	0.6307	0
CUE	-941.6 (100417)	0.0023	-0.0115	0.2311	0.0689	77.40	0.2276	0.4046	74
LIML	0.0697 (12.18)	0.0023	-0.0150	0.2313	0.0689	90.55	0.0943	0.2827	0
30 Instruments									
2SLS	0.5450 (0.1318)	0.0008	0.3009	0.3124	0.0088	35.75	0.6419	0.9673	0
GMM	0.5454 (0.1341)	0.0009	0.3015	0.3135	0.0085	34.92	0.6508	0.9693	0
CUE	-154.5 (138617)	0.0024	-0.0072	0.2406	0.0747	55.19	0.4507	0.5789	74
LIML	0.2975 (27.54)	0.0024	-0.0128	0.2367	0.0733	85.36	0.1463	0.3491	0
60 Instruments									
2SLS	0.6073 (0.0974)	0.0006	0.3629	0.3682	0.0074	5.24	0.9475	0.9998	0
GMM	0.6083 (0.1004)	0.0006	0.3639	0.3695	0.0072	5.18	0.9482	0.9997	0
CUE	-1209 (152709)	0.0028	0.0067	0.2791	0.0821	37.30	0.6269	0.7014	157
LIML	0.3010 (14.03)	0.0027	-0.0148	0.2709	0.0859	80.16	0.1983	0.3870	0
90 Instruments									
2SLS	0.6370 (0.0841)	0.0005	0.3924	0.3960	0.0095	0.66	0.9934	1.0000	0
GMM	0.6372 (0.0872)	0.0006	0.3925	0.3964	0.0089	0.65	0.9935	1.0000	0
CUE	5863 (301395)	0.0030	0.0267	0.3010	0.0929	29.02	0.7121	0.7709	234
LIML	-0.5102 (53.42)	0.0030	-0.0024	0.2954	0.1070	74.93	0.2501	0.4183	0

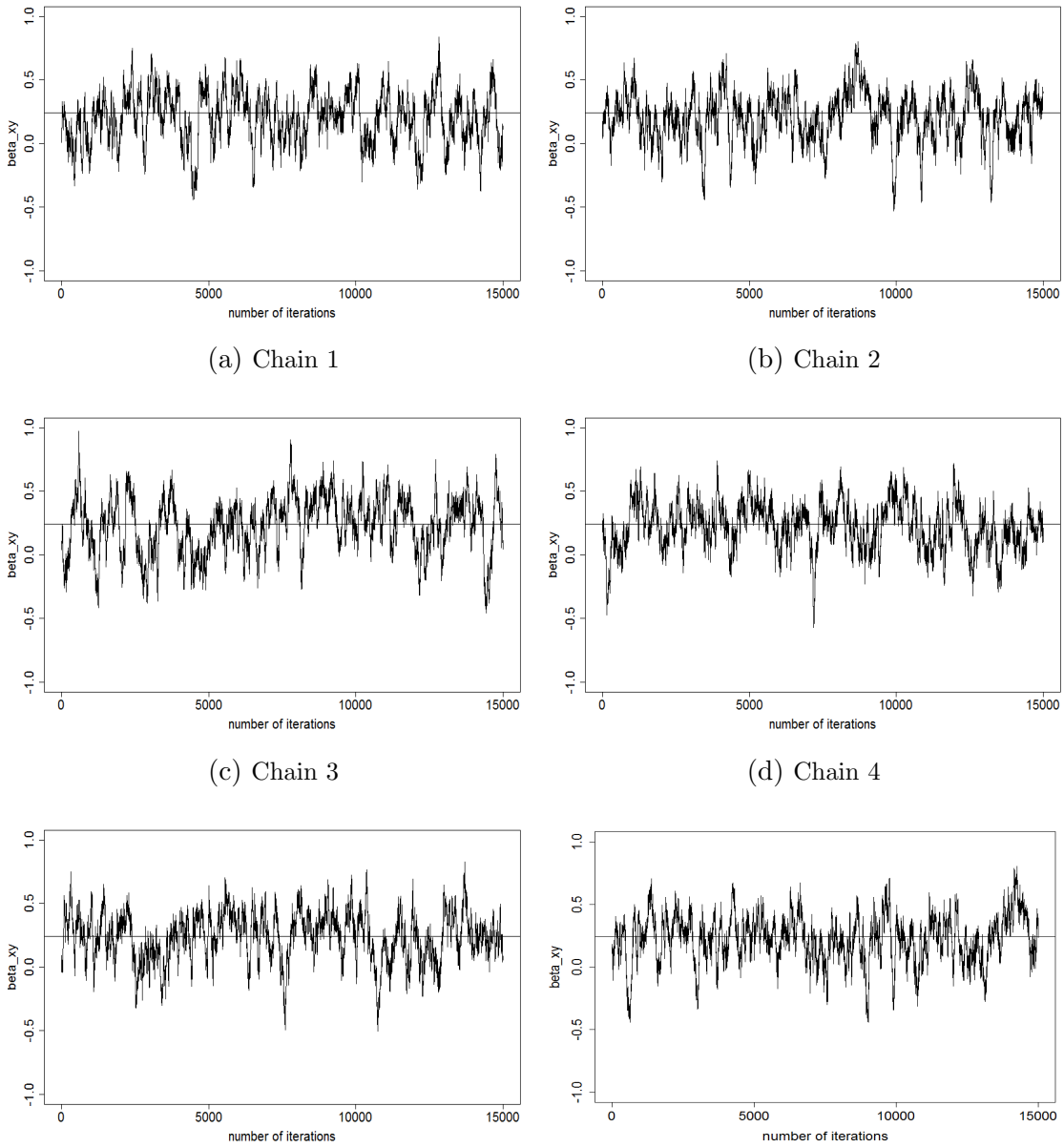
Appendix E

Bayesian Model Averaging

Figure E.1: Model code for `OpenBUGS` to thus compare with *ivbma*

```
model {

  for( i in 1 : N ) {

    mu[i,1]<-g[1]*alpha[1] + g[2]*alpha[2] * SNP1[i]
    + g[3]*alpha[3] * SNP2[i] + g[4]*alpha[4]* SNP3[i]
    mu[i,2] <- k[1]*beta[1] +  k[2]*beta[2]*XY[i,1]

    XY[i,1:2] ~ dmnorm(mu[i,1:2],Sigma.inv[1:2,1:2])
  }

  #priors for first-stage regression
   for (j in 1 : 4) {
    alpha[j] ~  dnorm(0,1);
  }

 #priors for second-stage regression
  for (j in 1 : 2) {
    beta[j] ~  dnorm(0,1);
  }

  #priors for variable indicators
  for (j in 1 : 4) {
```

```
    g[j] ~ dbern(0.5); }

  for (j in 1 : 2) {
    k[j] ~ dbern(0.5); }

  #priors for correlated errors
  Sigma.inv[1:2,1:2] ~ dwish(R[1:2,1:2],3)
  Sigma[1:2,1:2]<-inverse(Sigma.inv[1:2,1:2])

  #Defining Model Code
  mdl <- g[1]*1+g[2]*2+g[3]*4+g[4]*8

  #Defining vector with model indicators
  for (j in 1 : models) {
    pmdl[j]<-equals(mdl , j); }

}
```



(a) Chain 1

(b) Chain 2

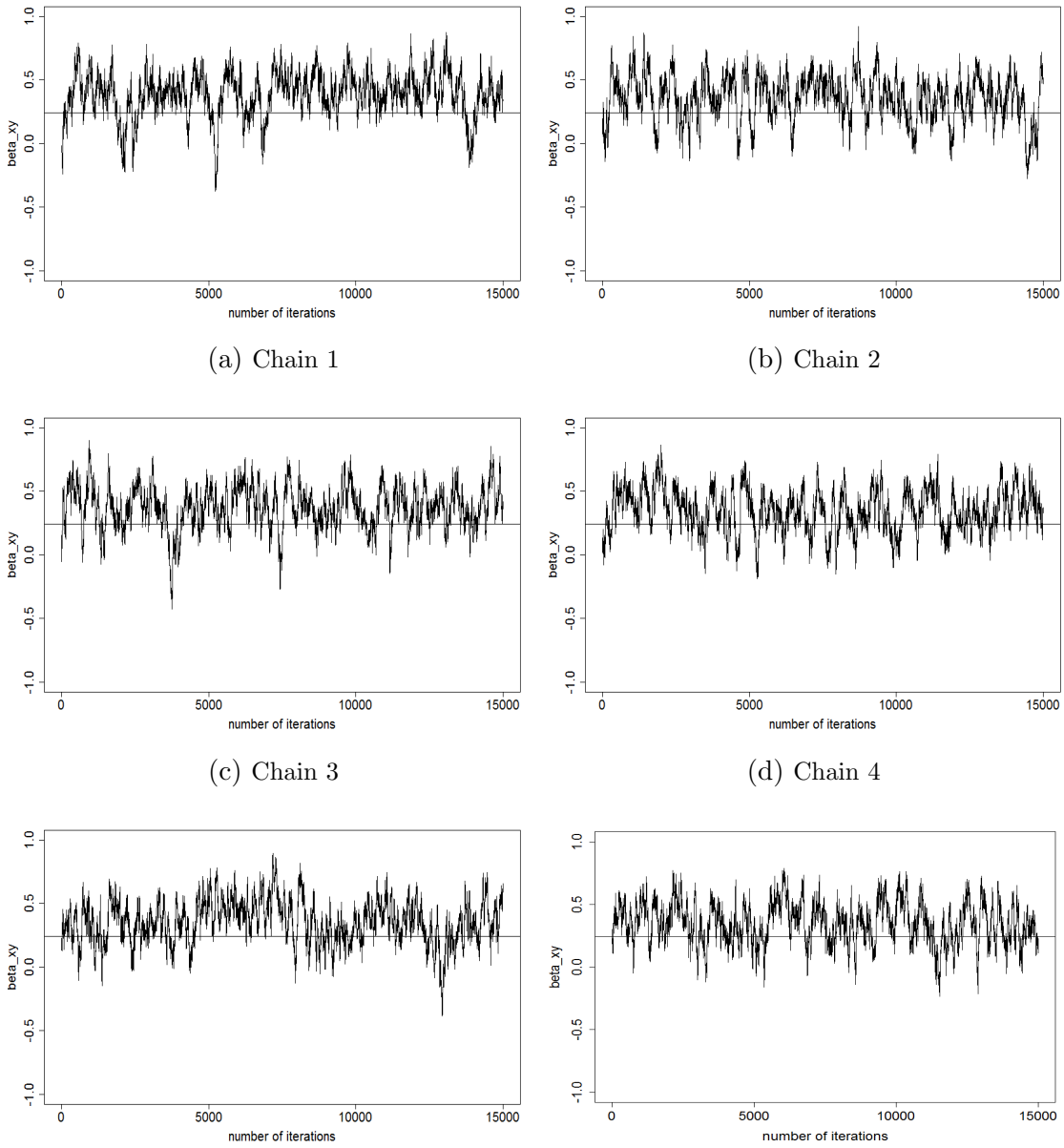
(c) Chain 3

(d) Chain 4

(e) Chain 5

(f) Long

Figure E.2: Trace plot of the causal effect estimate ($\hat{\beta}_{XY}$) from 10 instruments with five short and one long chain. Short and long chain consist of 50,000 and 500,000 iterations with 10,000 and 250,000 burn-in respectively. The horizontal line is the true β_{XY} (0.2449).



(a) Chain 1

(b) Chain 2

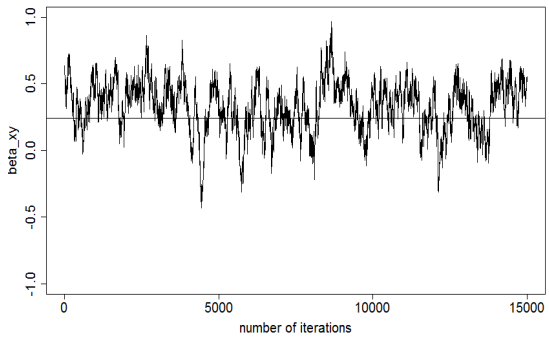
(c) Chain 3

(d) Chain 4

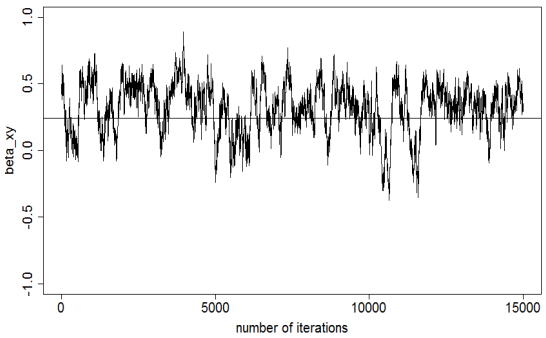
(e) Chain 5

(f) Long

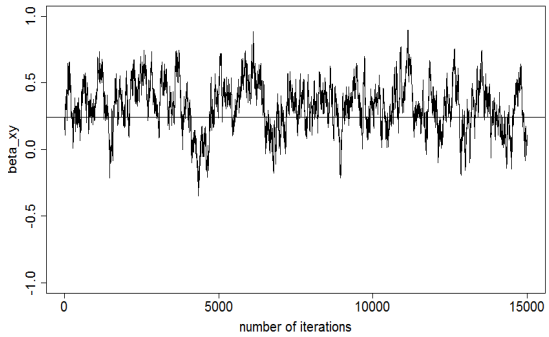
Figure E.3: Trace plot of the causal effect estimate ($\hat{\beta}_{XY}$) from 30 instruments with five short and one long chain. Short and long chain consist of 50,000 and 500,000 iterations with 10,000 and 250,000 burn-in respectively. The horizontal line is the true β_{XY} (0.2449).



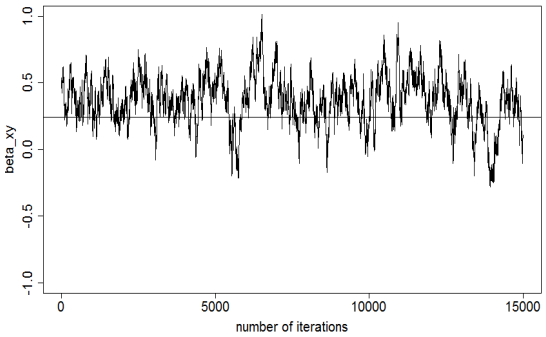
(a) Chain 1



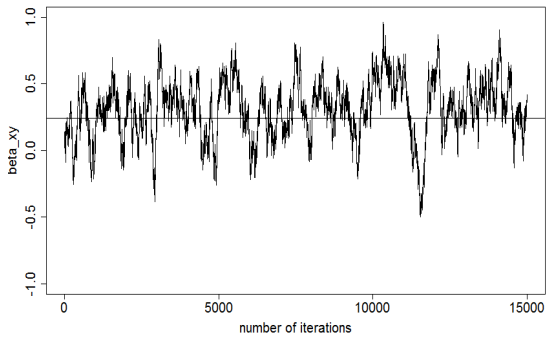
(b) Chain 2



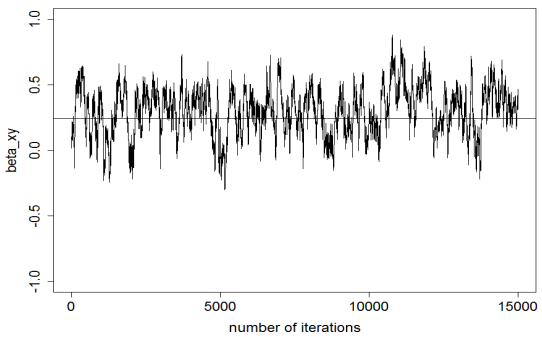
(c) Chain 3



(d) Chain 4

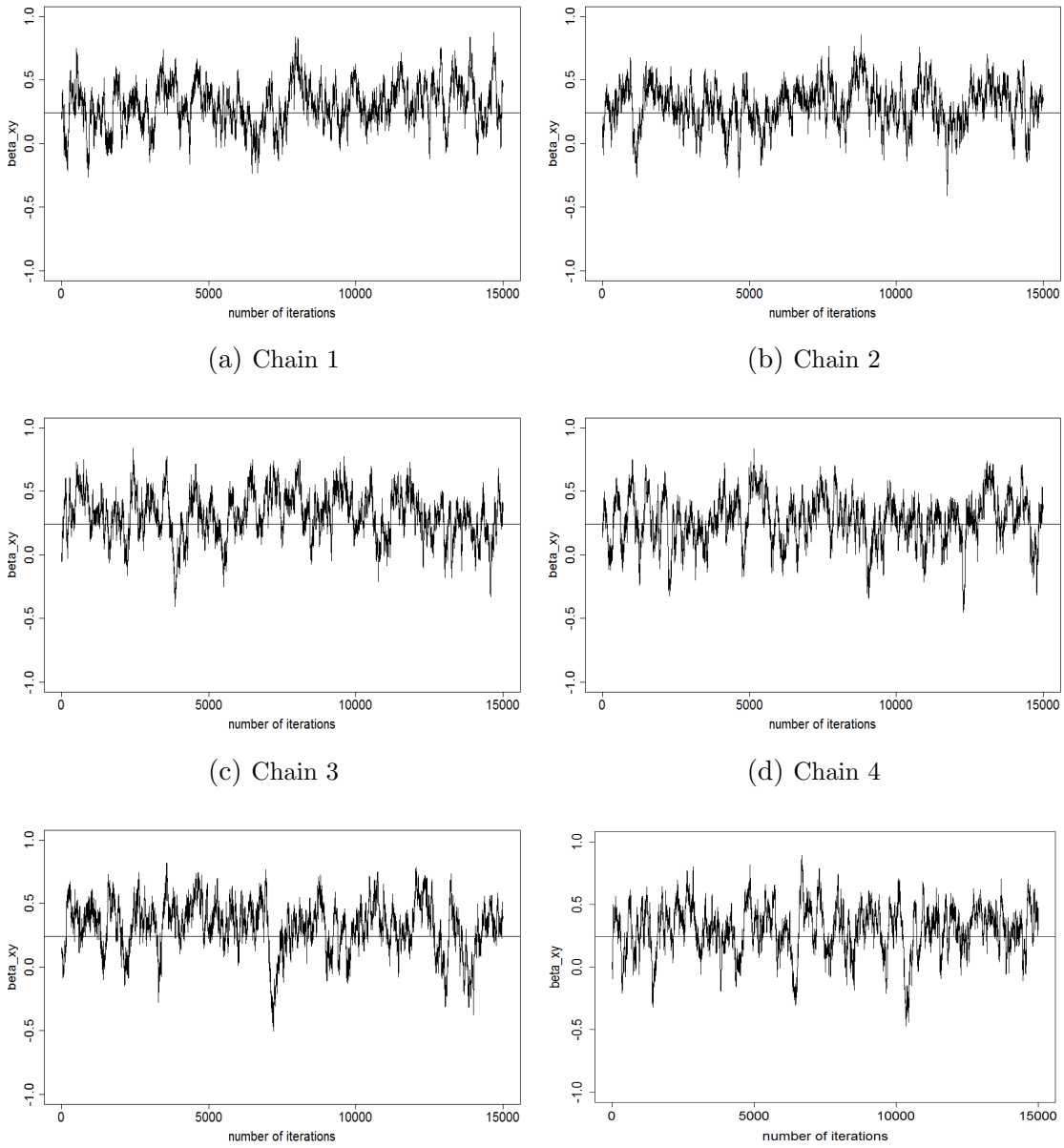


(e) Chain 5



(f) Long

Figure E.4: Trace plot of the causal effect estimate ($\hat{\beta}_{XY}$) from 60 instruments with five short and one long chain. Short and long chain consist of 50,000 and 500,000 iterations with 10,000 and 250,000 burn-in respectively. The horizontal line is the true β_{XY} (0.2449).



(a) Chain 1

(b) Chain 2

(c) Chain 3

(d) Chain 4

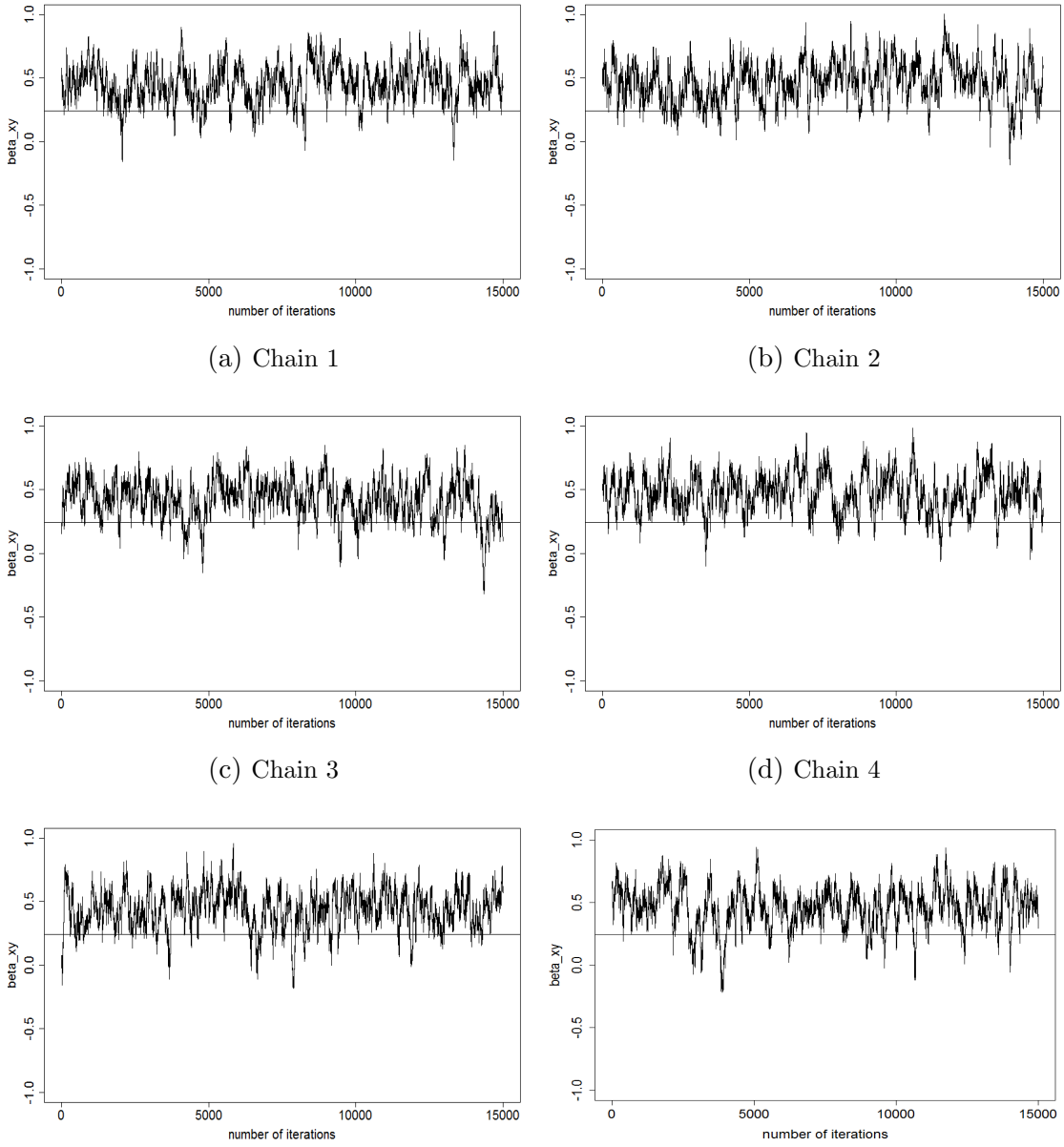
(e) Chain 5

(f) Long

Figure E.5: Trace plot of the causal effect estimate ($\hat{\beta}_{XY}$) from 90 instruments with five short and one long chain. Short and long chain consist of 50,000 and 500,000 iterations with 10,000 and 250,000 burn-in respectively. The horizontal line is the true β_{XY} (0.2449).

Table E.1: Convergence Diagnostic for instruments with low MAF: Comparing 5 short chains with 1 long chain. The short chain had 50,000 iterations with 10,000 burn in. The long chain had 500,000 iterations and 250,000 burn in. $\hat{\beta}_{XY}$ is the estimated β_{XY}. The true β_{XY} is 0.2449. Prob. X is the probability of X included in the second regression. Total Visit. Prob. is the visited probability of the set of models chosen in the first regression; the set consist of the top 10 models from the longer chain.

Instruments	Chain	Mean $\hat{\beta}_{XY}$	95% Credible Interval		Prob. X	Total Visit. Prob.
10	1	0.4145	0.0000	0.7490	0.9104	0.5053
	2	0.4226	0.0000	0.7543	0.9148	0.4861
	3	0.3855	0.0000	0.7225	0.8771	0.4956
	4	0.4338	0.0000	0.7367	0.9240	0.5132
	5	0.4167	0.0000	0.7340	0.9192	0.5056
	Long	0.4272	0.0000	0.7473	0.9252	0.5069
30	1	-0.0391	-0.4131	0.1232	0.2878	0.1037
	2	-0.0517	-0.4774	0.1170	0.3194	0.0982
	3	-0.0380	-0.4250	0.1209	0.2962	0.0871
	4	-0.0338	-0.3794	0.1596	0.3217	0.0974
	5	-0.0381	-0.4418	0.1472	0.3000	0.0928
	Long	-0.0405	-0.4057	0.1321	0.3139	0.0928
60	1	0.1495	-0.1593	0.5784	0.5694	0.0104
	2	0.2039	-0.0515	0.6030	0.6570	0.0141
	3	0.1198	-0.0704	0.5265	0.5038	0.0169
	4	0.1153	-0.0527	0.5302	0.4701	0.0148
	5	0.1617	-0.0800	0.5835	0.5659	0.0134
	Long	0.1480	-0.0901	0.5788	0.5450	0.0148
90	1	0.4345	0.0000	0.8143	0.8982	0.0022
	2	0.4464	0.0000	0.8433	0.9073	0.0015
	3	0.4293	0.0000	0.8017	0.9188	0.0000
	4	0.4128	0.0000	0.8325	0.8659	0.0004
	5	0.4583	0.0000	0.8247	0.9245	0.0024
	Long	0.4339	0.0000	0.8424	0.8821	0.0048



(a) Chain 1

(b) Chain 2

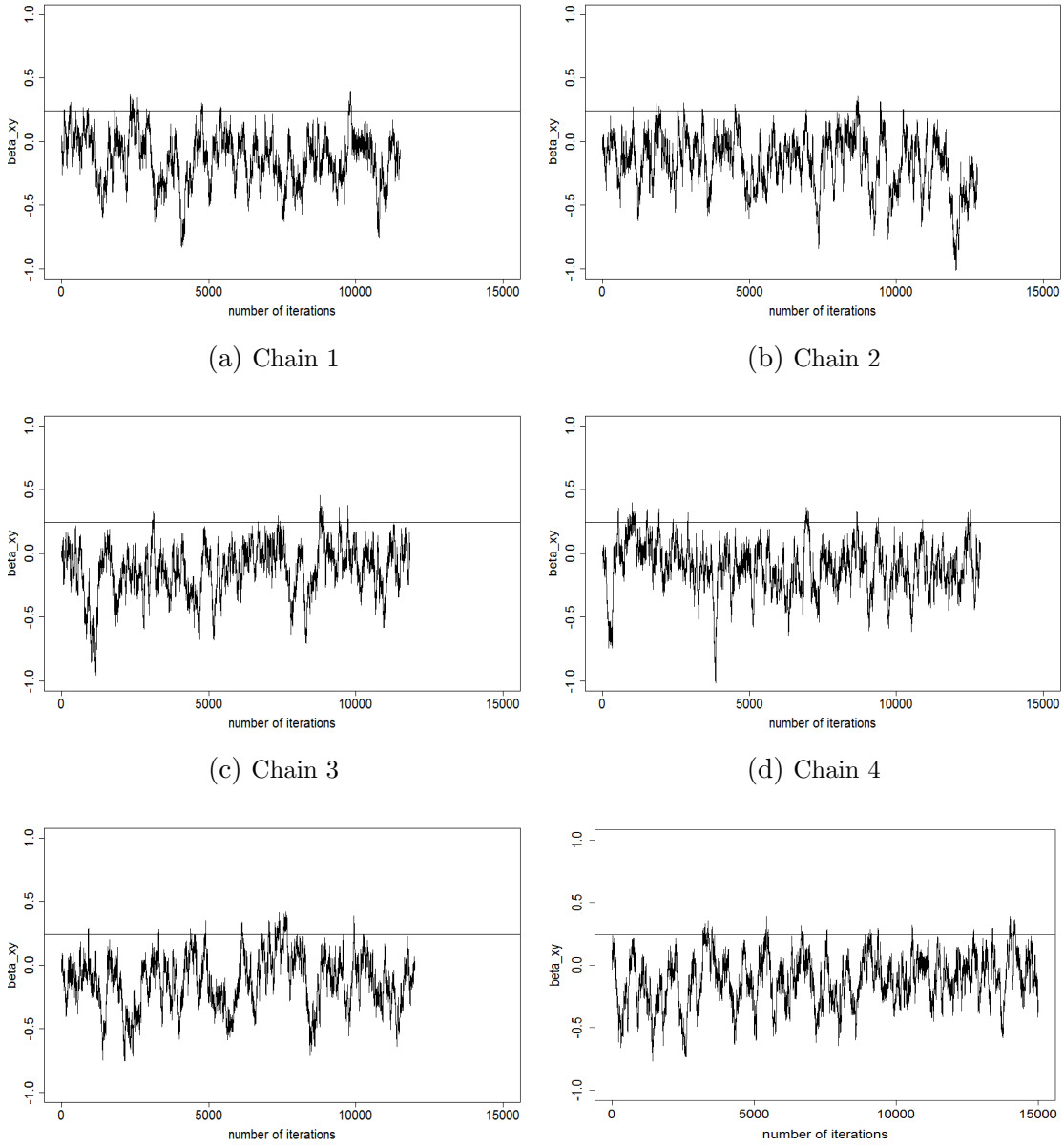
(c) Chain 3

(d) Chain 4

(e) Chain 5

(f) Long

Figure E.6: Trace plot of the causal effect estimate ($\hat{\beta}_{XY}$) from 10 instruments with low MAF. Short and long chain consist of 50,000 and 500,000 iterations with 10,000 and 250,000 burn-in respectively. The horizontal line is the true β_{XY} (0.2449).



(a) Chain 1

(b) Chain 2

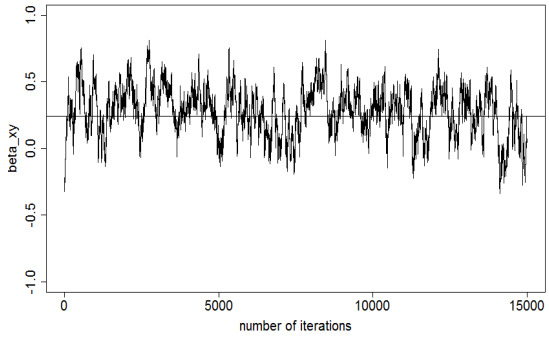
(c) Chain 3

(d) Chain 4

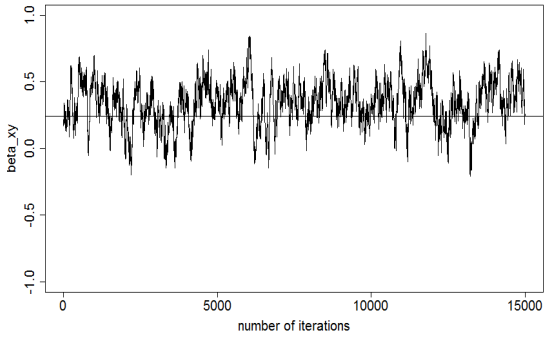
(e) Chain 5

(f) Long

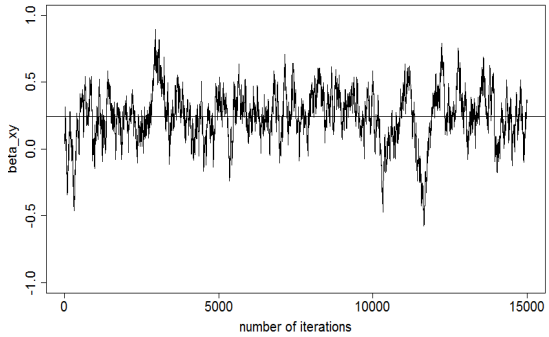
Figure E.7: Trace plot of the causal effect estimate ($\hat{\beta}_{XY}$) from 30 instruments with low MAF. Short and long chain consist of 50,000 and 500,000 iterations with 10,000 and 250,000 burn-in respectively. The horizontal line is the true β_{XY} (0.2449).



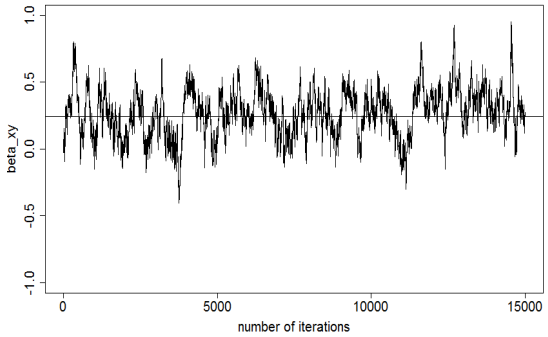
(a) Chain 1



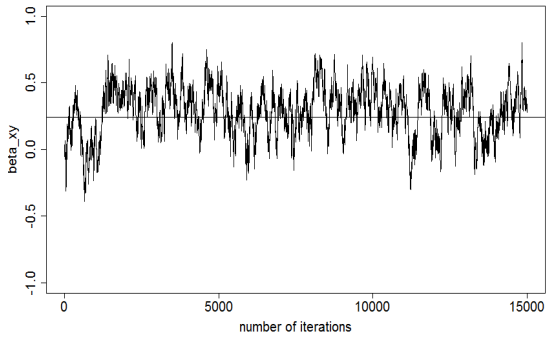
(b) Chain 2



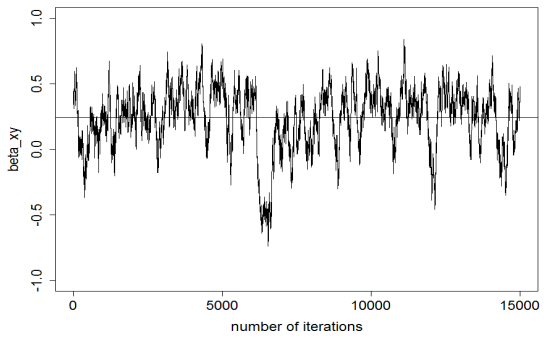
(c) Chain 3



(d) Chain 4

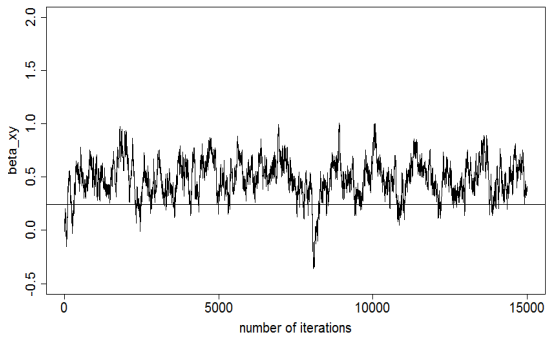


(e) Chain 5

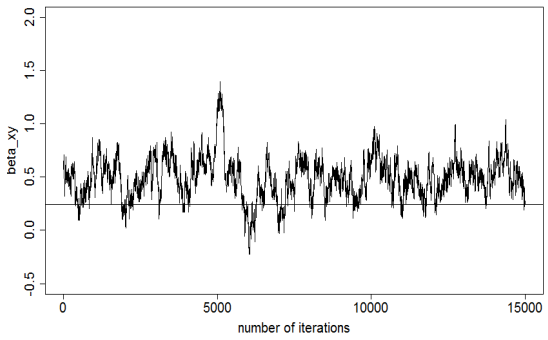


(f) Long

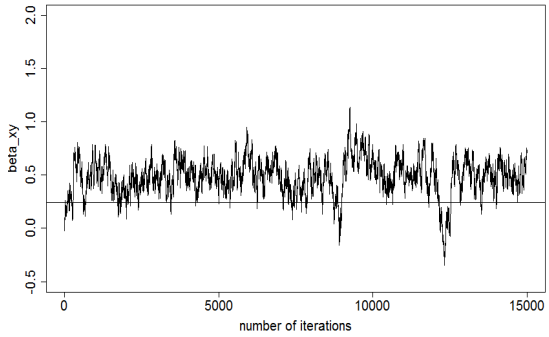
Figure E.8: Trace plot of the causal effect estimate ($\hat{\beta}_{XY}$) from 60 instruments with low MAF. Short and long chain consist of 50,000 and 500,000 iterations with 10,000 and 250,000 burn-in respectively. The horizontal line is the true β_{XY} (0.2449).



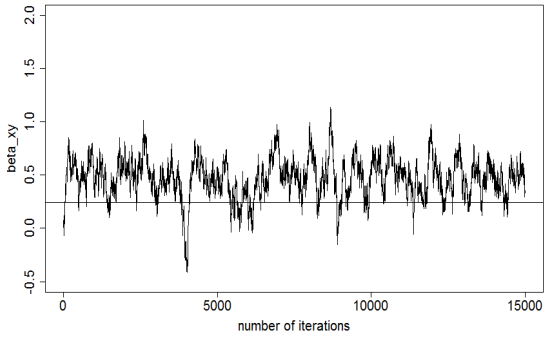
(a) Chain 1



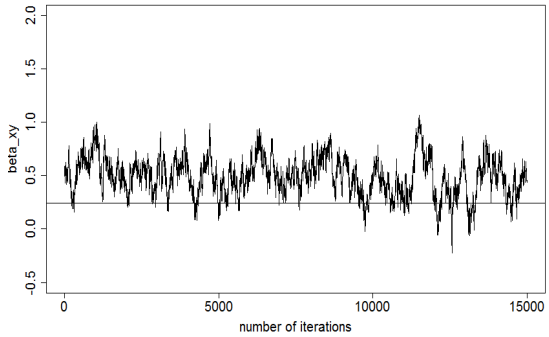
(b) Chain 2



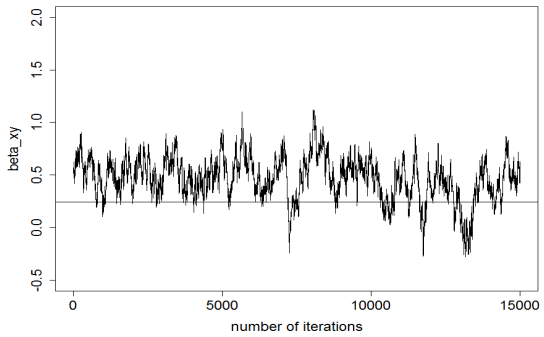
(c) Chain 3



(d) Chain 4

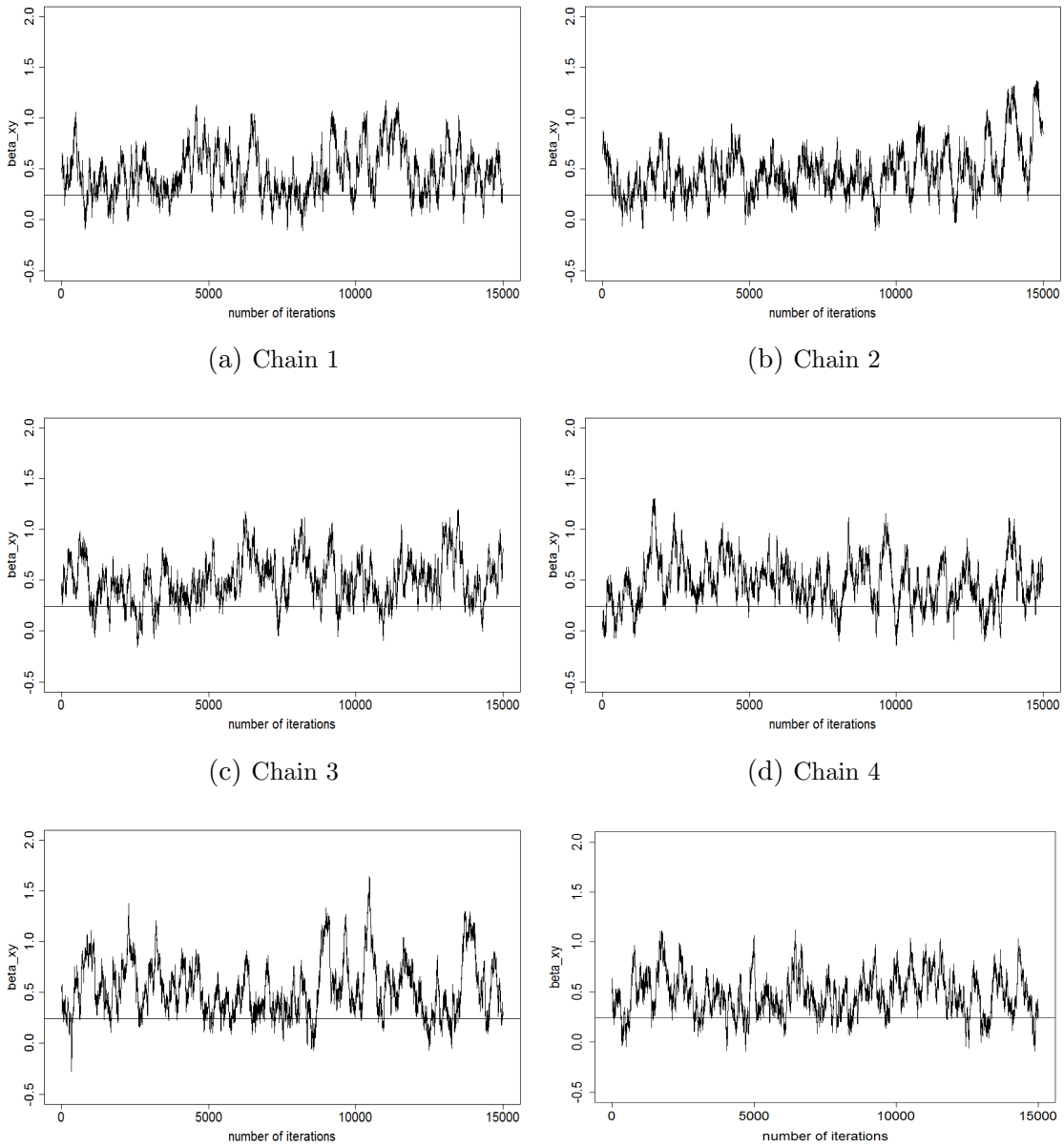


(e) Chain 5



(f) Long

Figure E.9: Trace plot of the causal effect estimate ($\hat{\beta}_{XY}$) from 90 instruments with low MAF. Short and long chain consist of 50,000 and 500,000 iterations with 10,000 and 250,000 burn-in respectively. Trace plots for chain 3 and 4 are not available, as the model was not selected in their iterations. The horizontal line is the true β_{XY} (0.2449).



(a) Chain 1

(b) Chain 2

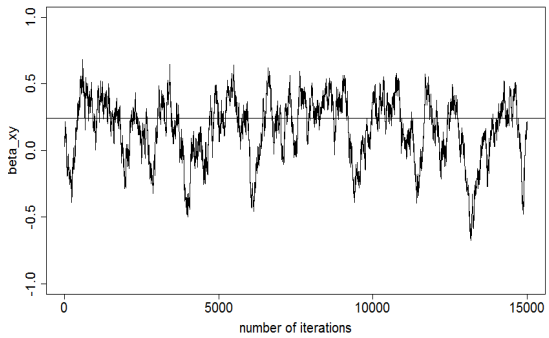
(c) Chain 3

(d) Chain 4

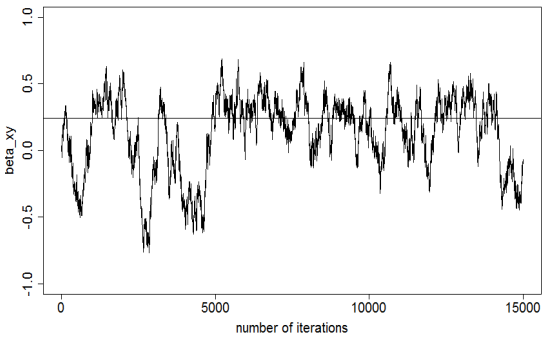
(e) Chain 5

(f) Long

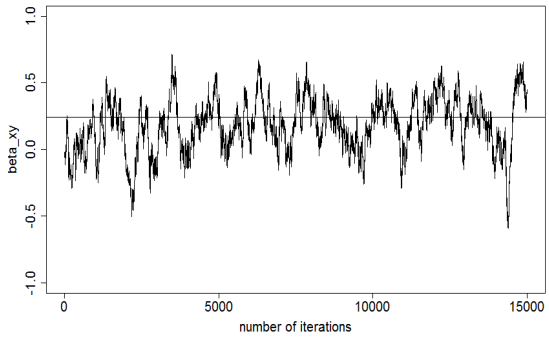
Figure E.10: Trace plot of the causal effect estimate ($\hat{\beta}_{XY}$) from 10 instruments with negative confounding effect. Short and long chain consist of 50,000 and 500,000 iterations with 10,000 and 250,000 burn-in respectively. The horizontal line is the true β_{XY} (0.2449).



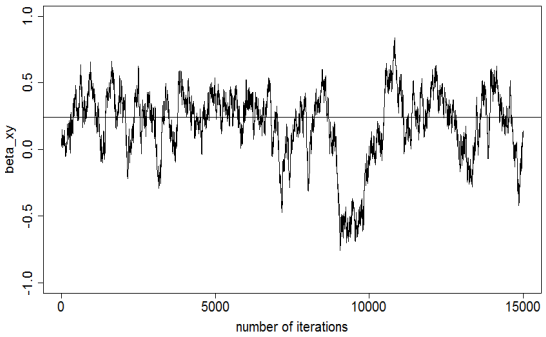
(a) Chain 1



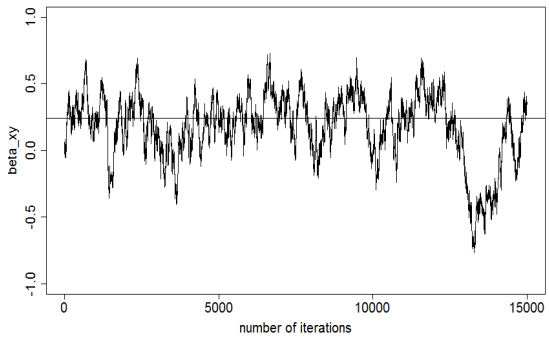
(b) Chain 2



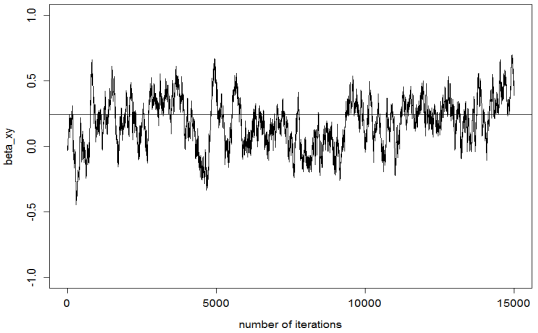
(c) Chain 3



(d) Chain 4



(e) Chain 5



(f) Long

Figure E.11: Trace plot of the causal effect estimate ($\hat{\beta}_{XY}$) from 10 instruments with strong confounding effect. Short and long chain consist of 50,000 and 500,000 iterations with 10,000 and 250,000 burn-in respectively. The horizontal line is the true β_{XY} (0.2449).

Appendix F

Bayesian vs Classic approaches to Mendelian randomisation

Table F.1: The evaluation criteria when including SNPs with different MAF, in 2SLS, LIML and IVBMA. Inst. is Instruments. Winsorised S.E. is the standard error from the standard deviation of Winsorised $\hat{\beta}_{XY}$.

	MAF	Inst.	2SLS	LIML	IVBMA
Winsorised Bias (S.E.)	**Common**	10	0.0892 (0.0066)	-0.0414 (0.0100)	-0.0581 (0.0071)
		30	0.2170 (0.0050)	-0.0154 (0.0097)	-0.0589 (0.0068)
		60	0.3089 (0.0042)	0.0016 (0.0115)	-0.0602 (0.0066)
		90	0.3379 (0.0042)	-0.0234 (0.0132)	-0.0242 (0.0076)
	Low	10	0.1457 (0.0069)	0.0049 (0.0092)	-0.0976 (0.0072)
		30	0.2481 (0.0057)	-0.0488 (0.0126)	-0.0675 (0.0080)
		60	0.3400 (0.0039)	-0.0071 (0.0147)	-0.0055 (0.0084)
		90	0.3689 (0.0035)	-0.0718 (0.0208)	0.0207 (0.0083)
	Variable	10	0.1639 (0.0087)	-0.0113 (0.0129)	-0.0978 (0.0067)
		30	0.2852 (0.0066)	-0.0219 (0.0161)	-0.0748 (0.0081)
		60	0.3419 (0.0051)	-0.0426 (0.0182)	-0.0436 (0.0089)
		90	0.3895 (0.0039)	0.0535 (0.0194)	0.0415 (0.0103)

Continued on next page

Table F.1 – *Continued from previous page*

	MAF	Inst.	2SLS	LIML	IVBMA
Winsorised RMSE	**Common**	10	0.1291	0.1470	0.1154
		30	0.2284	0.1370	0.1130
		60	0.3145	0.1623	0.1104
		90	0.3430	0.1873	0.1094
	Low	10	0.1756	0.1300	0.1410
		30	0.2606	0.1840	0.1318
		60	0.3444	0.2078	0.1186
		90	0.3721	0.3023	0.1187
	Variable	10	0.2043	0.1820	0.1362
		30	0.3000	0.2278	0.1370
		60	0.3495	0.2609	0.1328
		90	0.3934	0.2794	0.1505
Percentage of Outlier	**Common**	10	5.00	3.00	0.00
		30	1.00	1.50	0.00
		60	0.50	4.50	0.00
		90	0.00	7.00	0.00
	Low	10	1.00	5.50	3.50
		30	1.00	5.00	0.00
		60	1.00	4.00	0.00
		90	2.00	7.50	0.00
	Variable	10	0.50	5.00	1.50
		30	0.50	8.00	1.00
		60	0.50	7.00	0.50
		90	1.50	10.00	0.00

Continued on next page

Table F.1 – *Continued from previous page*

	MAF	Inst.	2SLS	LIML	IVBMA
Coverage	**Common**	10	86.50	92.00	87.50
		30	47.50	88.00	88.00
		60	10.00	83.00	90.00
		90	2.00	79.50	90.50
	Low	10	86.00	96.00	96.50
		30	45.00	91.50	93.00
		60	7.00	84.50	94.50
		90	0.50	76.00	92.00
	Variable	10	82.00	94.50	96.50
		30	41.50	89.50	95.00
		60	10.00	79.50	92.50
		90	0.50	72.50	91.00

Table F.2: The evaluation criteria when including SNPs with different MAF, in 2SLS, LIML and IVBMA. Inst. is Instruments. Winsorised S.E. is the standard error from the standard deviation of Winsorised $\hat{\beta}_{XY}$.

	MAF	Inst.	2SLS	LIML	IVBMA
Winsorised Bias (S.E.)	**Pattern I**	10	0.1639 (0.0087)	-0.0113 (0.0129)	-0.0978 (0.0067)
		30	0.2852 (0.0066)	-0.0219 (0.0161)	-0.0748 (0.0081)
		60	0.3419 (0.0051)	-0.0426 (0.0182)	-0.0436 (0.0089)
		90	0.3895 (0.0039)	0.0535 (0.0194)	0.0415 (0.0103)
	Pattern II	10	0.2322 (0.0103)	-0.0215 (0.0217)	-0.0586 (0.0083)
		30	0.3348 (0.0060)	0.0057 (0.0200)	-0.0505 (0.0094)
		60	0.3743 (0.0048)	-0.0007 (0.0226)	-0.0303 (0.0082)

Continued on next page

Table F.2 – *Continued from previous page*

	MAF	Inst.	2SLS	LIML	IVBMA
		90	0.3990 (0.0037)	-0.0388 (0.0263)	0.0028 (0.0102)
	Pattern III	10	0.1579 (0.0076)	-0.0377 (0.0139)	-0.1023 (0.0070)
		30	0.2663 (0.0062)	-0.0099 (0.0137)	-0.0611 (0.0087)
		60	0.3393 (0.0046)	0.0078 (0.0148)	-0.0272 (0.0091)
		90	0.3688 (0.0035)	-0.0581 (0.0173)	0.0261 (0.0099)
	Pattern IV	10	0.1889 (0.0089)	-0.0531 (0.0169)	-0.1063 (0.0068)
		30	0.3060 (0.0054)	0.0021 (0.0171)	-0.0636 (0.0086)
		60	0.3719 (0.0046)	0.0254 (0.0181)	-0.0074 (0.0093)
		90	0.3908 (0.0031)	-0.0351 (0.0218)	0.0261 (0.0091)
	Pattern I	10	0.2043	0.1820	0.1362
		30	0.3000	0.2278	0.1370
		60	0.3495	0.2609	0.1328
		90	0.3934	0.2794	0.1505
	Pattern II	10	0.2737	0.3067	0.1305
		30	0.3453	0.2825	0.1424
		60	0.3805	0.3193	0.1199
		90	0.4024	0.3724	0.1446
Winsorised RMSE	**Pattern III**	10	0.1909	0.2001	0.1420
		30	0.2805	0.1941	0.1365
		60	0.3454	0.2090	0.1307
		90	0.3722	0.2515	0.1416
	Pattern IV	10	0.2267	0.2447	0.1436
		30	0.3153	0.2406	0.1364

Continued on next page

Table F.2 – *Continued from previous page*

	MAF	Inst.	2SLS	LIML	IVBMA
		60	0.3775	0.2572	0.1308
		90	0.3933	0.3094	0.1308
Percentage of Outlier	**Pattern I**	10	0.50	5.00	1.50
		30	0.50	8.00	1.00
		60	0.50	7.00	0.50
		90	1.50	10.00	0.00
	Pattern II	10	1.50	11.00	1.00
		30	1.00	9.50	0.00
		60	1.50	14.00	0.50
		90	0.50	6.50	0.00
	Pattern III	10	0.50	5.00	3.00
		30	0.00	4.50	1.00
		60	0.00	6.50	0.00
		90	1.50	10.50	0.00
	Pattern IV	10	2.00	4.50	4.00
		30	2.50	8.00	0.50
		60	0.50	10.00	0.00
		90	2.50	9.50	0.00
Coverage	**Pattern I**	10	82.00	94.50	96.50
		30	41.50	89.50	95.00
		60	10.00	79.50	92.50
		90	0.50	72.50	91.00
	Pattern II	10	77.00	92.00	97.00
		30	30.00	83.50	97.50

Continued on next page

Table F.2 – *Continued from previous page*

MAF	Inst.	2SLS	LIML	IVBMA
	60	4.50	80.00	96.50
	90	0.00	70.00	91.50
Pattern III	10	82.50	93.50	92.50
	30	43.00	87.50	94.50
	60	5.00	82.50	93.00
	90	1.00	75.50	89.00
Pattern IV	10	79.50	91.50	95.50
	30	33.00	84.50	93.50
	60	5.50	79.00	94.50
	90	0.50	76.50	89.50

Table F.3: Individual SNP association with BMI for GRAPHIC study. The SNPs are all from chromosome 16. Coef. and F-stat. is coefficient and F-statistics from the genetic association with BMI respectively.

SNP	Position	Effect allele	Other allele	Coef.	F-stat.	R^2	p-value	Correlation with lead SNP
rs4783818	52295784	T	A	0.1481	0.3647	0.0004	0.5460	-0.0408
rs4389136	52305359	G	A	-0.1131	0.3127	0.0003	0.5761	-0.0123
rs7203521	52326794	G	A	-0.0890	0.1961	0.0002	0.6580	0.1105
rs16952482	52329084	C	T	0.1302	0.1766	0.0002	0.6744	-0.0769
rs6499641	52330127	T	A	-0.0390	0.0416	0.0000	0.8385	0.0581
rs5013514	52330605	G	C	0.1522	0.2498	0.0002	0.6173	0.0044
rs11861870	52343947	C	T	-0.1409	0.2553	0.0002	0.6135	0.0655
rs1861869	52347682	C	G	-0.3361	3.3023	0.0032	0.0695	0.0313

rs1075440	52348407	G	A	-0.5113	6.5449	0.0063	0.0107	0.0526
rs1077128	52349154	A	C	0.1621	0.5086	0.0005	0.4759	-0.0145
rs7184874	52349940	T	C	0.3225	3.0183	0.0029	0.0826	-0.0389
rs7186521	52350423	G	A	0.2945	2.4460	0.0024	0.1181	-0.0203
rs13333228	52351299	T	C	-0.5401	7.2163	0.0070	0.0073	0.0488
rs13334933	52353137	G	A	0.2253	0.9407	0.0009	0.3323	-0.0333
rs16952517	52354558	A	G	0.1118	0.1607	0.0002	0.6886	-0.0392
rs6499643	52355019	C	T	0.0259	0.0103	0.0000	0.9191	-0.0559
rs4784323	52355066	A	G	0.0179	0.0076	0.0000	0.9307	0.0483
rs7206790	52355409	G	C	0.1626	0.6941	0.0007	0.4050	0.0366
rs8047395	52356024	G	A	-0.1732	0.8065	0.0008	0.3694	-0.0504
rs1421085	52358455	C	T	0.2750	1.9674	0.0019	0.1610	0.0403
rs9923147	52359050	T	C	0.1276	0.4286	0.0004	0.5128	0.0520
rs2058908	52363646	T	C	-0.2208	1.0642	0.0010	0.3025	0.1185
rs17817288	52365265	G	A	0.2940	2.3079	0.0022	0.1290	0.0418
rs1477196	52365759	A	G	-0.1884	0.9161	0.0009	0.3387	0.0348
rs1121980	52366748	A	G	0.1205	0.3819	0.0004	0.5367	0.0505
rs7193144	52368187	C	T	0.2445	1.5242	0.0015	0.2173	0.0282
rs17817449	52370868	G	T	0.2531	1.6322	0.0016	0.2017	0.0267
rs11075987	52372662	G	T	0.0953	0.2403	0.0002	0.6241	0.0404
rs3751812	52375961	T	G	0.2569	1.6766	0.0016	0.1957	0.0312
rs11075989	52377378	T	C	0.2330	1.3824	0.0013	0.2400	0.0289
rs9939609	52378028	A	T	0.2330	1.3824	0.0013	0.2400	0.0289
rs7185735	52380152	G	A	0.2333	1.3851	0.0013	0.2395	0.0288
rs9941349	52382989	T	C	0.0968	0.2404	0.0002	0.6240	0.0425
rs7190492	52386253	A	G	-0.2007	1.0466	0.0010	0.3065	0.0388
rs9922708	52388647	T	C	0.1279	0.4365	0.0004	0.5090	0.0271
rs12149832	52400409	A	G	0.2723	1.9709	0.0019	0.1607	0.0165
rs17218700	52402080	A	G	0.2274	0.6003	0.0006	0.4386	-0.1243
rs11642841	52402988	A	C	0.2380	1.5702	0.0015	0.2105	0.0051
rs11075994	52407580	A	G	-0.4069	4.1245	0.0040	0.0425	0.1794

rs1421090	52407671	G	A	0.0610	0.0853	0.0001	0.7703	-0.0278
rs9939811	52408369	C	T	0.0709	0.0996	0.0001	0.7524	-0.2503
rs11075995	52412792	A	T	0.1164	0.2725	0.0003	0.6018	0.0617
rs17219084	52413101	G	A	0.0534	0.0699	0.0001	0.7915	-0.1794
rs8061518	52418525	G	A	-0.0957	0.2417	0.0002	0.6231	0.1623
rs7184573	52419093	A	G	0.0455	0.0507	0.0000	0.8219	-0.1728
rs10521307	52423202	G	A	-0.1961	1.0052	0.0010	0.3163	0.1540
rs2388405	52424960	T	C	0.0871	0.2119	0.0002	0.6454	-0.2144
rs16952577	52425817	T	G	0.0065	0.0008	0.0000	0.9770	-0.2426
rs17818866	52428675	G	A	0.0777	0.1188	0.0001	0.7304	-0.2509
rs17818997	52430484	C	G	0.0797	0.1247	0.0001	0.7241	-0.2510
rs17819063	52430929	A	G	-0.0851	0.0877	0.0001	0.7671	0.1164
rs2111115	52432517	G	A	0.0393	0.0429	0.0000	0.8360	-0.2075
rs7205213	52434067	T	C	0.0868	0.2094	0.0002	0.6474	-0.2121
rs6499653	52435093	T	C	0.1044	0.2256	0.0002	0.6349	0.0330
rs12447581	52439374	C	G	-0.1222	0.2406	0.0002	0.6239	0.0175
rs8061228	52439872	C	T	-0.1596	0.4153	0.0004	0.5195	0.0196
rs11075999	52440360	C	A	-0.1282	0.2653	0.0003	0.6066	0.0180
rs1344500	52440534	C	G	-0.1172	0.2198	0.0002	0.6393	0.0185
rs12596457	52442515	C	G	-0.0784	0.0998	0.0001	0.7522	0.0131
rs13337591	52444943	T	G	-0.2007	0.8479	0.0008	0.3574	0.1066
rs10521303	52466686	T	G	-0.4202	4.8941	0.0048	0.0272	0.1787
rs4784329	52467762	C	A	-0.4696	5.7435	0.0056	0.0167	0.1842
rs9934504	52474380	A	G	0.2728	1.0543	0.0010	0.3048	-0.1885
rs1558755	52478167	C	T	0.3718	3.6009	0.0035	0.0580	-0.0385
rs17820875	52484291	G	A	-0.0983	0.1687	0.0002	0.6814	-0.0004
rs11076008	52484824	A	G	-0.0762	0.0985	0.0001	0.7537	-0.0330
rs9926180	52486108	T	A	-0.0217	0.0092	0.0000	0.9237	-0.0830
rs7500562	52488391	C	G	-0.0337	0.0221	0.0000	0.8817	-0.0843
rs12933928	52488494	G	A	-0.0068	0.0008	0.0000	0.9775	0.1416
rs1362570	52491048	C	T	-0.1308	0.1680	0.0002	0.6820	0.0604

rs16952634	52492057	A	G	0.1870	0.3742	0.0004	0.5408	-0.1527
rs17222465	52493954	A	C	0.2093	1.1284	0.0011	0.2884	-0.0366
rs2111112	52495133	T	C	0.0517	0.0521	0.0001	0.8194	-0.0951
rs10852525	52496582	A	G	-0.1888	0.4205	0.0004	0.5168	0.0453
rs9929152	52496904	G	A	-0.0377	0.0288	0.0000	0.8654	-0.1029
rs8056040	52499645	G	A	-0.3336	1.1869	0.0012	0.2762	0.1723
rs12935710	52500306	T	C	-0.0166	0.0054	0.0000	0.9417	-0.0879
rs12708942	52503705	A	T	-0.2104	0.4335	0.0004	0.5104	0.0630
rs9806929	52507417	A	G	-0.2104	0.4335	0.0004	0.5104	0.0630
rs7197167	52508854	G	T	-0.0423	0.0337	0.0000	0.8545	-0.1027
rs1344503	52510447	A	G	-0.0857	0.0823	0.0001	0.7743	-0.1751
rs12232391	52510620	G	T	-0.1985	0.9767	0.0010	0.3233	0.2579
rs7193851	52510646	C	T	-0.0513	0.0350	0.0000	0.8516	-0.0624
rs8053966	52511499	C	T	-0.0861	0.1114	0.0001	0.7387	-0.0690
rs4784331	52519268	T	A	-0.2307	1.4658	0.0014	0.2263	0.1874
rs8061397	52522327	T	C	-0.1723	0.3185	0.0003	0.5726	-0.1614
rs4784333	52526589	G	C	0.2031	1.0373	0.0010	0.3087	-0.3327
rs7205426	52531308	A	C	0.3490	3.4514	0.0034	0.0635	-0.6736
rs12933996	52534163	G	A	-0.3856	3.9742	0.0039	0.0465	0.9346
rs12919344	52538175	C	A	-0.5233	7.2181	0.0070	0.0073	0.9712
rs9924877	52538922	A	G	-0.5631	8.2205	0.0079	0.0042	1.0000
rs7202360	52540292	T	G	-0.5614	8.1670	0.0079	0.0044	0.9958
rs7203181	52540981	A	C	-0.5495	7.9056	0.0076	0.0050	0.9719
rs12925189	52542774	G	A	-0.4804	6.2623	0.0061	0.0125	0.9029
rs6499656	52546456	C	G	-0.1313	0.2771	0.0003	0.5987	0.2099
rs7185301	52547167	G	A	0.0313	0.0162	0.0000	0.8988	-0.0009
rs7191513	52548024	A	G	-0.1704	0.8251	0.0008	0.3639	0.1296
rs11644943	52553085	A	T	-0.0444	0.0411	0.0000	0.8393	-0.0204
rs12446047	52554803	C	T	-0.1935	1.1141	0.0011	0.2914	0.0921
rs17823199	52556431	C	T	-0.1232	0.4349	0.0004	0.5097	0.0853
rs17823223	52557139	T	C	-0.1251	0.2128	0.0002	0.6447	-0.0438

rs1344502	52558293	G	A	-0.2715	2.2526	0.0022	0.1337	0.0325
rs8053888	52561306	T	C	0.2618	2.0202	0.0020	0.1555	-0.0597
rs9940629	52562312	A	G	-0.1165	0.3916	0.0004	0.5316	0.1682
rs9932411	52562664	C	T	-0.0790	0.1758	0.0002	0.6751	0.1512
rs7206456	52562990	A	G	0.0621	0.0888	0.0001	0.7657	0.0496
rs7185783	52565323	T	G	-0.2782	1.0584	0.0010	0.3038	0.1142
rs8057547	52567014	A	C	-0.1941	1.0660	0.0010	0.3021	-0.0664
rs9302654	52567046	T	C	-0.1545	0.3879	0.0004	0.5336	0.0186
rs4784335	52567189	T	G	-0.1915	1.0257	0.0010	0.3114	-0.0450
rs13337356	52574811	G	C	-0.1661	0.4711	0.0005	0.4926	0.0088
rs16952730	52576422	A	G	-0.1367	0.4281	0.0004	0.5131	-0.0821
rs12325409	52577134	G	T	-0.1909	0.3690	0.0004	0.5437	-0.0836
rs12324955	52577187	A	G	-0.1740	0.5713	0.0006	0.4499	0.0337
rs8049235	52578510	A	G	-0.2412	1.5887	0.0015	0.2078	0.0279
rs4784336	52580038	C	A	-0.5534	3.0345	0.0029	0.0818	0.0724
rs9933107	52581017	T	G	-0.1257	0.4030	0.0004	0.5257	-0.0072
rs9933805	52581085	C	T	-0.2685	1.8717	0.0018	0.1716	-0.0346
rs8056199	52582673	A	G	-0.4051	4.7214	0.0046	0.0300	0.0397
rs8056502	52582815	C	G	-0.2916	2.2271	0.0022	0.1359	-0.0236
rs6499660	52583705	A	T	0.4368	5.3494	0.0052	0.0209	-0.0658
rs1420571	52584994	G	A	-0.3775	1.5181	0.0015	0.2182	0.0796
rs11864881	52586805	A	C	-0.2799	2.1337	0.0021	0.1444	0.0077
rs11076013	52587029	A	G	0.4570	5.8883	0.0057	0.0154	-0.0270
rs1966435	52588027	T	C	-0.2204	1.3451	0.0013	0.2464	-0.0140
rs12385988	52589423	T	C	-0.1762	0.6177	0.0006	0.4321	-0.0532
rs7193917	52590586	T	C	-0.2887	2.2482	0.0022	0.1341	0.1285
rs7199716	52590749	T	C	-0.2454	1.6671	0.0016	0.1969	0.0107
rs13334214	52592442	T	C	-0.2554	1.1010	0.0011	0.2943	0.0675
rs4784338	52595472	T	G	0.4298	4.8454	0.0047	0.0279	-0.0896
rs7199363	52596515	C	T	-0.3976	3.6336	0.0035	0.0569	0.0769
rs4784339	52597179	G	A	0.4299	4.8513	0.0047	0.0278	-0.0943

rs8054364	52600117	G	C	0.0007	0.0000	0.0000	0.9979	0.0582
rs12600130	52600824	G	C	-0.5886	4.4352	0.0043	0.0354	0.0467
rs1990685	52608126	A	C	-0.1352	0.4992	0.0005	0.4800	0.1001
rs7186220	52611296	A	G	-0.3988	2.7571	0.0027	0.0971	0.0874
rs17825567	52615272	A	C	-0.0887	0.1234	0.0001	0.7254	-0.0103
rs1477094	52621119	A	T	0.4049	2.0539	0.0020	0.1521	0.0498
rs860713	52626966	G	A	-0.1111	0.3384	0.0003	0.5609	0.1099
rs2192872	52632128	C	T	-0.0917	0.2219	0.0002	0.6377	0.0941
rs4784351	52633199	G	A	-0.2814	1.9782	0.0019	0.1599	0.1134
rs1078013	52643195	A	T	0.4238	2.2351	0.0022	0.1352	0.0509
rs1076467	52643426	A	G	0.4249	2.1885	0.0021	0.1394	0.0474
rs940214	52648579	C	A	0.4012	1.9440	0.0019	0.1635	0.0491
rs856974	52652133	C	T	-0.1330	0.4858	0.0005	0.4859	0.1103
rs2003583	52657507	T	C	-0.4624	4.7252	0.0046	0.0300	0.0864
rs1108086	52657870	C	T	0.5118	3.2130	0.0031	0.0734	0.0578
rs12600060	52658962	T	G	0.0200	0.0085	0.0000	0.9263	-0.0126
rs1420318	52660267	A	G	0.5035	3.0267	0.0029	0.0822	0.0575
rs12149010	52665424	T	C	-0.0870	0.1710	0.0002	0.6793	-0.0189
rs2665275	52665624	T	C	0.1107	0.1381	0.0001	0.7103	0.0783
rs7206012	52671065	C	T	0.5784	3.6964	0.0036	0.0548	-0.0275
rs7206224	52671149	T	C	0.1987	0.9796	0.0010	0.3225	-0.0774
rs11646290	52671763	C	T	0.3686	3.4361	0.0033	0.0641	-0.0913
rs708262	52672030	T	G	-0.3325	2.4664	0.0024	0.1166	0.0420
rs16953002	52672325	A	G	0.0208	0.0064	0.0000	0.9364	0.0592
rs3928987	52675012	A	G	0.4220	4.9659	0.0048	0.0261	0.0011
rs1008400	52677393	T	C	-0.3574	3.4016	0.0033	0.0654	0.0269
rs9937121	52678345	T	A	-0.3964	4.1951	0.0041	0.0408	0.0228
rs697769	52679248	A	G	-0.3952	4.2150	0.0041	0.0403	0.0205
rs708255	52680014	A	G	0.4326	5.2334	0.0051	0.0224	-0.0319
rs8057572	52683399	C	A	0.5165	6.5266	0.0063	0.0108	-0.0856
rs2665272	52684118	C	T	0.4518	5.6769	0.0055	0.0174	-0.0335

rs2689248	52685410	T	G	0.2936	2.2925	0.0022	0.1303	-0.0366
rs16953047	52687671	T	G	-0.2574	0.9676	0.0009	0.3255	0.0429
rs741300	52691151	G	A	-0.3855	3.4921	0.0034	0.0619	-0.0051
rs12927155	52691301	T	C	-0.2323	0.7762	0.0008	0.3785	0.0373
rs12931414	52692494	A	G	-0.2140	0.6772	0.0007	0.4107	0.0347
rs2540776	52695363	A	G	-0.2233	0.7165	0.0007	0.3975	0.0365
rs2689264	52695481	T	C	-0.2445	0.8611	0.0008	0.3536	0.0325
rs2689269	52696065	G	A	0.3622	3.4926	0.0034	0.0619	-0.0452
rs17236863	52699591	T	G	0.4989	5.9543	0.0058	0.0149	-0.0703
rs708278	52702426	A	G	-0.1655	0.5708	0.0006	0.4501	-0.0099

References

[1] S. Adamsson Eryd, M. Sjögren, J. G. Smith, P. M. Nilsson, O. Melander, B. Hedblad, and G. Engström. Ceruloplasmin and atrial fibrillation: evidence of causality from a population-based Mendelian randomization study. *J Intern Med*, 275(2):164–171, Feb. 2014.

[2] S. Afzal, P. Brndum-Jacobsen, S. E. Bojesen, and B. G. Nordestgaard. Genetically low vitamin D concentrations and increased mortality: Mendelian randomisation analysis in three large cohorts. *BMJ*, 349:g6330, 2014.

[3] B. L. Aken, S. Ayling, D. Barrell, L. Clarke, V. Curwen, S. Fairley, et al. The Ensembl gene annotation system. *Database*, 2016:baw093, 2016.

[4] J. M. Alegret, G. Aragons, R. Elosua, R. Beltrn-Debn, A. Hernndez-Aguilera, C. Romero-Menor, J. Camps, and J. Joven. The relevance of the association between inflammation and atrial fibrillation. *Eur J Clin Invest*, 43(4):324–331, Apr 2013.

[5] K. H. Allin, B. G. Nordestgaard, J. Zacho, A. Tybjaerg-Hansen, and S. E. Bojesen. C-reactive protein and the risk of cancer: a Mendelian randomization study. *J Natl Cancer Inst*, 102(3):202–206, Feb 2010.

[6] C. A. Anderson, F. H. Pettersson, J. C. Barrett, J. J. Zhuang, J. Ragoussis, L. R. Cardon, and A. P. Morris. Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. *Am J Hum Genet*, 83(1):112–119, 2008.

[7] T. Anderson and H. Rubin. Estimators of the parameters of a single equation in a complete set of stochastic equations. *The Annals of Mathematical Statistics*, 21, 1949.

[8] T. Anderson, N. Kunitomo, and Y. Matsushita. On the asymptotic optimality of the LIML estimator with possibly many instruments. *Journal of Econometrics*, 157(2):191–204, 2010.

[9] T. W. Anderson. *An introduction to multivariate statistical analysis*. Wiley, 1958.

[10] D. Andrews. Consistent moment selection procedures for generalized method of moments estimation. *Econometrica*, 67(3):543–564, 1999.

[11] D. W. Andrews and J. H. Stock. Testing with many weak instruments. *Journal of Econometrics*, 138(1):24–46, 2007.

[12] D. W. Andrews, M. J. Moreira, and J. H. Stock. Performance of conditional Wald tests in IV regression with weak instruments. *Journal of Econometrics*, 139(1):116–132, 2007.

[13] J. D. Angrist and A. B. Krueger. Does compulsory school attendance affect schooling and earnings? Technical report, National Bureau of Economic Research, 1990.

[14] B. Aramini, C. Kim, S. Diangelo, E. Petersen, D. J. Lederer, L. Shah, H. Robbins, J. Floros, S. M. Arcasoy, J. R. Sonett, and F. D'Ovidio. Donor surfactant protein D (SP-D) polymorphisms are associated with lung transplant outcome. *Am J Transplant*, 13(8):2130–2136, Aug 2013.

[15] K. G. Ardlie, L. Kruglyak, and M. Seielstad. Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 3(4):299–309, 2002.

[16] I. K. S. array BMI Consortium et al. Gene-centric meta-analyses of 108,912 individuals confirm known body mass index loci and reveal three novel signals. *Hum Mol Genet*, page dds396, 2012.

[17] J. Attermann, C. Obel, N. Bilenberg, C. M. Nordenbæk, A. Skytthe, and J. Olsen. Traits of ADHD and autism in girls with a twin brother: a Mendelian randomization study. *Eur Child Adolesc Psychiatry*, 21(9):503–509, Sep 2012.

[18] A. Auton. *The Estimation of Recombination Rates from Population Genetic Data*. PhD thesis, University of Oxford, 2007.

[19] J. C. Barrett and L. R. Cardon. Evaluating coverage of genome-wide association studies. *Nat Genet*, 38(6):659–662, 2006.

[20] G. S. Barsh, G. P. Copenhaver, G. Gibson, and S. M. Williams. Guidelines for genome-wide association studies. *PLoS Genet*, 8(7):e1002812, 2012.

[21] S. Basu and W. Pan. Comparison of statistical tests for disease association with rare variants. *Genet Epidemiol*, 35(7):606–619, 2011.

[22] C. F. Baum, M. E. Schaffer, and S. Stillman. Instrumental variables and GMM: Estimation and testing. *Stata Journal*, 3(1):1–31, 2003.

[23] M. A. Beaumont and B. Rannala. The bayesian revolution in genetics. *Nature Reviews Genetics*, 5(4):251–261, 2004.

[24] P. A. Bekker. Alternative approximations to the distributions of instrumental variable estimators. *Econometrica: Journal of the Econometric Society*, pages 657–681, 1994.

[25] P. A. Bekker and F. Crudu. Jackknife instrumental variable estimation with heteroskedasticity. *Journal of Econometrics*, 185(2):332–342, 2015.

[26] A. G. Boef, O. M. Dekkers, and S. le Cessie. Mendelian randomization studies: a review of the approaches used and the quality of reporting. *Int J Epidemiol*, 44(2):496–511, 2015.

[27] C. E. Bolton, W. Schumacher, J. R. Cockcroft, N. J. Timpson, G. Davey Smith, J. Gallacher, and et al. The CRP genotype, serum levels and lung function in men: the caerphilly prospective study. *Clin Sci (Lond)*, 120(8): 347–355, Apr 2011.

[28] C. Bonilla, D. A. Lawlor, A. E. Taylor, D. J. Gunnell, Y. Ben-Shlomo, A. R. Ness, and et al. Vitamin B-12 status during pregnancy and child's IQ at age 8: a Mendelian randomization study in the avon longitudinal study of parents and children. *PLoS One*, 7(12):e51084, 2012.

[29] J. Bound, D. A. Jaeger, and R. M. Baker. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450, 1995.

[30] J. Bowden and S. Vansteelandt. Mendelian randomization analysis of case-control data using structural mean models. *Stat Med*, 30(6):678–694, Mar 2011.

[31] J. Bowden, G. Davey Smith, and S. Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol*, 44(2):512–525, 2015.

[32] J. Bowden, G. Davey Smith, P. C. Haycock, and S. Burgess. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genet Epidemiol*, 40(4):304–314, 2016.

[33] L. Breiman. The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association*, 87(419):738–754, 1992.

[34] L. P. Breitling, W. Koenig, M. Fischer, Z. Mallat, C. Hengstenberg, D. Rothenbacher, and H. Brenner. Type II secretory phospholipase A2 and prognosis in patients with stable coronary heart disease: Mendelian randomization study. *PLoS One*, 6(7):e22318, 2011.

[35] P. Brennan, J. McKay, L. Moore, D. Zaridze, A. Mukeria, N. Szeszenia-Dabrowska, and et al. Obesity and cancer: Mendelian randomization approach utilizing the fto genotype. *Int J Epidemiol*, 38(4):971–975, Aug 2009.

[36] M.-J. A. Brion, K. Shakhbazov, and P. M. Visscher. Calculating statistical power in Mendelian randomization studies. *Int J Epidemiol*, 42(5):1497–1501, 2013.

[37] P. Brøndum-Jacobsen, M. Benn, S. Afzal, and B. G. Nordestgaard. No evidence that genetically reduced 25-hydroxyvitamin D is associated with increased risk of ischaemic heart disease or myocardial infarction: a Mendelian randomization study. *Int J Epidemiol*, 44(2):651–661, Apr 2015.

[38] E. J. Brunner, M. Kivimki, D. R. Witte, D. A. Lawlor, G. Davey Smith, J. A. Cooper, and et al. Inflammation, insulin resistance, and diabetes–Mendelian randomization using CRP haplotypes points upstream. *PLoS Med*, 5(8):e155, Aug 2008.

[39] M. Bun and F. Windmeijer. A comparison of bias approximations for the two-stage least squares (2SLS) estimator. *Economics Letters*, 113(1):76–79, 2011.

[40] S. Burgess. Re: "credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions". *Am J Epidemiol*, 176(5): 456–457, Sep 2012.

[41] S. Burgess. Sample size and power calculations in Mendelian randomization with a single instrumental variable and a binary outcome. *Int J Epidemiol*, page dyu005, 2014.

[42] S. Burgess and S. Thompson. Improving bias and coverage in instrumental variable analysis with weak instruments for continuous and binary outcomes. *Stat Med*, 31(15):1582–1600, 2012.

[43] S. Burgess and S. G. Thompson. Bias in causal estimates from mendelian randomization studies with weak instruments. *Stat Med*, 30(11):1312–1323, May 2011.

[44] S. Burgess and S. G. Thompson. Use of allele scores as instrumental variables for Mendelian randomization. *Int J Epidemiol*, 42(4):1134–1144, Aug 2013.

[45] S. Burgess and S. G. Thompson. *Mendelian randomization: methods for using genetic variants in causal estimation*. CRC Press, 2015.

[46] S. Burgess and S. G. Thompson. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am J Epidemiol*, 181(4):251–260, 2015.

[47] S. Burgess, S. G. Thompson, C. R. P. C. H. D. G. Collaboration, et al. Bayesian methods for meta-analysis of causal relationships estimated using genetic instrumental variables. *Stat Med*, 29(12):1298–1311, May 2010.

[48] S. Burgess, S. G. Thompson, C. R. P. C. H. D. G. Collaboration, et al. Avoiding bias from weak instruments in Mendelian randomization studies. *Int J Epidemiol*, 40(3):755–764, Jun 2011.

[49] S. Burgess, A. Butterworth, and S. G. Thompson. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol*, 37(7):658–665, Nov 2013.

[50] S. Burgess, N. M. Davies, and S. G. Thompson. Instrumental variable analysis with a nonlinear exposure–outcome relationship. *Epidemiology (Cambridge, Mass.)*, 25(6):877, 2014.

[51] S. Burgess, R. M. Daniel, A. S. Butterworth, S. G. Thompson, E.-I. Consortium, et al. Network mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways. *Int J Epidemiol*, 44(2):484–495, 2015.

[52] S. Burgess, F. Dudbridge, and S. G. Thompson. Combining information on multiple instrumental variables in Mendelian randomization: comparison of allele score and summarized data methods. *Stat Med*, 2015.

[53] S. Burgess, D. S. Small, and S. G. Thompson. A review of instrumental variable estimators for Mendelian randomization. *Stat Methods Med Res*, page 0962280215597579, 2015.

[54] S. Burgess, A. S. Butterworth, and J. R. Thompson. Beyond Mendelian randomization: how to interpret evidence of shared genetic predictors. *Journal of Clinical Epidemiology*, 69:208–216, 2016.

[55] S. Burgess, S. G. Thompson, C. C. G. Collaboration, et al. Methods for meta-analysis of individual participant data from Mendelian randomisation studies with binary outcomes. *Stat Methods Med Res*, 25(1):272–293, 2016.

[56] A. Burton, D. G. Altman, P. Royston, and R. L. Holder. The design of simulation studies in medical statistics. *Stat Med*, 25(24):4279–4292, Dec 2006.

[57] R. M. Cantor, K. Lange, and J. S. Sinsheimer. Prioritizing gwas results: a review of statistical methods and recommendations for their application. *Am J Hum Genet*, 86(1):6–22, 2010.

[58] L. R. Cardon and L. J. Palmer. Population stratification and spurious allelic association. *Lancet*, 361(9357):598–604, Feb 2003.

[59] J. P. Casas, E. Ninio, A. Panayiotou, J. Palmen, J. A. Cooper, S. L. Ricketts, and et al. PLA2G7 genotype, lipoprotein-associated phospholipase A2 activity, and coronary heart disease risk in 10 494 cases and 15 624 controls of European Ancestry. *Circulation*, 121(21):2284–2293, Jun 2010.

[60] G. Chamberlain and G. Imbens. Random effects estimators with many instrumental variables. *Econometrica*, 72(1):295–306, 2004.

[61] J. C. Chao and N. R. Swanson. Consistent estimation with a large number of weak instruments. *Econometrica*, 73(5):1673–1692, 2005.

[62] J. C. Chao, N. R. Swanson, J. A. Hausman, W. K. Newey, and T. Woutersen. Asymptotic distribution of JIVE in a heteroskedastic IV regression with many instruments. *Econometric Theory*, 28(01):42–86, 2012.

[63] L. Chen, G. Davey Smith, R. M. Harbord, and S. J. Lewis. Alcohol intake and blood pressure: a systematic review implementing a Mendelian randomization approach. *PLoS Med*, 5(3):e52, Mar 2008.

[64] S. Chib and E. Greenberg. Semiparametric modeling and estimation of instrumental variable models. *Journal of Computational and Graphical Statistics*, 16(1):86–114, 2007.

[65] C.-M. Chung, T.-H. Lin, J.-W. Chen, H.-B. Leu, W.-H. Yin, H.-Y. Ho, S.-H. Sheu, W.-C. Tsai, J.-H. Chen, S.-J. Lin, et al. Common quantitative trait locus downstream of RETN gene identified by genome-wide association study is associated with risk of type 2 diabetes mellitus in Han Chinese: a Mendelian randomization effect. *Diabetes/Metabolism Research and Reviews*, 30(3):232–240, 2014.

[66] P. S. Clarke, T. M. Palmer, F. Windmeijer, et al. Estimating structural mean models with multiple instrumental variables using the generalised method of moments. *Statistical Science*, 30(1):96–117, 2015.

[67] R. Clarke, D. A. Bennett, S. Parish, P. Verhoef, M. Dtsch-Klerk, M. Lathrop, and et al. Homocysteine and coronary heart disease: meta-analysis of mthfr case-control studies, avoiding publication bias. *PLoS Med*, 9(2):e1001177, Feb 2012.

[68] D. Clayton and M. Hills. *Statistical models in epidemiology*. Oxford university press, 1993.

[69] D. Clayton, J. Chapman, and J. Cooper. Use of unphased multilocus genotype data in indirect association studies. *Genet Epidemiol*, 27(4):415–428, 2004.

[70] M. Clyde. Bayesian model averaging and model search strategies. *Bayesian statistics*, 6:157, 1999.

[71] C. B. Cole, M. Nikpay, A. F. R. Stewart, and R. McPherson. Increased genetic risk for obesity in premature coronary artery disease. *Eur J Hum Genet*, 24 (4):587–591, Apr 2016.

[72] C. R. P. C. H. D. G. Collaboration et al. Association between C reactive protein and coronary heart disease: Mendelian randomisation analysis based on individual participant data. *BMJ*, 342:d548, 2011.

[73] D. Conen, P. Vollenweider, V. Rousson, P. Marques-Vidal, F. Paccaud, G. Waeber, and M. Bochud. Use of a Mendelian randomization approach to assess the causal relation of gamma-Glutamyltransferase with blood pressure and serum insulin levels. *Am J Epidemiol*, 172(12):1431–1441, Dec 2010.

[74] T. Conley, C. Hansen, R. McCulloch, and P. Rossi. A semi-parametric Bayesian approach to the instrumental variable problem. *Journal of Econometrics*, 144(1):276–305, 2008.

[75] C. Cruchaga, J. S. K. Kauwe, P. Nowotny, K. Bales, E. H. Pickering, K. Mayo, and et al. Cerebrospinal fluid apoe levels: an endophenotype for genetic studies for alzheimer's disease. *Hum Mol Genet*, 21(20):4558–4571, Oct 2012.

[76] G. Cuellar-Partida, Y. Lu, P. F. Kho, A. W. Hewitt, H. Wichmann, D. Stambolian, et al. Assessing the genetic predisposition of education on myopia: a Mendelian randomization study. *Genet Epidemiol*, 40(1):66–72, 2016.

[77] M. Dahl, J. Vestbo, J. Zacho, P. Lange, A. Tybjrg-Hansen, and B. G. Nordestgaard. C reactive protein and chronic obstructive pulmonary disease: a Mendelian randomisation approach. *Thorax*, 66(3):197–204, Mar 2011.

[78] Z. Dastani, T. Johnson, F. Kronenberg, C. P. Nelson, T. L. Assimes, W. März, J. B. Richards, C. Consortium, A. Consortium, et al. The shared allelic architecture of adiponectin levels and coronary artery disease. *Atherosclerosis*, 229(1):145–148, 2013.

[79] G. Davey Smith and S. Ebrahim. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*, 32(1):1–22, Feb 2003.

[80] G. Davey Smith and S. Ebrahim. Mendelian randomization: prospects, potentials, and limitations. *Int J Epidemiol*, 33(1):30–42, Feb 2004.

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

[82] G. Davey Smith, S. Ebrahim, S. Lewis, A. L. Hansell, L. J. Palmer, and P. R. Burton. Genetic epidemiology and public health: hope, hype, and future prospects. *Lancet*, 366(9495):1484–1498, 2005.

[83] R. Davidson and J. MacKinnon. *Estimation and Inference in Econometrics*. Oxford University Press, 1993. ISBN 9780195060119.

[84] N. M. Davies, G. Davey Smith, F. Windmeijer, and R. M. Martin. Issues in the reporting and conduct of instrumental variable studies: a systematic review. *Epidemiology*, 24(3):363–369, May 2013.

[85] N. M. Davies, S. von Hinke Kessler Scholder, H. Farbmacher, S. Burgess, F. Windmeijer, and G. Davey Smith. The many weak instruments problem and Mendelian randomization. *Stat Med*, 34(3):454–468, Feb 2015.

[86] F. R. Day, D. A. Hinds, J. Y. Tung, L. Stolk, U. Styrkarsdottir, R. Saxena, et al. Causal mechanisms and balancing selection inferred from genetic associations with polycystic ovary syndrome. *Nat Commun*, 6:8464, 2015.

[87] F. R. Day, K. S. Ruth, D. J. Thompson, K. L. Lunetta, N. Pervjakova, D. I. Chasman, L. Stolk, H. K. Finucane, P. Sulem, B. Bulik-Sullivan, et al. Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nat Genet*, 47(11):1294–1303, 2015.

[88] F. R. Day, H. Helgason, D. I. Chasman, L. M. Rose, P.-R. Loh, R. A. Scott, et al. Physical and neurobehavioral determinants of reproductive onset and success. *Nat Genet*, 48(6):617–623, Jun 2016.

[89] P. I. de Bakker, R. Yelensky, I. Pe'er, S. B. Gabriel, M. J. Daly, and D. Altshuler. Efficiency and power in genetic association studies. *Nat Genet*, 37(11): 1217–1223, 2005.

[90] V. Didelez and N. Sheehan. Mendelian randomization as an instrumental variable approach to causal inference. *Stat Methods Med Res*, 16(4):309–330, 2007.

[91] E. L. Ding, Y. Song, J. E. Manson, D. J. Hunter, C. C. Lee, N. Rifai, J. E. Buring, J. M. Gaziano, and S. Liu. Sex hormone-binding globulin and risk of type 2 diabetes in women and men. *N Engl J Med*, 361(12):1152–1163, Sep 2009.

[92] S. G. Donald and W. K. Newey. Choosing the number of instruments. *Econometrica*, 69(5):1161–1191, 2001.

[93] F. Drenos, P. J. Talmud, J. P. Casas, L. Smeeth, J. Palmen, S. E. Humphries, and A. D. Hingorani. Integrated associations of genotypes with multiple blood biomarkers linked to coronary heart disease risk. *Hum Mol Genet*, 18(12): 2305–2316, Jun 2009.

[94] W. B. Drøyvold, K. Midthjell, T. I. L. Nilsen, and J. Holmen. Change in body mass index and its impact on blood pressure: a prospective population study. *Int J Obes*, 29(6):650–655, 2005.

[95] F. Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLoS Genet*, 9(3):e1003348, 2013.

[96] D. B. Dunson. Commentary: practical advantages of bayesian analysis of epidemiologic data. *Am J Epidemiol*, 153(12):1222–1226, 2001.

[97] M. Elovainio, J. E. Ferrie, A. Singh-Manoux, M. Shipley, G. D. Batty, J. Head, et al. Socioeconomic differences in cardiometabolic factors: social causation or health-related selection? evidence from the whitehall ii cohort study, 1991-2004. *Am J Epidemiol*, 174(7):779–789, Oct 2011.

[98] D. M. Evans and G. Davey Smith. Mendelian randomization: new applications in the coming age of hypothesis-free causality. *Annual review of genomics and human genetics*, 16:327–350, 2015.

[99] W. J. Ewens. *Mathematical Population Genetics 1: Theoretical Introduction*, volume 27. Springer Science & Business Media, 2012.

[100] T. Fall, S. Hgg, R. Mgi, A. Ploner, K. Fischer, M. Horikoshi, others, E. N. for Genetic, and G. E. E. consortium. The role of adiposity in cardiometabolic traits: a mendelian randomization analysis. *PLoS Med*, 10(6):e1001474, 2013.

[101] B. A. Ference, W. Yoo, I. Alesh, N. Mahajan, K. K. Mirowska, A. Mewada, J. Kahn, L. Afonso, K. A. Williams, Sr, and J. M. Flack. Effect of long-term exposure to lower low-density lipoprotein cholesterol beginning early in life on the risk of coronary heart disease: a mendelian randomization analysis. *J Am Coll Cardiol*, 60(25):2631–2639, Dec 2012.

[102] B. A. Ference, F. Majeed, R. Penumetcha, J. M. Flack, and R. D. Brook. Effect of naturally random allocation to lower low-density lipoprotein cholesterol on the risk of coronary heart disease mediated by polymorphisms in NPC1L1, HMGCR, or both: a 2 2 factorial Mendelian randomization study. *J Am Coll Cardiol*, 65(15):1552–1561, Apr 2015.

[103] E. Fisher, N. Stefan, K. Saar, D. Drogan, M. B. Schulze, A. Fritsche, et al. Association of AHSG gene polymorphisms with fetuin-A plasma levels and cardiovascular diseases in the EPIC-Potsdam study. *Circ Cardiovasc Genet*, 2(6):607–613, Dec 2009.

[104] T. M. Frayling, N. J. Timpson, M. N. Weedon, E. Zeggini, R. M. Freathy, C. M. Lindgren, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, 316 (5826):889–894, 2007.

[105] G. Freeman, B. J. Cowling, and C. M. Schooling. Power and sample size calculations for Mendelian randomization studies using one genetic instrument. *Int J Epidemiol*, 42(4):1157–1163, 2013.

[106] B. L. Fridley. Bayesian variable and model selection methods for genetic association studies. *Genet Epidemiol*, 33(1):27–37, 2009.

[107] S. Friedrichs, D. Malzahn, E. W. Pugh, M. Almeida, X. Q. Liu, and J. N. Bailey. Filtering genetic variants and placing informative priors based on putative biological function. *BMC Genet*, 17(2):33, 2016.

[108] S. B. Gabriel, S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, et al. The structure of haplotype blocks in the human genome. *Science*, 296 (5576):2225–2229, 2002.

[109] W. Gan, Y. Guan, Q. Wu, P. An, J. Zhu, L. Lu, L. Jing, Y. Yu, S. Ruan, D. Xie, et al. Association of TMPRSS6 polymorphisms with ferritin, hemoglobin, and type 2 diabetes risk in a Chinese Han population. *Am J Clin Nutr*, 95(3): 626–632, 2012.

[110] H. Gao, T. Fall, R. M. van Dam, A. Flyvbjerg, B. Zethelius, E. Ingelsson, and S. Hägg. Evidence of a causal relationship between adiponectin levels and insulin sensitivity: a Mendelian randomization study. *Diabetes*, 62(4): 1338–1344, Apr 2013.

[111] E. I. George and R. E. McCulloch. Approaches for Bayesian variable selection. *Statistica Sinica*, pages 339–373, 1997.

[112] R. A. Gibbs, J. W. Belmont, P. Hardenbol, T. D. Willis, F. Yu, H. Yang, L.-Y. Ch'ang, W. Huang, B. Liu, Y. Shen, et al. The International HapMap project. *Nature*, 426(6968):789–796, 2003.

[113] M. M. Glymour, E. J. Tchetgen Tchetgen, and J. M. Robins. Credible mendelian randomization studies: approaches for evaluating the instrumental variable assumptions. *Am J Epidemiol*, 175(4):332–339, Feb 2012.

[114] M. Gögele, C. Minelli, A. Thakkinstian, A. Yurkiewich, C. Pattaro, P. P. Pramstaller, J. Little, J. Attia, and J. R. Thompson. Methods for meta-analyses of genome-wide association studies: critical assessment of empirical evidence. *Am J Epidemiol*, 175(8):739–749, 2012.

[115] M. Greco, F. Del, C. Minelli, N. A. Sheehan, and J. R. Thompson. Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. *Stat Med*, 34(21):2926–2940, 2015.

[116] W. H. Greene. *Econometric analysis*. Pearson, Boston, Mass, 7th edition, 2012.

[117] S. Greenland. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol*, 29(4):722–729, 2000.

[118] S. Greenland. Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(2): 267–306, 2005.

[119] S. Greenland. Bayesian perspectives for epidemiological research: I. foundations and basic methods. *Int J Epidemiol*, 35(3):765–775, 2006.

[120] S. Greenland and H. Morgenstern. Confounding in health research. *Annual review of public health*, 22(1):189–212, 2001.

[121] S. Greenland and R. Neutra. Control of confounding in the assessment of medical technology. *Int J Epidemiol*, 9(4):361–367, 1980.

[122] S. Greenland and J. M. Robins. Identifiability, exchangeability, and epidemiological confounding. *Int J Epidemiol*, 15(3):413–419, 1986.

[123] I. Guessous, M. Dobrinas, Z. Kutalik, M. Pruijm, G. Ehret, M. Maillard, et al. Caffeine intake and CYP1A2 variants associated with high caffeine intake protect non-smokers from hypertension. *Hum Mol Genet*, 21(14):3283–3292, Jul 2012.

[124] J. Hahn, J. Hausman, and G. Kuersteiner. Estimation with weak instruments: Accuracy of higher-order bias and MSE approximations. *The Econometrics Journal*, 7(1):272–306, 2004.

[125] C. N. Hales, D. J. Barker, P. M. Clark, L. J. Cox, C. Fall, C. Osmond, and P. D. Winter. Fetal and infant growth and impaired glucose tolerance at age 64. *BMJ*, 303(6809):1019–1022, Oct 1991.

[126] C. Han and P. C. Phillips. GMM with many moment conditions. *Econometrica*, 74(1):147–192, 2006.

[127] C. Hansen and D. Kozbur. Instrumental variables estimation with many weak instruments using regularized JIVE. *Journal of Econometrics*, 182(2):290–308, 2014.

[128] C. Hansen, J. Hausman, and W. Newey. Estimation with many instrumental variables. *Journal of Business & Economic Statistics*, 26(4), 2008.

[129] L. P. Hansen, J. Heaton, and A. Yaron. Finite-sample properties of some alternative GMM estimators. *Journal of Business & Economic Statistics*, 14 (3):262–280, 1996.

[130] D. Hartl and E. Jones. *Genetics: Principles and Analysis*. Life Science Series. Jones and Bartlett Publishers, 1998.

[131] A. T. Hattersley and J. E. Tooke. The fetal insulin hypothesis: an alternative explanation of the association of low bir thweight with diabetes and vascular disease. *Lancet*, 353(9166):1789–1792, 1999.

[132] J. A. Hausman, W. K. Newey, T. Woutersen, J. C. Chao, and N. R. Swanson. Instrumental variable estimation with heteroskedasticity and many instruments. *Quantitative Economics*, 3(2):211–255, 2012.

[133] K. Heikkilä, K. Silander, V. Salomaa, P. Jousilahti, S. Koskinen, E. Pukkala, and M. Perola. C-reactive protein-associated genetic variants and cancer risk: findings from FINRISK 1992, FINRISK 1997 and Health 2000 studies. *Eur J Cancer*, 47(3):404–412, Feb 2011.

[134] C. Herder, N. Klopp, J. Baumert, M. Müller, N. Khuseyinova, C. Meisinger, et al. Effect of macrophage migration inhibitory factor (MIF) gene variants and MIF serum concentrations on the risk of type 2 diabetes: results from the MONICA/KORA Augsburg Case-Cohort Study, 1984-2002. *Diabetologia*, 51 (2):276–284, Feb 2008.

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

[136] A. B. Hill. The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*, 58(5):295, 1965.

[137] W. G. Hill and A. Robertson. Linkage disequilibrium in finite populations. *Theor Appl Genet*, 38(6):226–231, Jun 1968.

[138] J. N. Hirschhorn and M. J. Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108, 2005.

[139] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: a tutorial. *Statistical Science*, pages 382–401, 1999.

[140] M. V. Holmes, L. A. Lange, T. Palmer, M. B. Lanktree, K. E. North, B. Almoguera, et al. Causal effects of body mass index on cardiometabolic traits and events: a mendelian randomization analysis. *Am J Hum Genet*, 94(2): 198–208, Feb 2014.

[141] L. Hoogerheide, J. Kaashoek, and H. van Dijk. On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank: An application of flexible sampling methods using neural networks. *Journal of Econometrics*, 139(1):154–180, 2007.

[142] L. Hoogerheide, F. Kleibergen, and H. van Dijk. Natural conjugate priors for the instrumental variables regression model applied to the Angrist-Krueger data. *Journal of Econometrics*, 138(1):63–103, 2007.

[143] M. Horikoshi, H. Yaghootkar, D. O. Mook-Kanamori, U. Sovio, H. R. Taal, and e. a. Hennig, Branwen J. New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism. *Nat Genet*, 45(1):76–82, Jan 2013.

[144] M. Horikoshi, R. Mägi, M. van de Bunt, I. Surakka, A.-P. Sarin, A. Mahajan, L. Marullo, G. Thorleifsson, S. Hägg, J.-J. Hottenga, et al. Discovery and fine-mapping of glycaemic and obesity-related trait loci using high-density imputation. *PLoS Genet*, 11(7):e1005230, 2015.

[145] M. Horikoshi, R. N. Beaumont, F. R. Day, N. M. Warrington, M. N. Kooijman, J. Fernandez-Tajes, B. Feenstra, N. R. van Zuydam, K. J. Gaulton, N. Grarup, et al. Genome-wide associations for birth weight and correlations with adult disease. *Nature*, 538(7624):248–252, 2016.

[146] F. Hoti and M. Sillanpää. Bayesian mapping of genotype × expression interactions in quantitative and qualitative traits. *Heredity*, 97(1):4–18, 2006.

[147] Q. Huang, J. Mi, X. Wang, F. Liu, D. Wang, D. Yan, B. Wang, S. Zhang, and G. Tian. Genetically lowered concentrations of circulating sRAGE might cause an increased risk of cancer: Meta-analysis using Mendelian randomization. *J Int Med Res*, 44(2):179–191, Apr 2016.

[148] Y. Huang, M. Xu, L. Xie, T. Wang, X. Huang, X. Lv, Y. Chen, L. Ding, L. Lin, W. Wang, Y. Bi, Y. Sun, Y. Zhang, and G. Ning. Obesity and peripheral arterial disease: A Mendelian randomization analysis. *Atherosclerosis*, 247: 218–224, Apr 2016.

[149] M. Imamura, A. Takahashi, T. Yamauchi, K. Hara, K. Yasuda, N. Grarup, W. Zhao, X. Wang, A. Huerta-Chagoya, C. Hu, et al. Genome-wide association

studies in the Japanese population identify seven novel loci for type 2 diabetes. *Nat Commun*, 7, 2016.

[150] International HapMap Consortium et al. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005.

[151] V. E. Jackson. *Investigation of the influence of rare genetic variants on lung function and chronic obstructive pulmonary disease*. PhD thesis, University of Leicester, 2016.

[152] G. C. Johnson, L. Esposito, B. J. Barratt, A. N. Smith, J. Heward, G. Di Genova, H. Ueda, H. J. Cordell, I. A. Eaves, F. Dudbridge, et al. Haplotype tagging for the identification of common disease genes. *Nat Genet*, 29(2):233–237, 2001.

[153] E. M. Jones, J. R. Thompson, V. Didelez, and N. A. Sheehan. On the choice of parameterisation and priors for the Bayesian analyses of Mendelian randomisation studies. *Stat Med*, 31(14):1483–1501, Jun 2012.

[154] A. B. Jørgensen, R. Frikke-Schmidt, A. S. West, P. Grande, B. G. Nordestgaard, and A. Tybjærg-Hansen. Genetically elevated non-fasting triglycerides and calculated remnant cholesterol as causal risk factors for myocardial infarction. *Eur Heart J*, 34(24):1826–1833, Jun 2013.

[155] A. M. Jurek, G. Maldonado, S. Greenland, and T. R. Church. Exposure-measurement error is frequently ignored when interpreting epidemiologic study results. *Eur J Clin Nutr*, 21(12):871–876, 2006.

[156] P. R. Kamstrup and B. G. Nordestgaard. Elevated Lipoprotein(a) levels, LPA risk genotypes, and increased risk of heart failure in the general population. *JACC Heart Fail*, 4(1):78–87, Jan 2016.

[157] P. R. Kamstrup, A. Tybjærg-Hansen, and B. G. Nordestgaard. Elevated lipoprotein(a) and risk of aortic valve stenosis in the general population. *J Am Coll Cardiol*, 63(5):470–477, Feb 2014.

[158] G. Kapetanios and M. Marcellino. Cross-sectional averaging and instrumental variable estimation with many weak instruments. *Economics Letters*, 108(1): 36–39, 2010.

[159] G. Kapetanios and M. Marcellino. Factor-GMM estimation with large sets of possibly weak instruments. *Computational Statistics & Data Analysis*, 54(11): 2655–2675, 2010.

[160] A. Karl and A. Lenkoski. Instrumental variable Bayesian model averaging via conditional Bayes factors. *arXiv preprint arXiv:1202.5846*, 2012.

[161] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[162] M. B. Katan. Apolipoprotein E isoforms, serum cholesterol, and cancer. 1986. *Int J Epidemiol*, 33(1):9, Feb 2004.

[163] K. Kato. Quasi-Bayesian analysis of nonparametric instrumental variables models. *Annals of Statistics*, 41(5):2359–2390, 2013.

[164] J. F. C. Kingman. The coalescent. *Stochastic processes and their applications*, 13(3):235–248, 1982.

[165] M. Kivimäki, D. A. Lawlor, G. Davey Smith, M. Kumari, A. Donald, A. Britton, et al. Does high C-reactive protein concentration increase atherosclerosis? The Whitehall II Study. *PLoS One*, 3(8):e3013, 2008.

[166] M. Kivimäki, M. Jokela, M. Hamer, J. Geddes, K. Ebmeier, M. Kumari, A. Singh-Manoux, A. Hingorani, and G. D. Batty. Examining overweight and obesity as risk factors for common mental disorders using fat mass and obesity-associated (FTO) genotype-instrumented analysis: The Whitehall II Study, 1985-2004. *Am J Epidemiol*, 173(4):421–429, Feb 2011.

[167] M. Kivimäki, C. G. Magnussen, M. Juonala, M. Kähönen, J. Kettunen, B.-M. Loo, T. Lehtimäki, J. Viikari, and O. T. Raitakari. Conventional and Mendelian randomization analyses suggest no association between lipoprotein (a) and early atherosclerosis: the Young Finns Study. *Int J Epidemiol*, 40(2): 470–478, 2011.

[168] F. Kleibergen. Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica*, 70(5):1781–1803, 2002.

[169] F. Kleibergen and E. Zivot. Bayesian and classical approaches to instrumental variable regression. *Journal of Econometrics*, 114(1):29–72, 2003.

[170] J. Klovaite, B. G. Nordestgaard, A. Tybjærg-Hansen, and M. Benn. Elevated fibrinogen levels are associated with risk of pulmonary embolism, but not with deep venous thrombosis. *Am J Respir Crit Care Med*, 187(3):286–293, Feb 2013.

[171] W. C. Knowler, R. C. Williams, D. J. Pettitt, and A. G. Steinberg. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am J Hum Genet*, 43(4):520–526, Oct 1988.

[172] G. Koop, R. Leon-Gonzalez, and R. Strachan. Bayesian model averaging in the instrumental variable regression model. *Journal of Econometrics*, 171(2): 237–250, 2012.

[173] M. Korostishevsky, C. J. Steves, I. Malkin, T. Spector, F. M. Williams, and G. Livshits. Genomics and metabolomics of muscular mass in a community-based sample of UK females. *Eur J Hum Genet*, 24(2):277–283, 2016.

[174] A. Kraay. Instrumental variables regressions with uncertain exclusion restrictions: A Bayesian approach. *Journal of Applied Econometrics*, 27(1):108–128, 2012.

[175] L. Kruglyak, D. A. Nickerson, et al. Variation is the spice of life. *Nat Genet*, 27(3):234–235, 2001.

[176] A. M. Kueider, T. Tanaka, Y. An, M. H. Kitner-Triolo, E. Palchamy, L. Ferrucci, and M. Thambisetty. State- and trait-dependent associations of vitamin-d with brain function during aging. *Neurobiol Aging*, 39:38–45, Mar 2016.

[177] N. Kunitomo. An optimal modification of the LIML estimation for many instruments and persistent heteroscedasticity. *Annals of the Institute of Statistical Mathematics*, 64(5):881–910, 2012.

[178] T. Lancaster. *Introduction to Modern Bayesian Econometrics*. Wiley, 2004.

[179] J. M. Lane, I. Vlasac, S. G. Anderson, S. D. Kyle, W. G. Dixon, D. A. Bechtold, et al. Genome-wide association analysis identifies novel loci for chronotype in 100,420 individuals from the UK Biobank. *Nat Commun*, 7:10889, 2016.

[180] L. E. Laugsand, J. H. Ix, T. M. Bartz, L. Djousse, J. R. Kizer, R. P. Tracy, A. Dehghan, K. Rexrode, O. L. Lopez, E. B. Rimm, et al. Fetuin-A and risk of coronary heart disease: A Mendelian randomization analysis and a pooled analysis of AHSG genetic variants in 7 prospective studies. *Atherosclerosis*, 243(1):44–52, 2015.

[181] D. A. Lawlor, G. Davey Smith, K. R. Bruckdorfer, D. Kundu, and S. Ebrahim. Those confounded vitamins: what can we learn from the differences between observational versus randomised trial evidence? *Lancet*, 363(9422):1724–1727, 2004.

[182] D. A. Lawlor, R. M. Harbord, J. A. C. Sterne, N. Timpson, and G. Davey Smith. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med*, 27(8):1133–1163, Apr 2008.

[183] D. Lee and S.-A. Bacanu. Association testing strategy for data from dense marker panels. *PLoS One*, 8(11):e80540, 2013.

[184] A. Lenkoski, A. Karl, and A. Neudecker. *ivbma: Bayesian Instrumental Variable Estimation and Model Determination via Conditional Bayes Factors*, 2014. R package version 1.05.

[185] L. Liang, S. Zöllner, and G. R. Abecasis. Genome: a rapid coalescent-based whole genome simulator. *Bioinformatics*, 23(12):1565–1567, Jun 2007.

[186] Y. Liao and W. Jiang. Posterior consistency of nonparametric conditional moment restricted models. *Annals of Statistics*, 39(6):3003–3031, 2011.

[187] L. S. Lim, E.-S. Tai, T. Aung, W. T. Tay, S. M. Saw, M. Seielstad, and T. Y. Wong. Relation of age-related cataract with obesity and obesity genes in an Asian population. *Am J Epidemiol*, 169(10):1267–1274, 2009.

[188] Y.-P. Liu, Y.-M. Gu, L. Thijs, M. H. Knapen, E. Salvi, L. Citterio, T. Petit, S. D. Carpini, Z. Zhang, L. Jacobs, et al. Inactive Matrix Gla Protein is causally related to adverse health outcomes a Mendelian randomization study in a Flemish population. *Hypertension*, 65(2):463–470, 2015.

[189] G. Livshits, A. J. Macgregor, C. Gieger, I. Malkin, A. Moayyeri, H. Grallert, R. T. Emeny, T. Spector, G. Kastenmller, and F. M. K. Williams. An omics

investigation into chronic widespread musculoskeletal pain reveals epiandrosterone sulfate as a potential biomarker. *Pain*, 156(10):1845–1851, Oct 2015.

[190] H. Lopes and N. Polson. Bayesian instrumental variables: Priors and likelihoods. *Econometric Reviews*, 33(1-4):100–121, 2014.

[191] L. Love-Gregory, R. Sherva, T. Schappe, J.-S. Qi, J. McCrea, S. Klein, M. A. Connelly, and N. A. Abumrad. Common CD36 SNPs reduce protein expression and may contribute to a protective atherogenic profile. *Hum Mol Genet*, 20 (1):193–201, Jan 2011.

[192] D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter. *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2012.

[193] J. B. Maller, G. McVean, J. Byrnes, D. Vukcevic, K. Palin, Z. Su, J. M. Howson, A. Auton, S. Myers, A. Morris, et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet*, 44(12):1294–1301, 2012.

[194] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265): 747–753, 2009.

[195] J. I. Marden. Hypothesis testing: from p values to bayes factors. *Journal of the American Statistical Association*, 95(452):1316–1320, 2000.

[196] S. C. Marott, B. G. Nordestgaard, J. Zacho, J. Friberg, G. B. Jensen, A. Tybjærg-Hansen, and M. Benn. Does elevated C-Reactive protein increase atrial fibrillation risk?: A Mendelian randomization of 47,000 individuals from the general population. *J Am Coll Cardiol*, 56(10):789–795, 2010.

[197] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. Ioannidis, and J. N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5): 356–369, 2008.

[198] P. M. McKeigue, H. Campbell, S. Wild, V. Vitart, C. Hayward, I. Rudan, A. F. Wright, and J. F. Wilson. Bayesian methods for instrumental variable

analysis with genetic instruments ('Mendelian randomization'): example with urate transporter SLC2A9 as an instrumental variable for effect of urate levels on metabolic syndrome. *Int J Epidemiol*, 39(3):907–918, Jun 2010.

[199] C. Menzaghi, S. De Cosmo, M. Copetti, L. Salvemini, C. De Bonis, D. Mangiacotti, G. Fini, F. Pellegrini, and V. Trischitta. Relationship between ADIPOQ gene, circulating high molecular weight adiponectin and albuminuria in individuals with normal kidney function: evidence from a family-based study. *Diabetologia*, 54(4):812–818, Apr 2011.

[200] C. Minelli, A. De Grandi, C. X. Weichenberger, M. Gögele, M. Modenese, J. Attia, J. H. Barrett, M. Boehnke, G. Borsani, G. Casari, et al. Importance of different types of prior knowledge in selecting genome-wide findings for follow-up. *Genet Epidemiol*, 37(2):205–213, 2013.

[201] Y. Moreau and L.-C. Tranchevent. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature Reviews Genetics*, 13 (8):523–536, 2012.

[202] M. J. Moreira. A conditional likelihood ratio test for structural models. *Econometrica*, 71(4):1027–1048, 2003.

[203] A. P. Morris. Fine mapping type 2 diabetes susceptibility loci. In *Genetics in Diabetes*, volume 23, pages 14–28. Karger Publishers, 2014.

[204] N. J. Morris, C. Gray-McGuire, and C. M. Stein. Mendelian randomization in family data. In *BMC proceedings*, volume 3, page S45. BioMed Central Ltd, 2009.

[205] N. J. Morris, R. C. Elston, and C. M. Stein. A framework for structural equation models in general pedigrees. *Human Heredity*, 70(4):278–286, 2010.

[206] A. Nagar. The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica: Journal of the Econometric Society*, pages 575–595, 1959.

[207] W. K. Newey and F. Windmeijer. Generalized method of moments with many weak moment conditions. *Econometrica*, pages 687–719, 2009.

[208] K. Nimptsch, K. Aleksandrova, H. Boeing, J. Janke, Y.-A. Lee, M. Jenab, and et al. Association of CRP genetic variants with blood concentrations of C-reactive protein and colorectal cancer risk. *Int J Cancer*, 136(5):1181–1192, Mar 2015.

[209] A. T. Nordestgaard, M. Thomsen, and B. G. Nordestgaard. Coffee intake and risk of obesity, metabolic syndrome and type 2 diabetes: a Mendelian randomization study. *Int J Epidemiol*, 44(2):551–565, Apr 2015.

[210] I. Ntzoufras. *Bayesian modeling using WinBUGS*, volume 698. John Wiley & Sons, 2011.

[211] L. Oei, N. Campos-Obando, A. Dehghan, E. H. G. Oei, L. Stolk, J. B. J. van Meurs, et al. Dissecting the relationship between high-sensitivity serum C-reactive protein and increased fracture risk: the Rotterdam Study. *Osteoporos Int*, 25(4):1247–1254, Apr 2014.

[212] E. C. Oelsner, T. D. Pottinger, K. M. Burkart, M. Allison, S. G. Buxbaum, N. N. Hansel, and et al. Adhesion molecules, endothelin-1 and lung function in seven population-based cohorts. *Biomarkers*, 18(3):196–203, May 2013.

[213] R. B. O'Hara, M. J. Sillanpää, et al. A review of Bayesian variable selection methods: what, how and which. *Bayesian analysis*, 4(1):85–117, 2009.

[214] E. M. Ooi, S. Afzal, and B. G. Nordestgaard. Elevated remnant cholesterol in 25-hydroxyvitamin D deficiency in the general population: Mendelian randomization study. *Circ Cardiovasc Genet*, 7(5):650–658, Oct 2014.

[215] W. H. Organization. Obesity: preventing and managing the global epidemic: report of a WHO consultation. *WHO technical report series*, 894:253, 1999.

[216] L. Palmer, P. R. Burton, and G. Davey Smith. *An introduction to Genetic Epidemiology*. Policy Press, 2011.

[217] L. J. Palmer and L. R. Cardon. Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet*, 366(9492):1223–1234, 2005.

[218] T. M. Palmer, J. A. C. Sterne, R. M. Harbord, D. A. Lawlor, N. A. Sheehan, S. Meng, R. Granell, G. Davey Smith, and V. Didelez. Instrumental variable estimation of causal risk ratios and causal odds ratios in Mendelian randomization analyses. *Am J Epidemiol*, 173(12):1392–1403, Jun 2011.

[219] T. M. Palmer, D. A. Lawlor, R. M. Harbord, N. A. Sheehan, J. H. Tobias, N. J. Timpson, G. Davey Smith, and J. A. C. Sterne. Using multiple genetic variants as instrumental variables for modifiable risk factors. *Stat Methods Med Res*, 21(3):223–242, Jun 2012.

[220] T. M. Palmer, R. R. Ramsahai, D. A. Lawlor, N. A. Sheehan, and V. Didelez. Re: "credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions". *Am J Epidemiol*, 176(5):457–8; author reply 458, Sep 2012.

[221] N. M. Panduru, N. Sandholm, C. Forsblom, M. Saraheimo, E. H. Dahlström, L. M. Thorn, et al. Kidney injury molecule-1 and the loss of kidney function in diabetic nephropathy: a likely causal link in patients with type 1 diabetes. *Diabetes Care*, 38(6):1130–1137, Jun 2015.

[222] A. Parsa, E. Brown, M. R. Weir, J. C. Fink, A. R. Shuldiner, B. D. Mitchell, and P. F. McArdle. Genotype-based changes in serum uric acid affect blood pressure. *Kidney Int*, 81(5):502–507, Mar 2012.

[223] M. N. Passarelli, P. A. Newcomb, K. W. Makar, A. N. Burnett-Hartman, J. D. Potter, M. P. Upton, et al. Blood lipids and colorectal polyps: testing an etiologic hypothesis using phenotypic measurements and Mendelian randomization. *Cancer Causes Control*, 26(3):467–473, Mar 2015.

[224] I. Pe'er, Y. R. Chretien, P. I. de Bakker, J. C. Barrett, M. J. Daly, and D. M. Altshuler. Biases and reconciliation in estimates of linkage disequilibrium in the human genome. *Am J Hum Genet*, 78(4):588–603, 2006.

[225] B. Peng and M. Kimmel. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics*, 21(18):3686–3687, 2005.

[226] B. Peng, H.-S. Chen, L. E. Mechanic, B. Racine, J. Clarke, E. Gillanders, and E. J. Feuer. Genetic data simulators and their applications: an overview. *Genet Epidemiol*, 39(1):2–10, 2015.

[227] B. L. Pierce, H. Ahsan, and T. J. Vanderweele. Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *Int J Epidemiol*, 40(3):740–752, Jun 2011.

[228] B. L. Pierce, L. Tong, M. Argos, J. Gao, J. Farzana, S. Roy, et al. Arsenic metabolism efficiency has a causal role in arsenic toxicity: Mendelian randomization and gene-environment interaction. *Int J Epidemiol*, 42(6):1862–1871, Dec 2013.

[229] D. Poskitt and C. L. Skeels. Approximating the distribution of the two-stage least squares estimator when the concentration parameter is small. *Journal of Econometrics*, 139(1):217–236, 2007.

[230] J. K. Pritchard and M. Przeworski. Linkage disequilibrium in humans: models and data. *Am J Hum Genet*, 69(1):1–14, 2001.

[231] K. Qu, Q. Pang, T. Lin, L. Zhang, M. Gu, W. Niu, C. Liu, and M. Zhang. Circulating interleukin-10 levels and human papilloma virus and Epstein-Barr virus-associated cancers: evidence from a Mendelian randomization meta-analysis based on 11,170 subjects. *Onco Targets Ther*, 9:1251–1267, 2016.

[232] A. E. Raftery. Bayesian model selection in social research. *Sociological Methodology*, pages 111–163, 1995.

[233] B. Rannala and J. P. Reeve. High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. *Am J Hum Genet*, 69 (1):159–178, 2001.

[234] D. E. Reich, S. B. Gabriel, and D. Altshuler. Quality and completeness of SNP databases. *Nat Genet*, 33(4):457–458, 2003.

[235] C. L. Relton and G. Davey Smith. Two-step epigenetic mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *Int J Epidemiol*, 41(1):161–176, 2012.

[236] N. Rius-Ottenheim, A. J. M. de Craen, J. M. Geleijnse, P. E. Slagboom, D. Kromhout, R. C. van der Mast, et al. C-reactive protein haplotypes and dispositional optimism in obese and nonobese elderly subjects. *Inflamm Res*, 61(1):43–51, Jan 2012.

[237] C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer texts in statistics. Springer, 1999.

[238] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[239] N. A. Rosenberg, L. Huang, E. M. Jewett, Z. A. Szpiech, I. Jankovic, and M. Boehnke. Genome-wide association studies in diverse populations. *Nature Reviews Genetics*, 11(5):356–366, 2010.

[240] P. Rossi. *Bayesian Non-and Semi-parametric Methods and Applications*. Princeton University Press, 2014.

[241] K. J. Rothman. Epidemiologic methods in clinical trials. *Cancer*, 39(S4): 1771–1775, 1977.

[242] K. J. Rothman, S. Greenland, and T. L. Lash. *Modern Epidemiology*. Wolters Kluwer Health, 2008.

[243] H. Sallis, C. Steer, L. Paternoster, G. Davey Smith, and J. Evans. Perinatal depression and omega-3 fatty acids: a Mendelian randomisation study. *J Affect Disord*, 166:124–131, Sep 2014.

[244] M. J. Salois and K. G. Balcombe. A generalized Bayesian instrumental variable approach under student t-distributed errors with application. *The Manchester School*, 83(5):499–522, 2015.

[245] M. Sargolzaei and F. S. Schenkel. QMSim: a large-scale genome simulator for livestock. *Bioinformatics*, 25(5):680–681, 2009.

[246] H. Scharnagl, M. E. Kleber, B. Genser, S. Kickmaier, W. Renner, G. Weihrauch, et al. Association of myeloperoxidase with total and cardiovascular mortality in individuals undergoing coronary angiography–the luric study. *Int J Cardiol*, 174(1):96–105, Jun 2014.

[247] Schizophrenia Psychiatric Genome-Wide Association Study Consortium et al. Genome-wide association study identifies five new schizophrenia loci. *Nat Genet*, 43(10):969–976, 2011.

[248] N. A. Sheehan, S. Meng, and V. Didelez. Mendelian randomisation: a tool for assessing causality in observational epidemiology. *Methods Mol Biol*, 713: 153–166, 2011.

[249] J. S. Shoemaker, I. S. Painter, and B. S. Weir. Bayesian statistics in genetics: a guide for the uninitiated. *Trends in Genetics*, 15(9):354–358, 1999.

[250] K. C. Simon, S. Eberly, X. Gao, D. Oakes, C. M. Tanner, I. Shoulson, S. Fahn, M. A. Schwarzschild, A. Ascherio, and Parkinson Study Group. Mendelian randomization of serum urate and parkinson disease progression. *Ann Neurol*, 76(6):862–868, Dec 2014.

[251] M. Slatkin. Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485, 2008.

[252] Y. E. Song, C. M. Stein, and N. J. Morris. strum: an R package for structural modeling of latent variables for general pedigrees. *BMC Genet*, 16(1):35, 2015.

[253] D. Staiger and J. H. Stock. Instrumental variables regression with weak instruments. *Econometrica: Journal of the Econometric Society*, pages 557–586, 1997.

[254] J. Stamler. Epidemiologic findings on body mass and blood pressure in adults. *Annals of Epidemiology*, 1(4):347–362, 1991.

[255] B. H. Stegeman, F. M. Helmerhorst, H. L. Vos, F. R. Rosendaal, and A. van Hylckama Vlieg. Sex hormone-binding globulin levels are not causally related to venous thrombosis risk in women not using hormonal contraceptives. *Journal of Thrombosis and Haemostasis*, 10(10):2061–2067, 2012.

[256] M. Stephens and D. J. Balding. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681–690, 2009.

[257] J. H. Stock and M. Yogo. Asymptotic distributions of instrumental variables statistics with many instruments. *Identification and inference for econometric models: essays in honor of Thomas Rothenberg*, 2005.

[258] T. Strachan and A. Read. *Human Molecular Genetics 4*. Garland Science/Taylor & Francis Group, 2011.

[259] J. Sunyer, R. Pistelli, E. Plana, M. Andreani, F. Baldari, M. Kolz, W. Koenig, J. Pekkanen, A. Peters, and F. Forastiere. Systemic inflammation, genetic susceptibility and lung function. *Eur Respir J*, 32(1):92–97, Jul 2008.

[260] H. Tang. Confronting ethnicity-specific disease risk. *Nat Genet*, 38(1), 2006.

[261] M.-X. Tang, Y. Stern, K. Marder, K. Bell, B. Gurland, R. Lantigua, H. Andrews, L. Feng, B. Tycko, and R. Mayeux. The APOE-ϵ4 allele and the risk of Alzheimer disease among African Americans, whites, and Hispanics. *JAMA*, 279(10):751–755, 1998.

[262] E. J. T. Tchetgen, S. Walter, S. Vansteelandt, T. Martinussen, and M. Glymour. Instrumental variable estimation in a survival context. *Epidemiology (Cambridge, Mass.)*, 26(3):402, 2015.

[263] A. Thomas and R. B. O'Hara. Openbugs, 2004.

[264] D. C. Thomas and D. V. Conti. Commentary: the concept of 'Mendelian Randomization'. *Int J Epidemiol*, 33(1):21–25, Feb 2004.

[265] D. J. Thompson, T. A. O'Mara, D. M. Glubb, J. N. Painter, T. Cheng, E. Folkerd, and et al. CYP19A1 fine-mapping and Mendelian randomization: estradiol is causal for endometrial cancer. *Endocr Relat Cancer*, 23(2):77–91, Feb 2016.

[266] J. R. Thompson, C. Minelli, K. R. Abrams, M. D. Tobin, and R. D. Riley. Meta-analysis of genetic studies using Mendelian randomization–a multivariate approach. *Stat Med*, 24(14):2241–2254, 2005.

[267] J. R. Thompson, M. Gögele, C. X. Weichenberger, M. Modenese, J. Attia, J. H. Barrett, et al. SNP prioritization using a Bayesian probability of association. *Genet Epidemiol*, 37(2):214–221, 2013.

[268] M. Thomsen, A. Varbo, A. Tybjærg-Hansen, and B. G. Nordestgaard. Low nonfasting triglycerides and reduced all-cause mortality: a Mendelian randomization study. *Clin Chem*, 60(5):737–746, May 2014.

[269] A. P. Thrift, J. Gong, U. Peters, J. Chang-Claude, A. Rudolph, M. L. Slattery, and et al. Mendelian randomization study of body mass index and colorectal cancer risk. *Cancer Epidemiol Biomarkers Prev*, 24(7):1024–1031, Jul 2015.

[270] A. P. Thrift, J. Gong, U. Peters, J. Chang-Claude, A. Rudolph, M. L. Slattery, and et al. Mendelian randomization study of height and risk of colorectal cancer. *Int J Epidemiol*, 44(2):662–672, Apr 2015.

[271] N. J. Timpson, D. A. Lawlor, R. M. Harbord, T. R. Gaunt, I. N. M. Day, L. J. Palmer, et al. C-reactive protein and its role in metabolic syndrome: Mendelian randomisation study. *Lancet*, 366(9501):1954–1959, Dec 2005.

[272] N. J. Timpson, R. Harbord, G. Davey Smith, J. Zacho, A. Tybjaerg-Hansen, and B. G. Nordestgaard. Does greater adiposity increase blood pressure and hypertension risk?: Mendelian randomization using the FTO/MC4R genotype. *Hypertension*, 54(1):84–90, Jul 2009.

[273] N. J. Timpson, B. G. Nordestgaard, R. M. Harbord, J. Zacho, T. M. Frayling, A. Tybjrg-Hansen, and G. Davey Smith. C-reactive protein levels and body mass index: elucidating direction of causation through reciprocal mendelian randomization. *Int J Obes (Lond)*, 35(2):300–308, Feb 2011.

[274] M. D. Tobin, M. Tomaszewski, P. S. Braund, C. Hajat, S. M. Raleigh, T. M. Palmer, M. Caulfield, P. R. Burton, and N. J. Samani. Common variants in genes underlying monogenic hypertension and hypotension and blood pressure in the general population. *Hypertension*, 51(6):1658–1664, Jun 2008.

[275] Y. M. T. A. van Durme, L. Lahousse, K. M. C. Verhamme, L. Stolk, M. Eijgelsheim, D. W. Loth, et al. Mendelian randomization study of interleukin-6 in chronic obstructive pulmonary disease. *Respiration*, 82(6):530–538, 2011.

[276] A. Varbo, M. Benn, A. Tybjrg-Hansen, A. B. Jrgensen, R. Frikke-Schmidt, and B. G. Nordestgaard. Remnant cholesterol as a causal risk factor for ischemic heart disease. *J Am Coll Cardiol*, 61(4):427–436, Jan 2013.

[277] A. Varbo, M. Benn, A. Tybjrg-Hansen, and B. G. Nordestgaard. Elevated remnant cholesterol causes both low-grade inflammation and ischemic heart disease, whereas elevated low-density lipoprotein cholesterol causes ischemic heart disease without inflammation. *Circulation*, 128(12):1298–1309, Sep 2013.

[278] G. Veen, E. J. Giltay, I. M. van Vliet, R. H. Derijk, E. R. Klaassens, J. van Pelt, and F. G. Zitman. C-reactive protein polymorphisms are associated with the cortisol awakening response in basal conditions in human subjects. *Stress*, 14(2):128–135, Mar 2011.

[279] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002.

[280] L. A. Viikari, R. K. Huupponen, J. S. A. Viikari, J. Marniemi, C. Eklund, M. Hurme, T. Lehtimäki, M. Kivimäki, and O. T. Raitakari. Relationship between leptin and C-reactive protein in young Finnish adults. *J Clin Endocrinol Metab*, 92(12):4753–4758, Dec 2007.

[281] S. von Hinke, G. Davey Smith, D. A. Lawlor, C. Propper, and F. Windmeijer. Genetic markers as instrumental variables. *Journal of Health Economics*, 45: 131–148, 2016.

[282] K. N. Vu, C. M. Ballantyne, R. C. Hoogeveen, V. Nambi, K. A. Volcik, E. Boerwinkle, and A. C. Morrison. Causal role of alcohol consumption in an improved lipid profile: The Atherosclerosis Risk in Communities (ARIC) Study. *PLoS One*, 11(2):e0148765, 2016.

[283] F. Wang, N. J. Meyer, K. R. Walley, J. A. Russell, and R. Feng. Causal genetic inference using haplotypes as instrumental variables. *Genet Epidemiol*, 40(1): 35–44, 2016.

[284] T. Wang and R. C. Elston. Improved power by use of a weighted score test for linkage disequilibrium mapping. *Am J Hum Genet*, 80(2):353–360, 2007.

[285] T. Wang, C.-Y. Lin, Y. Zhang, R. Wen, and K. Ye. Design and statistical analysis of pooled next generation sequencing for rare variants. *Journal of Probability and Statistics*, 2012, 2012.

[286] G. L. Wehby, R. L. Ohsfeldt, and J. C. Murray. 'Mendelian randomization' equals instrumental variable analysis with genetic instruments. *Stat Med*, 27 (15):2745–2749, Jul 2008.

[287] P. Welsh, E. Polisecki, M. Robertson, S. Jahn, B. M. Buckley, A. J. M. de Craen, et al. Unraveling the directional link between adiposity and inflammation: a bidirectional mendelian randomization approach. *J Clin Endocrinol Metab*, 95(1):93–99, Jan 2010.

[288] M. Wiesenfarth, C. M. Hisgen, T. Kneib, and C. Cadarso-Suarez. Bayesian nonparametric instrumental variables regression based on penalized splines and dirichlet process mixtures. *Journal of Business & Economic Statistics*, 32 (3):468–482, 2014.

[289] R. R. Wilcox. *Introduction to robust estimation and hypothesis testing*. Academic Press, 2012.

[290] D. J. Wilkinson. Bayesian methods in bioinformatics and computational systems biology. *Brief Bioinform*, 8(2):109–116, 2007.

[291] M. K. Wium-Andersen, D. D. Ørsted, and B. G. Nordestgaard. Elevated C-reactive protein associated with late- and very-late-onset schizophrenia in the general population: a prospective study. *Schizophr Bull*, 40(5):1117–1127, Sep 2014.

[292] M. K. Wium-Andersen, D. D. Orsted, and B. G. Nordestgaard. Elevated C-reactive protein, depression, somatic diseases, and all-cause mortality: a Mendelian randomization study. *Biol Psychiatry*, 76(3):249–257, Aug 2014.

[293] M. K. Wium-Andersen, D. D. ørsted, and B. G. Nordestgaard. Elevated C-reactive protein and late-onset bipolar disorder in 78 809 individuals from the general population. *Br J Psychiatry*, 208(2):138–145, Feb 2016.

[294] J. M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

[295] J. M. Wooldridge. *Introductory Econometrics: A modern approach*. Nelson Education, 2015.

[296] Y. Wu, H. Li, R. J. F. Loos, Q. Qi, F. B. Hu, Y. Liu, and X. Lin. RBP4 variants are significantly associated with plasma RBP4 levels and hypertriglyceridemia risk in Chinese Hans. *J Lipid Res*, 50(7):1479–1486, Jul 2009.

[297] P. Würtz, Q. Wang, A. J. Kangas, R. C. Richmond, J. Skarp, M. Tiainen, and et al. Metabolic signatures of adiposity in young adults: Mendelian randomization analysis and effects of weight change. *PLoS Med*, 11(12):e1001765, Dec 2014.

[298] A. Xiong, Q. Yao, J. He, W. Fu, J. Yu, and Z. Zhang. No causal effect of serum urate on bone-related outcomes among a population of postmenopausal women and elderly men of Chinese Han ethnicity–a Mendelian randomization study. *Osteoporos Int*, 27(3):1031–1039, Mar 2016.

[299] H. Yaghootkar, C. Lamina, R. A. Scott, Z. Dastani, M.-F. Hivert, L. L. Warren, et al. Mendelian randomisation studies do not support a causal role for reduced circulating adiponectin levels in insulin resistance and type 2 diabetes. *Diabetes*, page DB_130128, 2013.

[300] D. Yan, J. Wang, F. Jiang, R. Zhang, T. Wang, S. Wang, et al. A causal relationship between uric acid and diabetic macrovascular disease in Chinese type 2 diabetes patients: A Mendelian randomization analysis. *Int J Cardiol*, 214:194–199, Jul 2016.

[301] J. Yang, T. Ferreira, A. P. Morris, S. E. Medland, P. A. Madden, A. C. Heath, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet*, 44(4):369–375, 2012.

[302] S. L. A. Yeung, K. K. Cheng, J. Zhao, W. Zhang, C. Jiang, T. H. Lam, G. M. Leung, and C. M. Schooling. Genetically predicted 17beta-estradiol and cardiovascular risk factors in women: a Mendelian randomization analysis using young women in Hong Kong and older women in the Guangzhou Biobank Cohort Study. *Annals of Epidemiology*, 26(3):171–175, 2016.

[303] N.-C. Y. You, B. H. Chen, Y. Song, X. Lu, Y. Chen, J. E. Manson, et al. A prospective study of leukocyte telomere length and risk of type 2 diabetes in postmenopausal women. *Diabetes*, 61(11):2998–3004, Nov 2012.

[304] J. Zacho, A. Tybjaerg-Hansen, and B. G. Nordestgaard. C-reactive protein and risk of venous thromboembolism in the general population. *Arterioscler Thromb Vasc Biol*, 30(8):1672–1678, Aug 2010.

[305] A. Zellner, T. Ando, N. Baştürk, L. Hoogerheide, and H. van Dijk. Bayesian analysis of instrumental variable models: Acceptance-rejection within direct Monte Carlo. *Econometric Reviews*, 33(1-4):3–35, 2014.

[306] B. Zhang, X.-O. Shu, R. J. Delahanty, C. Zeng, K. Michailidou, M. K. Bolla, et al. Height and breast cancer risk: evidence from prospective studies and Mendelian randomization. *J Natl Cancer Inst*, 107(11):djv219, 2015.

[307] G. Zhang, J. Bacelis, C. Lengyel, K. Teramo, M. Hallman, Ø. Helgeland, et al. Assessing the causal relationship of maternal height on birth size and

gestational age at birth: A Mendelian randomization analysis. *PLoS Med*, 12 (8):e1001865, Aug 2015.

[308] J. Zhao, C. Jiang, T. H. Lam, B. Liu, K. K. Cheng, L. Xu, S. L. A. Yeung, W. Zhang, G. M. Leung, and C. M. Schooling. Genetically predicted testosterone and systemic inflammation in men: a separate-sample Mendelian randomization analysis in older Chinese men. *PLoS One*, 10(5):e0126442, 2015.