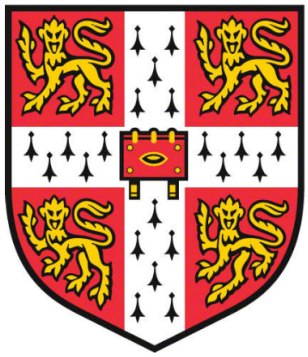
Towards More Accurate Causal Inference with Instrumental Variables and Mendelian Randomisation Analyses



Cunhao Liu

Selwyn College, University of Cambridge

February 2025

This thesis is submitted for the degree of

Doctor of Philosophy

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted, or is being concurrently submitted, for any degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Towards More Accurate Causal Inference with Instrumental Variables and Mendelian Randomisation Analyses

Abstract

Instrumental variable (IV) approach is a commonly used statistical approach to make causal inference in the presence of unmeasured confounding, and the use of genetic variants as instrumental variables in epidemiological studies is known as Mendelian Randomisation (MR). The conventional modelling assumption for using IVs is to assume a linear structural equation model between the exposure and the outcome. In recent years, there have been various developments in the literature that attempt to relax this strong linearity assumption and propose methods that estimate nonlinear causal effects using IVs. We review some of the important nonlinear IV methods, including approaches that nonparametrically generalise the usual 2SLS and GMM estimators using machine-learning algorithms. We also review stratified nonlinear MR, a widely used method for nonlinear causal effect estimation in MR that is based on dividing the population into strata with different average levels of the exposure and estimating average causal effect within each stratum. We present all the methods under a unified causal model and compare their performances through a simulation study.

Stratified nonlinear MR aggregates stratum-specific estimates using a fractional polynomial regression, which is unable to capture nonlinear exposure-outcome relationships beyond fractional polynomials. We propose a nonparametric extension to stratified nonlinear MR, which we call nonparametric MR, that can characterise more complex nonlinear structural functions. The estimand of nonparametric MR, called quantile average causal effect (QACE), approximates the derivative of the true structural function. We firstly develop nonparametric MR with discrete IVs, and then extend this framework to situations where the instrument is continuous. We demonstrate through extensive simulation studies that nonparametric

MR outperforms stratified nonlinear MR in a wide range of scenarios that cover various instrument-exposure relationships and different instrument strengths. We also illustrate our method using real data examples from UK Biobank.

Selection bias is a common issue in statistical applications which biases IV analyses. Starting with linear IV models, we firstly show that adjusting for measured confounders can reduce but cannot eliminate selection bias, and then provide asymptotic guarantees for using inverse probability weighting (IPW) to eliminate selection bias. We argue that in finite samples, IPW may fail to account for selection bias under severe selection even when the selection model is fully known. We demonstrate our findings through simulation studies and real data examples from UK Biobank that cover different selection mechanisms. We further consider selection bias for stratified nonlinear MR and argue that, unlike bias for the standard MR estimate, bias for stratum-specific estimates cannot be corrected using IPW if the we can only stratify on the selected sample. We also illustrate this argument through a simulation study.

Acknowledgement

I'm immensely grateful to my supervisor, Dr Stephen Burgess, who provides foundations for everything written in this thesis. Your rich knowledge and insightful comments are key to the completion of my academic work. You have also shown great patience and incredible flexibility of working, especially during the start (with the global pandemic situation) and the end (with busy writing-up) of my PhD, that I could never thank you enough for.

I'm also hugely indebted to MRC Biostatistics Unit, which provides me with funding through MRC studentship and research travel grants. I want to especially thank Dr Angela Talbot, who has always been responsive to my questions and has been incredibly supportive during difficult times. I also express my gratitudes to Selwyn College of Cambridge and my College tutor, Professor Robert Tasker, who provides me with relentless pastoral support that has been vital during my write-up.

I'm also grateful for being able to work with other incredible people at the BSU. I especially thank Haodong for his valuable insights on the selection bias topic with the doubly-ranked stratification method, and Shaun for his lively discussions on data adaptive methods and for being a nice and fun person who I always enjoy interacting with. My gratitude also goes to Peter - we were PhD representatives together and thank you for making this experience so memorable.

My biggest thanks should go to my parents who have given me undying support during and beyond the whole course of my PhD. And equally big thank you to Shiqing for always being there with me - thank you for everything.

Contents

List of Figures **xi**

List of Tables **xvi**

List of Abbreviations **xxx**

List of Notations **xxxii**

1 Introduction to Causal Inference, Instrumental Variables and Mendelian Randomisation **1**

1.1 Languages and Models for Causality 1

1.1.1 The potential outcomes framework 2

1.1.2 Pearl's *do* operator 3

1.1.3 Structural equation models 4

1.1.4 Graphical models 5

1.1.5 Common terminologies 6

1.2 Instrumental Variables 10

1.2.1 The core IV assumptions 10

1.2.2 Linearity assumption and two-stage-least-square (2SLS) 13

1.2.3 Linear generalised method of moments (Linear GMM) 17

1.2.4 Alternative identification assumptions 19

1.3 Mendelian Randomisation 20

1.3.1 What is Mendelian randomisation? 20

1.3.2 Issues specific to Mendelian randomisation 21

1.3.3 Mendelian randomisation with individual-level data versus summarised data 23

1.4 Overview of the Thesis 26

2 A Critical Review on Nonlinear Instrumental Variables and Mendelian Randomisation Methods **29**

2.1 Introduction 29

2.2 The Nonlinear Causal Model 31

2.3 Nonlinear 2SLS and Control Function Method 32

2.3.1 Nonlinear 2SLS 33

2.3.2 Control function method 34

2.4 Nonparametric 2SLS 36

2.4.1 Sieve IV 37
2.4.2 Kernel IV 39
2.4.3 Deep feature IV 45
2.4.4 Deep IV 48
2.4.5 Related work 50
2.5 Nonlinear GMM 50
2.5.1 Classic framework 51
2.5.2 DeepGMM 53
2.6 Stratified Nonlinear MR 55
2.6.1 Residual stratification 56
2.6.2 Doubly-ranked stratification 57
2.6.3 Estimation of the structural function 59
2.7 Polynomial MR 62
2.8 Simulation Study 63
2.8.1 Simulation design 63
2.8.2 Methods compared 65
2.8.3 Simulation results 67
2.9 Discussion and Concluding Remarks 71

3 Nonparametric Mendelian Randomisation for Characterising Nonlinear Exposure-Outcome Relationships with Discrete Instrumental Variables 74

3.1 Introduction 74

3.2 Related Work 76

3.3 Methodology 78

3.3.1 Rank-preserving assumption 78

3.3.2 Binary IV 79

3.3.3 General discrete IV 83

3.4 Simulation Study 86

3.4.1 Simulation design 86

3.4.2 Hyperparameter tuning for nonparametric regressions 92

3.4.3 Simulation results 94

3.5 Applied Examples from UK Biobank 115

3.5.1 Description of the examples 115

3.5.2 Results of the examples 116

3.6 Discussion and Concluding Remarks 121

4 Nonparametric Mendelian Randomisation for Characterising Nonlinear Exposure-Outcome Relationships: An Extension with Continuous Instrumental Variables 125

4.1 Introduction 125

4.2 Methodology 126

4.3 Simulation Study 130

4.3.1 Simulation design 130

4.3.2 Optimisation of the Gaussian mixture model 133

4.3.3 Hyperparameter tuning for nonparametric regressions 134

4.3.4 Simulation results 135

4.4 Applied Examples from UK Biobank 146

4.4.1 Description of the examples 146

4.4.2 Results of the examples 148

4.5 Discussion and Concluding Remarks 154

5 Investigating Selection Bias in Linear and Stratified Non-linear Mendelian Randomisation Analyses 157

5.1 Introduction 157

5.2 Related Works 160

5.3 Selection Mechanisms 162

5.4 Inverse Probability Weighted MR 164

5.4.1 Methodology 164

5.4.2 Discussion on large weights 171

5.4.3 Calculation of standard errors 173

5.5 Simulation Study 174

5.5.1 Simulation design 174

5.5.2 Simulation results 179

5.6 Applied Examples from UK Biobank 185

5.6.1 Description of the examples 185

5.6.2 Construction of the examples 186

5.6.3 Results of the examples 196

5.7 Stratified Nonlinear MR under Selection 204

5.7.1 Inverse probability weighting for stratified nonlinear MR 204

5.7.2 Simulation study 207

5.8 Discussion and Concluding Remarks 222

6 Summary and Future Prospects **226**

6.1 Summary of the Thesis 226

6.2 Prospects for Future Research 229

6.2.1 Bias from weak instruments for nonparametric 2SLS and GMM methods 229

6.2.2 Further development of nonparametric MR 230

6.2.3 Causal effect heterogeneity with stratified nonlinear MR and nonparametric MR . . . 230

6.2.4 Using inverse-probability weighting to account for selection bias with a binary outcome in the context of MR . . . 231

6.2.5 High dimensional propensity score estimation . . . 232

6.2.6 Further investigation of selection bias in the context of stratified nonlinear MR . . . 232

6.2.7 Nonlinear IV and MR methods robust to invalid instruments . . . 233

6.3 Conclusion . . . 234

Bibliography **235**

A Appendix to the Thesis **250**

A.1 Review on Kernel Methods . . . 250

A.1.1 Construction of kernels . . . 250

A.1.2 Examples of kernels . . . 252

A.1.3 Reproducing kernel Hilbert space (RKHS) . . . 252

A.1.4 Kernel ridge regression . . . 255

A.2 Additional Selection Mechanisms . . . 258

List of Figures

1.1 A very simple causal DAG. 5

1.2 A SWIG that corresponds to the DAG in Figure 1.1. 6

1.3 A standard IV model. A circle is drawn around the node U to emphasise that U is unmeasured. 11

1.4 Other versions of IV models. 12

1.5 A more detailed illustration of the *no unmeasured confounding for the instrument* condition for IV. 12

1.6 Comparison of a randomised controlled trial and a Mendelian randomisation study (adapted from Davey Smith and Ebrahim (2005) and Burgess and Thompson (2021)). 21

1.7 A DAG illustrating population stratification. The no unmeasured confounding for the instrument assumption is violated because of the confounding pathway through ancestry. 23

2.1 The RKHS model (adapted from Singh et al. (2019)) 41

2.2 A DAG illustrating the IV-free exposure. 57

2.3 A diagram that outlines the doubly-ranked stratification method (adapted from Tian et al. (2023)). 58

2.4 MSEs for the demand design. The MSEs are presented using standard boxplots. I.e. among results from 20 simulated datasets, the box represents the interquartile range (IQR), with the bottom of the box showing the 25th percentile, the top of the box showing the 75th percentile and the line inside the box showing the median. The length of the whiskers is 1.5 times the interquartile range. Data points beyond the whiskers are plotted as individual dots. 69

2.5 MSEs for the linear design, with and without KIV. The MSEs are presented using standard boxplots (same as in Figure 2.4). 70

3.1 The rank-preserving assumption in the case of a binary instrument. 79

3.2 Quantiles of counterfactual exposure distributions with a binary instrument. 81

3.3 Nonparametric MR and stratified nonlinear MR estimates for null and cubic functions with a binary instrument, constant genetic effect and 4% instrument strength. The black dots represent the estimates, the blue curves represent pointwise 95% confidence intervals and the red curve represents the true derivative of h. 99

3.4 Nonparametric MR and stratified nonlinear MR estimates for exponential and threshold functions with a binary instrument, constant genetic effect and 4% instrument strength. The black dots represent the estimates, the blue curves represent pointwise 95% confidence intervals and the red curve represents the true derivative of h. 100

3.5 Nonparametric MR and stratified nonlinear MR estimates for null and cubic functions with a three-value instrument, constant genetic effect and 4% instrument strength. The black dots represent the estimates, the blue curves represent pointwise 95% confidence intervals and the red curve represents the true derivative of h. 101

3.6 Nonparametric MR and stratified nonlinear MR estimates for exponential and threshold functions with a three-value instrument, constant genetic effect and 4% instrument strength. The black dots represent the estimates, the blue curves represent pointwise 95% confidence intervals and the red curve represents the true derivative of h. 102

3.7 Nonparametric MR and stratified nonlinear MR estimates for the relationship between BMI and average handgrip power and between BMI and pulse rate. The black curve represents the estimates and the blue curves represent pointwise 95% confidence intervals. The red curve indicates a null effect. . . 119

3.8 Strength of the genetic variant rs9939609 across the distribution of BMI for the subsamples of average handgrip power and pulse rate. The instrument strength is measured by the F-statistics of the instrument-exposure regressions within 100 strata defined by doubly-ranked stratification. 120

4.1 Nonparametric MR and stratified nonlinear MR estimates for null and cubic functions with a continuous instrument, constant genetic effect and 5% instrument strength. The black dots represent the estimates, the blue curves represent pointwise 95% confidence intervals and the red curve represents the true derivative of h. 138

4.2 Nonparametric MR and stratified nonlinear MR estimates for exponential and threshold functions with a continuous instrument, constant genetic effect and 5% instrument strength. The black dots represent the estimates, the blue curves represent pointwise 95% confidence intervals and the red curve represents the true derivative of h. 139

4.3 Histograms showing the distributions of 25[OH]D concentration (nmol/L) and alcohol intake (units per day) in the respective subsamples. 148

4.4 Nonparametric MR and stratified nonlinear MR estimates for the relationship between 25[OH]D and SBP and between alcohol intake and SBP. The black curve represents the estimates and the blue curves represent pointwise 95% confidence intervals. The red curve indicates a null effect. 152

4.5 Strength of the genetic risk scores across the distribution of 25[OH]D concentration and alcohol intake in their respective subsamples. The instrument strength is measured by the F-statistics of the instrument-exposure regressions within 100 strata defined by doubly-ranked stratification. 153

5.1 DAG for the instrument for selection model (Gkatzionis et al., 2024). Z denotes the instruments for the exposures, X denotes the exposures, Y denotes the outcome, U denotes unmeasured confounders, S denotes the selection indicator and $\widetilde{Z}$ denotes the instrument for selection. 161

5.2 DAG for the mediation model (Cai et al., 2022). Z denotes the instrument, G denotes the exposure or SNP of interest, X denotes the mediator, Y denotes the outcome and U denotes unmeasured confounders. β_{GX}, β_{GY}, β_{XY} are coefficients assuming linear SEMs. 162

5.3 DAGs illustrating four different selection mechanisms. Z denotes the instruments, X denotes the exposures, Y denotes the outcome, V denotes measured covariates, U denotes unmeasured confounders and S denotes the selection indicator. 163

5.4 Moderate and severe selection when the selection model is a logistic function of the exposure. For moderate selection, $\delta_0 = 0.5$ and $\delta_X = 1$; for severe selection, $\delta_0 = -3.85$ and $\delta_X = 3.6$. 170

A.1 DAGs illustrating four more selection mechanisms. Z denotes the instruments, X denotes the exposures, Y denotes the outcome, V denotes measured covariates, U denotes unmeasured confounders and S denotes the selection indicator. 258

List of Tables

2.1 A list of abbreviations for all the methods compared in the simulation study. 67

3.1 Model parameters for the data generating mechanisms for instrument-exposure relationships in scenarios with binary instruments. Instrument strength is the proportion of variance in the exposure explained by the instrument, measured by R^2 in the instrument-exposure regression. α_0 are chosen such that the mean of the exposure distribution is kept about the same in each scenario. 91

3.2 Model parameters for the data generating mechanisms for instrument-exposure relationships in scenarios with three-value instruments. Instrument strength is the proportion of variance in the exposure explained by the instrument, measured by R^2 in the instrument-exposure regression. α_0 are chosen such that the mean of the exposure distribution is kept about the same in each scenario. 92

3.3 Magnitude of the bias and standard deviation of the estimated derivatives with a binary instrument, constant genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$. 103

3.4 Magnitude of the bias and standard deviation of the estimated derivatives with a binary instrument, nonconstant but rank-preserving genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$. 104

3.5 Magnitude of the bias and standard deviation of the estimated derivatives with a binary instrument, rank-varying genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$. 105

3.6 Magnitude of the bias and standard deviation of the estimated derivatives with a three-value instrument, constant genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$. 106

3.7 Magnitude of the bias and standard deviation of the estimated derivatives with a three-value instrument, nonconstant but rank-preserving genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$. 107

3.8 Magnitude of the bias and standard deviation of the estimated derivatives with a three-value instrument, rank-varying genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$. 108

3.9 Comparison of bias of nonparametric MR estimates with respect to the derivative of the structural function h and with respect to the theoretical QACE, with a binary instrument, constant genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$. Results for the null function are omitted. 109

3.10 Comparison of bias of nonparametric MR estimates with respect to the derivative of the structural function h and with respect to the theoretical QACE, with a binary instrument, nonconstant but rank-preserving genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$. Results for the null function are omitted. 110

3.11 Comparison of bias of nonparametric MR estimates with respect to the derivative of the structural function h and with respect to the theoretical QACE, with a binary instrument, rank-varying genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$. Results for the null function are omitted. 111

3.12 Comparison of bias of nonparametric MR estimates with respect to the derivative of the structural function h and with respect to the theoretical QACE, with a three-value instrument, constant genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$. Results for the null function are omitted. 112

3.13 Comparison of bias of nonparametric MR estimates with respect to the derivative of the structural function h and with respect to the theoretical QACE, with a three-value instrument, nonconstant but rank-preserving genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$. Results for the null function are omitted. 113

3.14 Comparison of bias of nonparametric MR estimates with respect to the derivative of the structural function h and with respect to the theoretical QACE, with a binary instrument, rank-varying genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$. Results for the null function are omitted. 114

3.15 Stratum-specific (LACE) estimates using doubly-ranked stratification for estimating the relationship between BMI and average handgrip power in the randomly drawn subsample. 120

3.16 Stratum-specific (LACE) estimates using doubly-ranked stratification for estimating the relationship between BMI and pulse rate in the randomly drawn subsample. 121

4.1 Model parameters for the data generating mechanisms for instrument-exposure relationships. Instrument strength is the proportion of variance in the exposure explained by the instrument, measured by R^2 in the exposure-instrument regression. α_0 are chosen such that the mean of the exposure distribution is kept about the same in each scenario. 133

4.2 Magnitude of the bias and standard deviation of the estimated derivatives with a continuous instrument, constant genetic effect and instrument strength $\in \{5\%, 2.5\%, 1\%\}$. 140

4.3 Magnitude of the bias and standard deviation of the estimated derivatives with a continuous instrument, nonconstant but rank-preserving genetic effect and instrument strength $\in \{5\%, 2.5\%, 1\%\}$. 141

4.4 Magnitude of the bias and standard deviation of the estimated derivatives with a continuous instrument, rank-varying genetic effect and instrument strength $\in \{5\%, 2.5\%, 1\%\}$. 142

4.5 Comparison of bias of nonparametric MR estimates with respect to the derivative of the structural function h and with respect to the theoretical QACE, with a continuous instrument, constant genetic effect and instrument strength $\in \{5\%, 2.5\%, 1\%\}$. Results for the null function are omitted. 143

4.6 Comparison of bias of nonparametric MR estimates with respect to the derivative of the structural function h and with respect to the theoretical QACE, with a continuous instrument, nonconstant but rank-preserving genetic effect and instrument strength $\in \{5\%, 2.5\%, 1\%\}$. Results for the null function are omitted. 144

4.7 Comparison of bias of nonparametric MR estimates with respect to the derivative of the structural function h and with respect to the theoretical QACE, with a continuous instrument, rank-varying genetic effect and instrument strength $\in \{5\%, 2.5\%, 1\%\}$. Results for the null function are omitted. 145

4.8 Stratum-specific (LACE) estimates using the doubly-ranked stratification for estimating the relationship between 25(OH)D concentration (nmol/L) and systolic blood pressure (SBP) in the randomly drawn subsample. 153

4.9 Stratum-specific (LACE) estimates using doubly-ranked stratification for estimating the relationship between alcohol intake (units per day) and systolic blood pressure (SBP) in the randomly drawn subsample. 154

5.1 Model parameters for the data generating mechanisms. Instrument strength is the proportion of variance in the exposure explained by the instrument, measured by R^2 in the instrument-exposure regression; confounding level is the proportion of variance in the outcome not explained by the exposure, measured by $1 - R^2$ in the exposure-outcome regression. 176

5.2 Model parameters for selection scenarios where the selection depends on the exposure X only. Selection correlation is the correlation between S and X, and selection probability is the mean probability of getting selected into the sample. 177

5.3 Model parameters for selection scenarios where the selection depends on both the exposure X and the outcome Y. Selection correlation refers to both the correlation between S and X and the correlation between S and Y (e.g. in scenario B1, both corr(S, X) and corr(S, Y) are kept at 0.25), and selection probability is the mean probability of getting selected into the sample. . . . 177

5.4 Model parameters for selection scenarios where the selection depends on the outcome Y only. Selection correlation is the correlation between S and Y, and selection probability is the mean probability of getting selected into the sample. 178

5.5 Model parameters for selection scenarios where the selection depends on a downstream covariate V of the exposure X. Selection correlation is the correlation between S and V, and selection probability is the mean probability of getting selected into the sample. 178

5.6 Bias, standard error and coverage for inverse probability weighted MR estimator, scenarios A1 - A11. 181

5.7 Bias, standard error and coverage for standard MR estimator with and without covariate adjustment, scenarios A1 - A11. 181

5.8 Bias, standard error and coverage for inverse probability weighted MR estimator, scenarios B1 - B11. 182

5.9 Bias, standard error and coverage for standard MR estimator with and without covariate adjustment, scenarios B1 - B11. 182

5.10 Bias, standard error and coverage for inverse probability weighted MR estimator, scenarios C1 - C11. 183

5.11 Bias, standard error and coverage for standard MR estimator with and without covariate adjustment, scenarios C1 - C11. 183

5.12 Bias, standard error and coverage for inverse probability weighted MR estimator, scenarios D1 - D11. 184

5.13 Bias, standard error and coverage for standard MR estimator with and without covariate adjustment, scenarios D1 - D11. 184

5.14 Standard MR estimates in the full UK Biobank datasets for the causal effect of systolic blood pressure on concentration of sodium in the blood (sodium), concentration creatinine in the blood (creatinine), coronary heart disease and stroke. 188

5.15 Model parameters for selection scenarios where the selection depends on the exposure X only. The exposure is systolic blood pressure (SBP). Selection correlation is the correlation between S and X, and selection probability is the mean probability of getting selected into the sample. The same parameterisations for scenarios A1 - A11 are used for all four outcomes. 188

5.16 Model parameters for selection scenarios where the selection depends on both the exposure X and the outcome Y. The exposure is systolic blood pressure (SBP) and the outcome is concentration of sodium in the blood (sodium). Selection correlation refers to both the correlation between S and X and the correlation between S and Y (e.g. in scenario B1, $\text{corr}(S, X) = 0.25$ and $\text{corr}(S, X) = -0.1$), and selection probability is the mean probability of getting selected into the sample. 189

5.17 Model parameters for selection scenarios where the selection depends on the outcome Y only. The outcome is concentration of sodium in the blood (sodium). Selection correlation is the correlation between S and Y, and selection probability is the mean probability of getting selected into the sample. 190

5.18 Model parameters for selection scenarios where the selection depends on both the exposure X and the outcome Y. The exposure is systolic blood pressure (SBP) and the outcome is concentration of creatinine in the blood (creatinine). Selection correlation refers to both the correlation between S and X and the correlation between S and Y (e.g. in scenario B1, $\text{corr}(S, X) = 0.25$ and $\text{corr}(S, X) = -0.1$), and selection probability is the mean probability of getting selected into the sample. 191

5.19 Model parameters for selection scenarios where the selection depends on the outcome Y only. The outcome is concentration of creatinine in the blood (creatinine). Selection correlation is the correlation between S and Y, and selection probability is the mean probability of getting selected into the sample. 192

5.20 Model parameters for selection scenarios where the selection depends on both the exposure X and the outcome Y. The exposure is systolic blood pressure (SBP) and the outcome is coronary heart disease. Selection correlation is the correlation between S and X in the controls (i.e. individuals with $Y = 0$), and selection probability is the mean probability of getting selected into the sample. 193

5.21 Model parameters for selection scenarios where the selection depends on the outcome Y only. The outcome is coronary heart disease. Selection probability is the mean probability of getting selected into the sample. 194

5.22 Model parameters for selection scenarios where the selection depends on both the exposure X and the outcome Y. The exposure is systolic blood pressure (SBP) and the outcome is stroke. Selection correlation is the correlation between S and X in the controls (i.e. individuals with $Y = 0$), and selection probability is the mean probability of getting selected into the sample. . . . 195

5.23 Model parameters for selection scenarios where the selection depends on the outcome Y only. The outcome is stroke. Selection probability is the mean probability of getting selected into the sample. 196

5.24 Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is concentration of sodium, scenarios A1 - A11. 198

5.25 Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is concentration of sodium, scenarios B1 - B11. 198

5.26 Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is concentration of sodium, scenarios C1 - C11. 199

5.27 Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is concentration of creatinine, scenarios A1 - A11. 199

5.28 Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is concentration of creatinine, scenarios B1 - B11. 200

5.29 Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is concentration of creatinine, scenarios C1 - C11. 200

5.30 Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is coronary heart disease (CAD), scenarios A1 - A11. 201

5.31 Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is coronary heart disease (CAD), scenarios B1 - B11. 201

5.32 Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is coronary heart disease (CAD), scenarios C1 - C12. 202

5.33 Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is stroke, scenarios A1 - A11. 202

5.34 Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is stroke, scenarios B1 - B11. 203

5.35 Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is stroke, scenarios C1 - C12. 203

5.36 Bias, standard error and coverage for inverse probability weighted stratum-specific MR estimates with correct strata formed using the full exposure and instrument data and incorrect strata formed using the selected sample, in scenario A1 where selection correlation = 0.25 and selection probability = 60%.211

5.37 Bias, standard error and coverage for stratum-specific MR estimates with and without covariate adjustment with incorrect strata formed using the selected sample, in scenario A1 where selection correlation = 0.25 and selection probability = 60%. 211

5.38 Bias, standard error and coverage for inverse probability weighted stratum-specific MR estimates with correct strata formed using the full exposure and instrument data and incorrect strata formed using the selected sample, in scenario A2 where selection correlation = 0.25 and selection probability = 40%. 212

5.39 Bias, standard error and coverage for stratum-specific MR estimates with and without covariate adjustment with incorrect strata formed using the selected sample, in scenario A2 where selection correlation = 0.25 and selection probability = 40%. 212

5.40 Bias, standard error and coverage for inverse probability weighted stratum-specific MR estimates with correct strata formed using the full exposure and instrument data and incorrect strata formed using the selected sample, in scenario A3 where selection correlation = 0.25 and selection probability = 20%. 213

5.41 Bias, standard error and coverage for stratum-specific MR estimates with and without covariate adjustment with incorrect strata formed using the selected sample, in scenario A3 where selection correlation = 0.25 and selection probability = 20%. 213

5.42 Bias, standard error and coverage for inverse probability weighted stratum-specific MR estimates with correct strata formed using the full exposure and instrument data and incorrect strata formed using the selected sample, in scenario A4 where selection correlation = 0.25 and selection probability = 5%. 214

5.43 Bias, standard error and coverage for stratum-specific MR estimates with and without covariate adjustment with incorrect strata formed using the selected sample, in scenario A4 where selection correlation = 0.25 and selection probability = 5%. 214

5.44 Bias, standard error and coverage for inverse probability weighted stratum-specific MR estimates with correct strata formed using the full exposure and instrument data and incorrect strata formed using the selected sample, in scenario A5 where selection correlation = 0.45 and selection probability = 60%. 215

5.45 Bias, standard error and coverage for stratum-specific MR estimates with and without covariate adjustment with incorrect strata formed using the selected sample, in scenario A5 where selection correlation = 0.45 and selection probability = 60%. 215

5.46 Bias, standard error and coverage for inverse probability weighted stratum-specific MR estimates with correct strata formed using the full exposure and instrument data and incorrect strata formed using the selected sample, in scenario A6 where selection correlation = 0.45 and selection probability = 40%. 216

5.47 Bias, standard error and coverage for stratum-specific MR estimates with and without covariate adjustment with incorrect strata formed using the selected sample, in scenario A6 where selection correlation = 0.45 and selection probability = 40%. 216

5.48 Bias, standard error and coverage for inverse probability weighted stratum-specific MR estimates with correct strata formed using the full exposure and instrument data and incorrect strata formed using the selected sample, in scenario A7 where selection correlation = 0.45 and selection probability = 20%. 217

5.49 Bias, standard error and coverage for stratum-specific MR estimates with and without covariate adjustment with incorrect strata formed using the selected sample, in scenario A7 where selection correlation = 0.45 and selection probability = 20%. 217

5.50 Bias, standard error and coverage for inverse probability weighted stratum-specific MR estimates with correct strata formed using the full exposure and instrument data and incorrect strata formed using the selected sample, in scenario A8 where selection correlation = 0.45 and selection probability = 5%. 218

5.51 Bias, standard error and coverage for stratum-specific MR estimates with and without covariate adjustment with incorrect strata formed using the selected sample, in scenario A8 where selection correlation = 0.45 and selection probability = 5%. 218

5.52 Bias, standard error and coverage for inverse probability weighted stratum-specific MR estimates with correct strata formed using the full exposure and instrument data and incorrect strata formed using the selected sample, in scenario A9 where selection correlation = 0.65 and selection probability = 60%.219

5.53 Bias, standard error and coverage for stratum-specific MR estimates with and without covariate adjustment with incorrect strata formed using the selected sample, in scenario A9 where selection correlation = 0.65 and selection probability = 60%. 219

5.54 Bias, standard error and coverage for inverse probability weighted stratum-specific MR estimates with correct strata formed using the full exposure and instrument data and incorrect strata formed using the selected sample, in scenario A10 where selection correlation = 0.65 and selection probability = 40%.220

5.55 Bias, standard error and coverage for stratum-specific MR estimates with and without covariate adjustment with incorrect strata formed using the selected sample, in scenario A10 where selection correlation = 0.65 and selection probability = 40%. 220

5.56 Bias, standard error and coverage for inverse probability weighted stratum-specific MR estimates with correct strata formed using the full exposure and instrument data and incorrect strata formed using the selected sample, in scenario A11 where selection correlation = 0.65 and selection probability = 20%. 221

5.57 Bias, standard error and coverage for stratum-specific MR estimates with and without covariate adjustment with incorrect strata formed using the selected sample, in scenario A11 where selection correlation = 0.65 and selection probability = 20%. 221

List of Abbreviations

We list main abbreviations adopted in this thesis below:

2SLS	two-stage-least-square;
ACE	average causal effect;
BMI	body mass index;
CDF	cumulative density function;
DAG	directed acyclic graph;
GMM	generalised method of moments;
GWAS	genome-wide association studies;
IV	instrumental variable;
IVQR	instrumental variable quantile regression;
IPW	inverse probability weighting;
IVW	inverse variance weighting;
KKT conditions	Karush-Kuhn-Tucker conditions;
LACE	localised average causal effect;
LD	linkage disequilibrium;
MAR	missing at random;
MNAR	missing not at random;
MR	Mendelian randomisation;
MSE	mean-squared error;
NPSEM	nonparametric structural equation model;

OLS	ordinary least square;
OVB	omitted variable bias;
QACE	quantile average causal effect;
QTE	quantile treatment effect;
RCT	randomised controlled trial;
RKHS	reproducing kernel Hilbert space;
SBP	systolic blood pressure;
SCM	structural causal model;
SE	standard error;
SEM	structural equation model;
SNP	single nucleotide polymorphism;
SWIG	single world intervention graph

List of Notations

We list frequently used notations in this thesis below:

X	exposure, or the observed matrix of the exposure when we write formula for estimators in matrix notation;
$\mathcal{X}$	exposure space;
Y	outcome, or the observed vector of the outcome when we write formula for estimators in matrix notation;
$\mathcal{Y}$	outcome space;
Z	instrumental variable, or the observed matrix of the instrument when we write formula for estimators in matrix notation;
$\mathcal{Z}$	instrument space;
U	confounders, usually unmeasured;
V	measured covariates;
W	weights: either the weight matrix for GMM (Chapter 1 and 2) or the inverse probability weights (Chapter 5);
S	selection indicator (Chapter 5);
D	number of simulated datasets;
R	number of bootstrapped samples;
N	size of the underlying population;
n	size of the sample at hand;
i, j	subscript of individuals;

h	structural function;
t	quantiles of the counterfactual exposure distributions;
e	random noise (has mean zero, unit variance and is independent of other variables);
ϵ	error in our model (might be correlated with other variables in the model);
β	linear regression parameter (we will use β to denote the average causal effect where appropriate, but β does not have to be causal);
μ	mean parameter;
σ^2	variance parameter;
$\mathbb{E}$	expectation operator;
$\mathbb{P}$	probability operator;
$\mathbb{R}$	space of real numbers;
$\mathbb{R}^d$	d-dimensional real-valued space;
$\mathbb{I}$	indicator function;
$\mathbb{N}$	normal distribution;
Unif	uniform distribution;
logit	logistic function;
expit	expit function (i.e. the inverse of logistic function)

We follow the usual convention that we use capital letters for random variables/vectors and lower-case letters for observed values of random variables/vectors except situations where we specifically mention otherwise (N and n are an example of this exception).

Chapter 1

Introduction to Causal Inference, Instrumental Variables and Mendelian Randomisation

1.1 Languages and Models for Causality

The need to assess the causal effect of one variable (usually called an exposure, a risk factor, a treatment or an intervention) on an outcome of interest naturally arises in practice. As intelligent human beings, we all know that association is not causation. But what exactly is the so-called "causal effect"? To articulate concepts of causality, we need help from precise mathematical and statistical languages.

Because causal questions are present in almost every field of science, historically there have been various causal models and frameworks developed by researchers from different fields (statistics, econometrics, genetics, epidemiology, computer science and etc). In the following, we describe four mainstream frameworks for causality that we will rely upon throughout this

thesis. We will see that these frameworks are by no means distinct but closely related to each other, and we will bring them together to help us better understand causality.

Before we move on, it is useful to make clear of the basic notations that we adopt in this thesis. We assume that we have access to i.i.d copies of data $\{Y_i, X_i, U_i\}_{i=1}^n$, where $Y \in \mathcal{Y} \subset \mathbb{R}$ denotes the outcome, $X \in \mathcal{X} \subset \mathbb{R}^{d_X}$ denotes the exposure and $U \in \mathbb{R}^{d_U}$ denotes the confounders (could be measured or unmeasured). We will use, for example, x_i for an observation in the sample and x for any given value of X (usually out-of-sample).

1.1.1 The potential outcomes framework

The potential outcomes framework, first introduced by Jerzy Neyman in 1923 (Splawa-Newyman et al., 1990) and further developed by Rubin (1974), is one of the most important languages for causality that make casual effects precise. The potential outcome (also known as the counterfactual; we will discuss these terminologies in section 1.1.5) of an individual's outcome variable is written as $Y_i(x)$, defined as what would have been observed for the individual's outcome *if* this individual's exposure $X_i = x$. The (individualised) causal effect of the exposure can then be characterised by $Y_i(x) - Y_i(x')$ where $x \neq x'$. However, an obvious difficulty with the individualised causal effect is that only one of the potential outcomes can actually be observed. This is known as the fundamental problem of causal inference (Holland, 1986). Therefore, people usually hope to estimate some average measure of causal effect over the population, which we will introduce in section 1.1.5.

We make the following assumptions throughout this thesis so that these potential outcomes are trackable.

Assumption 1 (no interference)**.** *An individual's potential outcomes depend only on his/her exposure and are independent of other individuals' exposures.*

Assumption 2 (consistency)**.** *If* $X_i = x$, $Y_i = Y_i(x)$.

The above two assumptions together are known as the *stable unit treatment value assumption* (SUTVA) introduced by Rubin (1980). They say that one's potential outcome $Y_i(x)$ is characterised solely by X_i, and this potential outcome $Y_i(x)$ equals to the observed outcome Y_i when $X_i = x$. Without the no interference assumption, the potential outcomes are not uniquely defined. And the consistency assumption relates the hypothetical potential outcomes (i.e. counterfactuals) to quantities that can be observed (often called factuals, as opposed to counterfactuals). Because of the consistency assumption, many times we are safe to just write the observed outcome Y_i, and keep the potential outcomes under the hood.

We should note that the consistency assumption of the potential outcomes is not the same as the usual asymptotic consistency of statistical estimators. We should also note that, although the term is called "potential outcomes", we can define potential outcomes on any causal variables and not just the outcome. For example, we may write the potential outcomes of the exposure as $X_i(z)$, where Z is some cause of X, and SUTVA applies for $X_i(z)$ equally. We will see in the following that an instrumental variable Z is such an example.

1.1.2 Pearl's *do* operator

Pearl's *do* operator is another widely used notation to describe causality (e.g. Pearl, 2010). Under this framework, when we set the value of an individual's exposure X_i to x, his/her corresponding outcome value is denoted as $Y_i|do(X_i = x)$ or simply $Y_i|do(x)$. This notation emphasises on the idea of manipulation and experimentation by researchers, and is often known as "no causation without manipulation" (e.g. Holland, 1986). Although this view is a bit restrictive from modern perspective, it does emphasise on the fact that any exposures that are of interest in practice need to be modifiable.

There are philosophical debates over the equivalence and differences over $Y_i|do(x)$ and potential outcomes $Y_i(x)$. For example, Bareinboim et al. (2022) argue that the *do* operator implies the state of doing/intervening and the potential outcomes/counterfactuals implies the

state of imagining. For our purpose, we will treat them as equivalent (merely two different notations) and adopt the potential outcomes framework throughout this thesis.

1.1.3 Structural equation models

Structural equation models (SEMs) are handy tools to model the relationships between causal variables. A general nonparametric structural equation model (NPSEM) can be written as

$$Y_i = f(X_i, U_i, e_i), \quad e_i \perp\!\!\!\perp (X_i, U_i) \tag{1.1}$$

for some function f and random noises $\{e_i\}_{i=1}^n$ that are mutually independent with zero mean and finite variance and are independent of X_i and U_i. While this may look like a usual statistical model, it is called a "structural" model because it is a model for potential outcomes/counterfactual variables (e.g. Robins et al., 2000). This is to say that equation (1.1) should really be written as

$$Y_i(X_i) = f(X_i, U_i, e_i), \quad e_i \perp\!\!\!\perp (X_i, U_i)$$

and the potential outcomes $Y_i(x)$ are given by $f(x, U_i, e_i)$. But because of the consistency assumption made in section 1.1.1, people often just write the observed outcome Y_i for ease of notation.

A less general nonparametric structural equation model can be written as

$$Y_i = h(X_i) + \delta^T U_i + e_i, \quad e_i \perp\!\!\!\perp (X_i, U_i) \tag{1.2}$$

for some function h. Here we assume that the confounders U and random noise e are additive. We will see in later chapters that equation (1.2) forms the basis for our analysis of nonlinear causal effects.

The commonly used linear structural equation model (linear SEM) can be written as

$$Y_i = \beta_0 + \beta_X^T X_i + \delta^T U_i + e_i, \quad e_i \perp\!\!\!\perp (X_i, U_i) \tag{1.3}$$

This model can be estimated using a simple linear regression if the confounders U are fully measured.

1.1.4 Graphical models

Graphs are widely applied in almost every subject of science to describe the relationships between variables of interest. In causal inference, a graphical model usually refers to a directed acyclic graph (DAG) that models the causal structure of the variables. Figure 1.1 shows one of the simplest DAGs that describes the causal structure of an outcome Y, an exposure X and confounders U.

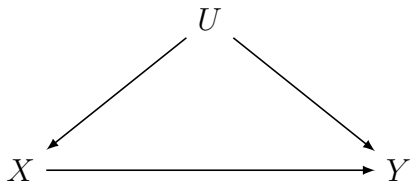


Figure 1.1: A very simple causal DAG.

One useful way, developed by Richardson and Robins (2013), to make potential outcomes explicit in graphical models is to modify the DAG into a single world intervention graph (SWIG). The approach splits the nodes of exposures into a random and a fixed component, where the random component inherits all edges that goes into the node and the fixed component inherits all edges that goes out of the node, and replaces the observed outcomes by potential outcomes. For example, Figure 1.2 shows the SWIG adapted from Figure 1.1. The usual graph theories, such as the famous back-door criterion, can then be applied to SWIGs to determine conditional independence structures of potential outcomes.

We can easily define structural equation models according to the DAGs (and equivalently SWIGs), where all the parents of a node go to the right hand side of the structural equation.

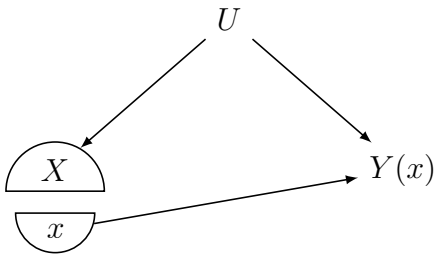


Figure 1.2: A SWIG that corresponds to the DAG in Figure 1.1.

For example, a NPSEM according to Figure 1.1 can be defined as

$$Y_i = f_Y(X_i, U_i, e_{Y,i}), \quad e_{Y,i} \perp\!\!\!\perp (X_i, U_i)$$
$$X_i = f_X(U_i, e_{X,i}), \quad e_{X,i} \perp\!\!\!\perp (U_i, e_{Y,i})$$

for some functions f_Y, f_X and random noises e_Y, e_X. Similarly a linear SEM can be defined as

$$Y_i = \beta_0 + \beta_X^T X_i + \delta^T U_i + e_{Y,i}, \quad e_{Y,i} \perp\!\!\!\perp (X_i, U_i) \tag{1.4}$$
$$X_i = \alpha_0 + \Gamma^T U_i + e_{X,i}, \quad e_{X,i} \perp\!\!\!\perp (U_i, e_{Y,i}) \tag{1.5}$$

A NPSEM defined based on a causal DAG is sometimes also referred to as a structural causal model (SCM) (e.g. Peters et al., 2017).

1.1.5 Common terminologies

Because causal inference has multidisciplinary origins, there are many different terminologies, developed and adopted by researchers from various fields, that relate to similar causal concepts. In the following we briefly introduce and clarify some of the commonly used terminologies.

(1). Exposure, risk factor, treatment and intervention

The above terms are four of the most commonly used terms that refer to the variable(s) that we wish to know the causal effect of, which we denote by X. There is a philosophical difference between these terms that *exposure* and *risk factor* usually refer to observational (but modifiable) variables, such as body-mass index (BMI), while *treatment* and *intervention* usually refers to something that the researchers actively assign, such as a drug-placebo pair. In this thesis, we will use *exposure* for most of the time as the analyses presented are mostly observational.

(2). Potential outcomes, counterfactuals and factuals

Potential outcomes and *counterfactuals* both refer to the causal concept of what would have been observed if the cause takes a certain value. As introduced in section 1.1.1, they both imply hypothetical situations (mind experiments) that only one of the potential outcomes/counterfactuals can be observed in reality. There may be a subtle difference between these two terms that *potential outcomes* are prospective (what could potentially happen) and *counterfactuals* are retrospective (what could have happened), though in most literatures and throughout this thesis, *potential outcomes* and *counterfactuals* are used interchangeably.

As opposed to *counterfactuals*, *factuals* simply refer to quantities that are observable (such as Y_i instead of $Y_i(x_i)$).

(3). Randomisation, ignorability, exchangeability and no unmeasured confounding

One of the most important concepts in causal inference is *randomisation* of the exposure (or perhaps treatment is more proper here). If the treatment is randomly assigned, then the treatment X is independent of the potential outcome $Y(x)$ and we can relate counterfactuals to factuals by (we suppress the subscript i for simplicity):

$$\mathbb{E}\left[Y(x)\right] = \mathbb{E}\left[Y(x) \middle| X = x\right] = \mathbb{E}\left[Y \middle| X = x\right] \tag{1.6}$$

where the latter equation is due to the consistency assumption of potential outcomes, and $\mathbb{E}\left[Y \middle| X = x\right]$ is something we can estimate in practice. This is the reason why randomised

controlled trials (RCTs) are usually considered as the gold standard of causal inference.

Without randomised controlled trials, if the independence condition

$$Y(x) \perp\!\!\!\perp X \quad \forall x \tag{1.7}$$

is satisfied, then equation (1.6) can be equally applied. Alternatively, if we impose a weaker conditional independence condition

$$Y(x) \perp\!\!\!\perp X \mid U \quad \forall x \tag{1.8}$$

then we similarly have

$$\mathbb{E}\left[Y(x)\right] = \mathbb{E}\Big(\mathbb{E}\left[Y(x)\middle| U\right]\Big) = \mathbb{E}\Big(\mathbb{E}\left[Y(x)\middle| X = x, U\right]\Big) = \mathbb{E}\Big(\mathbb{E}\left[Y\middle| X = x, U\right]\Big) \tag{1.9}$$

where $\mathbb{E}\left[Y\middle| X = x, U\right]$ is also estimable as long as U is measured.

The (conditional) independence condition (1.7) or (1.8) is known as the (conditional) *ignorability* or (conditional) *exchangeability* condition (e.g. Zhao, 2021; Hernán and Robins, 2024). One useful result of checking these (conditional) independence conditions is known as the back-door criterion, which states that X is independent of $Y(x)$ given U if U contains no descendant of X and blocks all the back-door path (i.e. path with an arrow going into the given node) from X to Y. By using back-door criterion, we immediately see that the conditional exchangeability condition is satisfied for the DAG in Figure 1.1 if the confounders U are fully measured. Therefore, *conditional ignorability/exchangeability* is often also called *no unmeasured confounding*. If this condition is satisfied, usually some (average) measure of causal effect is identified.

Note that in some literatures, there are subtle distinctions between *randomisation, ignorability, exchangeability* and *no unmeasured confounding*, depending on how these terms are defined. Greenland and Robins (2009), for example, provide a detailed discussion on these concepts. For our purpose, these terms refer to the same condition, i.e. equation (1.7) or (1.8).

(4). Exogeneity, endogeneity and omitted variable bias (OVB)

These three terms stem from economics and econometrics (e.g. Imbens and Angrist, 1994; Burgess and Thompson, 2021) and are usually used in combination with linear structural equation models. *Exogenity*, meaning something given/taken from outside of the model, refers to the condition that the variable of interest is uncorrelated with the error term of the regression. Oppositely, *endogeneity*, meaning something coming from within the model, refers to the phenomenon that the variable of interest is correlated with the error term.

Take the following linear SEM as an example:

$$Y_i = \beta_0 + \beta_X^T X_i + \delta^T U_i + e_i$$

The exposure X is (weakly) exogenous if X is uncorrelated with e. Otherwise, X is endogenous. We see immediately that if the no unmeasured confounding assumption is violated, X will be endogenous because U will be absorbed into the error term. However, endogeneity is more general than mere violation of the no unmeasured confounding assumption, because other issues such as measurement error, reverse causation and simultaneous equation models can also lead to endogeneity (e.g. Hausman, 1977; Burgess and Thompson, 2021).

On the other hand, *omitted variable bias (OVB)* is just the economics term for unmeasured confounding. It is called OVB since the ordinary least square (OLS) estimator for β will be biased if U is unmeasured (i.e. omitted). Again, OVB is only one potential reason for endogeneity.

(5). Average causal effect (ACE)

As mentioned in section 1.1.1, the individualised causal effect $Y_i(x) - Y_i(x')$ cannot be calculated because only one of the potential outcomes can be observed. Instead, the most common target of causal estimation is the so-called *average causal effect (ACE)*, defined as

$$\mathbb{E}\left[Y(x) - Y(x')\right] \tag{1.10}$$

With the no unmeasured confounding assumption (equation (1.8)), (1.10) becomes

$$\begin{aligned}\mathbb{E}\left[Y(x) - Y(x')\right] &= \mathbb{E}\Big(\mathbb{E}\left[Y(x) - Y(x')|U\right]\Big) \\ &= \mathbb{E}\Big(\mathbb{E}\left[Y|X = x, U\right]\Big) - \mathbb{E}\Big(\mathbb{E}\left[Y|X = x', U\right]\Big) \end{aligned} \tag{1.11}$$

where (1.11) can be estimated in practice.

If we additionally assume a linear SEM (equation (1.3)) and take $x' = x - 1$, we see that our average causal effect becomes simply β_X. I.e. ACE can be estimated using standard OLS with linear SEM and the no unmeasured confounding assumption.

1.2 Instrumental Variables

1.2.1 The core IV assumptions

We have seen in section 1.1 that the no unmeasured confounding assumption enables us to estimate the average causal effect. However, with observational data, we can almost never be sure that we have measured all relevant confounders. The question then becomes: what else can we do?

The instrumental variables approach is a well-established method used in econometrics, epidemiology and biostatistics that enables us to validly estimate causal effects even in the presence of unmeasured confounding. An instrumental variable (instrument or IV for simplicity) Z is usually assumed to satisfy the following three conditions:

(1). *Relevance*: Z is correlated with the exposure X;
(2). *No unmeasured confounding for the instrument*: there exists no confounding pathway between Z and the outcome Y;
(3). *Exclusion restriction*: Z affects the outcome Y only through the exposure X.

These three conditions can be formalised using the potential outcomes framework:

(1). *Relevance*: the conditional density function $f(x|z)$ is a non-trivial function of z;
(2). *No unmeasured confounding for the instrument*: $X(z), Y(x,z) \perp\!\!\!\perp Z \ \ \forall x, z$;
(3). *Exclusion restriction*: $Y(x,z) = Y(x) \ \ \forall x, z$.

The above three conditions are usually referred to as the core IV assumptions (e.g. Angrist et al., 1996; Burgess and Thompson, 2021; Hernán and Robins, 2024). Figure 1.3 shows a DAG that describes a standard IV model, where the (non)existence of arrows 1)-3) corresponds to the above assumptions.

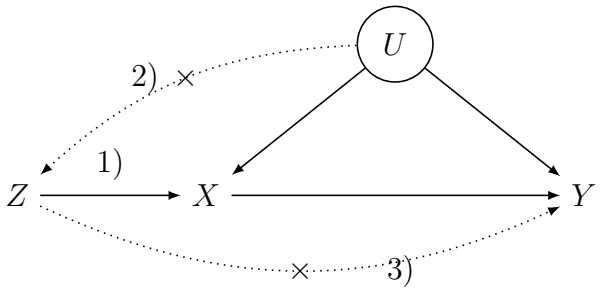


Figure 1.3: A standard IV model. A circle is drawn around the node U to emphasise that U is unmeasured.

However, note that although Figure 1.3 is the way most literatures introduce the IV model, it implicitly makes an unnecessary assumption that Z causes X. Indeed, other versions of IV models exist (e.g. Didelez and Sheehan, 2007; Hernán and Robins, 2024), as illustrated by Figure 1.4, where in (a) the correlation between X and Z is induced by another confounder U_z and in (b) the correlation is induced by conditioning on a collider S. The instrument described in Figure 1.3 is sometimes also called a causal instrument.

Moreover, we should note that the nonexistence of arrow 2) is in fact not sufficient for the no unmeasured confounding for the instrument condition, as it does not exclude the possibility of other confounding pathways. To be more specific, Figure 1.5 shows two different confounding pathways (arrows 4) and 5)) than arrow 2). The no unmeasured confounding

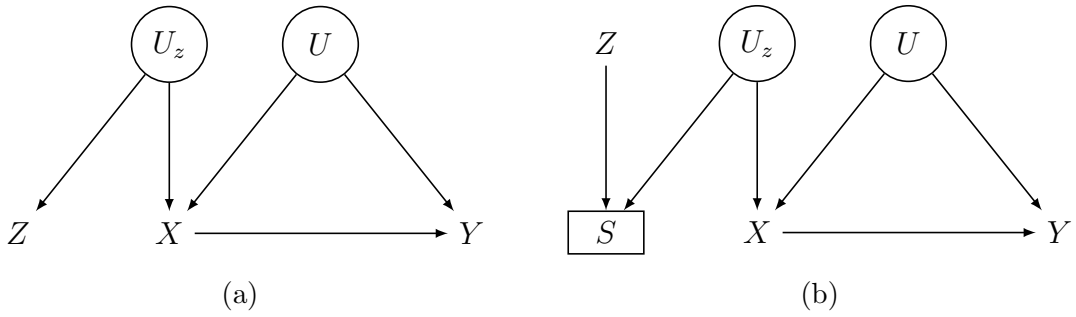


Figure 1.4: Other versions of IV models.

for the instrument condition requires that arrows 4) and 5) also cannot exist.

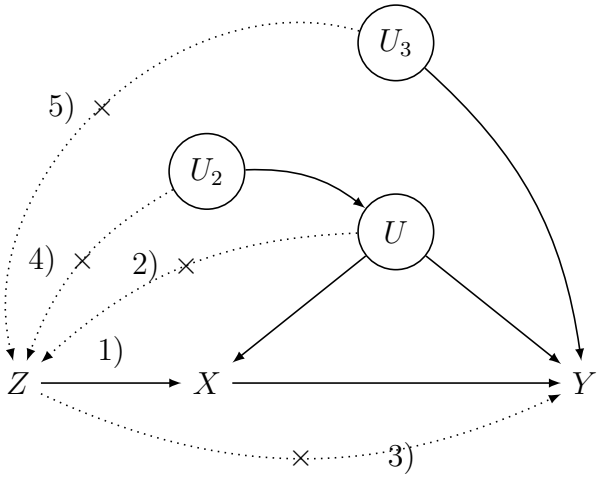


Figure 1.5: A more detailed illustration of the *no unmeasured confounding for the instrument* condition for IV.

In this thesis, we will not distinguish between these IV models, as long as the IV assumptions are satisfied. In addition, we should note that the three core IV assumptions are often stronger than what is actually assumed for various IV applications. We will make clear of the relevant assumptions when we introduce different IV models in the rest of the thesis. We will call an instrument a *valid* instrumental variable if the instrument satisfies the IV assumptions for the corresponding model. Therefore, the three core IV assumptions are sufficient for an instrument to be valid.

1.2.2 Linearity assumption and two-stage-least-square (2SLS)

Unfortunately, having valid instrumental variables alone is not enough to identify causal effects. Additional modelling assumptions are required, depending on what type of causal effects we would like to estimate. The most common choice is to assume a linear structural equation model between the exposure and the outcome, and the IV will enable us to estimate the average causal effect.

Specifically, assume the following linear SEM between the exposure and the outcome:

$$\begin{aligned} Y_i &= \beta_0 + \beta_X^T X_i + \delta^T U_i + e_{Y,i}, \quad e_{Y,i} \perp\!\!\!\perp (X_i, U_i) \\ &= \beta^T X_i + \epsilon_{Y,i} \end{aligned} \tag{1.12}$$

where the unmeasured confounders U is incorporated in the error term ϵ_Y and therefore X is generally correlated with ϵ_Y. This is the endogeneity issue we mentioned in section 1.1. For simplicity of notation, we write $\beta = (\beta_0, \beta_X^T)^T$ and use X_i to also denote the design vector that includes the exposure and a constant of 1 (with slight abuse of notation). Recall in section 1.1.5 we have established that with linear SEM, the average causal effect is simply β, which is our target of estimation. Moreover, we should note that by assuming the the linear SEM (1.12), we implicitly assume that the causal parameter β is constant across individuals, as there is not subscript i in β. This is often also referred to as the homogeneity assumption in the literature.

Similarly, we can specify a linear model between the exposure and the instruments:

$$X_i = \Gamma^T Z_i + \epsilon_{X,i} \tag{1.13}$$

where $Z \in \mathcal{Z} \subset \mathbb{R}^{d_Z}$ and ϵ_X is the i.i.d noise variable that has zero mean, finite variance and is uncorrelated with Z. Recall that $X \in \mathcal{X} \subset \mathbb{R}^{d_X}$ and we assume $d_Z \geq d_X$. I.e. we have at least as many instruments as the number of exposures, as otherwise we won't be able to identify β.

We note that (1.13) is usually called the first stage model and (1.12) is usually called the second stage model.

We make the following assumptions about the instrument Z (we suppress the subscript i for simplicity):

Assumption 3 (rank)**.** *The expectation* $\mathbb{E}[ZX^T]$ *is of full column rank.*

Assumption 4 (IV exogeneity)**.** *The expectation* $\mathbb{E}[Z\epsilon_Y] = 0$*.*

Note that the rank assumption is usually considered the same as the relevance condition, but the IV exogeneity assumption is actually weaker than the no unmeasured confounding for the instrument and the exclusion restriction conditions, which, together, impose that Z is independent of ϵ_Y.

The IV exogeneity assumption and equation (1.12) imply the following moment equation:

$$\mathbb{E}[Z(Y - \beta^T X)] = 0 \tag{1.14}$$

where the rank assumption ensures that there is a unique β that solves equation (1.14). I.e. β is identified. An IV estimator tries to solve the empirical analogue of (1.14):

$$\frac{1}{n}\sum_{i=1}^{n} z_i (y_i - \hat{\beta}_{IV}^T x_i) = 0 \tag{1.15}$$

When $d_Z = d_X$, i.e. we have the same number of instruments as the number of exposures, (1.15) is solved by

$$\hat{\beta}_{IV} = \Big(\sum_{i=1}^{n} z_i x_i^T\Big)^{-1} \sum_{i=1}^{n} z_i y_i \tag{1.16}$$

In this case, we say that our model is just-identified.

When $d_Z > d_X$, i.e. we have more instruments than the number of exposures, however, equation (1.15) generally does not yield a solution due to sampling error, as we have more equations than the unknown. In this case, we say that our model is over-identified. An

efficient way to circumvent this problem is to project X onto the space of Z, and use the instrument-predicted X, which is of the same dimension as β but also satisfies the IV assumptions, as the instruments in (1.15). This is the idea of two-stage-least-square (2SLS).

Therefore, 2SLS involves first regressing X_i on instruments Z_i and obtaining the fitted values $\widehat{X}_i = \hat{\Gamma}^T Z_i$. Then Y_i are regressed against the fitted values $\widehat{X}_i$, and the estimated coefficients give us the 2SLS estimates. The 2SLS estimator is given by

$$\hat{\beta}_{2SLS} = \underset{\beta \in \mathbb{R}^{d_X}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \beta^T (\hat{\Gamma}^T z_i) \right)^2 \tag{1.17}$$

where $\hat{\Gamma}$ are the first stage estimates

$$\hat{\Gamma} = \underset{\Gamma \in \mathbb{R}^{d_Z \times d_X}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left\| x_i - \Gamma^T z_i \right\|_2^2 \tag{1.18}$$

Using the standard regression formula, $\hat{\beta}_{2SLS}$ is given by

$$\hat{\beta}_{2SLS} = \left(X^T P_Z X \right)^{-1} X^T P_Z Y \tag{1.19}$$

where $P_Z = Z(Z^T Z)^{-1} Z^T$ is the projection matrix and X, Z, Y denote the corresponding observed matrix/vector (with a slight abuse of notation). Note that when the exposure and the instrument are both just one-dimensional variables, $\hat{\beta}_{2SLS}$ reduces to the famous Wald estimator (Wald, 1940):

$$\hat{\beta}_{Wald} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(z_i - \bar{z})}{\sum_{i=1}^{n} (x_i - \bar{x})(z_i - \bar{z})} = \frac{\hat{\beta}_Y}{\hat{\beta}_X}$$

where $\hat{\beta}_Y$ is the regression coefficient for regressing Y on Z and $\hat{\beta}_X$ is the regression coefficient for regressing X on Z. Therefore, this estimator is sometimes also called the ratio estimator (e.g. Burgess and Thompson, 2021).

The following theorem establishes the asymptotics of the 2SLS estimator.

Theorem 1 (Asymptotics of the 2SLS estimator)**.** *Under standard regularity conditions, we have*

$$\sqrt{n} \left(\hat{\beta}_{2SLS} - \beta \right) \xrightarrow{d} \mathbb{N}\left(0, \Sigma\right)$$

where $\Sigma = (\Sigma_{zx}^T \Sigma_{zz}^{-1} \Sigma_{zx})^{-1} \Sigma_{zx}^T \Sigma_{zz}^{-1}\, \Omega\, \Sigma_{zz}^{-1} \Sigma_{zx} (\Sigma_{zx}^T \Sigma_{zz}^{-1} \Sigma_{zx})^{-1}, \Sigma_{zx} = \mathbb{E}\left[ZX^T\right], \Sigma_{zz} = \mathbb{E}\left[ZZ^T\right]$ *and* $\Omega = \mathbb{E}\left[\epsilon_Y^2 ZZ^T\right]$.

Proof. Plug the structural equation (1.12) into equation (1.19), we have

$$
\begin{aligned}
\hat{\beta}_{2SLS} &= \beta + \left(X^T P_Z X\right)^{-1} X^T P_Z \epsilon_Y \\
&= \beta + \left(\frac{1}{n} \sum_{i=1}^{n} x_i z_i^T \big(\frac{1}{n} \sum_{i=1}^{n} z_i z_i^T \big)^{-1} \frac{1}{n} \sum_{i=1}^{n} z_i x_i^T \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} x_i z_i^T \big(\frac{1}{n} \sum_{i=1}^{n} z_i z_i^T \big)^{-1} \frac{1}{n} \sum_{i=1}^{n} z_i \epsilon_{Y,i} \\
&\xrightarrow{p} \beta + (\Sigma_{zx}^T \Sigma_{zz}^{-1} \Sigma_{zx})^{-1} \Sigma_{zx}^T \Sigma_{zz}^{-1}\, \mathbb{E}[Z\epsilon_Y] \\
&= \beta
\end{aligned}
$$

where the last equality is due to the IV exogeneity assumption. This establishes the asymptotic consistency of $\hat{\beta}_{2SLS}$. Similarly, we have

$$
\begin{aligned}
&\sqrt{n}\,(\hat{\beta}_{2SLS} - \beta) \\
&= \left(\frac{1}{n} \sum_{i=1}^{n} x_i z_i^T \big(\frac{1}{n} \sum_{i=1}^{n} z_i z_i^T \big)^{-1} \frac{1}{n} \sum_{i=1}^{n} z_i x_i^T \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} x_i z_i^T \big(\frac{1}{n} \sum_{i=1}^{n} z_i z_i^T \big)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_i \epsilon_{Y,i} \\
&\xrightarrow{d} (\Sigma_{zx}^T \Sigma_{zz}^{-1} \Sigma_{zx})^{-1} \Sigma_{zx}^T \Sigma_{zz}^{-1}\, \mathbb{N}\Big(0, \mathbb{E}\left[\epsilon_Y^2 ZZ^T\right]\Big) \\
&= \mathbb{N}\big(0,\, \Sigma\big)
\end{aligned}
$$

This completes the proof. □

An important note here is that, the first stage linear model (1.13) is irrelevant for the proof of Theorem 1. This is to say that we do not need to specify a linear structural equation model for the first stage to obtain $\hat{\beta}_{2SLS}$, as long as we have valid instruments and the second stage structural model is truly linear. We will call assuming a linear structural equation model between the exposure and the outcome as the *linearity assumption*.

1.2.3 Linear generalised method of moments (Linear GMM)

As discussed above, the problem with the simple IV estimator is that the empirical moment equations (1.15) do not generally have a solution when $d_Z > d_X$. Generalised method of moments (GMM) provides an alternative framework to 2SLS for estimating β with multiple instruments (Hansen and Singleton, 1982).

With the linearity assumption and Assumptions 3 - 4, the linear GMM estimator aims to minimise a weighted Euclidean norm of the empirical moment analogue of (1.15):

$$\hat{\beta}_{GMM} = \underset{\beta \in \mathbb{R}^{d_X}}{\operatorname{argmin}} \left[\frac{1}{n}\sum_{i=1}^{n} z_i(y_i - \beta^T x_i)\right]^T W_n \left[\frac{1}{n}\sum_{i=1}^{n} z_i(y_i - \beta^T x_i)\right] \tag{1.20}$$

where W_n is a sequence of weight matrices that converges in probability to some W. A closed form solution for (1.20) is given by

$$\hat{\beta}_{GMM} = \left(X^T Z W_n Z^T X\right)^{-1} X^T Z W_n Z^T Y \tag{1.21}$$

where X, Z, Y denote the corresponding observed matrix/vector (with a slight abuse of notation) and W_n is the matrix we choose. The following theorem establishes the asymptotics of the linear GMM estimator.

Theorem 2 (Asymptotics of the linear GMM estimator)**.** *Under standard regularity conditions, we have*

$$\sqrt{n}\left(\hat{\beta}_{GMM} - \beta\right) \xrightarrow{d} \mathbb{N}\big(0,\, \Sigma\big) \tag{1.22}$$

where $\Sigma = (\Sigma_{zx}^T W \Sigma_{zx})^{-1} \Sigma_{zx}^T W \, \Omega \, W \Sigma_{zx} (\Sigma_{zx}^T W \Sigma_{zx})^{-1}, \Sigma_{zx} = \mathbb{E}\left[Z X^T\right]$ *and* $\Omega = \mathbb{E}\left[\epsilon_Y^2 Z Z^T\right]$.

The proof is very similar to the proof of Theorem 1 so we will not repeat it here.

It is clear from Theorem 2 that any W will produce a consistent GMM estimator, but the choice of W affects GMM's asymptotic efficiency. As a well-known result in Hansen (1982), minimal asymptotic variance is achieved when we choose $W = \Omega^{-1}$. In practice, we may

choose

$$W_n^{-1} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}^T x_i)^2 z_i z_i^T \tag{1.23}$$

where $\hat{\beta}$ can be any consistent estimators of β, such as $\hat{\beta}_{2SLS}$, so that $W_n \xrightarrow{p} \Omega^{-1}$. The GMM estimator with this choice of weight matrix is known as the optimally weighted GMM, or $\widehat{\beta}_{OWGMM}$.

Hansen et al. (1996) suggest an iterative procedure for estimating the optimal weight matrix. Namely, we can calculate an initial GMM estimator $\hat{\beta}_{1st}$ by using an arbitrary weight matrix, say the matrix in (1.23) or just an identity matrix I_n. We can then use $\hat{\beta}_{1st}$ to estimate the optimal weight by

$$(\widehat{W}_n)^{-1} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_{1st}^T x_i)^2 z_i z_i^T$$

and use $\widehat{W}_n$ to calculate an updated GMM estimator $\hat{\beta}_{2nd}$, and repeat this procedure until a level of convergence is achieved. This is sometimes referred to as the feasible optimally weighted GMM, or $\widehat{\beta}_{FOWGMM}$.

An important observation here is that, when the error ϵ_Y is conditionally homoskedastic (i.e. $\mathbb{E}[\epsilon_Y^2 | Z] = \sigma^2$), $\Omega = \mathbb{E}[\epsilon_Y^2 Z Z^T] = \sigma^2 \mathbb{E}[Z Z^T]$. Therefore, we may choose $W_n = \left(\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^T\right)^{-1}$ in (1.20) as the constant σ^2 does not change the argmin. Plug this into (1.21) we see that the optimally weighted GMM estimator $\hat{\beta}_{OWGMM}$ is exactly the same as $\hat{\beta}_{2SLS}$ in (1.19).

Therefore, 2SLS is a special case of linear GMM in the sense that, under conditional homoskedasticity, the 2SLS estimator is the same as the optimally weighted GMM estimator. But if the error ϵ_Y is conditionally heteroskedastic (i.e. the variance $\mathbb{E}[\epsilon_Y^2 | Z]$ depends on Z), optimal GMM will be more efficient than 2SLS.

1.2.4 Alternative identification assumptions

For some practical applications, the linearity assumption may seem too strong. For example, with a binary exposure such as a drug, the linearity assumption implies (often implausibly) that the effect of the drug is exactly the same for all individuals in the population.

Researchers have proposed other assumptions that also enable us to identify causal effects using IVs. For example, with a binary instrument and a binary exposure/treatment, a weaker homogeneity assumption (without assuming linear SEM) is that the average causal effect is the same for the treated and the untreated groups across different levels of the instrument (Hernán and Robins, 2024):

$$\mathbb{E}[Y(1) - Y(0) | X = x, Z = 1] = \mathbb{E}[Y(1) - Y(0) | X = x, Z = 0] \text{ for } x = 0, 1$$

A different (and perhaps more well-known) condition is called the *monotonicity* assumption (e.g. Angrist and Imbens, 1995; Baiocchi et al., 2010) and is usually used in encouragement designs to solve the non-compliance issue. For example, with a binary instrument and a binary exposure, we can classify individuals based on their counterfactual exposure distributions:

(1). *Always-takers*: $X_i(1) = X_i(0) = 1$;
(2). *Never-takers*: $X_i(1) = X_i(0) = 0$;
(3). *Compliers*: $X_i(1) = 1$ and $X_i(0) = 0$;
(4). *Defiers*: $X_i(1) = 0$ and $X_i(0) = 1$

The monotonicity assumption requires that there are no defiers in the population. With this assumption, it can be shown that the simple IV estimator (i.e. the Wald or the ratio estimator) converges to the so-called compliers average causal effect (e.g. Zhao, 2021; Hernán and Robins, 2024), i.e. the average causal effect for the group of compliers. With a continuous exposure and a continuous causal instrument, the monotonicity assumption can also be

defined as $X(z)$ being a non-decreasing function of Z (Angrist and Imbens, 1995; Heckman and Vytlacil, 1999).

While assuming monotonicity instead of the strong linearity assumption may seem appealing, there are a couple of criticisms on this assumption. A prominent disadvantage of assuming monotonicity is that the group of compliers is not known, so we do not know who will benefit from the treatment. In practice, the majority of the population may be always-takers and never-takers, so the compliers average causal effect might be practically irrelevant. Another issue with monotonicity occurs when we use a binary proxy instrument in replacement of the underlying continuous causal instrument. This can happen, for example, if the instrument is physician's preference but we only have some binary proxy measurement of the preference. In this case, the standard IV estimator will not converge to compliers average causal effect but will converge to a weighted average of effects across the whole population, which makes the estimand hard to interpret (Hernán and Robins, 2024).

1.3 Mendelian Randomisation

1.3.1 What is Mendelian randomisation?

In one sentence, Mendelian randomisation (MR) is an epidemiological method that uses genetic variants as instrumental variables to estimate the causal effect of an exposure on an outcome (e.g. Davey Smith and Ebrahim, 2003; Burgess and Thompson, 2021). Genetic variants are good candidates of instruments as individuals' genetic variants can be largely considered as random with respect to environmental confounders due to Mendelian inheritance and random mating. There are also empirical studies that demonstrate the credibility of genetic variants being independently distributed from potential confounders (e.g. Taylor et al., 2017).

MR (in fact, general IV analysis) is an example of a so-called natural experiment, mimicking a randomised controlled trial (RCT) in the sense that the randomisation is ensured by the randomness of the genetic variants. Figure 1.6 gives a simple illustration of the similarities and differences between a randomised controlled trial and a Mendelian randomisation study.

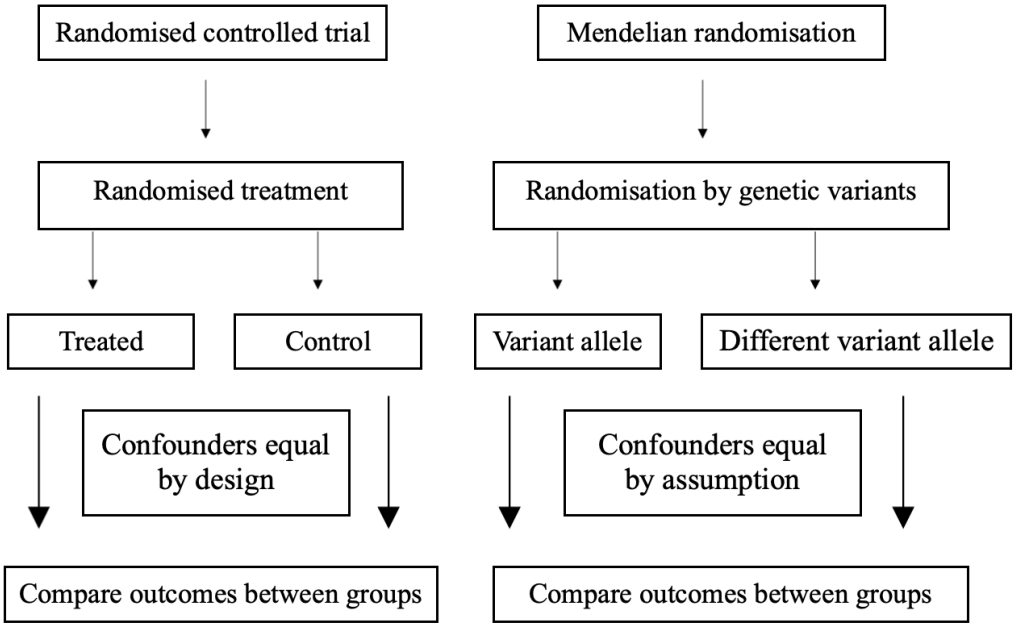


Figure 1.6: Comparison of a randomised controlled trial and a Mendelian randomisation study (adapted from Davey Smith and Ebrahim (2005) and Burgess and Thompson (2021)).

We should note that in MR literatures, the instruments, i.e. genetic variants, are often denoted as G. But for coherence of presentations we will use Z to denote the instruments (genetic or not) throughout this thesis.

1.3.2 Issues specific to Mendelian randomisation

Although a MR analysis is essentially an IV analysis, there are specific issues involved with using genetic variants as instrumental variables that we need to take into account when performing MR. A prominent issue (or feature) of genetic variants is that they are pretty

weak instruments. This means that genetic variants can only explain a very small proportion of the variance of the exposure distribution, typically only 1% - 5% as measured by R^2 in the instrument-exposure regression (Ebrahim and Davey Smith, 2008). Consequently, MR analyse are more prone to weak instrument bias, and some methods (especially complex machine-learning based methods) that work for problems with strong IVs (e.g. an encouragement design) may not work well in MR.

Another important issue with using genetic variants as IVs is that a genetic variant may be correlated with multiple variables via different causal pathways and not just associated with the exposure. This phenomenon is known as horizontal pleiotropy. On the contrary, if there is a causal effect of the exposure on the outcome, it is known as vertical pleiotropy. A horizontally pleiotropic variant is not a valid instrument, as either the no unmeasured confounding for the instrument assumption (if the variant is correlated with unmeasured confounders) or the exclusion restriction assumption (if the variant directly affects the outcome) or both assumptions will be violated. A potential reason for pleiotropy is called linkage disequilibrium (LD), which means that variants from proximate regions are correlated with each other. For example, if a causal variant of the exposure and a causal variant of the outcome are in LD, then this variant of the exposure will not be a valid instrument as the exclusion restriction assumption will be violated.

Another relevant issue in MR is known as population stratification, which refers to the phenomenon that the analysed population can be divided into distinct subpopulations. This can happen if the analysed population consists of individuals of different ethnicities. If distributions of the variants and distributions of the outcome are different across subpopulations, the variants and the outcome will be correlated because of differences in the subpopulations rather than the true causal effect (Burgess and Thompson, 2021). This will violate the no unmeasured confounding for the instrument assumption. Figure 1.7 provides a simple illustration of population stratification.

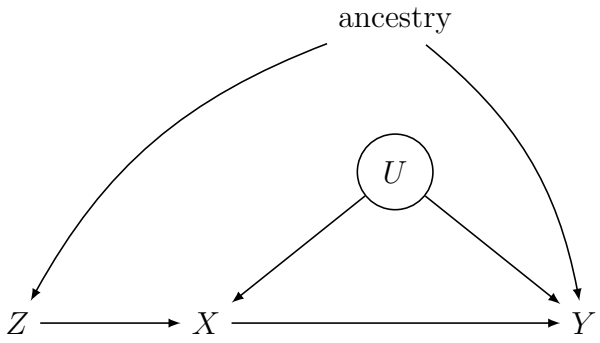


Figure 1.7: A DAG illustrating population stratification. The no unmeasured confounding for the instrument assumption is violated because of the confounding pathway through ancestry.

Throughout the rest of this thesis, we will assume that the genetic variants are all valid instrumental variables (e.g. there is no pleiotropy or population stratification), so weak instrument will be our primary concern, but researchers will need to keep these issues in mind when performing MR analyses.

1.3.3 Mendelian randomisation with individual-level data versus summarised data

MR performed using individual-level data is essentially the same as a standard IV analysis. Therefore, 2SLS, GMM and everything we have discussed so far can all be applied for MR with individual-level data.

With summarised data, on the other hand, we only have access to estimated (linear) genetic associations $\hat{\beta}_Y$'s between variants and the outcome, estimated genetic associations $\hat{\beta}_X$'s between variants and the exposure, and their respective standard errors. Summarised data are more easily accessible than individual-level data, such as in the form of a genome-wide association study (GWAS).

With a single variant j, we can obtain the ratio estimate $\hat{\theta}_j$ as

$$\hat{\theta}_j = \frac{\hat{\beta}_{Y,j}}{\hat{\beta}_{X,j}}$$

This is also referred to as the variant-specific estimate.

With multiple variants, 2SLS or GMM can no longer be applied as we do not have individual-level data. Instead, we combine the variant-specific estimates together by weighting the estimates by the inverse of their variances. This is the famous inverse-variance weighted (IVW) estimator, which, when the genetic variants are uncorrelated, can be written as (e.g. Johnson, 2012; Burgess and Thompson, 2021):

$$\hat{\theta}_{IVW} = \frac{\sum_{j=1}^{J} \hat{\theta}_j \, \mathrm{se}(\hat{\theta}_j)^{-2}}{\sum_{j=1}^{J} \mathrm{se}(\hat{\theta}_j)^{-2}} = \frac{\sum_{j=1}^{J} \hat{\beta}_{Y,j}\hat{\beta}_{X,j} \, \mathrm{se}(\hat{\beta}_{Y,j})^{-2}}{\sum_{j=1}^{J} \hat{\beta}_{X,j}^2 \, \mathrm{se}(\hat{\beta}_{Y,j})^{-2}} \tag{1.24}$$

where the latter equation is obtained using the first-order delta method approximation of the variance of the ratio estimator

$$\mathrm{se}\left(\hat{\theta}_j\right) = \mathrm{se}\left(\frac{\hat{\beta}_{Y,j}}{\hat{\beta}_{X,j}}\right) \approx \left|\frac{\mathrm{se}(\hat{\beta}_{Y,j})}{\hat{\beta}_{X,j}}\right| \tag{1.25}$$

It can be shown that this IVW estimator with first-order weights can also be obtained using an inverse-variance weighted regression of the following model (Burgess et al., 2016; Burgess and Thompson, 2021):

$$\hat{\beta}_{Y,j} = \theta\hat{\beta}_{X,j} + \epsilon_j \tag{1.26}$$

where $\epsilon_j \sim \mathbb{N}\big(0, \mathrm{se}(\hat{\beta}_{Y,j})^2\big)$.

The following theorem establishes conditions under which the IVW estimator is equivalent to 2SLS estimator with individual-level data.

Theorem 3 (Equivalence of IVW and 2SLS estimator)**.** *When the instruments are uncorrelated and the errors in the instrument-outcome regressions are conditionally homoskedastic, the IVW estimator with first-order weights is equivalent to the 2SLS estimator with corresponding individual-level data.*

Proof. Using the standard regression formulation, we have

$$\hat{\beta}_Y = (Z^T Z)^{-1} Z^T Y \tag{1.27}$$

$$\hat{\beta}_X = (Z^T Z)^{-1} Z^T X \tag{1.28}$$

where X, Z, Y denote the corresponding observed matrix/vector. When the errors are conditionally homoskedastic, the variance of $\hat{\beta}_Y$ (conditional on the instruments Z) is given by

$$\mathrm{Var}(\hat{\beta}_Y) = \sigma^2 (Z^T Z)^{-1}$$

Therefore, the IVW estimator with first-order weights, obtained by inverse-variance weighted regression of $\hat{\beta}_Y$ on $\hat{\beta}_X$, is given by

$$\begin{aligned}
\hat{\theta}_{IVW} &= \left(\hat{\beta}_X^T (\sigma^{-2}(Z^T Z))\hat{\beta}_X\right)^{-1} \hat{\beta}_X^T \, \sigma^{-2}(Z^T Z)\hat{\beta}_Y \\
&= \left(X^T Z (Z^T Z)^{-1} Z^T X\right)^{-1} X^T Z (Z^T Z)^{-1} Z^T Y \\
&= \left(X^T P_Z X\right)^{-1} X^T P_Z Y \\
&= \hat{\theta}_{2SLS}
\end{aligned}$$

where the last equation is exactly the same as the formula for the 2SLS estimator in (1.19).

Note that genetic associations for the IVW estimator are estimated using univariable regressions of the outcome (and also the exposures) on each genetic variant, rather than using multivariable regressions on all genetic variants. The assumption of uncorrelated instruments ensures that the estimated genetic associations from univariable regressions are equal to those from multivariable regressions as in (1.27) and (1.28).

This completes the proof. □

We should also note that in practice, the IVW estimator will differ slightly from the 2SLS estimator even if all the conditions in Theorem 3 are satisfied. This is because the genetic

variants will be correlated in finite samples even if they are truly uncorrelated in the population. Nevertheless, the differences are likely to be negligible when we have a large enough sample size (Burgess and Thompson, 2021).

When the instruments are correlated, we need to take this correlation into account when combing variant-specific estimates. Specifically, we can specify a weighted regression model (Burgess and Thompson, 2021):

$$\hat{\beta}_Y = \theta^T \hat{\beta}_X + \epsilon \tag{1.29}$$

where $\epsilon \sim \mathbb{N}(0, \Omega)$, Ω is the covariance matrix that $\Omega_{j1,j2} = \text{se}(\hat{\beta}_{Y,j1})\text{se}(\hat{\beta}_{Y,j2})\rho_{j1,j2}$ denotes the $j1, j2_{\text{th}}$ element of Ω and $\rho_{j1,j2}$ denotes the correlation between variants $j1$ and $j2$. The IVW estimator obtained from this weighted regression is therefore given by

$$\hat{\theta}_{IVW} = \left(\hat{\beta}_X^T \Omega^{-1} \hat{\beta}_X\right)^{-1} \hat{\beta}_X^T \Omega^{-1} \hat{\beta}_Y \tag{1.30}$$

It can be shown that the IVW estimator given in (1.30) is also equivalent to the 2SLS estimator with correlated variants obtained from the corresponding individual-level data.

All the theories, results and applications presented in the following chapters require the availability of individual-level data. Therefore, we will assume we have access to individual-level data throughout the rest of this thesis unless we specifically mention otherwise.

1.4 Overview of the Thesis

Before we move on to the next chapter, we should give an overview of this thesis.

In Chapter 1, we have introduced essential languages and models for causality. We have also introduced the three core IV assumptions, and reviewed the widely used 2SLS and GMM methods applied to linear IV models. In addition, we have briefly introduced MR by identifying common issues that are specific to MR analyses and distinguishing between MR with individual-level data and summarised data.

In Chapter 2, we will go beyond linear IV models and consider estimating nonlinear causal effects using IVs. We give a critical review of recently developed nonlinear IV methods, starting from the more conventional nonlinear 2SLS and control function approaches, to more complex, machine-learning based nonparametric 2SLS and GMM methods, and to nonlinear approaches that are more tailored for MR analyses. We present all the methods under a unified causal model, and compare the empirical performances of the methods reviewed via a simulation study.

In Chapter 3, we will introduce a novel nonlinear MR method with discrete instrumental variables, which we call nonparametric MR, that improves upon existing nonlinear MR approaches in the sense that it requires weaker assumptions and is able to estimate more complicated causal relationships. We demonstrate the empirical performance of our approach using simulation studies and real-data examples from UK Biobank that consider the causal effect of body-mass index (BMI) on average handgrip power and pulse rate.

In Chapter 4, we will continue our methodological development in Chapter 3, extending our nonparametric MR method to situations where the instrument is continuous. We demonstrate the empirical performance of our approach using simulation studies and real-data examples from UK Biobank that consider the causal effect of vitamin D level and alcohol intake on systolic blood pressure (SBP).

In Chapter 5, we will consider selection bias in IV and MR analyses. We start from the standard linear IV models and consider covariate adjustment and inverse probability weighting (IPW) as possible approaches to reduce selection bias. We provide a formal account of the inverse probability weighted MR estimator, and argue that IPW fails to correct for selection bias under severe degrees of selection even if the selection model is fully specified. We also argue that the commonly used method for calculating standard errors is inappropriate for inverse probability weighted MR, and give a formula for calculating robust standard error. We demonstrate our findings with simulation studies and real data examples from UK Biobank

that look at the causal effect of SBP on four different clinical outcomes.

In Chapter 5, we also examine how selection bias affects a popular nonlinear MR method, called stratified nonlinear MR, that we will introduce in Chapter 2. We argue that IPW applied to stratum-specific estimates cannot reduce selection bias if the stratification is performed on the selected sample instead of the overall population. We also demonstrate our argument through a simulation study.

In Chapter 6, we will conclude this thesis by discussing contributions and limitations of the work presented, and also giving prospects for future research.

We should also mention that the majority of the simulation experiments and real data analyses in this thesis were run using R. The only exceptions are the simulation exercises in Chapter 2 for the machine-learning based algorithms (we will make a clear introduction to them in Chapter 2), for which the authors' codes were written in Python. All of the codes we use, except for the methods that have already been published with codes (so we adopt the authors' codes), will be made publicly available through https://github.com/cunhaobruceliu/thesis_code.

Chapter 2

A Critical Review on Nonlinear Instrumental Variables and Mendelian Randomisation Methods

2.1 Introduction

We have seen in Chapter 1 (section 1.2) that the linearity assumption enables us to estimate the average causal effect (ACE) using IVs. However, situations where the linearity assumption is unlikely to hold are widespread, such as the effect of education on earnings (Card, 1994), the effect of systolic blood pressure on BMI (Sulc et al., 2022), and the effect of BMI on all-cause mortality (Burgess et al., 2024). In such situations, investigating the nonlinear relationship between the exposure and the outcome is of primary interest, as it could be used to guide thresholds for interventions (Staley and Burgess, 2017), but classic tools such as the famous two-stage-least-square (2SLS), are inappropriate for this purpose.

Literatures regarding nonlinear/nonparametric causal estimation using IVs have only begun

to arise very recently. Newey and Powell (2003) propose one of the initial ideas of extending 2SLS nonparametrically, following which several further extensions have been suggested since 2010. In the past couple of years, with increasing availability of huge datasets, machine learning researchers have proposed various flexible deep learning algorithms that are able to capture highly nonlinear structures even in the presence of unmeasured confounding with the help of IVs. At the same time, epidemiologists and biostatisticians have developed MR methods that can characterise nonlinear exposure-outcome relationships using genetic variants. However, in the literature there lacks a systematic comparison between the methods with respect to their assumptions, underlying causal models and statistical properties. Moreover, for some of the methods it is unclear what exactly the causal quantity that they are trying to estimate is or what the causal interpretation of the models they propose is. Nevertheless, although many of the nonlinear IV estimation are not very well-established, they are exciting because they show that nonlinearity should not be an obstacle for causal inference.

This chapter aims to provide a relatively comprehensive account of recent advances on estimating nonlinear causal effects using instrumental variables. Since existing methods are developed by researchers from different areas (econometricians, epidemiologists, biostatisticians and computer scientists), they naturally consider models under different scenarios and focus on different issues: some discuss theoretical soundness of the approaches, while others are more concerned about their empirical performances. We attempt to bring existing methods under a unified causal model, and compare the methods in terms of assumptions, identification strategies, estimation and inference to see their connections and differences. By systematic comparison, we hope to understand which method works better under which context, and to shed light on directions to which we can develop some of the ideas further.

In the following, we firstly introduce the nonlinear model of interest. We then discuss some of the important methods for nonlinear IV and MR analyses, and compare their performances through a simulation study.

2.2 The Nonlinear Causal Model

We now introduce the nonlinear causal model of interest. This will be the default nonlinear causal model considered for the rest of this thesis unless we specifically mention different models.

We assume a nonlinear structural equation model between the exposure and the outcome, represented by:

$$Y_i = h(X_i) + \epsilon_{Y,i} \tag{2.1}$$

where, similar to the construction for the linear IV model in Chapter 1 (section 1.2.2), the unmeasured confounder U is incorporated in the error term ϵ_Y and therefore $\mathbb{E}\left[\epsilon_Y \,|X\right] \neq 0$ in general. Our causal quantity of interest is the (potentially) nonlinear structural function $h(\cdot)$. For simplicity, we will just write h when there is no ambiguity. Similar to section 1.2.2, by assuming structural equation model (2.1), we implicitly impose the homogeneity assumption that the structural function h is deterministic and does not vary across individuals.

We make the following assumptions about the instrument Z.

Assumption 5 (relevance)**.** *The conditional density $F(x|z)$ is non-constant in z.*

Assumption 6 (conditional mean independence)**.** *The conditional expectation $\mathbb{E}\left[\epsilon_Y \,|Z\right] = 0$.*

Note that the three core IV assumptions are sufficient for the above assumptions made on the instrument Z. Assumption 5 is the usual relevance condition. Assumption 6 follows from the no unmeasured confounding for the instrument and the exclusion restriction conditions, though it is stronger than the IV exogeneity assumption (Assumption 4) we impose for the linear IV model.

It is also worth mentioning that this formation is general enough to include exogenous measured covariates $\{V_i\}_{i=1}^n$ (i.e. they do not suffer from unmeasured confounding) by simply

replacing X_i with $(X_i^T, V_i^T)^T$ and Z_i with $(Z_i^T, V_i^T)^T$. For simplicity, we do not write out V_i explicitly.

We should highlight that what makes this model causal is the *structural relationship* we impose between the exposure and the outcome. If we recall the potential outcomes framework introduced in Chapter 1 (section 1.1.1), this is to say that if we observe one's exposure $X_i = x$, then his/her potential outcome is given by $Y_i(x) = h(x) + \epsilon_{Y,i}$, and this potential outcome $Y_i(x)$ coincides with the observed outcome Y_i by the consistency assumption. This is not to say that, if we have an (out-of-sample) x, the model can give us a corresponding y. In fact, such prediction is precisely not what the model does. Instead, if we observe another exposure value x', then the model tells us that the causal effect of the exposure can be characterised by $Y_i(x) - Y_i(x') = h(x) - h(x')$. And because of the homogeneous structural equation assumption, this causal effect is the same across all individuals.

2.3 Nonlinear 2SLS and Control Function Method

We start from the simpler case where we assume the nonlinear structural function h is known in the sense that it can be represented as a sum of J known basis functions:

$$h(x) = \sum_{j=1}^{J} \beta_j p_j(x)$$

so the structural equation (2.1) becomes:

$$Y_i = \sum_{j=1}^{J} \beta_j p_j(X_i) + \epsilon_{Y,i} \tag{2.2}$$

Our task simply becomes to estimate the coefficients β_j's.

In a similar fashion, let

$$p_j(X_i) = \sum_{k=1}^{K} \gamma_{j,k}\, q_k(Z_i) + \epsilon_{X,j}, \quad K \geq J \tag{2.3}$$

where $\{q_k(\cdot)\}_{k=1}^K$ is another set of known basis functions assumed to be linearly independent. Assumption 6 ensures that $\{q_k(Z)\}_{k=1}^K$ are valid instruments. We assume $K \geq J$ because we need at least many instruments as the number of exposures.

2.3.1 Nonlinear 2SLS

We can proceed to estimate β_j's following the usual 2SLS framework, with $\{p_j(X)\}_{j=1}^J$ as the exposure and $\{q_k(Z)\}_{k=1}^K$ as the instruments. This is the nonlinear 2SLS estimator for $\beta = (\beta_1, ..., \beta_J)^T$, given by

$$\hat{\beta}_{2SLS} = \underset{\beta \in \mathbb{R}^J}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{J} \beta_j \sum_{k=1}^{K} \hat{\gamma}_{j,k}^T \, q_k(z_i) \right)^2 \tag{2.4}$$

where $\hat{\gamma}_j = (\hat{\gamma}_{j,1}^T, ..., \hat{\gamma}_{j,K}^T)^T$ are the first stage estimates given by

$$\hat{\gamma}_j = \underset{\gamma_j \in \mathbb{R}^{K d_Z}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(p_j(x_i) - \sum_{k=1}^{K} \gamma_{j,k}^T \, q_k(z_i) \right)^2 \tag{2.5}$$

Using the standard regression formula, the nonlinear 2SLS estimator for the structural function h is given by

$$\hat{h}(x) = \hat{\beta}_{2SLS}^T \, p(x) \tag{2.6}$$

where

$$\hat{\beta}_{2SLS} = \left[P^T Q (Q^T Q)^{-1} Q^T P \right]^{-1} P^T Q (Q^T Q)^{-1} Q^T Y$$

$$P = \left[p(x_1), ..., p(x_n) \right]^T, Q = \left[q(z_1), ..., q(z_n) \right]^T, p(x) = \left[p_1(x), ..., p_J(x) \right]^T, q(z) = \left[q_1(z)^T, ..., q_K(z)^T \right]^T$$

Recall that the first stage model (equation (2.3)) needs not to be correctly specified for $\hat{\beta}_{2SLS}$ to be consistent, as long as the relevance assumption is satisfied and $\{q_k(Z)\}_{k=1}^K$ are valid instruments, which is ensured by Assumption 6. Inference about $\hat{\beta}_{2SLS}$ can be done using the usual normal approximation.

2.3.2 Control function method

We know that the usual 2SLS framework, applied to the nonlinear model in (2.2), will still give us valid estimates. The question now becomes: is it still efficient?

The answer is not necessarily, because there exists a more efficient estimator, known as the control function estimator, that is more efficient than nonlinear 2SLS if we are willing to make additional modelling assumptions (Guo and Small, 2016). The control function method is also referred to as the two stage residual inclusion method in some medical statistics literature (e.g. Nagelkerke et al. (2000)).

The control function estimator for $\beta = (\beta_1, ..., \beta_J)^T$ is given by (e.g. Guo and Small, 2016):

$$\hat{\beta}_{CF}, \hat{\delta} = \underset{\beta \in \mathbb{R}^J, \delta \in \mathbb{R}^{d_X}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{J} \beta_j p_j(x_i) - \delta^T \hat{\epsilon}_{X,i} \right)^2 \tag{2.7}$$

where $\hat{\epsilon}_{X,i}$ is the residual in the first stage regression. I.e.

$$\hat{\epsilon}_{X,i} = x_i - \sum_{k=1}^{K} \hat{\Gamma}_k^T \, q_k(z_i) \tag{2.8}$$

where

$$\hat{\Gamma}_k = \underset{\Gamma_k \in \mathbb{R}^{d_Z \times d_Z}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(x_i - \sum_{k=1}^{K} \Gamma_k^T \, q_k(z_i) \right)^2 \tag{2.9}$$

The residuals $\hat{\epsilon}_{X,i}$ are called the control functions. A closed form solution for $\hat{\beta}_{CF}$ can be obtained using standard regression formula. Similar to $\hat{\beta}_{2SLS}$, inference about $\hat{\beta}_{CF}$ can be done using the usual normal approximation.

To investigate the control function estimator further, we observe that, due to the residual-based regression formula (i.e. partitioned regression formula), $\hat{\beta}_{CF}$ in (2.7) is also given by

$$\hat{\beta}_{CF} = \underset{\beta \in \mathbb{R}^J}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{J} \beta_j \widetilde{p_j(x_i)} \right)^2 \tag{2.10}$$

where $\widetilde{p_j(x_i)}$ are the residuals obtained by regressing $p_j(X_i)$ on $\hat{\epsilon}_{X,i}$. In this sense, we can also view (2.10) as the second stage regression of the control function approach.

Note that when the model is linear, $p_j(x) = x_j$ which is simply the j_{th} element of x. I.e.

$$h(x) = \sum_{j=1}^{J} \beta_j p_j(x) = \sum_{j=1}^{d_X} \beta_j x_j = \beta^T x$$

Then it is straightforward to show that

$$\widetilde{p(x_i)} = \left[\widetilde{p_1(x_i)}, ..., \widetilde{p_J(x_i)}\right]^T = \widetilde{x}_i = \widehat{x}_i = \sum_{k=1}^{K} \hat{\Gamma}_k \, q_k(z_i)$$

which is simply the fitted value from the first stage regression. Therefore, (2.10) is identical to the second stage regression in the 2SLS framework and the control function estimator will be equivalent to the 2SLS estimator.

However, when the model is nonlinear as in (2.2), we define

$$e_{j,i} = \widetilde{p_j(x_i)} - \widehat{p_j(x_i)} \tag{2.11}$$

where $\widehat{p_j(x_i)}$ is fitted value obtained by regressing $p_j(X_i)$ on instruments $\{q_k(Z_i)\}_{k=1}^{K}$.

Let E_j denotes the random variable whose realisation is given by $e_{j,i}$. A caveat here is that these $\{E_j\}_{j=1}^{J}$ may not be orthogonal to each other. This correlation may be removed by sequentially regressing E_{j+1} on E_j and set E_{j+1} equal to the residual $\tilde{E}_{j+1}$ of the regression (with a slight abuse of notation). Guo and Small (2016) show that the control function estimator is equivalent to the 2SLS estimator with an augmented set of instruments $\{q_1(Z), ..., q_K(Z), E_1, ..., E_J\}$.

Therefore, the control function estimator is consistent and more efficient than the nonlinear 2SLS estimator provided $\{E_1, ..., E_J\}$ are valid instrumental variables. However, if $\{E_1, ..., E_J\}$ are not valid instruments, the control function estimator will be inconsistent.

Unfortunately, Assumptions 5 and 6 are not enough to ensure that $\{E_1, ..., E_J\}$ are valid

instruments. Additional modelling assumptions are required. A set of sufficient conditions that have been proposed in the literature is the following:

(1). ϵ_Y and ϵ_X are independent of the instrument Z;
(2). $\mathbb{E}[\epsilon_Y|\epsilon_X] = \rho^T \epsilon_X$ where $\rho = \left(\mathbb{E}[\epsilon_X \epsilon_X^T]\right)^{-1} \mathbb{E}[\epsilon_Y \epsilon_X]$.

ϵ_Y and ϵ_X are the second and first stage errors as in (2.2) and (2.3). Note in particular that ϵ_X being independent of the instrument Z requires that the first stage model (2.3) to be correctly specified. Recall we have discussed in section 2.3.1 that for nonlinear 2SLS to be consistent, the first stage model does not need to be correctly specified. This illustrates one of the stronger conditions imposed by the control function method.

We can test the validity of the augmented set of instruments using, for example, the Hausman test (Hausman, 1978), whose idea is to compare $\hat{\beta}_{CF}$ and $\hat{\beta}_{2SLS}$ asymptotically. Under the null that $\{E_1, ..., E_J\}$ are valid instruments, both estimators are consistent though $\hat{\beta}_{CF}$ is more efficient. Under the alternative, only $\hat{\beta}_{2SLS}$ will be consistent. The test statistic is given by

$$H(\hat{\beta}_{CF}, \hat{\beta}_{2SLS}) = (\hat{\beta}_{CF} - \hat{\beta}_{2SLS})^T \Big[Var(\hat{\beta}_{2SLS}) - Var(\hat{\beta}_{CF}) \Big]^{-1} (\hat{\beta}_{CF} - \hat{\beta}_{2SLS}) \tag{2.12}$$

which is asymptotically $\mathcal{X}_1^2$ distributed under the null (Guo and Small, 2016).

2.4 Nonparametric 2SLS

We now move to the more difficult case where the structural function h is unknown. In this section we review some of the recent developments in generalising two-stage-least-square (2SLS) nonparametrically. The term "nonparametric" means that we do not assume any parametric form for h.

When h is unknown, the assumptions in section 2.2 are not enough to identify h. Newey and

Powell (2003) propose the following completeness condition that ensures identification.

Assumption 7 (Completeness (Newey and Powell, 2003; Newey, 2013))**.** *Expectation of the outcome conditional on the instrument is injective. I.e. there is a unique solution h that satisfies*

$$\mathbb{E}\left[Y\,|Z=z\right]=\mathbb{E}\left[h(X)\,|Z=z\right]=\int h(x)dF(x|z) \tag{2.13}$$

Equation (2.13) can be obtained by plugging the structural equation $Y=h(X)+\epsilon_Y$ into the conditional expectation $\mathbb{E}\left[Y\,|Z=z\right]$. If h_1 and h_2 both solve the integral equation, then $\mathbb{E}\left[h_1(X)-h_2(X)\,|Z=z\right]=0$. So the completeness condition requires that $q(X)=0$ is the only solution to $\mathbb{E}\left[q(X)\,|Z=z\right]=0$.

Newey (2013) note that for the linear IV model, the completeness condition is the same as the usual rank condition (the Gram matrix X^TX is nonsingular). However, Canay et al. (2013) point out that the completeness condition is not empirically testable in a fully nonparametric model.

Equation (2.13), or just $\mathbb{E}\left[Y\,|Z\right]=\mathbb{E}\left[h(X)\,|Z\right]$, is known as the *reduced form* in econometrics literatures, and as the *operator equation* in machine learning literatures. After specifying the causal model and establishing identification, our task simply becomes to solve equation (2.13) to obtain h. This is, however, a well-known ill-posed inverse problem (e.g. Nashed and Wahba, 1974; Kress, 1989). To get a well-posed problem, researchers adopt a number of regularisation conditions, which we will see when reviewing specific methods in the following.

2.4.1 Sieve IV

Motivated by the simpler case in section 2.3, when the structural function h is unknown, a natural solution is to hope to approximate h using linear combinations of J known basis

functions:

$$h(x) = \sum_{j=1}^{J} \beta_j p_j(x) + \epsilon_h$$

where ϵ_h is asymptotically ignorable error for large enough J. This is a sensible approach as most function can be approximated arbitrarily well as J grows (Newey, 2013). An estimator for $\beta = (\beta_1, ..., \beta_J)^T$ is given by

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^J}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{J} \beta_j \, \widehat{\mathbb{E}} \left[p_j(x_i) | z_i \right] \right)^2 \tag{2.14}$$

where $\widehat{\mathbb{E}} \left[p_j(x_i) | z_i \right]$ is some nonparametric estimator of the conditional expectation. In a similar fashion, let

$$p_j(x) = \sum_{k=1}^{K} \gamma_{j,k} \, q_k(z) + \epsilon_{X,j}, \quad K \geq J \tag{2.15}$$

where $\{q_k(\cdot)\}_{k=1}^{K}$ is another set of known basis functions defined on $\mathcal{Z}$. $\widehat{\mathbb{E}} \left[p_j(x_i) | z_i \right]$ can then be obtained by running OLS regression on (2.15).

This way of estimating coefficients β_j's via basis function approximation is simply the same as the nonlinear 2SLS method discussed in section 2.3.1. It was first introduced by Newey and Powell (2003) as the series estimator, and then further discussed by Newey (2013). It has been subsequently referred to as the sieve IV estimator (e.g. Chen and Christensen, 2018; Singh et al., 2019; Xu et al., 2021), which we adopt in this thesis.

The sieve IV estimator for the structural function h is therefore of the same form as the nonlinear 2SLS estimator given by (2.6).

Regularised versions of sieve IV have also been proposed by adding a ridge-type penalty to (2.14) (e.g. Blundell et al., 2007; Newey, 2013; Singh et al., 2019):

$$\hat{\beta}_\lambda = \underset{\beta \in \mathbb{R}^J}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{J} \beta_j \, \widehat{\mathbb{E}} \left[p_j(x_i) | z_i \right] \right)^2 + \lambda \, \beta^T \Sigma \beta \tag{2.16}$$

where Σ is some weight matrix. As noted by Newey (2013), the penalty tends to reduce the variance of the estimator, and hence helps stabilise the estimator especially when J is large,

allowing for high-order nonlinearities. Singh et al. (2019) demonstrate through simulation studies that the regularisation is able to improve the performance of sieve IV estimator massively.

Newey and Powell (2003) provide a consistency result for $\hat{\beta}$ by placing standard regularity conditions (e.g. compactness and boundedness) on the parameter space with additional technical constraints on convergence. Their argument applies for both unpenalised and penalised verions of sieve IV.

An advantage of sieve IV is that it allows us to conduct inference simply by estimating the standard error of β_j's using normal approximation in the usual 2SLS framework. Newey (1997) prove that the usual OLS asymptotic results are valid for single stage sieve estimators, while Newey (2013) show through simulation studies that the usual 2SLS asymptotics work reasonably well for the sieve IV estimator.

However, an obvious difficulty with sieve IV is how to choose the basis functions. Newey and Powell (2003) consider using Hermite polynomials:

$$g_j(x) = \exp(-x^T x) \prod_{t=1}^{d_X} x_t^{\lambda_t(j)}$$

where $\lambda_t(j) \in \mathbb{N}$ and $\sum_{t=1}^{d_X} \lambda_t(j)$ is increasing in j, while Singh et al. (2019) use B-spline in their experiments. In practice, which or how many basis functions to choose will largely depend on the question at hand.

2.4.2 Kernel IV

A different nonparametric IV approach is the kernel instrumental variable regression (KIV) method proposed by Singh et al. (2019), who cast the IV problem into a reproducing kernel Hilbert space (RKHS) framework to overcome the ill-posed functional equation problem. We give a quick overview on basic theories of kernel methods and kernel ridge regression in

Appendix A.1.

Singh et al. (2019) define the following RKHS model:

Let $k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be positive definite kernels corresponding to RKHSs $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Z}}$. Let $\psi : \mathcal{X} \to \mathcal{H}_{\mathcal{X}},\ x \mapsto k_{\mathcal{X}}(\cdot, x)$ and $\phi : \mathcal{Z} \to \mathcal{H}_{\mathcal{Z}},\ z \mapsto k_{\mathcal{Z}}(\cdot, z)$ be feature maps. We assume $h \in \mathcal{H}_{\mathcal{X}}$.

Note that by assuming h belongs to the RKHS, we can write h as linear combinations of feature maps (see Appendix A.1.3 for justification):

$$h(\cdot) = \sum_{i=1}^{n} \alpha_i \psi(x_i) \tag{2.17}$$

Therefore, kernel IV extends the idea of sieve IV in the sense that it allows for (potentially) infinite number of basis functions using feature maps, as we can allow $n \to \infty$. Moreover, because the algorithm for learning h is going to be defined in terms of the kernel, there is no need to specify the feature maps (i.e. basis functions) explicitly.

Define the conditional expectation operator E that

$$E : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Z}}, \quad [Eh](z) = \mathbb{E}\left[h(X) \,|Z = z\right] \tag{2.18}$$

which is a natural objective for estimation in the first stage. Also define the conditional mean embedding μ that

$$\mu : Z \to \mathcal{H}_{\mathcal{X}}, \quad \mu(z) = \mathbb{E}\left[\psi(X) \,|Z = z\right] \tag{2.19}$$

which is perhaps a more popular object in kernel literatures. Notice that E gives the same information as μ, since $\mu(z) = E^*\phi(z)$, where $E^* : \mathcal{H}_{\mathcal{Z}} \to \mathcal{H}_{\mathcal{X}}$ is the adjoint of E.

The structural function $h : \mathcal{X} \to \mathcal{Y}$ is the object of interest in the second stage. Note that h provides the same information as the so-called structural operator $H : \mathcal{H}_{\mathcal{X}} \to \mathcal{Y}$, where $h(x) = H\psi(x)$. In their original construction, Singh et al. (2019) first propose an estimator for H, which implies an estimator for h. But for our purpose, we will define an estimator for h directly.

We assume, following Singh et al. (2019), that $H \in \mathcal{H}_\Theta$ and $E \in \mathcal{H}_\Sigma$, where $\mathcal{H}_\Theta$ is a RKHS isometrically isomorphic to $\mathcal{H}_\mathcal{X}$, and $\mathcal{H}_\Sigma$ is the vector-valued RKHS of operators mapping from $\mathcal{H}_\mathcal{X}$ to $\mathcal{H}_\mathcal{Z}$. We can cleverly choose an appropriate kernel Σ such that $\mathcal{H}_\Sigma = \mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{Z})$, the space of Hilbert-Schmidt operators from $\mathcal{H}_\mathcal{X}$ to $\mathcal{H}_\mathcal{Z}$, and they share the same norm.

Figure 2.1 provides a graphical demonstration of the RKHS model described.

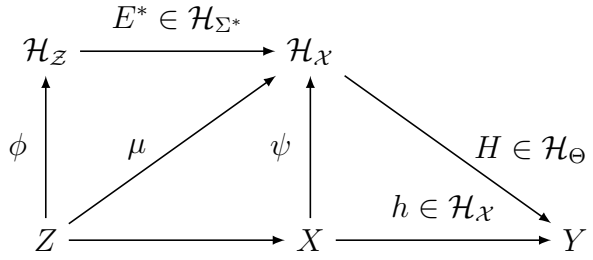


Figure 2.1: The RKHS model (adapted from Singh et al. (2019))

Following the standard formation of kernel ridge regression, stage 1 objective function for learning E and its empirical analogue are given by

$$\begin{aligned} E_\lambda &= \operatorname*{argmin}_{E \in \mathcal{H}_\Sigma} \mathbb{E}\|\psi(X) - E^*\phi(Z)\|^2_{\mathcal{H}_\mathcal{X}} + \lambda \|E\|^2_{\mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{Z})} \\ \widehat{E}_\lambda &= \operatorname*{argmin}_{E \in \mathcal{H}_\Sigma} \frac{1}{n} \sum_{i=1}^n \|\psi(x_i) - E^*\phi(z_i)\|^2_{\mathcal{H}_\mathcal{X}} + \lambda \|E\|^2_{\mathcal{L}_2(\mathcal{H}_\mathcal{X}, \mathcal{H}_\mathcal{Z})} \end{aligned} \tag{2.20}$$

Similarly, stage 2 objective function for learning h and its empirical analogue are given by

$$\begin{aligned} h_\kappa &= \operatorname*{argmin}_{h \in \mathcal{H}_\mathcal{X}} \mathbb{E}\Big(Y - [Eh](Z)\Big)^2 + \kappa \|h\|^2_{\mathcal{H}_\mathcal{X}} \\ \widetilde{h}_\kappa &= \operatorname*{argmin}_{h \in \mathcal{H}_\mathcal{X}} \frac{1}{n} \sum_{i=1}^n \Big(y_i - [Eh](z_i)\Big)^2 + \kappa \|h\|^2_{\mathcal{H}_\mathcal{X}} \end{aligned} \tag{2.21}$$

Equation (2.21) cannot be solved because we do not know the conditional expectation operator E. As in the usual 2SLS approach, we replace it with the first stage estimate. So the kernel IV estimator for h is given by

$$\widehat{h}_{\kappa,\lambda} = \operatorname*{argmin}_{h \in \mathcal{H}_\mathcal{X}} \frac{1}{n} \sum_{i=1}^n \Big(y_i - [\widehat{E}_\lambda h](z_i)\Big)^2 + \kappa \|h\|^2_{\mathcal{H}_\mathcal{X}} \tag{2.22}$$

Since both the first and second stages for obtaining the kernel IV estimator are just kernel ridge regressions, we can repeatedly apply the famous representer theorem (e.g. Scholkopf et al., 2001; stated in Appendix A.1.4 as Theorem 13) to obtain a closed form solution. This is stated below as a theorem.

Theorem 4 (Kernel IV estimator (Singh et al., 2019))**.** *Let y denotes vector of the outcome, K_{XX} and K_{ZZ} denote the kernel matrices associated with kernels $k_{\mathcal{X}}$ and $k_{\mathcal{Z}}$. Assume $k_{\mathcal{X}}$ and $k_{\mathcal{Z}}$ are bounded and continuous, ψ and ϕ are measurable and $k_{\mathcal{X}}$ is characteristic (Sriperumbudur et al., 2010)). Then the kernel IV estimator for the structural function h is given by*

$$\widehat{h}_{\kappa,\lambda}(x) = \hat{\alpha}^T K_{Xx}$$

where $\hat{\alpha} = (WW^T + n\kappa K_{XX})^{-1} W y$, $W = K_{XX}(K_{ZZ} + n\lambda I)^{-1} K_{ZZ}$, and $K_{Xx} = \big(k(x, x_1), ..., k(x, x_n)\big)^T$ is the kernel vector.

Sketch of proof. We firstly explain what we mean by a characteristic kernel.

Definition 1 (Sriperumbudur et al. (2010))**.** Let k be a measurable and bounded kernel defined on $\mathcal{X}$, and $\mathbb{P}$ be some Borel probability measure. k is called a characteristic kernel if the map

$$\mathbb{P} \to \int_{\mathcal{X}} k(\cdot, x)\, d\mathbb{P}(x)$$

is injective.

To prove Theorem 4, we will also need the following lemma:

Lemma 1 (Singh et al. (2019))**.** *Assume $k_{\mathcal{X}}$ and $k_{\mathcal{Z}}$ are bounded and continuous, ψ and ϕ are measurable and $k_{\mathcal{X}}$ is characteristic. We have*

$$\mathbb{E}\left[h(X)\,|Z = z\right] = [Eh](z) = \langle h, \mu(z)\rangle_{\mathcal{H}_{\mathcal{X}}} = H\mu(z)$$

We will present key steps of the proof, as they provide useful insights into how kernel IV generalises the idea of basis function approximation to allow infinite number of basis functions in the RKHS. For full derivation, see Appendix A.5.1 of Singh et al. (2019).

Stage 1 optimisation, i.e. equation (2.20) can be rewritten as

$$\widehat{E}_\lambda = \underset{E\in\mathcal{H}_\Sigma}{\operatorname{argmin}} \frac{1}{n}\left\|\Psi_X - E^*\Phi_Z\right\|^2_{\mathcal{H}_\mathcal{X}} + \lambda\|E\|^2_{\mathcal{L}_2(\mathcal{H}_\mathcal{X},\mathcal{H}_\mathcal{Z})} \tag{2.23}$$

where the i_{th} column of Ψ_X is $\psi(x_i)$, and the same goes for Φ_Z. By the standard regression formula we have

$$(\widehat{E}_\lambda)^* = \Psi_X(K_{ZZ} + n\lambda I_n)^{-1}\Phi_Z^T \tag{2.24}$$

Therefore we have

$$\begin{aligned}
\hat{\mu}_\lambda(z) = (\widehat{E}_\lambda)^*\phi(z) &= \Psi_X(K_{ZZ} + n\lambda I_n)^{-1}\Phi_Z^T\phi(z) \\
&= \Psi_X(K_{ZZ} + n\lambda I_n)^{-1}K_{Zz} \\
&= \Psi_X\varphi(z) \\
&= \sum_{i=1}^n \varphi_i(z)\psi(x_i)
\end{aligned}$$

where $\varphi(z) = (K_{ZZ} + n\lambda I_n)^{-1}K_{Zz}$.

For the second stage, by the representer theorem we know that the optimal h is a linear combination of feature maps with n observations where n is the sample size:

$$h(\cdot) = \sum_{i=1}^n \alpha_i\psi(x_i) \tag{2.25}$$

Then by Lemma 1 we have

$$\begin{aligned}
[\widehat{E}_\lambda h](z) = \langle h, \hat{\mu}(z)\rangle_{\mathcal{H}_\mathcal{X}} &= \left\langle \sum_{i=1}^n \alpha_i\psi(x_i), \sum_{j=1}^n \varphi_j(z)\psi(x_j)\right\rangle_{\mathcal{H}_\mathcal{X}} \\
&= \sum_{i=1}^n\sum_{j=1}^n \alpha_i\varphi_j(z)k_\mathcal{X}(x_i,x_j) \\
&= \alpha^T K_{XX}\varphi(z) \\
&= \alpha^T w(z)
\end{aligned} \tag{2.26}$$

where $\alpha = (\alpha_1, ..., \alpha_n)^T \in \mathbb{R}^n$ and $w(z) = K_{XX}(K_{ZZ} + n\lambda I_n)^{-1} K_{Zz}$.

Plug (2.25), (2.26) into (2.22) the objective function becomes

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \alpha^T w(z_i)\right)^2 + \kappa \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k(x_i, x_j) \\ &= \frac{1}{n}\|y - W^T\alpha\|_2^2 + \kappa\alpha^T K_{XX}\alpha \end{aligned} \tag{2.27}$$

where W is the matrix whose i_{th} column is $w(z_i)$. I.e. $W = K_{XX}(K_{ZZ} + n\lambda I_n)^{-1}K_{ZZ}$.

By standard optimisation formula, the solution to minimising (2.27) is

$$\hat{\alpha} = (WW^T + n\kappa K_{XX})^{-1} W y \tag{2.28}$$

and hence the optimal solution to (2.22) is given by

$$\widehat{h}_{\kappa,\lambda}(x) = \sum_{i=1}^{n}\hat{\alpha}_i k_{\mathcal{X}}(x, x_i) = \hat{\alpha}^T K_{Xx} \tag{2.29}$$

This completes the proof. □

Because of the inherent "missing data" problem of the potential outcomes (only one of the potential outcomes can be observed), causal inference literatures (e.g. Shalit et al., 2016) suggest the standard hyperparameter tuning methods, such as cross-validation, should not be used for counterfactual inference/prediction. In practice, the regularisation parameters (λ, κ) can be determined by a validation procedure on a held-out dataset. Alternatively, we can split the sample into S_1 and S_2. We can then run the first stage regression on S_1 and select λ based on S_2; similarly, we run the second stage regression on S_2 and select κ based on S_1. Singh et al. (2019) use the latter in their experiments.

Note that the above principle of hyperparameter selection applies to all machine learning type methods based on 2SLS that are going to be reviewed later (i.e. deep feature IV and deep IV), so we will not repeat it in the rest of this chapter.

By placing regularity conditions on spaces $\mathcal{Y}, \mathcal{X}, \mathcal{Z}$, kernels $k_{\mathcal{X}}, k_{\mathcal{Z}}$, feature maps ψ, ϕ, and imposing smoothness conditions on distributions $p(x, z), p(\psi(x), y)$, Singh et al. (2019)) derive convergence results for the kernel IV estimator in terms of the expected discrepancy:

$$\mathcal{E}(h) = \mathbb{E}\Big(Y - [Eh](Z)\Big)^2$$

Singh et al. (2019) also discuss the idea of sample splitting. They consider the scenario of having access to separate datasets $\{X_i, Z_i\}_{i=1}^n$ and $\{Y_i, X_i\}_{i=1}^m$ where n and m might be different, and in the case of a single dataset $\{Y_i, X_i, Z_i\}_{i=1}^n$, splitting the sample as above. They derive an asymmetric sample splitting ratio, where $n > m$, and claim that this ratio enables the kernel IV estimator to achieve the minimax optimal rate derived by Caponnetto and De Vito (2007). I.e. the learning rate of the (two-stage) kernel IV estimator with confounded data is as good as that of the single-stage kernel ridge regression without confounding.

2.4.3 Deep feature IV

More recent approaches for nonlinear IV estimation involve modelling basis functions via neural networks. In this section we review the deep feature instrumental variable regression (DFIV) method proposed by Xu et al. (2021). Specifically, they define the following model:

Let

$$h(x) = \beta^T \psi_{\theta_X}(x) \tag{2.30}$$

$$\mathbb{E}[\psi_{\theta_X}(X) \,|Z = z] = \Omega\, \phi_{\theta_Z}(z) \tag{2.31}$$

where $\beta \in \mathbb{R}^{d_1}$, $\Omega \in \mathbb{R}^{d_1 \times d_2}$ are weight parameters, and $\psi_{\theta_X}(x) \in \mathbb{R}^{d_1}$, $\phi_{\theta_Z}(z) \in \mathbb{R}^{d_2}$ are neural networks parametrised by θ_X and θ_Z. This is probably a more natural extension of the 2SLS framework than the kernel IV method, since the usual linear 2SLS will simply use the identity map $\psi_{\theta_X}(x) = x$ and $\phi_{\theta_Z}(z) = z$. Here, the feature maps are parametrised as neural networks to allow highly nonlinear structures.

As usual, the objective of the first stage is to learn the conditional expectation $\mathbb{E}[\psi_{\theta_X}(X) \,|Z = z]$ by estimating Ω and θ_Z. This is done by performing the following optimisation (Xu et al., 2021):

$$\widehat{\Omega}, \hat{\theta}_Z = \underset{\Omega \in \mathbb{R}^{d_1 \times d_2}, \theta_Z}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left\| \psi_{\hat{\theta}_X}(x_i) - \Omega \, \phi_{\theta_Z}(z_i) \right\|_2^2 + \lambda \|\Omega\|_2^2 \tag{2.32}$$

where θ_X is fixed at some initial level $\hat{\theta}_X$.

Note that if θ_Z is also kept fixed, (2.32) reduces to a standard ridge regression that gives a closed-form solution for $\widehat{\Omega} = \widehat{\Omega}(\hat{\theta}_X, \theta_Z; \lambda)$ that

$$\widehat{\Omega}(\hat{\theta}_X, \theta_Z; \lambda) = \widehat{\Psi}_X^T \Phi_Z \big(\Phi_Z^T \Phi_Z + n\lambda I_n\big)^{-1} \tag{2.33}$$

where $\widehat{\Psi}_X^T = \big[\psi_{\hat{\theta}_X}(x_1), ..., \psi_{\hat{\theta}_X}(x_n)\big]^T \in \mathbb{R}^{n \times d_1}$, $\Phi_Z^T = \big[\phi_{\theta_Z}(z_1), ..., \phi_{\theta_Z}(z_n)\big]^T \in \mathbb{R}^{n \times d_2}$ are feature matrices. We can then plug (2.33) into (2.32) and estimate θ_Z by gradient descent, and write $\widehat{\Omega} = \widehat{\Omega}(\hat{\theta}_X, \hat{\theta}_Z; \lambda)$.

The objective of the second stage is to learn the structural function h by estimating β and θ_X, substituting $\hat{\theta}_Z$ from the first stage. This is done by performing the following optimisation (Xu et al., 2021):

$$\hat{\beta}, \hat{\theta}_X = \underset{\beta \in \mathbb{R}^{d_1}, \theta_X}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - \beta^T \widehat{\Omega}(\theta_X, \hat{\theta}_Z; \lambda) \, \phi_{\hat{\theta}_Z}(z_i) \Big)^2 + \kappa \|\beta\|_2^2 \tag{2.34}$$

where $\widehat{\Omega}(\theta_X, \hat{\theta}_Z; \lambda) = \Psi_X^T \hat{\Phi}_Z \big(\hat{\Phi}_Z^T \hat{\Phi}_Z + n\lambda I_n\big)^{-1}$ is simply rewriting the analytical solution (2.33) with $\hat{\theta}_Z$ being the estimate from the first stage.

Similarly, if we fix θ_X, (2.34) reduces to a standard ridge regression that gives a closed-form solution for $\hat{\beta} = \hat{\beta}(\theta_X, \hat{\theta}_Z; \lambda, \kappa)$ that

$$\hat{\beta}(\theta_X, \hat{\theta}_Z; \lambda, \kappa) = \Big(\widehat{\Omega}(\theta_X, \hat{\theta}_Z; \lambda) \, \hat{\Phi}_Z^T \, \hat{\Phi}_Z \, \widehat{\Omega}(\theta_X, \hat{\theta}_Z; \lambda)^T + n\kappa I_n \Big)^{-1} \widehat{\Omega}(\theta_X, \hat{\theta}_Z; \lambda) \, \hat{\Phi}_Z^T \, y \tag{2.35}$$

where $y = [y_1, ..., y_n] \in \mathbb{R}^n$. We can then plug (2.35) into (2.34) and estimate θ_X via gradient descent, and write $\hat{\beta} = \hat{\beta}(\hat{\theta}_X, \hat{\theta}_Z; \lambda, \kappa)$.

The estimated $\hat{\theta}_X$ can then be used as an updated "initial" value of θ_X in the first stage optimisation (2.32), and the whole procedure above will be repeated. I.e. we can alternate between stage 1 and stage 2 until a level of convergence is achieved. The deep feature IV estimator for the structural function h is then given by

$$\hat{h}(x) = \hat{\beta}(\hat{\theta}_X, \hat{\theta}_Z; \lambda, \kappa)^T \psi_{\hat{\theta}_X}(x) \tag{2.36}$$

This algorithm is summarised below.

Algorithm 1 (Deep feature IV estimator (Xu et al., 2021)). *Let the first and second stage loss functions in (2.32) and (2.34) be denoted as $\mathcal{L}_1, \mathcal{L}_2$ respectively. Let $\hat{\theta}_X, \hat{\theta}_Z$ be some initial values, λ, κ be regularisation parameters, and T_1, T_2 be chosen steps in gradient descent. Do the following:*

1. **repeat**
2. return $\widehat{\Omega}(\hat{\theta}_X, \theta_Z; \lambda)$ as in (2.33);
3. **for** $t = 1$ to T_1:
4. update $\hat{\theta}_Z \leftarrow \hat{\theta}_Z - \alpha \nabla_{\theta_Z} \mathcal{L}_1\big(\widehat{\Omega}(\hat{\theta}_X, \theta_Z; \lambda), \theta_Z\big)|_{\theta_Z = \hat{\theta}_Z}$
5. return $\widehat{\Omega}(\theta_X, \hat{\theta}_Z; \lambda)$ and $\hat{\beta}(\theta_X, \hat{\theta}_Z; \lambda, \kappa)$ as in (2.35);
6. **for** $t = 1$ to T_2:
7. update $\hat{\theta}_X \leftarrow \hat{\theta}_X - \alpha \nabla_{\theta_X} \mathcal{L}_2\big(\hat{\beta}(\theta_X, \hat{\theta}_Z; \lambda, \kappa), \theta_X\big)|_{\theta_X = \hat{\theta}_X}$
8. **until convergence**
9. compute $\hat{\beta} = \hat{\beta}(\hat{\theta}_X, \hat{\theta}_Z; \lambda, \kappa)$;
10. **return** $\hat{h}(x) = \hat{\beta}^T \psi_{\hat{\theta}_X}(x)$

Xu et al. (2021) comment that the computational speed of the deep feature IV algorithm is $O(nd_1d_2 + d_2^3)$, which can be smaller to that of the kernel IV algorithm that takes $O(n^3)$. They also suggest a mini-batch type training algorithm that may speed up the computation in practice.

Let

$$\Psi_X = \{\psi_{\theta_X} : \mathcal{X} \to \mathbb{R}^{d_1};\ \theta_X \in \Theta_X\}$$
$$\mathcal{H} = \{\beta^T \psi_{\theta_X} : \mathcal{X} \to \mathcal{Y};\ \beta \in \mathbb{R}^{d_1}, \psi_{\theta_X} \in \Psi_X\}$$

be the hypothesis spaces for ψ_{θ_X} and h respectively. In Appendix B of the original paper, Xu et al. (2021) establish consistency of the deep feature IV estimator in terms of the L_2-norm with respect to $F(X)$:

$$\|h(X)\|_{F(X)} = \Big(\int \|h(X)\|_2^2 \, dF(X) \Big)^{\frac{1}{2}}$$

by placing boundedness constraint on spaces $\{\mathcal{Y}, \mathcal{Z}, \Psi_X, \mathcal{H}\}$ and identification conditions on hypothesis spaces $\{\Psi_X, \mathcal{H}\}$ (e.g. we need to assume the true $h \in \mathcal{H}$). They bound the L_2 norm using vanishing Rademacher complexities (e.g. Bartlett and Mendelson, 2002; Mohri et al., 2012).

2.4.4 Deep IV

Another framework for nonlinear IV estimation using neural networks, named deep IV, is developed by Hartford et al. (2017), who, instead of estimating the conditional expectation, propose to estimate the conditional density in the first stage. The motivation also comes from the operator equation

$$\mathbb{E}\left[Y \,|Z = z\right] = \int h(x) dF(x|z) \tag{2.37}$$

For the first stage, Hartford et al. (2017) suggest modelling the conditional density by a so-called mixture density network (e.g. Bishop, 2006), which can approximate smooth densities arbitrarily well with enough mixtures. Adapting from their constructions, we let

$$F(x|z) = F_\phi(x|z) \tag{2.38}$$

be a neural network parametrised by $\phi \in \mathbb{R}^{d_1}$. For discrete exposures, F_ϕ is modelled as a probability mass function

$$p(x = x_k|z) = \pi_k(x, z; \phi) \tag{2.39}$$

output by a neural network. For continuous exposures, F_ϕ is modelled as a Gaussian mixture distribution where component weights $\pi_k(x, z; \phi)$ and Gaussian parameters $\{\mu_k(x, z; \phi), \sigma_k(x, z; \phi)\}$ form the final layer of the neural network. F_ϕ can then trained as a standard supervised learning problem using maximum likelihood.

For the second stage, it is natural to optimise the following:

$$\widehat{\psi} = \underset{\psi \in \mathbb{R}^{d_2}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \int h_\psi(x_i) dF_{\widehat{\phi}}(x_i|z_i) \right)^2 \tag{2.40}$$

where h_ψ is a neural network parametrised by ψ and $F_{\widehat{\phi}}$ is the estimated density from the first stage. Therefore, the deep IV estimator is given by $h_{\hat{\psi}}(x)$.

This neural network regression can then be trained via (stochastic) gradient descent (SGD), but we need to account for the integral (expectation) in (2.40). To do so, Hartford et al. (2017) suggest estimating the integral with an average of samples drawn from the first stage model

$$\widehat{\psi} = \underset{\psi \in \mathbb{R}^{d_2}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \frac{1}{B} \sum_{b=1, x_{i,b} \overset{iid}{\sim} F_{\widehat{\phi}}(x_i|z_i)}^{B} h_\psi(x_{i,b}) \right)^2 \tag{2.41}$$

and unbiasedly estimate the gradient of (2.41) by drawing two independent samples from the first stage estimates, as the convergence of SGD requires that each sampled gradient is unbiased (Hartford et al., 2017). Alternatively, it is also possible to minimise an upper bound of (2.41), given by

$$\widetilde{\psi} = \underset{\psi \in \mathbb{R}^{d_2}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \sum_{b=1, x_{i,b} \overset{iid}{\sim} F_{\widehat{\phi}}(x_i|z_i)}^{B} \left(y_i - h_\psi(x_{i,b}) \right)^2 \tag{2.42}$$

where a single draw of sample is enough to calculate the gradient. This may speed up convergence of SGD, but will end up with suboptimal asymptotics as the minimisation is

only performed on an upper bound. Note that if the exposure is discrete, we will no longer have such complication as the gradient of (2.40) can then be computed exactly. We will not elaborate on the technical details of gradient calculation.

Hartford et al. (2017) provide no theoretical guarantees on the consistency of the deep IV estimator.

2.4.5 Related work

There are many other methods proposed in the literature that try to generalise 2SLS non-parametrically. For example, Darolles et al. (2011) implement the first stage by estimating the conditional density $F(x|z)$ using the Nadaraya-Watson density estimation technique and then perform a ridge regression in the second stage; Dual IV, proposed by Muandet et al. (2020), generalises 2SLS via a dual formulation and performs kernel ridge regression on this novel formulation; BoostIV, developed by Bakhitov and Singh (2022), extends the usual 2SLS in a data-driven way using gradient boosting algorithms.

Since these approaches do not differ considerably from what is already discussed, we do not elaborate further on them.

2.5 Nonlinear GMM

We have discussed in Chapter 1 (section 1.2.3) that GMM is a more general framework than 2SLS for causal estimation using IVs in the sense that, for linear IV models, 2SLS is a special case of optimally weighted GMM when the error ϵ_Y is conditionally homoskedastic given the instrument Z. Note that we have made no assumption on the conditional variance $\mathbb{E}\left[\epsilon_Y^2 \,|Z\right]$. Therefore, it is reasonable to try generalising GMM, which might provide more efficient estimators than the 2SLS framework when there exists conditional heteroskedasticity,

although systematic comparison between GMM and 2SLS in nonlinear setting is lacking in existing literatures.

In this section, we first provide an introduction to the nonlinear GMM framework, and then discuss a method proposed by Bennett et al. (2019) that attempts to generalise GMM to estimate the structural function h using neural networks.

2.5.1 Classic framework

Let the structural function $h(x) = h(x;\theta)$ be parametrised by $\theta \in \mathbb{R}^d$. Assumption 6 implies the following moment condition:

$$\mathbb{E}\left[f(Z)\,\epsilon\right] = \mathbb{E}\left[f(Z)\big(Y - h(X;\theta)\big)\right] = 0 \tag{2.43}$$

where f can be any functions defined on $\mathcal{Z}$. Therefore, we can potentially identify an infinite number of such moment conditions.

Suppose m functions $f_1, ..., f_m$ and their associated m moment conditions are sufficient to identify θ, where $m \geq d$ necessarily. Note that identification requires the following assumption:

Assumption 8. *There is a unique θ such that (2.43) is satisfied for all $f \in \{f_1, ..., f_m\}$*

Assumption 8 is the same as the usual rank condition (Assumption 3) when h is linear.

Let

$$q(R;\theta) = \big[f_1(Z)\big(Y - h(X;\theta)\big), ..., f_m(Z)\big(Y - h(X;\theta)\big)\big]^T \tag{2.44}$$

be a vector-valued function, where $R = (Y, X^T, Z^T)^T$. We have:

$$\mathbb{E}\left[q(R;\theta)\right] = 0 \tag{2.45}$$

A method of moment estimator for θ involves solving the empirical analogue of moment equation (2.45):

$$\frac{1}{n}\sum_{i=1}^{n} q(r_i; \widehat{\theta}_{MM}) = 0 \tag{2.46}$$

Similar to linear GMM discussed in section 1.2.3, the system of equations (2.46) may not yield a solution when $m > d$. The idea of GMM is to minimise a weighted Euclidean norm of the empirical moment conditions

$$\widehat{\theta}_{GMM} = \underset{\theta\in\mathbb{R}^d}{\operatorname{argmin}} \Big[\frac{1}{n}\sum_{i=1}^{n} q(r_i;\theta)\Big]^T W_n \Big[\frac{1}{n}\sum_{i=1}^{n} q(r_i;\theta)\Big] \tag{2.47}$$

where W_n is a sequence of weight matrices that converges in probability to some W. It can be shown that, under standard regularity conditions,

$$\sqrt{n}\,(\widehat{\theta}_{GMM} - \theta) \xrightarrow{d} \mathbb{N}\big(0,\, \Sigma\big) \tag{2.48}$$

where $\Sigma = (Q^T W Q)^{-1} Q^T W \Omega W Q (Q^T W Q)^{-1}, Q = \mathbb{E}\left[\nabla_\theta q(R;\theta)\right], \Omega = \mathbb{E}\left[q(R;\theta)\big(q(R;\theta)\big)^T\right]$.

Similar to linear GMM, any W will produce a consistent nonlinear GMM estimator, but the choice of W affects GMM's asymptotic efficiency. and minimal asymptotic variance is achieved when we choose $W = \Omega^{-1}$. In practice, we may choose $W_n = W_n(\tilde{\theta})$, given by

$$\big[W_n(\tilde{\theta})\big]^{-1} = \frac{1}{n}\sum_{i=1}^{n} q(r_i;\tilde{\theta})\big(q(r_i;\tilde{\theta})\big)^T \tag{2.49}$$

where $\tilde{\theta}$ can be any consistent estimate of θ, so that $W_n(\tilde{\theta})$ converges in probability to Ω^{-1} under standard regularity conditions. Note that the $(j,k)_{th}$ element of $W_n(\tilde{\theta})$ is given by

$$(W_{j,k})^{-1} = \frac{1}{n}\sum_{i=1}^{n} f_j(z_i) f_k(z_i)\big(y_i - h(x_i;\tilde{\theta})\big)^2 \tag{2.50}$$

This is the optimally weighted nonlinear GMM. The iterative procedure discussed in section 1.2.3 can be similarly applied to obtain the feasible optimally weighted nonlinear GMM (FOWGMM).

2.5.2 DeepGMM

We would like to extend the above framework to allow modelling $h(x;\theta)$ as a flexible neural network with high-dimensional θ. However, the classic framework of FOWGMM will fail when we attempt to use too many moment conditions, as, for example, the optimal weight in (2.49) will necessarily be singular (Bennett et al., 2019).

To overcome this problem, Bennett et al. (2019) propose a framework called DeepGMM based on a variational reformulation of the objective function of the optimally weighted GMM. Specifically, consider the following objective function

$$\Lambda(\theta;\mathcal{F},\tilde{\theta}) = \sup_{f\in\mathcal{F}} \left\{ \frac{1}{n}\sum_{i=1}^{n} f(z_i)\big(y_i - h(x_i;\theta)\big) - \frac{1}{4n}\sum_{i=1}^{n}\Big[f(z_i)\big(y_i - h(x_i;\tilde{\theta})\big)\Big]^2 \right\} \tag{2.51}$$

where $\mathcal{F}$ is the linear span of $\{f_1,...,f_m\}$. Also, rewrite the optimal GMM objective as

$$\mathcal{L}(\theta;q,\tilde{\theta}) = \Big[\frac{1}{n}\sum_{i=1}^{n} q(r_i;\theta)\Big]^T W_n(\tilde{\theta})\Big[\frac{1}{n}\sum_{i=1}^{n} q(r_i;\theta)\Big] \tag{2.52}$$

where $q(R;\theta)$ is defined in (2.44) and $W_n(\tilde{\theta})$ is defined as in (2.49). The following theorem establishes the equivalence of these two objectives.

Theorem 5. *(Bennett et al., 2019) Let $\mathcal{F} = span(f_1,...,f_m)$, q be defined as in (2.44) and $W_n(\tilde{\theta})$ be defined as in (2.49). Then*

$$\Lambda(\theta;\mathcal{F},\tilde{\theta}) = \mathcal{L}(\theta;q,\tilde{\theta})$$

With a slight abuse of notation, let $\mathcal{F} = \{f(z;\eta) : \eta \in \mathbb{H}\}$ be the set of neural networks of a given architecture parametrised by $\eta \in \mathbb{H}$ for some parameter space $\mathbb{H}$. Note that $\mathcal{F}$ is not a span of the neural networks, as otherwise the framework will fail due to the same reason discussed previously. Likewise, let $h(x) = h(x;\theta)$ be a neural network parametrised by $\theta \in \Theta$ for some parameter space Θ and let $\mathcal{H} = \{h(x;\theta) : \theta \in \Theta\}$.

Based on theorem 5, Bennett et al. (2019) suggest simply replacing $\mathcal{F}$ in (2.51) with this rich set of functions $\mathcal{F}$ defined by neural networks. Note that in order to guarantee identification, we need to assume the following:

Assumption 9. *There is a unique θ such that $\mathbb{E}\left[f(Z)\big(Y - h(X;\theta)\big)\right] = 0 \;\; \forall f \in \mathcal{F}$.*

The above is the same as Assumption 8 with $\{f_1, ..., f_m\}$ replaced by this rich set $\mathcal{F}$. Note that, however, this can be a stronger assumption than the completeness condition (Assumption 7), since for example, if we consider hierarchical models, there can be many parameters expressing the same model. Therefore, this is in fact a quite strong assumption for neural networks.

The DeepGMM estimator is given by

$$\widehat{\theta}_{DeepGMM} = \underset{\theta \in \Theta}{\operatorname{argmin}} \, \Lambda(\theta; \eta, \tilde{\theta}) \tag{2.53}$$

where

$$\Lambda(\theta; \eta, \tilde{\theta}) = \sup_{\eta \in \mathbb{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f(z_i; \eta)\big(y_i - h(x_i; \theta)\big) - \frac{1}{4n} \sum_{i=1}^{n} \left[f(z_i; \eta)\big(y_i - h(x_i; \tilde{\theta})\big) \right]^2 \right\} \tag{2.54}$$

and $\tilde{\theta}$ can be chosen in a data-driven way. As discussed in section 2.5.1, $\tilde{\theta}$ does not affect consistency but only asymptotic efficiency of the DeepGMM estimator, and minimal asymptotic variance is achieved if we choose $\tilde{\theta}$ to be a consistent estimate of θ. Therefore, similar to the iterative procedure discussed in section 1.2.3, Bennett et al. (2019) also suggest an iterative procedure to choose $\tilde{\theta}$. Specifically, we can obtain an initial DeepGMM estimate by choosing $\tilde{\theta}$ to be some initial value θ_0, denoted by $\widehat{\theta}_{DeepGMM1st}$. We can then obtain an updated DeepGMM estimate by choosing $\tilde{\theta}$ to be $\widehat{\theta}_{DeepGMM1st}$, as $\widehat{\theta}_{DeepGMM1st}$ is a consistent estimate of θ. Obviously, we can repeat this process to continuously update $\tilde{\theta}$ until a level of convergence is achieved.

Bennett et al. (2019) implement the optimisation (2.53) as solving a smooth zero-sum game, and use the OAdam algorithm (Daskalakis et al., 2017) in their experiments. See section 4 of the original paper for details.

Similar to the consistency result of the deep feature IV estimator, Bennett et al. (2019) provide consistency guarantees of the DeepGMM estimator by placing continuity constraint on $\{f(z;\eta), h(x;\theta)\}$, boundedness constraint on $\{Y, f(Z;\eta), h(X;\theta)\}$ and assuming vanishing Rademacher complexities (Bartlett and Mendelson, 2002; Mohri et al., 2012) on $\mathcal{F}$ and $\mathcal{H}$.

2.6 Stratified Nonlinear MR

We have so far discussed various methods that try to flexibly generalise 2SLS or GMM. However, a general problem about these two stage methods is that the estimated structural function will only be valid on the range of the exposure predicted by the instrument, and anything outside this range will be extrapolation (Burgess et al., 2014; Staley and Burgess, 2017). This is fine under the linearity assumption as the extrapolation will be valid, but it becomes problematic when the structural function is nonlinear. If the instrument is weak in the sense that it only explains a small amount of the exposure variance, our ability to detect nonlinearity is limited. This problem is particularly relevant in MR, where genetic variants can only explain a small proportion of the exposure distribution.

To overcome this limitation, Burgess et al. (2014) develop an alternative approach, which we call stratified nonlinear MR, to characterise the nonlinear structural function h. The idea is to stratify on the distribution of the exposure, estimate an average causal effect within each stratum, known as the localised average causal effect (LACE), and trace back h by treating the LACEs as the local derivatives of the structural function h (Small, 2014; Staley and Burgess, 2017). If we believe the structural function in a stratum can be approximated by a linear function well enough when the size of the stratum is small, the LACE estimates will be valid in general and can be treated as good approximates for the derivative of h if we have large enough sample size. The LACE estimates are also referred to as the stratum-specific estimates. A complication with this approach is that we cannot stratify on the exposure directly, as the exposure X is a collider on the path from Z to Y (recall the DAGs

in Figures 1.3 and 1.4). If we stratify on X directly, an association between the instrument and the unmeasured confounders will be induced and the no unmeasured confounding for the instrument assumption will be violated within strata (Didelez and Sheehan, 2007). This is known as the collider bias.

There have been two stratification methods proposed in the literature that avoid collider bias, which we call residual stratification (Burgess et al., 2014; Staley and Burgess, 2017) and doubly-ranked stratification (Tian et al., 2023). We will introduce these two methods in the following.

We note that the stratified nonlinear MR methods are only feasible when the exposure is low-dimensional. Throughout this section we will assume the exposure is a one-dimensional variable, and discuss this limitation in section 2.9.

2.6.1 Residual stratification

The first stratification method, proposed by Burgess et al. (2014), avoids collider bias by creating a new variable X_0, obtained by performing regression of the exposure X on the instrument Z and setting X_0 equal to the residual of this regression, and stratifying on X_0. This method is therefore known as the residual stratification approach. Note that X_0 is the same as the control function mentioned in section 2.3.2.

If we assume that the following first stage instrument-exposure model:

$$X_i = \gamma^T Z_i + \epsilon_{X,i} \tag{2.55}$$

where ϵ_X is the noise variable that has zero mean, finite variance and is uncorrelated with Z, is correctly specified, i.e. the effect of the instrument on the exposure is truly linear and homogeneous for all individuals (Staley and Burgess, 2017), X_0 is guaranteed to be a non-collider. This is because, under this assumption, X_0 can be interpreted as the potential

outcome $X_i(0)$ where we set the instrument value to be zero, so X_0 cannot be a child of X on the DAG. See Figure 2.2 for illustration. Because of this interpretation, X_0 is also referred to as the IV-free exposure (Burgess et al., 2014; Staley and Burgess, 2017) This linearity and homogeneity assumption on the instrument-exposure model is also referred to as the constant genetic effect assumption (Burgess, 2023).

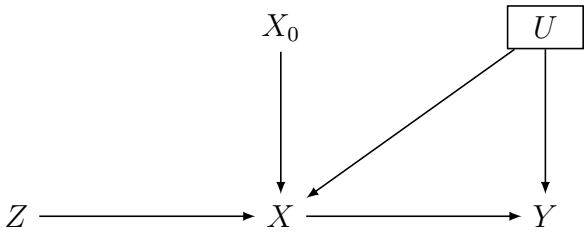


Figure 2.2: A DAG illustrating the IV-free exposure.

Nevertheless, as noted by Burgess et al. (2014), X_0 needs not to have a counterfactual interpretation for it to be a valid stratifying variable. All we need for valid stratification is a covariate that is correlated (strongly) with the exposure X but independent of the instrument Z, and this covariate needs not even have a place on the DAG. Therefore, the constant genetic effect assumption is a sufficient condition.

2.6.2 Doubly-ranked stratification

Recent literatures have suggested that the constant genetic effect assumption is a strong assumption, and is in fact violated in many practical scenarios. For example, when the exposure takes discrete values and the instrument is continuous (e.g. a genetic risk score), the constant genetic effect assumption is almost surely violated. A relevant example can be the case where the exposure values are coarsened, such as alcohol intake, which is typically measured in categories instead of the precise consumption, and sleep duration, which is usually reported in units of hours rather than minutes or seconds (Tudball et al., 2021; Tian et al., 2023). When the constant genetic effect assumption is violated, residual stratification

largely results in biased estimates (Burgess, 2023; Tian et al., 2023).

Tian et al. (2023) relax the constant genetic effect assumption by proposing an alternative method of stratification, known as the doubly-ranked stratification approach. The method firstly ranks individuals according to their values of the instrument, and divides individuals into pre-strata of a pre-specified size (usually the number of strata we want). The method then ranks individuals in each pre-stratum based on their values of the exposure, and forms the strata by grouping individuals with the same rank in each pre-stratum together. Figure 2.3 illustrates the doubly-ranked stratification method.

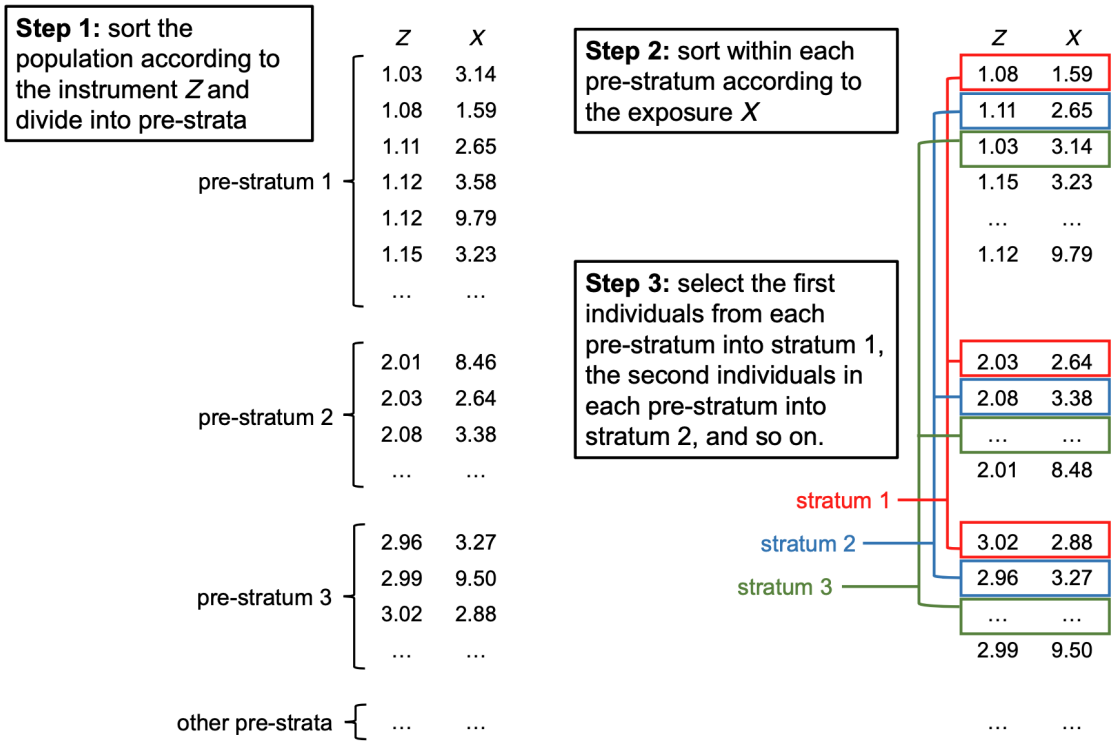


Figure 2.3: A diagram that outlines the doubly-ranked stratification method (adapted from Tian et al. (2023)).

Because the rankings of individuals in each pre-stratum do not depend on the instrument values, the doubly-ranked stratification method will not introduce collider bias and therefore the constant genetic effect assumption is not required. Instead, the method requires a strictly

weaker assumption, known as the rank-preserving assumption on counterfactual exposure distributions, that the ranks of an individual's counterfactual exposure values are identical for all values of the instrument Tian et al. (2023). This is so that the data in each stratum can be regarded as realisations of valid counterfactual variables. We will revisit and explain this assumption in more details in Chapter 3 (section 3.3.1).

One disadvantage of the doubly-ranked method is that it will only work for a single instrument. However, this is surmountable in MR applications as multiple genetic variants are typically combined into one genetic risk score. A more relevant weakness is that unlike residual stratification, the strata from doubly-ranked stratification usually cannot be created according to clinically meaningful cut-offs of the exposure values, so it is harder to interpret the stratum-specific estimates (Burgess, 2023).

2.6.3 Estimation of the structural function

Within each stratum created by either residual or doubly-ranked stratification, the localised average causal effect (LACE) is estimated via the usual Wald estimator, or the usual 2SLS estimator when we have multiple instruments. Staley and Burgess (2017) propose two methods to estimate h from the LACE estimates, including the fractional polynomial method and the piecewise linear method.

Specifically, fractional polynomials of degree 1 are given by

$$h(x) = \alpha_0 + \alpha_1 x^p \tag{2.56}$$

and fractional polynomials of degree 2 are given by

$$h(x) = \begin{cases} \alpha_0 + \alpha_1 x^{p_1} + \alpha_2 x^{p_2}, & \text{if } p_1 \neq p_2 \\ \alpha_0 + \alpha_1 x^p + \alpha_2 x^p \log(x), & \text{if } p_1 = p_2 = p \end{cases} \tag{2.57}$$

where $p, p_1, p_2 \in \{-2, -1, -\frac{1}{2}, 0, \frac{1}{2}, 1, 2, 3\}$ are standard powers for modelling fractional polynomials and x^0 is defined to be $\log(x)$ (Royston and Altman, 1994). Fractional polynomials

of degrees larger than 2 are usually not considered in practice.

Assuming the structural function h can be written as a fractional polynomial of either degree 1 or 2, we can then meta-regress LACE estimates against the mean of the exposure in each stratum using the derivative of the fractional polynomials given above (Staley and Burgess, 2017). Specifically, this involves solving the following optimisation problem:

$$\hat{\alpha} = \underset{\alpha \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{s=1}^{S} \Big(\hat{\beta}_s - h'(\bar{x}_s; \alpha) \Big)^2 \tag{2.58}$$

where $\hat{\beta}_s$ is the LACE estimate in stratum s, $\bar{x}_s$ is the mean of the exposure in stratum s, and $h(\cdot)$ is given by either (2.56) or (2.57). Note that the derivative of a fractional polynomial is just another fractional polynomial, so (2.58) is nothing more than fitting a standard fractional polynomial regression. The best-fitting fractional polynomial can then be selected using standard model selection criterions such as maximum likelihood, with standard assumption that the errors in fractional polynomial regressions are normally distributed.

A caveat here is that, since we are fitting the derivative of the fractional polynomials, the intercept term (i.e. α_0) does not appear in the derivative and hence cannot be estimated. In practice, we can only choose it arbitrarily. Nevertheless, α_0 is irrelevant for the evaluation of causal effects as any causal interpretation must be based on a difference $h(x) - h(x')$ (recall section 2.2). In this sense, this nonidentifiability issue of α_0 is not essential.

Likewise, if we assume the structural function h is continuous and piecewise linear, and our stratification correctly accounts for each change point, we can estimate h by treating each LACE estimate as the gradient of the line segment in that stratum. Because of continuity, each line segment begins where the previous segment ends (Staley and Burgess, 2017). Similar to above, the overall intercept is not identified and can only be set arbitrarily. In practice, the piecewise linear method is less popular than the fractional polynomials method. Therefore, we consider only the fractional polynomials method for stratified nonlinear MR in the rest of this thesis unless specifically mention otherwise.

The confidence interval for $\hat{h}$ can be calculated using either standard normal approximation for the estimated coefficients in the meta-regression (for the fractional polynomial method), or by bootstrapping within each stratum and re-estimating h (Staley and Burgess, 2017). One disadvantage of using normal approximation for the estimated coefficients of fractional polynomial regression is that the standard error is calculated after selecting the degree and estimating the powers of the fractional polynomial, and therefore does not consider the possibility of choosing a fractional polynomial of a different degree or with different powers. This means that this approach does not take into account the uncertainty of model selection, so the confidence interval will be overly optimistic. However, we note that the model selection uncertainty will be taken into account if we calculate the standard error using bootstrap, although bootstrapping will be computationally more intensive.

Another potential disadvantage of using the fractional polynomial method, which is common with either using normal approximation for the estimated coefficients or using bootstrap, is that the confidence interval for $\hat{h}$ might be different from the confidence intervals for the LACE estimates. This is because the fractional polynomial model imposes a smoothness constraint and looks for the best global fit across strata. The standard error of the fractional polynomial regression quantifies the uncertainty of this global fit and is propagated across the exposure distribution in a way that depends on the estimated powers of the fractional polynomials. This "global" standard error for $\hat{h}$, when interpreted locally for a given value of the exposure, might not be consistent with the standard error for the corresponding LACE estimate. When the smoothness assumption holds (i.e. the uncertainty of estimates in a given stratum is not only informed by local data, but also affected by observations in other strata in a way that can be modelled by fractional polynomials), the change in the confidence interval could be seen as a benefit of making this assumption. However, if this smoothness assumption is violated, the change in the confidence interval will give misleading conclusions for the actual uncertainties of estimated nonlinear causal effects. We will see examples of this limitation in Chapter 3 (section 3.5) and Chapter 4 (section 4.4).

Another obvious difficulty with stratified nonlinear MR is how to choose the number of strata. This is a classic bias-variance trade-off type problem that if the number of strata is too large, the sample size within each stratum will be small and the LACE estimates will be imprecise and subject to weak instrument bias (Burgess et al., 2011; Staley and Burgess, 2017). On the other hand, if the number of strata is too small, the estimation of h might be far off. In this sense, the number of strata may be regarded as a model tuning parameter. In practice, this will largely depend on the sample size at hand and can be guided by field knowledge. For example, Burgess et al. (2024) use WHO guidelines to categorise BMI into underweight, normal weight, overweight and obese. In the absence of field knowledge, the number of strata can only be set arbitrarily by the researchers (often 10 strata in practice).

2.7 Polynomial MR

Another method that tries to characterise nonlinear structural functions in the context of MR is the polynomial MR method (or simply PolyMR) proposed by Sulc et al. (2022). The idea is similar to sieve IV discussed in section 2.4.1, choosing polynomials as the basis functions. However, PolyMR adopts the construction of the control functions framework (recall section 2.3.2) instead of nonlinear 2SLS, using polynomials to approximate both $h(x)$ and ϵ_X. This is a natural extension of sieve IV, since the control functions method can be more efficient than nonlinear 2SLS as we have discussed in section 2.3.

Specifically, Sulc et al. (2022) postulate a linear first stage model

$$X_i = \gamma_0 + \gamma_1 Z_i + \epsilon_{X,i} \tag{2.59}$$

and implement the second stage in the control functions manner with polynomial approximations:

$$\hat{\beta}_{polyIV}, \hat{\delta} = \underset{\beta \in \mathbb{R}^{J+1},\, \delta \in \mathbb{R}^J}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{J} \beta_j x_i^j - \sum_{j=1}^{J} \delta_j \hat{\epsilon}_{X,i}^j \right)^2 \tag{2.60}$$

where $\hat{\epsilon}_{X,i}$ is the residual of the first stage regression (2.59) and Y, X, Z are all one-dimensional random variables centred around 0 and standardised to have variance 1 for simplicity.

Although it is not clear how polynomial approximations can be applied for exposures of higher dimensions, the same idea equally applies with polynomials replaced by other basis functions, such as those mentioned in section 2.4.1. Sulc et al. (2022) provide no theoretical justification for replacing $\hat{\epsilon}_X$ in the control functions framework with polynomials $\sum_{j=1}^{J} \delta_j \hat{\epsilon}_X^j$, but the method works reasonably well in their simulations.

Sulc et al. (2022) does not discuss either modelling assumptions or consistency results for PolyMR, but we anticipate the same assumptions for the consistency of control functions method should equally apply here, including the first stage model (2.59) being correctly specified.

2.8 Simulation Study

2.8.1 Simulation design

We now perform a simple simulation study to examine the empirical performances of the methods discussed. We only consider two designs for illustration purpose.

The first is a demand design introduced by Hartford et al. (2017) and also used by both Singh et al. (2019) and Xu et al. (2021) in their experiments. In this design, the exposure is price P, the outcome is sales Y, and the instrument is a cost shifter C. There are also two covariates, time T and customer sensitivity S, that are free from confounding. So, in our notation, $X = (P, T, S)^T$ and $Z = (C, T, S)^T$. The structural function is given by

$$h(p, t, s) = 100 + (10 + p)sf(t) - 2p$$

where $f(T)$ is a complex nonlinear function given by

$$f(t) = 2\Big(\frac{(t-5)^4}{600} + \exp\big(-4(t-5)^2\big) + \frac{t}{10}\Big)$$

We have modified $f(t)$ slightly from what is originally proposed by Hartford et al. (2017) to ensure that the sales Y is always nonnegative. The first stage exposure-instrument relationship is specified nonlinearly as

$$P_i = 25 + (C_i + 3)f(T_i) + e_i$$

where e is a standard normal random noise.

The second is a simple linear design, in which we modified the above structural function into

$$h(p, t, s) = 100 - 2p + s + \frac{t}{2}$$

The first stage is also specified linearly as

$$P_i = 25 + C_i + \frac{T_i}{5} + e_i$$

Everything else is kept unchanged. The goal of having a linear design is to see what the costs are of using the flexible nonparametric methods when the true structural relationship is simple.

We generate both designs according to the following:

$$\begin{aligned}
T_i &\overset{i.i.d}{\sim} \text{unif}\,[0, 10] \\
S_i &\overset{i.i.d}{\sim} \text{unif}\,\{1, 2, 3, 4, 5, 6, 7\} \\
C_i &\overset{i.i.d}{\sim} \mathbb{N}\big(0,\ 1\big) \\
e_i &\overset{i.i.d}{\sim} \mathbb{N}\big(0,\ 1\big)
\end{aligned}$$

and T, S, C, e are mutually independent.

We generate the output Y according to

$$Y_i = h(P_i, T_i, S_i) + \epsilon_i$$

where $\epsilon_i \overset{i.i.d}{\sim} \mathbb{N}\big(\rho e,\, 1-\rho^2\big)$, $\epsilon_i \perp\!\!\!\perp (T, S, C)$, $h(P_i, T_i, S_i)$ corresponds to either the demand function or the linear function specified before and ρ is the parameter for the level of confounding. In all simulated datasets, we take $\rho = 0.5$. The above are mostly taken from Hartford et al. (2017), who originally propose the demand design for simulation.

2.8.2 Methods compared

Specifically, we compare sieve IV (SIV), control function estimator with sieve approximation (SIV-CF), kernel IV (KIV), deep feature IV (DFIV), deep IV (DIV), DeepGMM (DGMM), doubly-ranked nonlinear MR with fractional polynomials (DR) and polynomial MR (PolyMR). For each method we consider three sample sizes: $n \in \{1000, 5000, 10000\}$ and for each sample size we average results across 20 simulated datasets due to large computational time for the deep learning algorithms (this is consistent with the simulation experiments conducted by Xu et al. (2021). Due to instability of the training procedure in DeepGMM for small samples, DeepGMM is not performed for $n = 1000$. The evaluation metric is out-of-sample mean squared error (MSE) between $\hat{h}$ and the true h.

For sieve IV, we choose a variation of the Hermite polynomials as the basis functions, consider polynomial degrees up to 6 and choose the best one based on likelihood. We consider both the penalised (simple L_2 norm penalty) and unpenalised versions of the sieve IV and report the minimum MSE. We also apply the same basis function approximation in the control functions framework (SIV-CF), but penalisation is not considered there.

For kernel IV, deep feature IV and deep IV, we train the algorithm by splitting the data equally for two stages. We use the Gaussian kernel in both stages for kernel IV.

For stratified nonlinear MR, we only implement doubly-ranked stratification as it is more robust than residual stratification. We choose the number of strata to be 10, which is what people commonly choose in practice. For simplicity, we adjust for the covariates in each

stratum but ignore the coefficients of the covariates when fitting the fractional polynomials regression. The estimated structural function h can be interpreted as the estimated h conditioning on the covariates (and assuming the effects of the covariates are also linear within strata, since the adjustment of covariates mentioned above assumes linear effects). Since we are only interested in the causal effect of the exposure, this approach is sound.

However, in order to calculate MSE, we also need to choose the intercept correctly (recall that the fractional polynomials method cannot estimate the intercept). We do this in two ways. The first way is to consider h conditioning on the covariates as above, choose the intercept as the average of the difference between the estimated conditional h and the true conditional h, and report the MSE conditioning on the covariates (this is DR-c). However, this will not be a fair comparison to MSEs from other methods, as the uncertainties of estimating the covariates are somewhat ignored. The second way is to regress the true h on the estimated conditional h and the covariates. I.e. we assume the covariates enter linearly in h for the whole distribution of the exposure (this is DR-l).

For PolyMR, as the original method only works for one-dimensional exposure, we handle the covariates by including them linearly in both the first and second stages. Similar to sieve IV, we consider polynomial degrees up to 6 and choose the best one based on likelihood.

Table 2.1 lists the abbreviations for all the above methods that we compare in our simulation study.

We use the code made publicly available by Xu et al. (2021) for deep feature IV, deep IV and DeepGMM, and use the SUMnlmr package (Mason and Burgess, 2022) for doubly-ranked nonlinear MR. We code the rest of the methods according to their constructions in their original papers.

Abbreviations for methods

Method	Abbreviation
sieve IV	SIV
control function estimator with sieve approximation	SIV-CF
kernel IV	KIV
deep IV	DIV
deep feature IV	DFIV
DeepGMM	DGMM
polynomial MR	PolyMR
doubly-ranked nonlinear MR conditioning on the covariates	DR-c
doubly-ranked nonlinear MR assuming linear effects of the covariates	DR-l

Table 2.1: A list of abbreviations for all the methods compared in the simulation study.

2.8.3 Simulation results

Figure 2.4 presents MSEs for the demand design. We note that all non-machine-learning type methods perform poorly, and increasing the sample size does not help. This is expected, as the demand structural function is neither a polynomial nor a fractional polynomial, and it cannot be well-approximated using pre-specified basis functions. It might be possible to improve the performance of sieve IV by choosing a different set of basis functions, but in practice it is not feasible to exhaust all possibilities, and from the simulation results in Singh et al. (2019), the performance of B-splines is not better than what is presented here in any notable scale.

We also note that the conditional MSE for the doubly-ranked nonlinear MR is the second smallest among all. This is expected, because conditioning on the covariates, the demand function becomes linear so the intercept (conditional on the covariates) is chosen very precisely. It is not the smallest, however, because the conditional h is estimated assuming the covariates enter linearly in h within each stratum, which is not satisfied for only 10 strata.

Among the machine learning methods, DeepGMM performs poorly in the demand design. Xu et al. (2021) observe the same, and they suggest this might be due to the way DeepGMM handles covariates. Kernel IV, deep IV and deep feature IV perform reasonably well, and increasing the sample size notably reduces the error. Deep feature IV outperforms all other methods in the demand design.

Figure 2.5 presents MSEs for the linear design. We firstly observe that the errors are in general on a much smaller scale compared to the demand design. We note that except kernel IV, all methods perform reasonably well when both the first and second stages are truly linear (except perhaps Deep IV when sample size is not large). This shows that it will not be too much a cost if we misjudge the simple situation, though the machine learning type algorithms are more costly than the rest. In particular, deep feature IV performs quite well when the sample size is large. Note that PolyMR and doubly-ranked nonlinear MR perform especially well in the linear design. This is expected, since the linear structural function belongs to the set of polynomials/fractional polynomials.

It is curious that kernel IV performs poorly in the linear design, considering it can be interpreted as a generalisation (infinite basis function approximation) of the sieve IV method. Moreover, in the simulation conducted by Singh et al. (2019), kernel IV works reasonably well in a linear design without covariates. This could be due to the asymmetric sample splitting technique that we did not implement, but more likely this shows that the way we handle covariates (simply including covariates in both the exposure vector and the instrument vector) is not ideal for kernel IV.

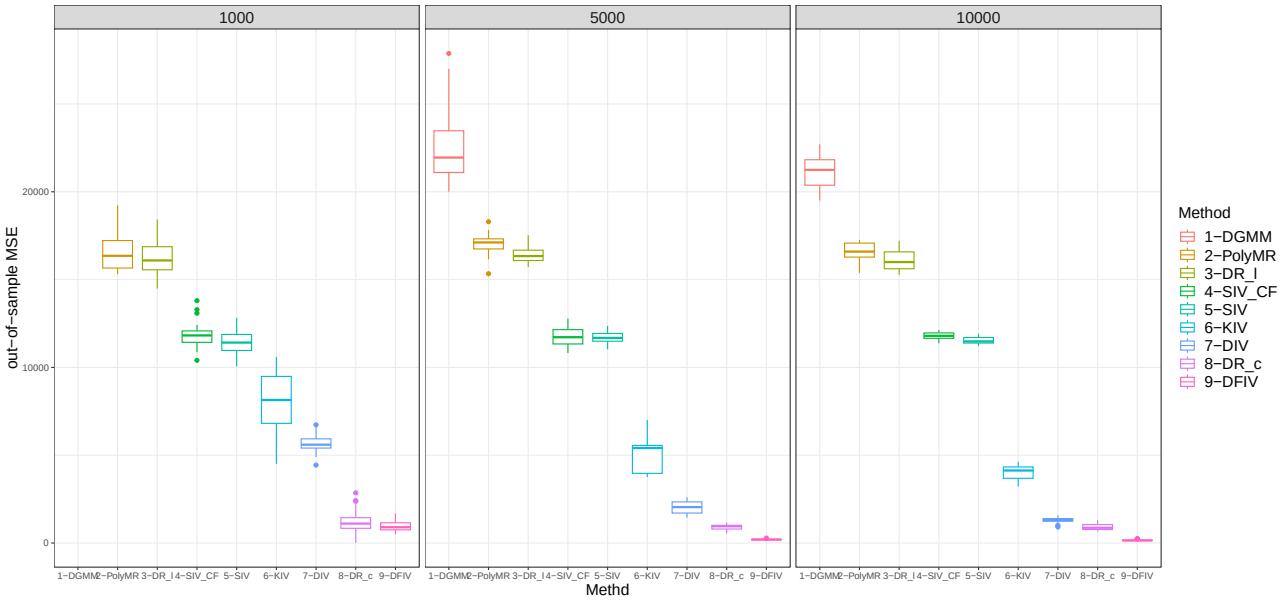
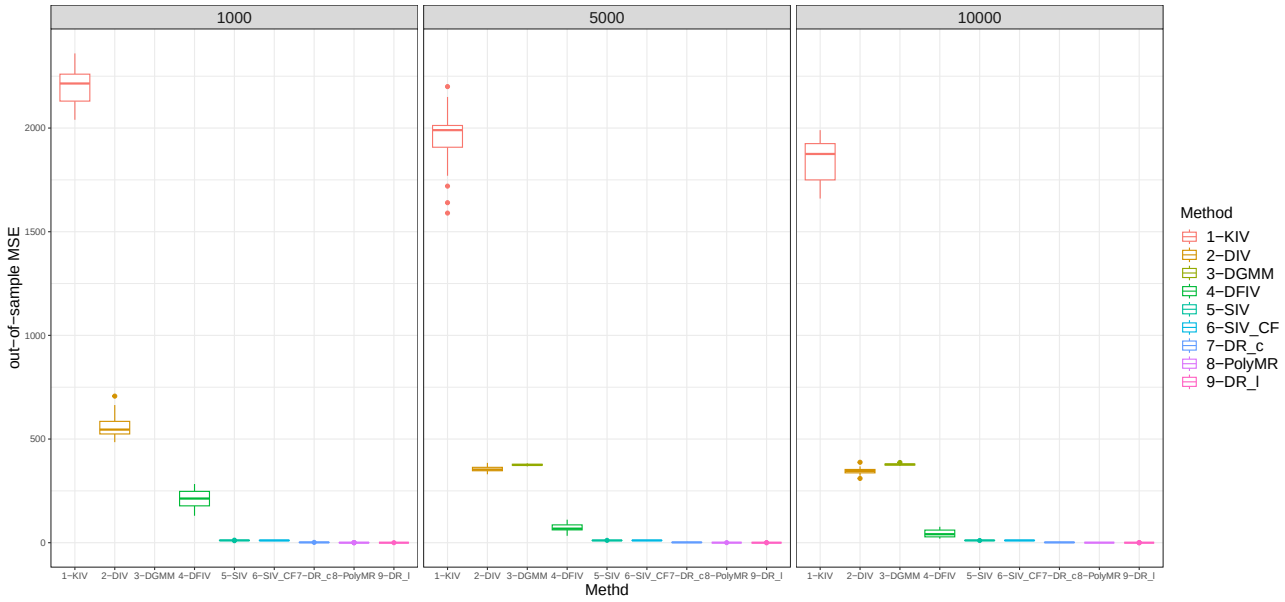
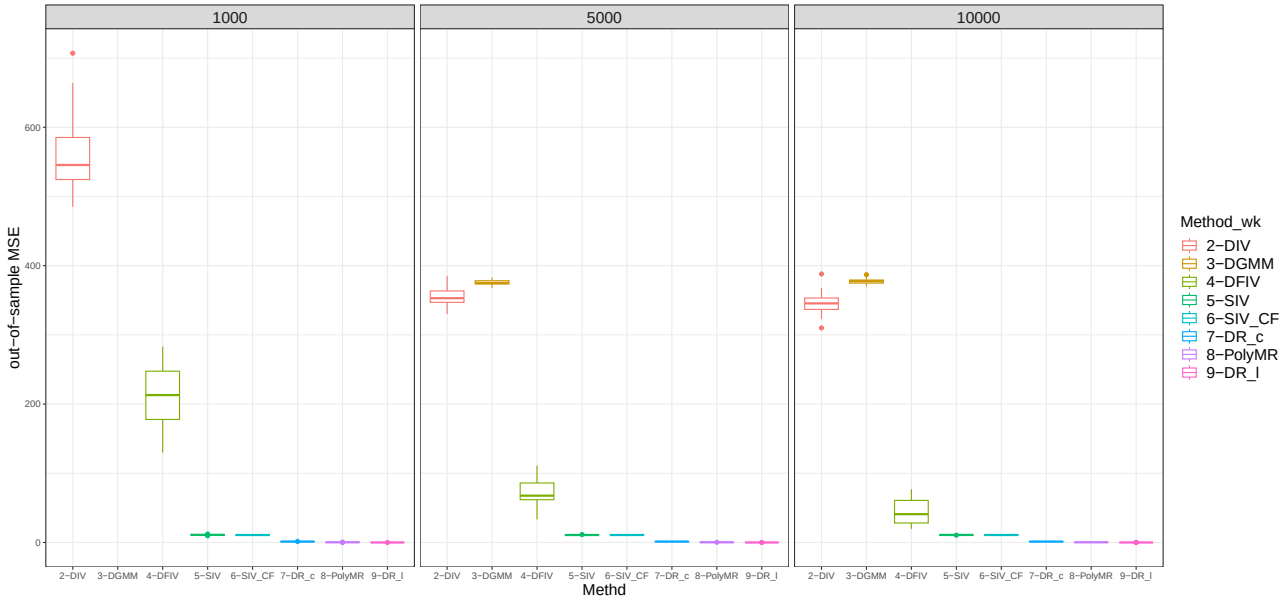


Figure 2.4: MSEs for the demand design. The MSEs are presented using standard boxplots. I.e. among results from 20 simulated datasets, the box represents the interquartile range (IQR), with the bottom of the box showing the 25th percentile, the top of the box showing the 75th percentile and the line inside the box showing the median. The length of the whiskers is 1.5 times the interquartile range. Data points beyond the whiskers are plotted as individual dots.



(a)



(b)

Figure 2.5: MSEs for the linear design, with and without KIV. The MSEs are presented using standard boxplots (same as in Figure 2.4).

2.9 Discussion and Concluding Remarks

In this chapter we provide a relatively comprehensive review of existing methods for estimating nonlinear causal effects using instrumental variables. We present all the methods under a general nonlinear causal model, which we introduce in section 2.2, so that the assumptions, estimation strategies, consistency results (if any) and limitations for each method can be compared relatively straightforwardly. We also show how we interpret the model using the potential outcomes framework, and compare the discussed methods empirically through a simulation study.

Except stratified nonlinear MR, all other approaches are based on generalising the classic 2SLS, control functions or GMM framework. Sieve IV and polynomial MR try to approximate the true structural function using pre-specified basis functions, which are not flexible enough and perform poorly in the demand design simulation where the structural function is a complex nonlinear function. Kernel IV, though also uses pre-specified features (i.e. the kernel we choose), is more flexible as it uses infinite number of feature maps, and it indeed performs much better in the demand design.

The deep learning type algorithms (deep feature IV, deep IV and DeepGMM) are much more flexible and are able to handle very high-dimensional exposure or instrument vectors such as images, as they aim to learn the feature maps adaptively from the data. Nevertheless, deep IV involves conditional density estimation in the first stage, which is a more difficult problem than estimating conditional expectations. DeepGMM uses the OAdam algorithm, which is originally developed for generative adversarial networks (GAN) and can exhibit oscillating behaviour in practice (Wiatrak and Nystrom, 2019). Although deep feature IV is free from these problems and indeed performs the best in the simulation study among the three, it suffers from general limitations of applying neural networks to IV problems. Firstly, it is not clear when we can expect the method to be consistent. For deep IV, there is no consistency result, while for deep feature IV and DeepGMM, although the authors are able to establish

consistency by assuming vanishing Rademacher complexity, this is a very strong assumption and is in fact violated for many neural networks (Zhang et al., 2017). Secondly, it is also unclear how we can perform inference using the deep learning algorithms, which is critical for causal interpretation.

A third problem with the machine learning algorithms, as mentioned in section 2.6, is that they typically do not generalise well for exposure values beyond the range of instrument-predicted exposure values. This is a general problem about any two stage methods, and is particularly relevant in MR applications where the genetic variants explain only a small amount of exposure variance. We should highlight that this is not simply a finite sample problem. When instruments are weak, two stage methods will not be able to extrapolate to wide range of exposure values even with infinite sample size. Therefore, the ability of machine learning type methods to uncover nonlinearities is limited with weak instruments.

Stratified nonlinear MR adopts a very different philosophy than generalising 2SLS or GMM and relies on validly stratifying the exposure distribution. Between the two stratification approaches, doubly-ranked stratification is preferred to residual stratification as it relaxes the constant genetic effect assumption and requires instead the weaker rank-preserving assumption. Moreover, Tian et al. (2023) demonstrate through their simulation studies that doubly-ranked nonlinear MR also outperforms PolyMR for rejecting/not rejecting the causal null hypothesis under various scenarios. Indeed, doubly-ranked nonlinear MR is currently one of the widely applied methods for understanding nonlinear causal effects in the context of MR.

However, stratified nonlinear MR also has a number of disadvantages. Firstly, if the structural function is not a fractional polynomial, estimation for the structural function will be problematic, as shown in the demand design simulation. Secondly, it cannot handle high-dimensional exposure. If there are two or three exposures, we might still stratify on them and trace back the shape of h using, for example, multivariate fractional polynomials, but

stratification is not feasible when the dimension is too large.

Thirdly and perhaps most importantly, a number of recent investigations using the negative control outcome framework suggest that there might be other problems with stratified nonlinear MR (e.g. Wade et al., 2023; Hamilton et al., 2024). These investigations chose an outcome (called a negative control outcome) that cannot possibly be affected by the exposure of interest, and discovered nonlinear causal effects using stratified nonlinear MR that are difficult to interpret. For example, Wade et al. (2023) found that both residual and doubly-ranked nonlinear MR indicate that lower BMI is causal for higher probability of being a female in UK Biobank. Although there is currently no conclusive explanation for such nonsensical results, one potential reason is differential selection bias across the strata of the exposure distribution. This motivates us to investigate the effect of selection bias on stratified nonlinear MR estimates in Chapter 5, and we will return to this discussion in later chapters.

In summary, existing methods work in different situations, but they are by no means exhaustive and future endeavours are very much needed to bridge the gaps in the area of nonlinear IV and MR methods.

Chapter 3

Nonparametric Mendelian Randomisation for Characterising Nonlinear Exposure-Outcome Relationships with Discrete Instrumental Variables

3.1 Introduction

We have discussed in Chapter 2 that stratified nonlinear MR is useful for estimating nonlinear causal effects when the true structural function can be well-approximated by fractional polynomials. We have also discussed its disadvantages, including not able to characterise complex nonlinear relationships beyond fractional polynomials. One example of a structural function that cannot be characterised by fractional polynomials is a threshold function that

involves a sudden non-differentiable change point. In addition, the number of strata is often determined arbitrarily (usually 10 strata) in practice.

In this chapter, we propose a nonparametric extension to stratified nonlinear MR with discrete instrumental variables, which we call nonparametric MR, that overcomes these disadvantages. We argue that our nonparametric MR estimator converges to a different version of localised average causal effect (LACE), which we call quantile average causal effect (QACE), that approximates the derivative of the structural function. Similar to doubly-ranked stratification, nonparametric MR does not rely on the constant genetic effect assumption and instead requires only the weaker rank-preserving assumption. More importantly, the nonparametric nature of our approach means that it is able to flexibly characterise a wide range of nonlinear structural functions, including non-smooth (i.e. non-differentiable) functions, and there is no need to pre-specify the number or the size of the strata.

We demonstrate through simulation studies that nonparametric MR outperforms stratified nonlinear MR in various scenarios, notably those in which the structural function is not a fractional polynomial. We also show, when the rank-preserving assumption is satisfied, that nonparametric MR is robust to instruments of very weak strength, a situation which most of the existing nonparametric IV methods struggle to deal with. Lastly, we apply our method to real data examples from UK Biobank to investigate the causal effect of BMI on handgrip power and on pulse rate.

Recall that the nonlinear causal model we are interested in is introduced in Chapter 2 (section 2.2). We also follow from the construct of stratified nonlinear MR in section 2.6 to assume that the exposure X and the instrument Z are both one-dimensional variables. This is common with MR applications.

3.2 Related Work

Before we introduce the nonparametric MR methodology, it is worth reviewing another field of IV literatures that aim to estimate the so-called quantile treatment effect (QTE), a concept that is in fact distinct from the quantile average causal effect (QACE) we propose. We will see in section 3.3.2 that, QACE refers to the average causal effect on the outcome (i.e. mean of counterfactual outcome distributions) at the quantiles of counterfactual exposure distributions. Analogous to localised average causal effects (LACEs), QACEs at all quantiles allow us to characterise the structural function h.

Distinct from QACE, QTE refers to the treatment (i.e. exposure) effect on the quantiles of counterfactual outcome distributions. For example, in a study that examines the effect of a job-training program on future earnings, researchers might be more interested in the causal effect of training on lower tail of the earnings distribution than the effect on the mean of earnings (Abadie et al., 2002). In practice, when estimating QTEs, people often look at effects of a binary treatment.

Chernozhukov and Hansen (2005) provide one of the initial frameworks in estimating QTE using IVs, known as the instrumental quantile regression (IVQR) model. Notationally, they use D for treatment and X for measured covariates. The quantity of interest is the τ-th quantile of potential outcome $Y(d)$ conditional on measured covariates $X = x$, denoted by

$$q(d, x, \tau)$$

This is also known as the quantile treatment response (QTR) function. The QTE can then be defined as the difference in the QTR functions

$$q(1, x, \tau) - q(0, x, \tau)$$

considering a binary treatment D.

Chernozhukov and Hansen (2005; 2006; 2008; 2013) model the potential outcome $Y(d)$ con-

ditional on $X = x$ by

$$Y(d) = q(d, x, U_d)$$

where $U_d \sim U(0,1)$ is the structural error term and represents the rank in the distribution of $Y(d)$. Under the core IV assumptions, SUTVA of potential outcomes and an additional rank similarity assumption that $\{U_d\}$ are identically distributed, the observed outcomes can be related as

$$Y = q(D, X, U), \quad U \sim U(0,1)|X, Z$$

and identification is guaranteed by the following quantile conditional moment restriction:

$$\text{for} \quad \tau \in (0,1), \quad \mathbb{P}\left[Y \leq q(D, X, \tau)|X, Z\right] = \tau \quad a.s. \tag{3.1}$$

for any quantile function $q(d, x, \tau)$ strictly increasing in τ.

Estimation of QTE can be quite challenging, because the sample analogue to the quantile moment condition in (3.1) is generally nonsmooth even for parametric quantile functions (Chernozhukov and Hansen, 2013). Chernozhukov and Hong (2003) estimate the model using MCMC. For linear quantile models and low-dimensional treatments, Chernozhukov and Hansen (2006) provide an estimation procedure named inverse quantile regression using grid search.

Recently, Kook and Pfister (2024) provide a different IV framework with binary treatment than the IVQR model for estimating QTE, which they call distributional causal effect (DCE). They provide a more general identification strategy based on simple structural equation models, and provide a distributional IV estimator based on some criterions of distributional loss.

There have also been efforts that attempt to estimate QTE beyond binary treatment. For example, both Lee (2007) and Imbens and Newey (2009) propose a control function approach for estimating QTE with continuous treatment. Lee (2007) use linear quantile regression in the first stage to get estimated residuals as control functions. He then uses a series estimator

in the second stage, similar to what we have discussed about sieve IV in Chapter 2 (section 2.4.1). Imbens and Newey (2009) adopt a fully nonparametric approach, using conditional density estimation in the first stage and plug-in estimators in the second stage.

3.3 Methodology

3.3.1 Rank-preserving assumption

Similar to doubly-ranked stratification, our nonparametric MR method will also require the rank-preserving assumption on counterfactual exposure distributions, which states that the ranks of an individual's counterfactual exposure values are identical for all values of the instrument.

Figure 3.1 illustrates this assumption in the case of a binary instrument. For example, for an individual with $Z = 0$ whose exposure value is at the 20th percentile of the counterfactual exposure distribution with $Z = 0$ ($X_1(0)$ on the red curve), this individual's counterfactual exposure value had $Z = 1$ is also at the 20th percentile of the counterfactual exposure distribution with $Z = 1$ ($X_1(1)$ on the blue curve). This similarly holds for all other percentiles of the counterfactual exposure distributions. Graphically speaking, this means that the arrows in Figure 3.1 are all horizontal.

If the instrument is discrete or continuous, the rank-preserving assumption requires that the ranks for all of the individual's counterfactual exposure values at each value of the instrument are the same. We will see in the following that the rank-preserving assumption allows us to validly estimate quantile average causal effect (QACE) at each quantile of the counterfactual exposure distributions. As noted by Tian et al. (2023), the rank-preserving assumption is strictly weaker than the constant genetic effect assumption required by residual stratification.

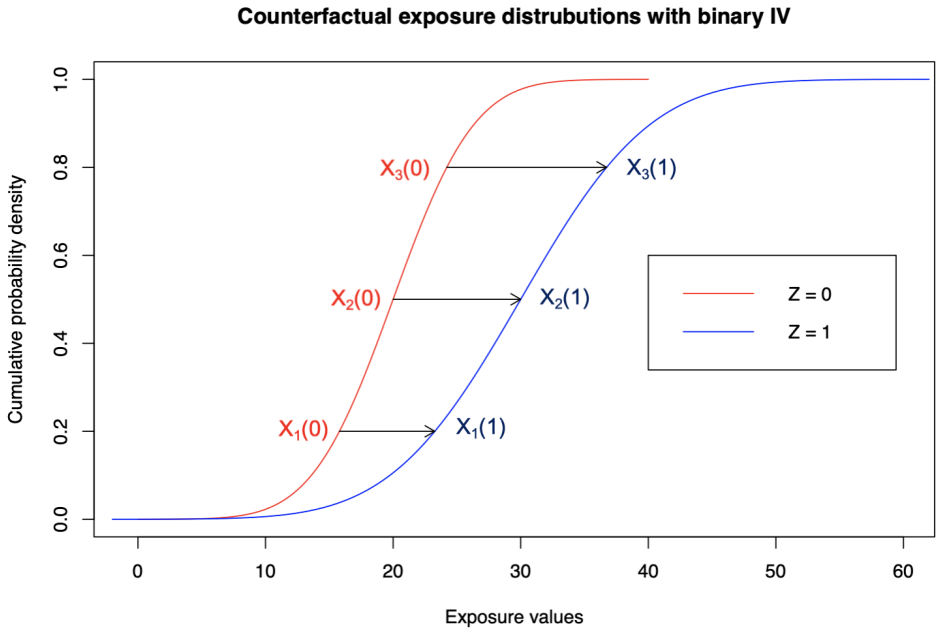


Figure 3.1: The rank-preserving assumption in the case of a binary instrument.

3.3.2 Binary IV

We now introduce our nonparametric MR method. We will use T to denote the quantile of the counterfactual exposure distribution, i.e. distribution of $X(z)$, and t to denote realisations/observations of T. Note that T is a random variable that is a deterministic function of (Z, X) and conversely X is a deterministic function of (T, Z).

We will start from the simpler case where the instrument is binary, i.e. $Z \in \{0, 1\}$.

Nonparametric MR does the following:

STEP 1. Conditioning on $Z_i = z$, rank the exposure values x_i and obtain $t_i \in [0, 1]$ as the empirical quantiles of the counterfactual exposure distribution, i.e. the distribution of $X_i(z)$;

STEP 2. Conditioning on $Z_i = z$, nonparametrically estimate the relationships between Y and T and X and T respectively, denoted as $\hat{Y}(t\,|Z = z)$ and $\hat{X}(t\,|Z = z)$;

Note that $\hat{Y}(t\,|Z = z)$ is the estimated expectation of Y given $T = t$ and $Z = z$, i.e. $\hat{Y}(t\,|Z = z) = \widehat{\mathbb{E}}\big[Y\ |T = t, Z = z\big]$. On the other hand, $\hat{X}(t\,|Z = z)$ is the estimated deterministic function of X given $T = t$ and $Z = z$. We write $\hat{Y}(t\,|Z = z)$ and $\hat{X}(t\,|Z = z)$ to emphasise that we run the nonparametric regressions conditioning on $Z = z$.

STEP 3. Estimate an average causal effect based on $T = t$, which we call quantile average causal effect (QACE):

$$\widehat{QACE}(t) = \frac{\hat{Y}(t\,|Z = 1) - \hat{Y}(t\,|Z = 0)}{\hat{X}(t\,|Z = 1) - \hat{X}(t\,|Z = 0)} \tag{3.2}$$

Equation (3.2) gives the nonparametric MR estimator with a binary instrument. Note that (3.2) looks similar to the usual Wald estimator (with a binary instrument), whose estimand is given by

$$\beta_{Wald} = \frac{\mathbb{E}(Y\,|Z = 1) - \mathbb{E}(Y\,|Z = 0)}{\mathbb{E}(X\,|Z = 1) - \mathbb{E}(X\,|Z = 0)}$$

This is indeed the motivation. Similar to stratified nonlinear MR, the idea is to replace the population mean of Y and X conditional on Z by something "local". For nonparametric MR, the "localness" is given by the quantiles of the counterfactual exposure distributions. I.e. we compare individuals at the same quantile of counterfactual exposure distributions under different values of the instrument. The rank-preserving assumption ensures that we are comparing valid counterfactuals. Figure 3.2 gives a graphical illustration.

The quantile T can essentially be regarded as a stratification variable, similar to the IV-free exposure in residual stratification but the constant genetic effect assumption is relaxed to the weaker rank-preserving assumption. This is due to the key observation that T is the quantile of the counterfactual exposure distribution, i.e. the distribution of $X(z)$, where z is fixed (recall that this is the definition of counterfactuals/potential outcomes). Therefore, conditioning on T will not introduce dependency between the instrument Z and the confounder U, and hence there is no collider bias involved.

Nevertheless, we should note that we do not directly observe the true values of T. Instead, we

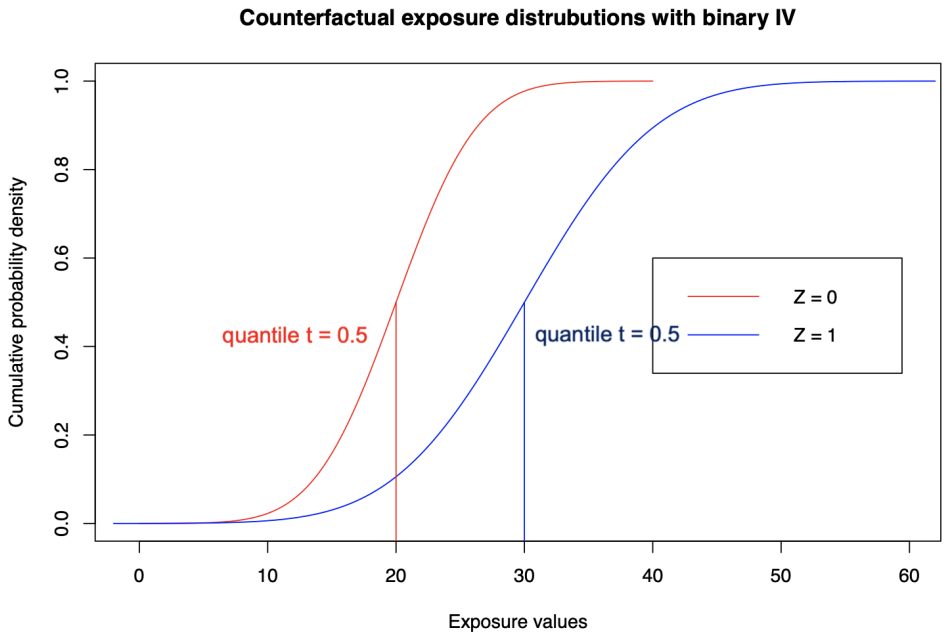


Figure 3.2: Quantiles of counterfactual exposure distributions with a binary instrument.

have a procedure to calculate/estimate the values of T as described in **STEP 1**. Therefore, in practice, there could be some empirical correlation between t_i and z_i and consequently between z_i and u_i. However, since in **STEP 1** we get t_i by directly fixing $Z_i = z$ as Z is discrete, this empirical correlation is very close to zero and hence negligible.

The following theorem establishes the asymptotics of the nonparametric MR estimator.

Theorem 6 (Asymptotics of the nonparametric MR estimator with binary instruments)**.** *Assume the nonparametric regressions in **STEP 2** are asymptotically consistent. Under the usual SUTVA of potential outcomes and valid IV assumptions, we have*

$$\widehat{QACE}(t) \xrightarrow{p} \frac{h\left(x^{(t)}(1)\right) - h\left(x^{(t)}(0)\right)}{x^{(t)}(1) - x^{(t)}(0)} \tag{3.3}$$

where $x^{(t)}(z)$ denotes the counterfactual exposure value at quantile t with $Z = z$, $z \in \{0, 1\}$. h denotes the true structural function of the exposure-outcome relationship.

Proof. The asymptotic consistency of the nonparametric regression implies that

$$\hat{Y}(t \,|Z = z) \xrightarrow{p} \mathbb{E}\big[Y \mid T = t, Z = z\big] \tag{3.4}$$

Plugging in the structural equation $Y = h(X) + \epsilon_Y$, we see the right hand side of (3.4) becomes

$$\begin{aligned}
\mathbb{E}\big[Y \mid T = t, Z = z\big] &= \mathbb{E}\big[h(X) + \epsilon_Y \mid T = t, Z = z\big] \\
&= \mathbb{E}\big[h(X) \mid T = t, Z = z\big] \\
&= \mathbb{E}\big[h(X(z)) \mid T = t\big] \\
&= h\left(x^{(t)}(z)\right)
\end{aligned}$$

where the second line comes from the conditional mean independence assumption (Assumption 6) and the fact that conditioning on T will not introduce collider bias, the third line is due to the consistency assumption of counterfactual exposures and the fourth line is due to the fact that the counterfactual value $X(z)$ is fixed given $T = t$.

Similarly, we have

$$\hat{X}(t \,|Z = z) \xrightarrow{p} x^{(t)}(z)$$

due to the asymptotic consistency of the nonparametric regression and the consistency assumption of counterfactual exposures. Combining the above results with equation (3.2), we complete the proof. □

The right hand side of (3.3) in Theorem 6 gives a formal definition of the quantile average causal effect (QACE) with binary instrument, which refers to the average causal effect of X on Y at quantile t, induced by a change in the instrument value from $Z = 0$ to $Z = 1$. It is easy to see that QACE approximates the derivative of the structural function h with respect to x if the change $x^{(t)}(1) - x^{(t)}(0)$ is small. How well the approximation is depends on the size of the genetic effect on the exposure and the shape of the structural function. QACE can be of particular interest when h involves a non-differential change point, such as

a threshold function, and t is the quantile such that a change from $Z = 0$ to $Z = 1$ induces a jump around the change point. In this case, QACE will be a causal quantity of interest in practice, instead of the undefined derivative of h around the change point.

The performance of nonparametric MR also depends strongly on the nonparametric regression method we choose in **STEP 2**. In this chapter, we consider two nonparametric methods that are relatively easy to implement in practice: kernel ridge regression (e.g. Scholkopf et al., 2001) and local linear kernel estimation (Li and Racine, 2007). Technical details of kernel ridge regression are given in the Appendix A.1. As for local linear kernel estimation, we direct interested readers to Li and Racine (2007).

3.3.3 General discrete IV

For instruments that take general discrete values, we decompose the Wald estimand

$$\beta_{Wald} = \frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)}$$

into conditional expectations so that we can use the same idea as in the binary instrument case to replace expectations of Y and X conditional on Z by nonparametric regressions of Y and X on quantiles t of the counterfactual exposure distributions. We will need the following result.

Lemma 2. *For random variables Y and Z where $Z \in \{0, 1, ..., k\}$, we have*

$$Cov(Y, Z) = \sum_{i=0}^{k} \sum_{j=1, i \neq j}^{k} \big(\mathbb{E}[Y|Z = j] - \mathbb{E}[Y|Z = i]\big) j \mathbb{P}(Z = i) \mathbb{P}(Z = j)$$

Proof. We write the covariance in terms of conditional expectations and expand the expec-

tations over the support of Z:

$$\begin{aligned}
\mathrm{Cov}(Y,Z) &= \mathbb{E}[YZ] - \mathbb{E}[Y]\,\mathbb{E}[Z] \\
&= \mathbb{E}\Big[Z\,\mathbb{E}[Y|Z]\Big] - \mathbb{E}\Big[\mathbb{E}[Y|Z]\Big]\mathbb{E}[Z] \\
&= \sum_{i=1}^{k} i\mathbb{E}[Y|Z=i]\,\mathbb{P}(Z=i) - \left(\sum_{j=0}^{k}\mathbb{E}[Y|Z=j]\,\mathbb{P}(Z=j)\right)\left(\sum_{i=1}^{k} i\mathbb{P}(Z=i)\right) \\
&= \sum_{i=1}^{k}\mathbb{E}[Y|Z=i]\,i\mathbb{P}(Z=i) - \sum_{i=1}^{k}\left(\sum_{j=0}^{k}\mathbb{E}[Y|Z=j]\,\mathbb{P}(Z=j)\right) i\mathbb{P}(Z=i) \\
&= \sum_{i=1}^{k}\left(\mathbb{E}[Y|Z=i] - \sum_{j=0}^{k}\mathbb{E}[Y|Z=j]\,\mathbb{P}(Z=j)\right) i\mathbb{P}(Z=i) \\
&= \sum_{i=1}^{k}\left(\sum_{j=0}^{k}\mathbb{E}[Y|Z=i]\,\mathbb{P}(Z=j) - \sum_{j=0}^{k}\mathbb{E}[Y|Z=j]\,\mathbb{P}(Z=j)\right) i\mathbb{P}(Z=i) \\
&= \sum_{i=0}^{k}\sum_{j=1,i\neq j}^{k}\big(\mathbb{E}[Y|Z=j] - \mathbb{E}[Y|Z=i]\big) j\mathbb{P}(Z=i)\mathbb{P}(Z=j)
\end{aligned}$$

This completes the proof. □

We now generalise the nonparametric MR method to the case where the instrument takes discrete values. For $Z \in \{0,1,...,k\}$, nonparametric MR does the following:

STEP 1. Conditioning on $Z_i = z$, rank the exposure values x_i and obtain $t_i \in [0,1]$ as the empirical quantiles of the counterfactual exposure distribution, i.e. the distribution of $X_i(z)$;

STEP 2. Conditioning on $Z_i = z$, nonparametrically estimate the relationships between Y and T and X and T respectively, denoted as $\hat{Y}(t\,|Z=z)$ and $\hat{X}(t\,|Z=z)$;

STEP 3. Estimate the probabilities $\mathbb{P}(Z=z), z \in \{0,1,...,k\}$ by standard maximum likelihood estimation;

STEP 4. Estimate an average causal effect based on $T=t$, which we call quantile average

causal effect (QACE):

$$\widehat{QACE}(t) = \frac{\sum_{i=0}^{k}\sum_{j=1,i\neq j}^{k}\big(\hat{Y}(t|Z=j)-\hat{Y}(t|Z=i)\big)j\hat{\mathbb{P}}(Z=i)\hat{\mathbb{P}}(Z=j)}{\sum_{i=0}^{k}\sum_{j=1,i\neq j}^{k}\big(\hat{X}(t|Z=j)-\hat{X}(t|Z=i)\big)j\hat{\mathbb{P}}(Z=i)\hat{\mathbb{P}}(Z=j)} \tag{3.5}$$

Equation (3.5) gives the nonparametric MR estimator with a discrete instrument. Similar to the case with binary instruments, we have the following theorem that establishes the asymptotics of the nonparametric MR estimator with discrete instruments.

Theorem 7 (Asymptotics of the nonparametric MR estimator with discrete instruments)**.** *Assume the nonparametric regressions in* ***STEP 2*** *are asymptotically consistent. Under the usual SUTVA of potential outcomes and valid IV assumptions, we have*

$$\widehat{QACE}(t) \xrightarrow{p} \frac{Cov^{(t)}\Big(h(X),Z\Big)}{Cov^{(t)}\left(X,Z\right)} \tag{3.6}$$

where $Cov^{(t)}\left(X,Z\right)$ *denotes the covariance between* X *and* Z *conditional on* $T=t$*, i.e.* $Cov^{(t)}\left(X,Z\right)=Cov\left(X,Z\,|\,T=t\right)$*, and similarly* $Cov^{(t)}\left(h(X),Z\right)$ *denotes the covariance between* $h(X)$ *and* Z *conditional on* $T=t$*, i.e.* $Cov^{(t)}\left(h(X),Z\right)=Cov\left(h(X),Z\,|\,T=t\right)$*.*

Proof. Similar to the proof of Theorem 6, the asymptotic consistency of the nonparametric regressions implies that

$$\hat{Y}(t\,|Z=j) \xrightarrow{p} \mathbb{E}\big[Y\,|\,T=t,Z=j\big], j\in\{0,1,...,k\} \tag{3.7}$$

and in **STEP 3**, we have

$$\hat{\mathbb{P}}(Z=j) \xrightarrow{p} \mathbb{P}(Z=j), j\in\{0,1,...,k\} \tag{3.8}$$

Combing Lemma 2, (3.7) and (3.8), we see that the numerator in (3.5) converges to $\mathrm{Cov}^{(t)}\left(Y,Z\right)$.

Plugging in the structural equation $Y=h(X)+\epsilon_Y$, using valid IV assumptions and the fact that conditioning on T will not introduce collider bias, we have

$$\mathrm{Cov}^{(t)}\left(Y,Z\right)=\mathrm{Cov}^{(t)}\Big(h(X),Z\Big)$$

Similarly, using the asymptotic consistency of the nonparametric regressions, Lemma 2, and (3.8), we see that the denominator in (3.5) converges to $\text{Cov}^{(t)}(X, Z)$. This completes the proof. □

The right hand side of (3.6) in Theorem 7 gives a formal definition of the quantile average causal effect (QACE) with discrete IVs, which is the average causal effect of X on Y at quantile t, induced by the average change in the instrument value from $Z = 0$ to $Z = k$. This is a natural extension of QACE with binary IVs, and similarly approximates the derivative of h. However, note that the approximation will be less precise as k increases, because a larger k means that QACEs are estimated over a wider range of instrument-predicted exposure values.

For MR applications, the most relevant case is where $k = 2$. This can be a single genetic variant, which typically takes three values 0, 1 and 2. Therefore, in the rest of this chapter, we will focus on the binary instrument case ($k = 1$) and the case where the instrument takes three values ($k = 2$).

3.4 Simulation Study

3.4.1 Simulation design

We design a simulation study that intends to demonstrate the empirical performances of nonparametric MR by comparing it against stratified nonlinear MR under various exposure-outcome relationships and different instrument-exposure models. Similar to the simulation study in Chapter 2 (section 2.8), we perform stratified nonlinear MR using doubly-ranked stratification as it has been shown to be more robust than residual stratification. Doubly-ranked nonlinear MR is implemented using the SUMnlmr package (Mason and Burgess, 2022) with the default choice of 10 strata.

We simulate both cases of binary instrument and instruments taking three values.

We consider the following four structural functions h:

(1). Null function: $h(x) = 0$;

(2). Cubic function: $h(x) = 0.05x^3$;

(3). Exponential function: $h(x) = 0.8\exp(x/4)$;

(4). Threshold function: $h(x) = \begin{cases} 0.5(x-15), & \text{if } x > 15 \\ 0, & \text{if } x \leq 15 \end{cases}$

We generate Y according to the structural equation

$$Y_i = h(X_i) + U_i + e_{Y,i}$$

where $U_i \overset{i.i.d}{\sim} \mathbb{N}(0,1)$, $e_{Y,i} \overset{i.i.d}{\sim} \mathbb{N}(0, 0.1^2)$ and $e_{Y,i} \perp\!\!\!\perp (U_i, X_i)$.

X is generated according to the following three different structural instrument-exposure models that correspond to different assumptions on the instrumental effect:

(i). *Constant genetic effect*:

$$X_i = \alpha_0 + \alpha_Z Z_i + \alpha_U U_i + e_{X,i}$$

where $e_{X,i} \overset{i.i.d}{\sim} \mathbb{N}(0,1)$ and $e_{X,i} \perp\!\!\!\perp (U_i, Z_i, e_{Y,i})$ for both binary instrument and three-value instrument cases.

(ii). *Non-constant but rank-preserving genetic effect*:

$$X_i = \alpha_0 + \alpha_Z Z_i + \alpha_L L(Z_i) + \alpha_U U_i + e_{X,i}$$

where $L(Z_i) = \begin{cases} \mu_i, & \text{if } Z_i = 0 \\ \exp(\mu_i), & \text{if } Z_i = 1 \end{cases}$, $\mu_i \overset{i.i.d}{\sim} \text{Unif}(0,1)$ for binary instrument case,

$$L(Z_i) = \begin{cases} \mu_i, & \text{if } Z_i = 0 \\ \frac{1}{2}\exp(\mu_i), & \text{if } Z_i = 1, \ \mu_i \overset{i.i.d}{\sim} \text{Unif}(0,1) \\ \exp(\mu_i) + \mu_i, & \text{if } Z_i = 2 \end{cases} \text{ for three-value instrument case,}$$

and $\mu_i \perp\!\!\!\perp (Z_i, U_i, e_{X,i}, e_{Y,i})$.

I.e. we define $L(Z_i)$ as different functions of random variable μ_i for different values of Z_i, so the genetic effect for different individuals will be different. However, because for each value of Z_i the function is monotonically increasing, ranks of the counterfactual exposure values are preserved.

(iii). *Rank-varying genetic effect*:

$$X_i = \alpha_0 + \alpha_Z Z_i + \alpha_L L(Z_i) + \alpha_M M(Z_i) + \alpha_U U_i + e_{X,i}$$

$$\text{where } M(Z_i) = \begin{cases} 0, & \text{if } Z_i = 0 \\ \xi_i \overset{i.i.d}{\sim} \mathbb{N}(0,1), & \text{if } Z_i = 1 \end{cases} \text{ for binary instrument case,}$$

$$M(Z_i) = \begin{cases} 0, & \text{if } Z_i = 0 \\ \xi_{1,i} \overset{i.i.d}{\sim} \mathbb{N}(0, 0.5^2), & \text{if } Z_i = 1 \\ \xi_{2,i} \overset{i.i.d}{\sim} \mathbb{N}(0,1) & \text{if } Z_i = 2 \end{cases} \text{ for three-value instrument case,}$$

and $(\xi_i, \xi_{1,i}, \xi_{2,i}) \perp\!\!\!\perp (Z_i, U_i, \mu_i, e_{X,i}, e_{Y,i})$, $\xi_{1,i} \perp\!\!\!\perp \xi_{2,i}$.

I.e. we define $M(Z_i)$ as different random variables for different values of Z_i. Because the mapping is random for each individual, the rank-preserving assumption will be violated. We design this scenario in order to investigate how sensitive nonparametric MR is to the violation of the rank-preserving assumption.

For all three instrument-exposure models above, we generate $Z_i \overset{i.i.d}{\sim}$ Bernoulli(0.5) for a binary instrument, $Z_i \overset{i.i.d}{\sim}$ Binomial$(2, 0.5)$ for a three-value instrument.

In addition to different effects of the genetic instrument on the exposure, we would also like to examine how sensitive nonparametric MR is to the strength of the genetic instrument.

Therefore, for each instrument-exposure model, we consider three different levels of instrument strength $\in \{4\%, 1\%, 0.3\%\}$, characterised by the proportion of variance in the exposure explained by the instruments and measured by R^2 in the instrument-exposure regression. The particularly weak 0.3% instrument strength case is motivated by the real data examples from UK Biobank that we will present in section 3.5.

Tables 3.1 and 3.2 summarise the parameterisations of the data-generating models for all simulated scenarios.

In all simulated scenarios, we generate $D = 100$ simulated datasets and set the sample size to be $N = 10,000$ for each simulation. The main metrics we use for comparison include bias and standard deviation of the estimates. Since our estimand is the (derivative of) the structural function h (instead of a fixed quantity β), we report the magnitude of bias between the estimated derivative and the true derivative of the structural function h, averaged across the exposure distribution. Note that we cannot directly average bias across the exposure distribution. For example, consider an estimator of h' that consistently underestimates $h'(x)$ at lower tails of the exposure distribution and consistently overestimates $h'(x)$ at upper tails of the exposure distribution. Then the average bias across the exposure distribution could be zero, which is not a fair representation of the bias property of this estimator.

Therefore, we define and present the bias between $\hat{h'}$ and h' as the magnitude of bias averaged across the exposure distribution, given by

$$\text{Bias}(\hat{h'}, h') = \frac{1}{m} \sum_{j=1}^{m} \left| \text{bias}\Big(\hat{h'}(x_j), h'(x_j) \Big) \right| \tag{3.9}$$

where we set $m = 100$ for evaluation of the bias, x_j's are out-of-sample values of X across the distribution of X and

$$\text{bias}\Big(\hat{h'}(x_j), h'(x_j) \Big) = \frac{1}{D} \sum_{d=1}^{D} \big(\hat{h'_d}(x_j) - h'(x_j) \big)$$

is the bias of $\hat{h'}(x_j)$ across the 100 simulated datasets. $\hat{h'_d}$ denotes the estimator of h' for the d-th round of the simulation,

Similarly, we define and present the standard deviation of $\hat{h}'$ as standard deviation of the estimates averaged across the exposure distribution, given by

$$\mathrm{SD}(\hat{h}') = \frac{1}{m}\sum_{j=1}^{m} \mathrm{s.d.}\big(\hat{h}'(x_j)\big) \tag{3.10}$$

where

$$\mathrm{s.d.}\big(\hat{h}'(x_j)\big) = \frac{1}{D-1}\sum_{d=1}^{D}\left(\hat{h}'_d(x_j) - \frac{1}{D}\sum_{d=1}^{D}\hat{h}'_d(x_j)\right)^2$$

is the standard deviation of $\hat{h}'(x_j)$ across the 100 simulated datasets.

For nonparametric MR, we regard the estimated QACE as $\hat{h}'$ and for stratified nonlinear MR, we get $\hat{h}'$ by differentiating $\hat{h}$ obtained from the fractional polynomial regression. Because QACE is estimated based on quantiles t of the counterfactual exposure distributions, for nonparametric MR we get back to the exposure values x by averaging across $\hat{X}(t\,|Z = z), z \in \{0,1\}$ or $z \in \{0,1,2\}$. I.e. we let

$$x_j = \frac{1}{k+1}\sum_{z=0}^{k}\hat{X}(t_j\,|Z_j = z), \quad k = 1 \text{ or } k = 2$$

where t_j is a given quantile of the counterfactual exposure distribution. This is similar to the approach used in stratified nonlinear MR, where the LACE estimates are regressed against the mean of the exposure in each stratum (Staley and Burgess, 2017). However, because nonparametric regressions at tails of the counterfactual exposure distribution are prone to weak instrument bias (e.g. Burgess et al., 2011), we trim the tails of QACE estimates at 5% level in all simulated scenarios for consistent comparison of the bias and standard deviation.

In addition, as we have seen in section 3.3 that QACE is only an approximation of the derivative of h, for nonparametric MR we also present average magnitude of bias between estimated QACE and the theoretical QACE. I.e. We replace h' by the theoretical QACE in equation (3.9). The theoretical QACE is given by equation (3.3) in Theorem 6 for binary instruments and equation (3.6) in Theorem 7 for three-value instrument.

Lastly, we should mention that we choose $m = 100$ (instead of a bigger number such as

$m = 1000$) due to computational constraint of the nonparametric regression methods. This is because for each round of the simulation, we need to calculate $\hat{h}'(x_j)$ for $j = 1, ..., m$ from the fitted nonparametric regressions, which is in fact computationally demanding. We generate $D = 100$ simulated datasets (instead of a bigger number such as $D = 1000$) for the same reason. We will discuss again the computational limitations of our approach at the end of this chapter.

Data generating model for simulated scenarios with binary instruments

Instrument-exposure model	IV strength	α_0	α_Z	α_L	α_M	α_U
constant genetic effect	4%	15	1	0	0	2
constant genetic effect	1%	15	0.5	0	0	2
constant genetic effect	0.3%	15	0.25	0	0	2
non-constant but rank-preserving genetic effect	4%	14	0	0.85	0	2
non-constant but rank-preserving genetic effect	1%	14.5	0	0.45	0	2
non-constant but rank-preserving genetic effect	0.3%	15	0	0.25	0	2
rank-varying genetic effect	4%	14	0	1	1	2.5
rank-varying genetic effect	1%	14.5	0	0.5	1	2.5
rank-varying genetic effect	0.3%	14.5	0	0.38	1	2.5

Table 3.1: Model parameters for the data generating mechanisms for instrument-exposure relationships in scenarios with binary instruments. Instrument strength is the proportion of variance in the exposure explained by the instrument, measured by R^2 in the instrument-exposure regression. α_0 are chosen such that the mean of the exposure distribution is kept about the same in each scenario.

Data generating model for simulated scenarios with three-value instruments

Instrument-exposure model	IV strength	α_0	α_Z	α_L	α_M	α_U
constant genetic effect	4%	14	0.85	0	0	2.5
constant genetic effect	1%	14.5	0.45	0	0	2.5
constant genetic effect	0.3%	14.5	0.25	0	0	2.5
non-constant but rank-preserving genetic effect	4%	13.5	0	1	0	2.5
non-constant but rank-preserving genetic effect	1%	14.5	0	0.5	0	2.5
non-constant but rank-preserving genetic effect	0.3%	14.5	0	0.25	0	2.5
rank-varying genetic effect	4%	14	0	1	1	2.5
rank-varying genetic effect	1%	14.5	0	0.5	1	2.5
rank-varying genetic effect	0.3%	14.5	0	0.32	1	2.5

Table 3.2: Model parameters for the data generating mechanisms for instrument-exposure relationships in scenarios with three-value instruments. Instrument strength is the proportion of variance in the exposure explained by the instrument, measured by R^2 in the instrument-exposure regression. α_0 are chosen such that the mean of the exposure distribution is kept about the same in each scenario.

3.4.2 Hyperparameter tuning for nonparametric regressions

As mentioned in section 3.3, we use kernel ridge regression (e.g. Scholkopf et al., 2001) and local linear kernel estimation (Li and Racine, 2007) as the nonparametric regression methods in **STEP 2** of nonparametric MR and compare their performances in our simulation studies. An important step in these data-driven methods is selecting hyperparameters: for kernel ridge regression these are the penalisation parameter λ and Gaussian kernel width parameter σ^2 (we choose the Gaussian kernel for all kernel ridge regressions in this chapter); for local linear kernel estimation this is the bandwidth parameter h.

As discussed in Chapter 2 (section 2.4.2), there is a general understanding that the standard hyperparameter tuning methods, such as cross-validation, are not suitable for causal inference problems. However, since we are not directly estimating any causal quantity in **STEP 2** of nonparametric MR, the nonparametric regressions can essentially be regarded as prediction

tasks. Therefore, cross-validation is valid for our purpose. Nevertheless, in order to avoid overfitting, we perform cross-validation on held-out datasets in our simulations.

It is further noted that nonparametric MR using kernel ridge regressions still tends to overfit if we use the standard leave-one-out cross-validation for each of the regressions between Y and T and between X and T (recall that we need to perform four different kernel ridge regressions if the instrument is binary and six different kernel ridge regressions if the instrument takes three values) on held-out datasets. To overcome this issue, we impose an additional constraint that we choose the same hyperparameter for all kernel ridge regressions. A justification for this constraint is that we want the kernel ridge regressions to be similarly penalised. Otherwise, for example, if $\hat{Y}(t\,|Z=1)$ is undersmoothed but $\hat{Y}(t\,|Z=0)$ is oversmoothed, our QACE estimates will be quite unstable. However, a limitation of this constraint is that it is not scale invariant. If, for example, Y is inflated in scale (e.g. multiply by 100) and X is left unchanged, the hyperparameter of the Y - T regression should increase and the hyperparameter of the X - T regression should not change. However, since we impose the constraint of choosing the same hyperparameter, either the Y - T regression or the X - T regression will not be correctly penalised. We leave it for future research to consider more sophisticated hyperparameter tuning procedures.

In order to speed up computation, we select the hyperparameters using one held-out dataset and keep these selected hyperparameters fixed for 100 simulated datasets to calculate bias and standard deviation. We also construct pointwise 95% confidence intervals for estimated QACEs using nonparametric bootstrap with hyperparameters fixed. We choose the number of bootstrapped sample to be $R = 1,000$. However, a cost of this is that the bootstrapped confidence intervals will be overly-optimistic, as the uncertainties of hyperparameter selection will not be taken into account.

3.4.3 Simulation results

Figures 3.3 - 3.6 provide visualisation of nonparametric MR estimates with both kernel ridge regression and local linear kernel estimation and doubly-ranked nonlinear MR estimates using fractional polynomials. Figures 3.3 and 3.4 visualise scenarios where the instrument is binary and Figures 3.5 and 3.6 visualise scenarios where the instrument takes three values. As mentioned in section 3.4.2, for nonparametric MR, we use nonparametric bootstrap to construct pointwise 95% confidence intervals (the blue curves). For doubly-ranked nonlinear MR, we construct pointwise 95% confidence intervals using asymptotic standard errors from the fractional polynomial regression (this is the default option in the SUMnlmr package (Mason and Burgess, 2022)). We also plot the true derivatives of the structural functions (the red curves) as comparisons. For illustration purpose, we plot the simple case where the genetic effect is constant and the instrument strength is 4%. Each figure is generated using one single dataset.

We observe that for the null and the cubic relationships (Figure 3.3 and 3.5), which both belong to the class of fractional polynomials, all three methods perform reasonably well. For the exponential and the threshold relationships (Figure 3.4 and 3.6) that are not fractional polynomial functions, on the other hand, both nonparametric MR methods continue to perform well, but doubly-ranked nonlinear MR fails to estimate the correct shape. For the exponential function, the 95% confidence interval of doubly-ranked nonlinear MR fails to include the truth for the majority of the curve; for the threshold function, doubly-ranked nonlinear MR provides an estimated function that is completely non-threshold. This confirms our argument that stratified nonlinear MR cannot characterise complex exposure-outcome relationships beyond fractional polynomials, and demonstrates the flexibility of the nonparametric MR method.

Between kernel ridge regression and local linear kernel estimation for nonparametric MR, we observe that local linear kernel estimation gives rather non-smooth estimates. This is

because there is no smoothing involved with local linear kernel estimation, so the estimates are quite variable. Kernel ridge regression, on the other hand, imposes an implicit smoothness constraint by assuming the function belongs to a RKHS (reproducing kernel Hilbert space; i.e. the function can be represented as linear combinations of potentially infinite numbers of feature maps defined in terms of the kernel, as we have also discussed in Chapter 2 (section 2.4.2) and that $\|f\|^2_{\mathcal{H}} \leq \infty$ (see Appendix A.1.3 for more details). Consequently, nonparametric MR with kernel ridge regression gives smooth and accurate estimates for all four structural functions.

Due to the discussion above, we can also compare the three methods in terms of the amount of smoothing they impose. Stratified nonlinear MR imposes high degree (polynomial level) of smoothing, nonparametric MR with kernel ridge regression imposes small degree of smoothing while nonparametric MR with local linear kernel estimation admits no smoothing.

Note that for the threshold function, nonparametric MR gives QACE estimates that are different from the derivative of the function around the non-differential change point. This is expected, as QACE is the average causal effect induced by the change in the instrument values. When such a change includes a jump at the change point, the QACE represents a relevant causal quantity that can be of interest in practice.

Tables 3.3 - 3.8 present average magnitude of bias and standard deviation of the estimates for all simulated scenarios. We firstly note that doubly-ranked nonlinear MR is more advantageous than nonparametric MR in estimating the null function both in terms of bias and standard deviation. This is not surprising, as the null function is the simplest relationship for which nonparametric methods are probably not needed in the first place. Nevertheless, nonparametric MR still performs comparatively well for estimating the null, especially when the instrument strength is either 4% or 1%.

We then note that nonparametric MR outperforms doubly-ranked nonlinear MR for estimating the cubic, the exponential and the threshold functions in almost all of the simulated

scenarios in terms of bias, with only very few exceptions (e.g. the bias of nonparametric MR with local linear kernel estimation exceeds the bias of doubly-ranked nonlinear MR for estimating the cubic function when the instrument strength is only 0.3% with constant genetic effect and a binary instrument). The exceptions only occur for binary instrument cases, either when the instrument is as weak as 0.3% or when the rank-preserving assumption is violated. This demonstrates that nonparametric MR is more accurate for estimating complex functions, especially when the instrument is not binary, when the instrument strength is at least 1%, or when the rank-preserving assumption holds.

In terms of standard deviations of the estimates for the cubic, the exponential and the threshold functions, nonparametric MR with kernel ridge regression also outperforms doubly-ranked nonlinear MR for the majority of the simulated scenarios. Similar to the discussion about bias, exceptions (i.e. nonparametric MR with kernel ridge regression results in larger standard deviation of the estimates) occur when the instrument is weak and/or when the rank-preserving assumption does not hold. Nonparametric MR with local linear kernel estimation, on the other hand, often produces estimates that are more variable than nonparametric MR with kernel ridge regression and doubly-ranked nonlinear MR. Notably, the standard deviation of nonparametric MR estimates with local linear kernel estimation quickly gets large as the instrument strength decreases, especially for the cubic and the threshold functions.

Between the two nonparametric regression methods, we note that local linear kernel estimation can often give smaller bias than kernel ridge regression, though the improvement is typically not substantial. However, it is noticeable that local linear kernel estimation results in larger standard deviation of the estimates than kernel ridge regression in all simulated scenarios. As mentioned above, the large variability issue with local linear kernel estimation is especially obvious when the instrument is weak. This is likely due to the different degree of smoothing imposed by these two methods as discussed previously. Therefore, from a bias-variance tradeoff point of view, kernel ridge regression is preferred.

Tables 3.9 - 3.14 present average magnitude of bias with respect to the theoretical QACEs for both nonparametric MR methods. As a comparison, biases with respect to the derivatives of the structural functions are also duplicated. By definition, for the null function, the derivative is the same as the theoretical QACE (both are null), so the null function is omitted in these tables.

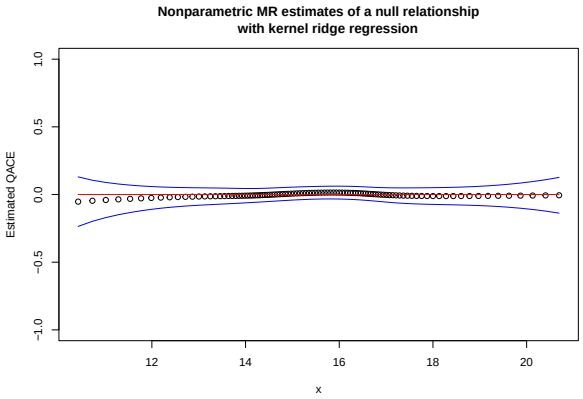
For binary instruments (Tables 3.9 - 3.11), we observe that the biases with respect to the derivative of h are very similar to the biases with respect to the theoretical QACEs for the cubic and the exponential functions. This suggests that the change in the exposure values induced by the change in the binary instrument is small, and therefore the approximation of theoretical QACE to the derivative of h is relatively accurate (in many scenarios, the difference in bias is not observable in three decimal places). The difference in bias is noticeable for the threshold function, as expected, due to the existence of the non-differentiable change point.

On the other hand, for three-value instruments (Tables 3.12 - 3.14), the differences between the biases with respect to the derivative of h and the biases with respect to the theoretical QACEs are more obvious, especially when the genetic effect is not constant. This is expected, as the change in the exposure values induced by the change in a three-value instrument will be bigger than that induced by the change in a binary instrument. Also, when the genetic effect is not constant, we specify a bigger size of genetic effect, so the approximation of QACEs to the derivative of h is less precise (recall our discussion on this in section 3.3.2).

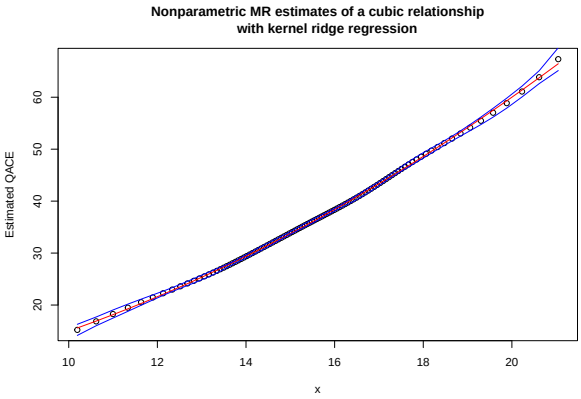
Moreover, careful readers must have noticed that some results with respect to the derivative of h with three-value instruments seem slightly odd. For example, in Tables 3.7 and 3.8, the bias of both nonparametric MR methods for the cubic and the exponential functions with 1% instrument strength are smaller than those with 4% instrument strength. This seems contradictory to the general understanding that weak instrument increases finite sample bias of MR estimates (Burgess et al., 2011). However, if we examine Tables 3.13 and 3.14,

we observe that the biases with respect to the theoretical QACEs for the cubic and the exponential functions with 4% instrument strength are similar to those with 1% instrument strength. I.e. much of the bias with respect to the derivative of h comes from errors due to using QACE to approximate the derivative of h.

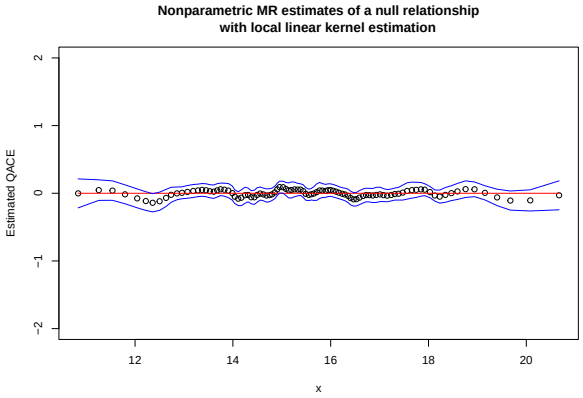
Additionally, by comparing Table 3.9 against Table 3.10 and Table 3.12 against Table 3.13, we observe that the biases with respect to the theoretical QACEs for constant genetic effect are very similar to those for nonconstant but rank-preserving genetic effect. This confirms our argument that nonparametric MR is robust to the violation of the constant genetic effect assumption, as long as the rank-preserving assumption holds.



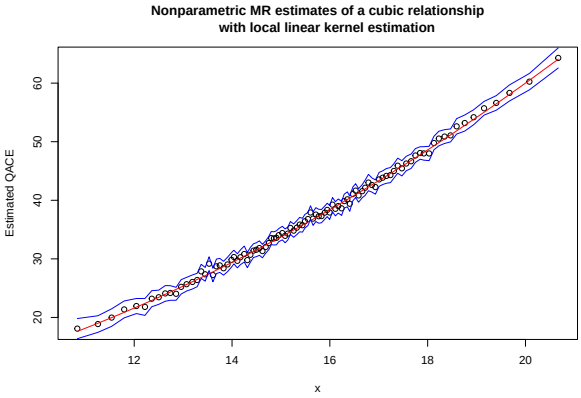
(a)



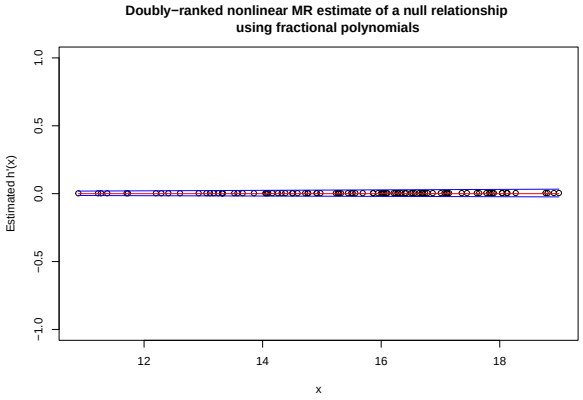
(d)



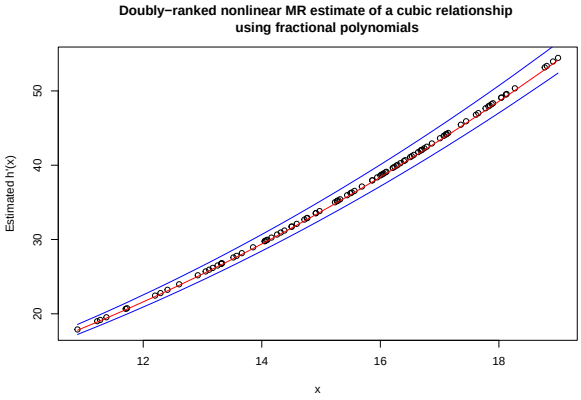
(b)



(e)

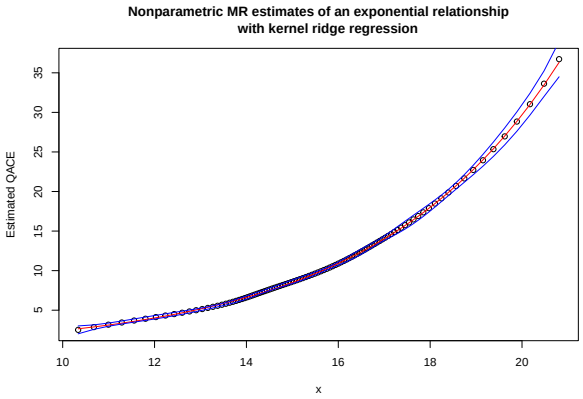


(c)

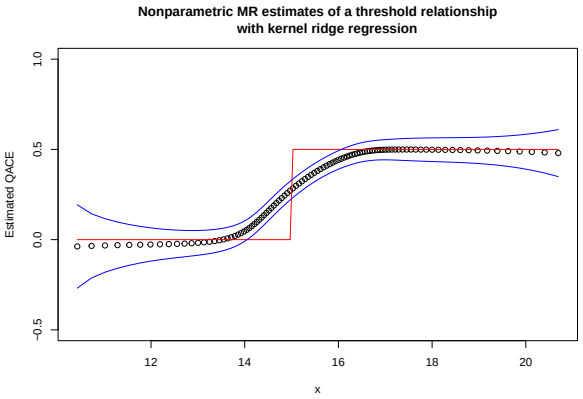


(f)

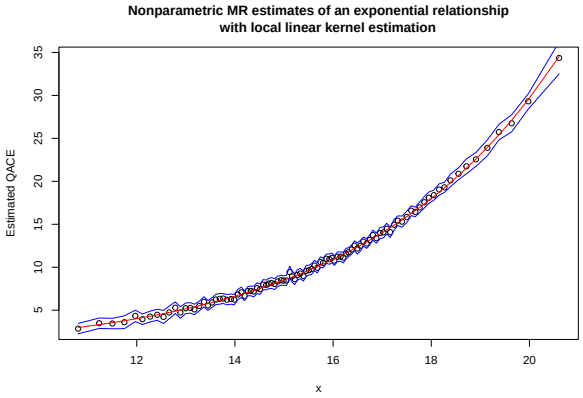
Figure 3.3: Nonparametric MR and stratified nonlinear MR estimates for null and cubic functions with a binary instrument, constant genetic effect and 4% instrument strength. The black dots represent the estimates, the blue curves represent pointwise 95% confidence intervals and the red curve represents the true derivative of h.



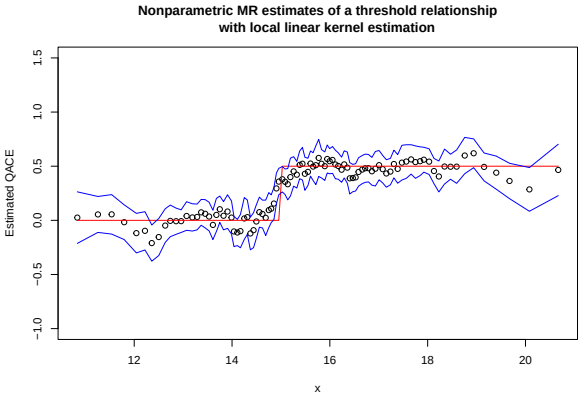
(a)



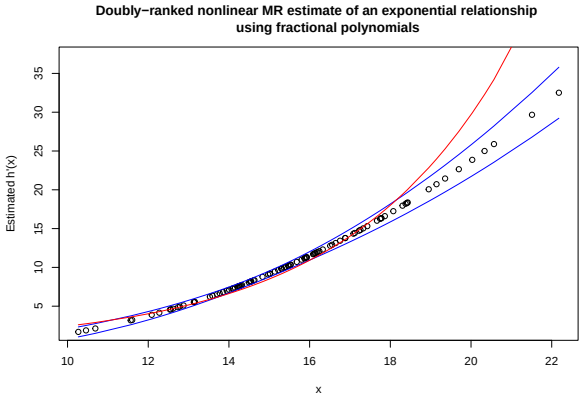
(d)



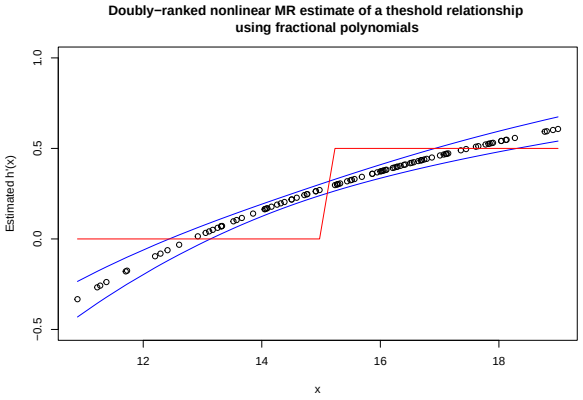
(b)



(e)

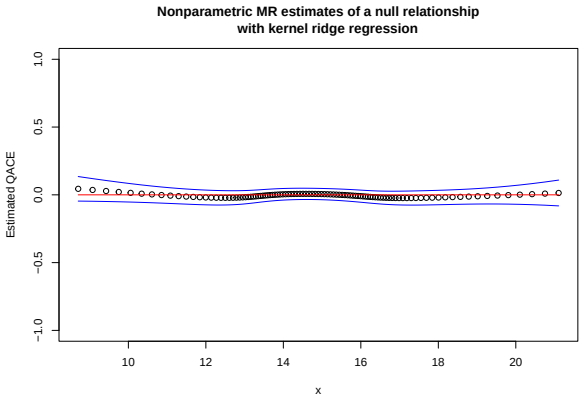


(c)

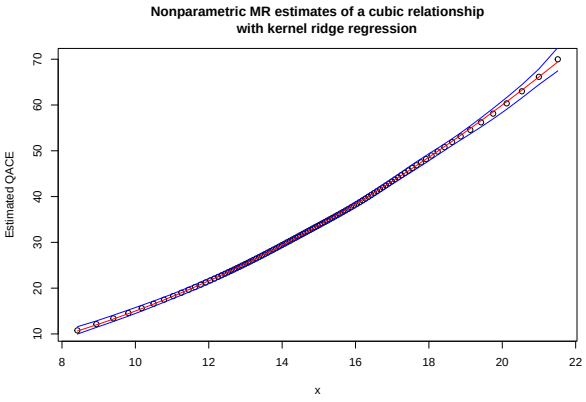


(f)

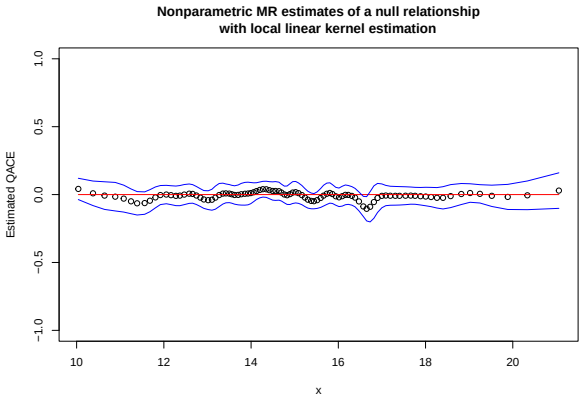
Figure 3.4: Nonparametric MR and stratified nonlinear MR estimates for exponential and threshold functions with a binary instrument, constant genetic effect and 4% instrument strength. The black dots represent the estimates, the blue curves represent pointwise 95% confidence intervals and the red curve represents the true derivative of h_{400}



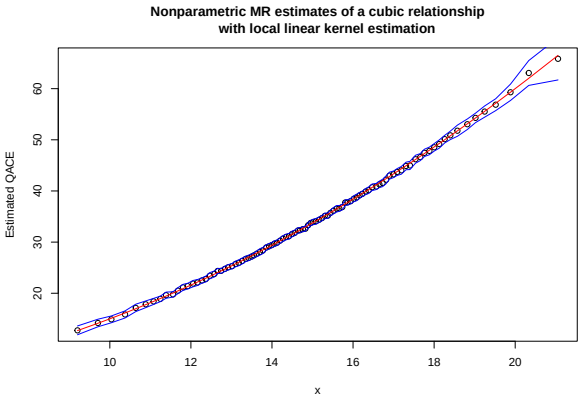
(a)



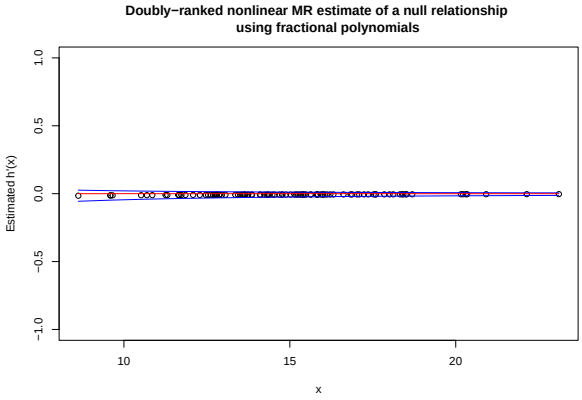
(d)



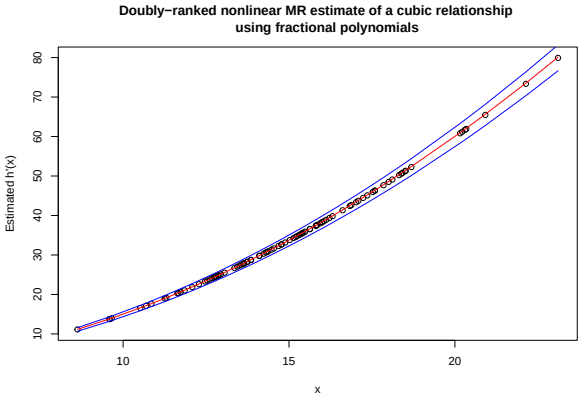
(b)



(e)

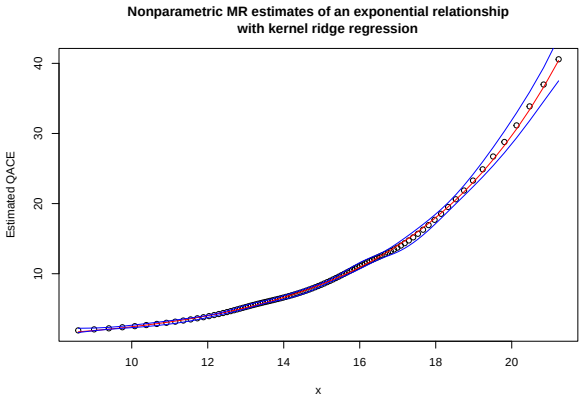


(c)

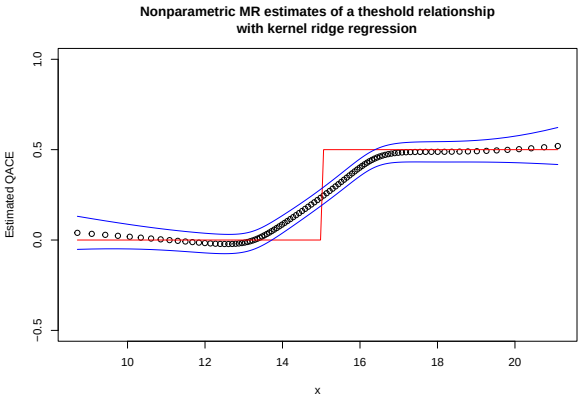


(f)

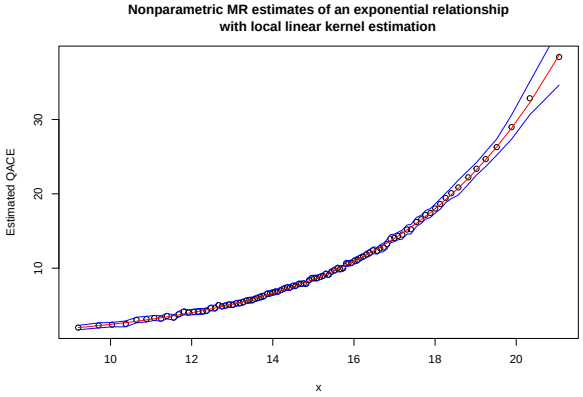
Figure 3.5: Nonparametric MR and stratified nonlinear MR estimates for null and cubic functions with a three-value instrument, constant genetic effect and 4% instrument strength. The black dots represent the estimates, the blue curves represent pointwise 95% confidence intervals and the red curve represents the true derivative of h.



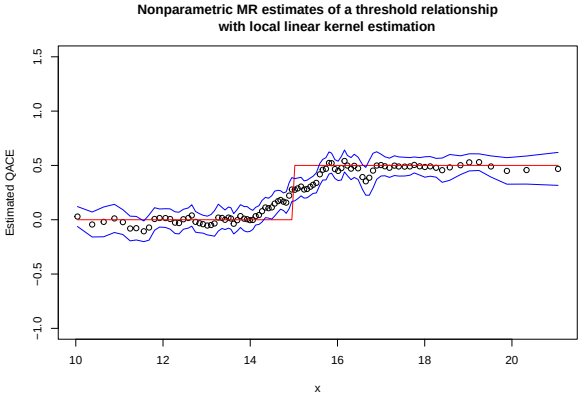
(a)



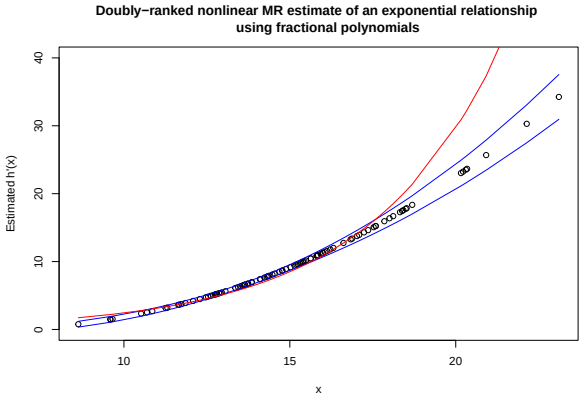
(d)



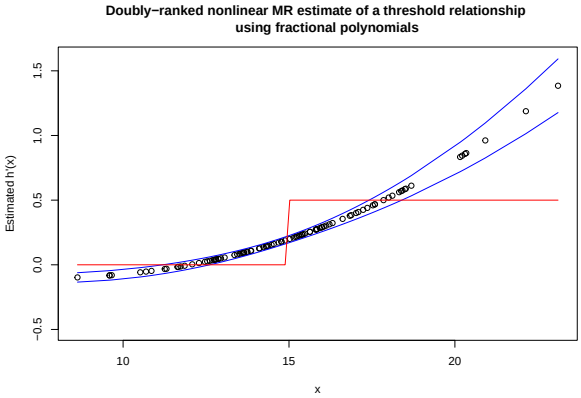
(b)



(e)



(c)



(f)

Figure 3.6: Nonparametric MR and stratified nonlinear MR estimates for exponential and threshold functions with a three-value instrument, constant genetic effect and 4% instrument strength. The black dots represent the estimates, the blue curves represent pointwise 95% confidence intervals and the red curve represents the true derivative of h.

Bias and standard deviation with a binary instrument and constant genetic effect

		Nonparametric MR with kernel ridge regression		Nonparametric MR with local linear kernel estimation		Stratified nonlinear MR with doubly-ranked stratification	
Structural function h	IV strength	Bias	SD	Bias	SD	Bias	SD
null	4%	0.003	0.033	0.005	0.053	0.002	0.023
null	1%	0.009	0.069	0.012	0.111	0.004	0.046
null	0.3%	0.058	0.470	0.085	0.710	0.010	0.107
cubic	4%	0.080	0.213	0.031	0.369	0.133	1.681
cubic	1%	0.091	0.353	0.060	0.707	0.152	1.545
cubic	0.3%	0.173	1.350	0.195	2.209	0.175	1.534
exponential	4%	0.201	0.189	0.139	1.344	0.683	0.864
exponential	1%	0.191	0.327	0.253	2.613	0.614	0.815
exponential	0.3%	0.253	1.180	0.526	5.553	1.558	1.113
threshold	4%	0.047	0.033	0.025	0.065	0.100	0.046
threshold	1%	0.051	0.069	0.023	0.134	0.104	0.071
threshold	0.3%	0.096	0.483	0.092	0.776	0.135	0.143

Table 3.3: Magnitude of the bias and standard deviation of the estimated derivatives with a binary instrument, constant genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$.

Bias and standard deviation with a binary instrument and nonconstant but rank-preserving genetic effect

		Nonparametric MR with kernel ridge regression		Nonparametric MR with local linear kernel estimation		Stratified nonlinear MR with doubly-ranked stratification	
Structural function h	IV strength	Bias	SD	Bias	SD	Bias	SD
null	4%	0.007	0.032	0.007	0.052	0.004	0.022
null	1%	0.011	0.063	0.011	0.102	0.005	0.040
null	0.3%	0.048	0.185	0.073	0.644	0.008	0.085
cubic	4%	0.105	0.200	0.032	0.378	0.181	1.520
cubic	1%	0.107	0.308	0.045	0.670	0.145	1.616
cubic	0.3%	0.155	0.819	0.252	2.872	0.153	1.614
exponential	4%	0.201	0.155	0.300	3.165	0.604	0.703
exponential	1%	0.188	0.265	0.391	4.011	0.580	0.735
exponential	0.3%	0.233	1.003	0.361	4.107	1.550	1.143
threshold	4%	0.048	0.032	0.029	0.064	0.105	0.041
threshold	1%	0.048	0.062	0.022	0.122	0.106	0.058
threshold	0.3%	0.060	0.194	0.053	0.467	0.133	0.127

Table 3.4: Magnitude of the bias and standard deviation of the estimated derivatives with a binary instrument, nonconstant but rank-preserving genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$.

Bias and standard deviation with a binary instrument and rank-varying genetic effect

		Nonparametric MR with kernel ridge regression		Nonparametric MR with local linear kernel estimation		Stratified nonlinear MR with doubly-ranked stratification	
Structural function h	IV strength	Bias	SD	Bias	SD	Bias	SD
null	4%	0.041	0.027	0.039	0.042	0.040	0.028
null	1%	0.085	0.071	0.081	0.109	0.075	0.062
null	0.3%	0.145	0.399	0.119	0.223	0.101	0.095
cubic	4%	0.323	0.268	0.314	0.743	0.657	2.123
cubic	1%	0.605	1.678	0.464	3.234	0.897	2.093
cubic	0.3%	0.667	1.429	0.533	6.010	1.034	2.049
exponential	4%	0.642	0.280	0.158	1.599	1.095	1.059
exponential	1%	0.819	0.636	0.523	5.910	1.044	1.096
exponential	0.3%	1.361	1.359	0.641	6.574	1.038	1.109
threshold	4%	0.071	0.027	0.061	0.054	0.115	0.038
threshold	1%	0.104	0.070	0.093	0.138	0.157	0.041
threshold	0.3%	0.173	0.421	0.137	0.300	0.169	0.063

Table 3.5: Magnitude of the bias and standard deviation of the estimated derivatives with a binary instrument, rank-varying genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$.

Bias and standard deviation with a three-value instrument and constant genetic effect

		Nonparametric MR with kernel ridge regression		Nonparametric MR with local linear kernel estimation		Stratified nonlinear MR with doubly-ranked stratification	
Structural function h	IV strength	Bias	SD	Bias	SD	Bias	SD
null	4%	0.003	0.025	0.003	0.038	0.004	0.020
null	1%	0.006	0.049	0.007	0.075	0.008	0.039
null	0.3%	0.023	0.182	0.023	0.220	0.017	0.082
cubic	4%	0.106	0.308	0.033	0.299	0.186	1.972
cubic	1%	0.103	0.434	0.023	0.383	0.189	1.952
cubic	0.3%	0.259	2.136	0.125	1.379	0.182	1.748
exponential	4%	0.302	0.260	0.210	2.185	0.879	0.905
exponential	1%	0.294	0.428	0.231	2.481	0.873	0.924
exponential	0.3%	0.316	1.504	0.948	9.260	1.132	1.383
threshold	4%	0.046	0.026	0.034	0.047	0.100	0.046
threshold	1%	0.049	0.050	0.023	0.091	0.101	0.060
threshold	0.3%	0.063	0.183	0.034	0.270	0.109	0.127

Table 3.6: Magnitude of the bias and standard deviation of the estimated derivatives with a three-value instrument, constant genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$.

Bias and standard deviation with a three-value instrument and nonconstant but rank-preserving genetic effect

		Nonparametric MR with kernel ridge regression		Nonparametric MR with local linear kernel estimation		Stratified nonlinear MR with doubly-ranked stratification	
Structural function h	IV strength	Bias	SD	Bias	SD	Bias	SD
null	4%	0.014	0.026	0.013	0.037	0.008	0.022
null	1%	0.010	0.055	0.008	0.076	0.005	0.038
null	0.3%	0.037	0.149	0.031	0.213	0.005	0.082
cubic	4%	0.818	0.330	0.800	0.335	1.508	1.850
cubic	1%	0.403	0.485	0.396	0.415	0.817	1.830
cubic	0.3%	0.165	0.883	0.163	1.325	0.473	1.816
exponential	4%	0.577	0.269	0.417	2.541	1.083	0.847
exponential	1%	0.459	0.447	0.514	3.361	0.977	0.886
exponential	0.3%	0.554	2.820	0.559	5.480	1.406	1.305
threshold	4%	0.052	0.027	0.045	0.047	0.104	0.038
threshold	1%	0.050	0.055	0.024	0.094	0.106	0.058
threshold	0.3%	0.063	0.149	0.039	0.255	0.130	0.123

Table 3.7: Magnitude of the bias and standard deviation of the estimated derivatives with a three-value instrument, nonconstant but rank-preserving genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$.

Bias and standard deviation with a three-value instrument and rank-varying genetic effect

		Nonparametric MR with kernel ridge regression		Nonparametric MR with local linear kernel estimation		Stratified nonlinear MR with doubly-ranked stratification	
Structural function h	IV strength	Bias	SD	Bias	SD	Bias	SD
null	4%	0.037	0.027	0.035	0.040	0.032	0.028
null	1%	0.060	0.058	0.057	0.097	0.045	0.055
null	0.3%	0.100	0.208	0.102	0.321	0.066	0.093
cubic	4%	0.866	0.341	0.823	0.433	1.827	2.135
cubic	1%	0.502	0.636	0.409	0.526	1.407	2.058
cubic	0.3%	0.621	2.138	0.401	2.135	1.330	2.021
exponential	4%	0.925	0.332	0.478	2.225	1.366	1.118
exponential	1%	0.823	0.494	0.379	2.331	1.166	1.110
exponential	0.3%	0.907	1.095	0.875	8.142	1.110	1.264
threshold	4%	0.072	0.028	0.067	0.050	0.111	0.038
threshold	1%	0.087	0.059	0.074	0.122	0.146	0.049
threshold	0.3%	0.128	0.217	0.104	0.419	0.158	0.080

Table 3.8: Magnitude of the bias and standard deviation of the estimated derivatives with a three-value instrument, rank-varying genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$.

Bias w.r.t QACE with a binary instrument and constant genetic effect

		Nonparametric MR with kernel ridge regression		Nonparametric MR with local linear kernel estimation	
Structural function h	IV strength	w.r.t. h'	w.r.t. QACE	w.r.t. h'	w.r.t. QACE
cubic	4%	0.080	0.080	0.031	0.029
cubic	1%	0.091	0.091	0.060	0.059
cubic	0.3%	0.173	0.173	0.195	0.195
exponential	4%	0.302	0.294	0.139	0.130
exponential	1%	0.294	0.292	0.253	0.252
exponential	0.3%	0.253	0.253	0.526	0.526
threshold	4%	0.047	0.028	0.025	0.006
threshold	1%	0.051	0.042	0.023	0.013
threshold	0.3%	0.096	0.093	0.092	0.089

Table 3.9: Comparison of bias of nonparametric MR estimates with respect to the derivative of the structural function h and with respect to the theoretical QACE, with a binary instrument, constant genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$. Results for the null function are omitted.

Bias w.r.t QACE with a binary instrument and nonconstant but rank-preserving genetic effect

		Nonparametric MR with kernel ridge regression		Nonparametric MR with local linear kernel estimation	
Structural function h	IV strength	w.r.t. h'	w.r.t. QACE	w.r.t. h'	w.r.t. QACE
cubic	4%	0.105	0.105	0.032	0.029
cubic	1%	0.107	0.107	0.045	0.045
cubic	0.3%	0.155	0.155	0.252	0.252
exponential	4%	0.201	0.199	0.300	0.298
exponential	1%	0.188	0.188	0.391	0.388
exponential	0.3%	0.229	0.229	0.361	0.361
threshold	4%	0.048	0.030	0.029	0.008
threshold	1%	0.048	0.040	0.022	0.013
threshold	0.3%	0.060	0.056	0.053	0.050

Table 3.10: Comparison of bias of nonparametric MR estimates with respect to the derivative of the structural function h and with respect to the theoretical QACE, with a binary instrument, nonconstant but rank-preserving genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$. Results for the null function are omitted.

Bias w.r.t. QACE with a binary instrument and rank-varying genetic effect

		Nonparametric MR with kernel ridge regression		Nonparametric MR with local linear kernel estimation	
Structural function h	IV strength	w.r.t. h'	w.r.t. QACE	w.r.t. h'	w.r.t. QACE
cubic	4%	0.323	0.323	0.314	0.314
cubic	1%	0.605	0.605	0.464	0.466
cubic	0.3%	0.667	0.666	0.533	0.533
exponential	4%	0.642	0.634	0.158	0.151
exponential	1%	0.819	0.817	0.523	0.527
exponential	0.3%	1.361	1.359	0.641	0.649
threshold	4%	0.071	0.053	0.061	0.054
threshold	1%	0.112	0.104	0.093	0.084
threshold	0.3%	0.173	0.168	0.137	0.130

Table 3.11: Comparison of bias of nonparametric MR estimates with respect to the derivative of the structural function h and with respect to the theoretical QACE, with a binary instrument, rank-varying genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$. Results for the null function are omitted.

Bias w.r.t QACE with a three-value instrument and constant genetic effect

		Nonparametric MR with kernel ridge regression		Nonparametric MR with local linear kernel estimation	
Structural function h	IV strength	w.r.t. h'	w.r.t. QACE	w.r.t. h'	w.r.t. QACE
cubic	4%	0.105	0.102	0.033	0.011
cubic	1%	0.103	0.103	0.023	0.023
cubic	0.3%	0.259	0.259	0.125	0.119
exponential	4%	0.201	0.199	0.210	0.196
exponential	1%	0.191	0.191	0.231	0.234
exponential	0.3%	0.316	0.316	0.526	0.526
threshold	4%	0.046	0.018	0.034	0.008
threshold	1%	0.049	0.035	0.023	0.008
threshold	0.3%	0.063	0.056	0.034	0.026

Table 3.12: Comparison of bias of nonparametric MR estimates with respect to the derivative of the structural function h and with respect to the theoretical QACE, with a three-value instrument, constant genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$. Results for the null function are omitted.

Bias w.r.t QACE with a three-value instrument and nonconstant but rank-preserving genetic effect

		Nonparametric MR with kernel ridge regression		Nonparametric MR with local linear kernel estimation	
Structural function h	IV strength	w.r.t. h'	w.r.t. QACE	w.r.t. h'	w.r.t. QACE
cubic	4%	0.818	0.154	0.800	0.017
cubic	1%	0.403	0.119	0.396	0.025
cubic	0.3%	0.165	0.130	0.163	0.110
exponential	4%	0.577	0.391	0.417	0.231
exponential	1%	0.459	0.394	0.514	0.322
exponential	0.3%	0.554	0.545	0.559	0.546
threshold	4%	0.052	0.026	0.045	0.013
threshold	1%	0.050	0.037	0.023	0.009
threshold	0.3%	0.063	0.057	0.039	0.033

Table 3.13: Comparison of bias of nonparametric MR estimates with respect to the derivative of the structural function h and with respect to the theoretical QACE, with a three-value instrument, nonconstant but rank-preserving genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$. Results for the null function are omitted.

Bias w.r.t. QACE with a three-value instrument and rank-varying genetic effect

		Nonparametric MR with kernel ridge regression		Nonparametric MR with local linear kernel estimation	
Structural function h	IV strength	w.r.t. h'	w.r.t. QACE	w.r.t. h'	w.r.t. QACE
cubic	4%	0.866	0.282	0.823	0.038
cubic	1%	0.502	0.371	0.409	0.067
cubic	0.3%	0.621	0.577	0.401	0.232
exponential	4%	0.925	0.717	0.477	0.238
exponential	1%	0.823	0.754	0.379	0.212
exponential	0.3%	0.907	0.870	0.875	0.794
threshold	4%	0.072	0.045	0.067	0.037
threshold	1%	0.087	0.075	0.074	0.060
threshold	0.3%	0.128	0.120	0.113	0.104

Table 3.14: Comparison of bias of nonparametric MR estimates with respect to the derivative of the structural function h and with respect to the theoretical QACE, with a binary instrument, rank-varying genetic effect and instrument strength $\in \{4\%, 1\%, 0.3\%\}$. Results for the null function are omitted.

3.5 Applied Examples from UK Biobank

3.5.1 Description of the examples

We now illustrate nonparametric MR using real data examples from UK Biobank, a prospective cohort study of around half a million UK residents recruited from 2006 to 2010 (Sudlow et al., 2015). We consider the causal effect of BMI on two clinical outcomes, including average handgrip power and pulse rate. As for the instrument, we choose FTO (fat mass and obesity-associated) gene rs9939609, a well-characterised SNP (single nucleotide polymorphism) that shows the strongest association with BMI and obesity. The effect of rs9939609 on BMI is thought to be mediated via effects on satiety (i.e. how full up people feel), such as loss of control eating behaviour and selection of energy-dense foods (e.g. Tanofsky-Kraff et al., 2009).

A very small number of missing individuals for BMI and for both outcomes are dropped for complete-case analyses. For simplicity, we do not consider any covariates in our analysis (the estimates will still valid without adjustment for covariates as long as we are using valid instruments). The standard MR estimate for the causal effect of BMI on average handgrip power is 0.286 (with standard error 0.0770) and for the causal effect of BMI on pulse rate is null (the estimate is 0.00158 with standard error 0.0794) on the full UK Biobank sample.

Because bootstrapping the nonparametric MR estimator is computationally demanding, we apply our method to a randomly drawn subsample of size 10,000 for each of the outcomes. The strength of the instrument, as measured by R^2 of the instrument-exposure regression, is about 0.45% for the subsample of average handgrip power and about 0.3% for the subsample of pulse rate. The standard MR estimates for both subsamples are null (for average handgrip power the estimate is 0.0202 with standard error 0.345; for pulse rate the estimate is -0.641 with standard error 0.467).

3.5.2 Results of the examples

Figure 3.7 presents nonparametric MR estimates with both kernel ridge regression and local linear kernel estimation and doubly-ranked nonlinear MR estimates. Consistent with our approach in the simulation studies, we show QACE estimates for nonparametric MR and show the derivative of the estimated structural function for doubly-ranked nonlinear MR. We construct pointwise 95% confidence intervals (the blue curves) using nonparametric bootstrap with $R = 1,000$ bootstrapped samples for nonparametric MR, and using asymptotic standard errors of fractional polynomial regression for doubly-ranked nonlinear MR. We also plot the null effects (the red curves) for reference.

For average handgrip power, all three methods give clear null estimates for the causal effect (subfigures (a), (b) and (c) in Figure 3.7). This is consistent with the standard MR estimate of the subsample. However, since the standard MR estimate on the full UK Biobank sample is significant, the null estimates can simply be due to lack of power. Nevertheless, we note that the confidence intervals of nonparametric MR estimates (subfigures (a) and (b) in Figure 3.7) are very wide for small values of BMI, and become tight for large values of BMI. This suggests that if there is positive average causal effect of BMI on handgrip power, the causal effect is possibly nonlinear and likely to be null for large values of BMI.

The varying width of confidence intervals for the nonparametric MR estimates is likely due to varying instrument strength across the exposure distribution. Subfigure (a) in Figure 3.8 shows how the strength of the genetic variant rs9939609 changes across the distribution of BMI for the subsample of average handgrip power. We use doubly-ranked stratification to form 100 strata and use F-statistics of the instrument-exposure regressions to measure the instrument strength within strata. F-statistics, another commonly used measure of instrument strength, has an advantage over R^2 that it takes into account the sample size, so it is more appropriate to measure the strength of the genetic variants within strata. We see that the genetic variant is the weakest at the lower tails of BMI distribution, gradually increases

as the value of BMI increases and is the strongest at BMI values about 35. This is consistent with the change of the width of confidence intervals for the nonparametric MR estimates.

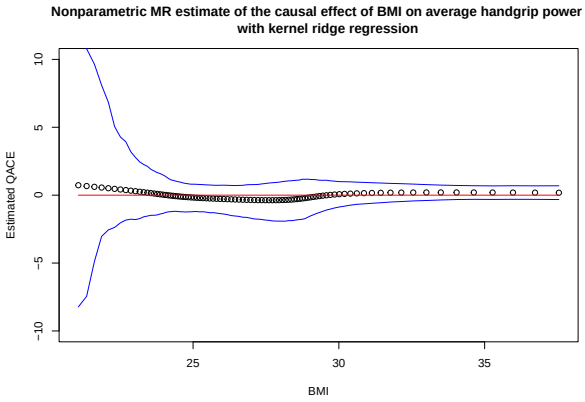
The confidence intervals for stratified nonlinear MR estimates (subfigure (c) in Figure 3.7), on the other hand, are tight for small values of BMI and become increasingly wide for large values of BMI. This is contradictory to what we have observed with nonparametric MR. However, if we examine the localised average causal effect (LACE) estimates from the doubly-ranked stratification (Table 3.15), we observe that the standard errors are the biggest for stratum 1, 2 and 3, and gradually decreases from stratum 3 to stratum 10. This agrees with the observations from nonparametric MR. As we have discussed in Chapter 2 (section 2.6.3), this is an example of the disadvantage of using fractional polynomial regression to estimate the structural function that the standard error(s) of the estimated fractional polynomial coefficient(s) is a measure of uncertainty for the best global fit across strata, which can be different from the standard errors of LACE estimates (and also nonparametric MR estimates) that measure the uncertainties locally. In this case, the standard errors of the fractional polynomial regression for stratified nonlinear MR will not reflect the actual uncertainties of estimating causal effects at the lower tails of BMI distribution.

For pulse rate, both nonparametric MR methods provide null estimates (subfigures (d) and (e) in Figure 3.7). This agrees with the standard MR estimates of both the subsample and the full UK Biobank sample. Nonparametric MR with kernel ridge regression shows a smoother and clearer estimate, which is consistent with our observations in the simulation study. The confidence intervals of nonparametric estimates are similar to those for average handgrip power that the confidence intervals are wide for small values of BMI and become tight in general as BMI value increases. Again, this is consistent with the varying instrument strength as shown by subfigure (b) in Figure 3.8.

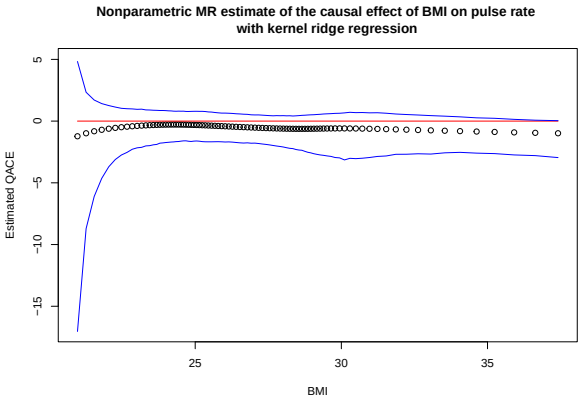
However, stratified nonlinear MR gives non-null (negative) estimates for the causal effect, and the magnitude of the estimated effect increases as BMI increases (subfigure (f) in Figure 3.7).

The estimated exposure-outcome relationship is a cubic function, with estimated coefficient equal to -0.000275 (with standard error 0.000139) that is significant at 5% level. The pointwise 95% confidence interval slightly excludes the null (which may not be obvious from the graph).

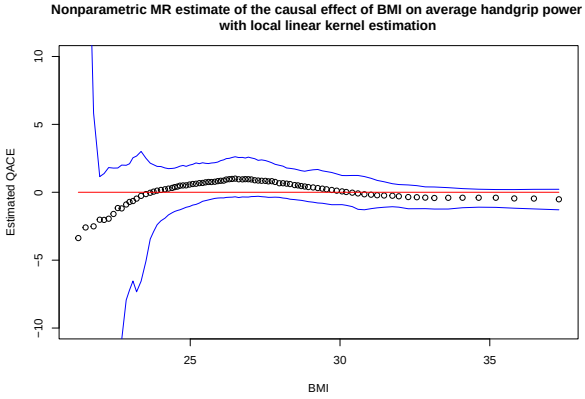
Nevertheless, if we examine the localised average causal effect (LACE) estimates from the doubly-ranked stratification (Table 3.16), we observe that none of the stratum-specific estimates are significant. The confidence intervals for stratified nonlinear MR estimates also contradict with the standard errors of the respective LACE estimates. This shows that the LACE estimates agree with the nonparametric MR estimates, but the fractional polynomial regression incorrectly extrapolates the LACE estimates with an overly optimistic standard error. This again corroborates our discussions before that estimating the structural function from LACE estimates using fractional polynomials can result in misleading causal conclusions if the estimates are interpreted locally.



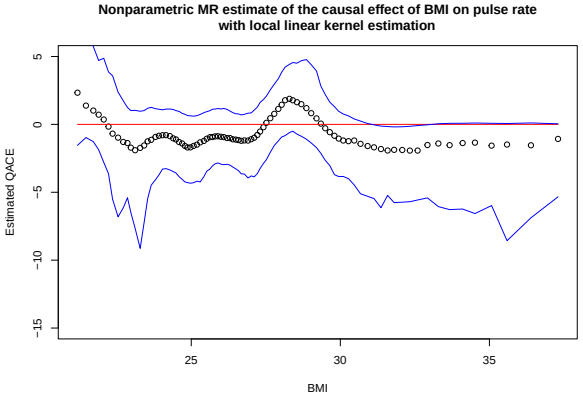
(a)



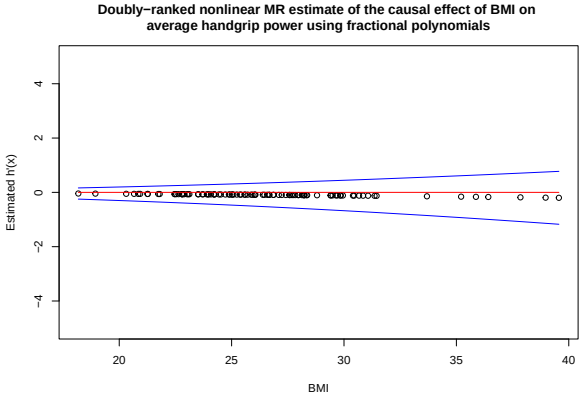
(d)



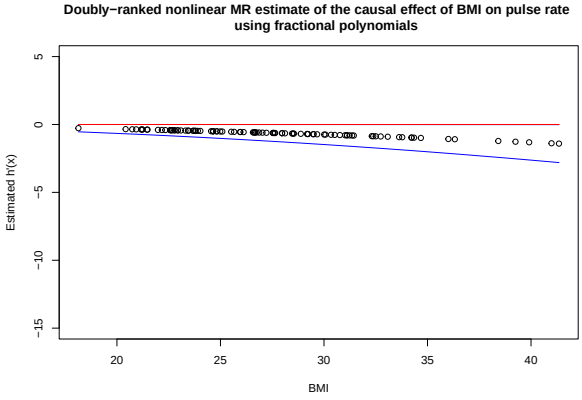
(b)



(e)

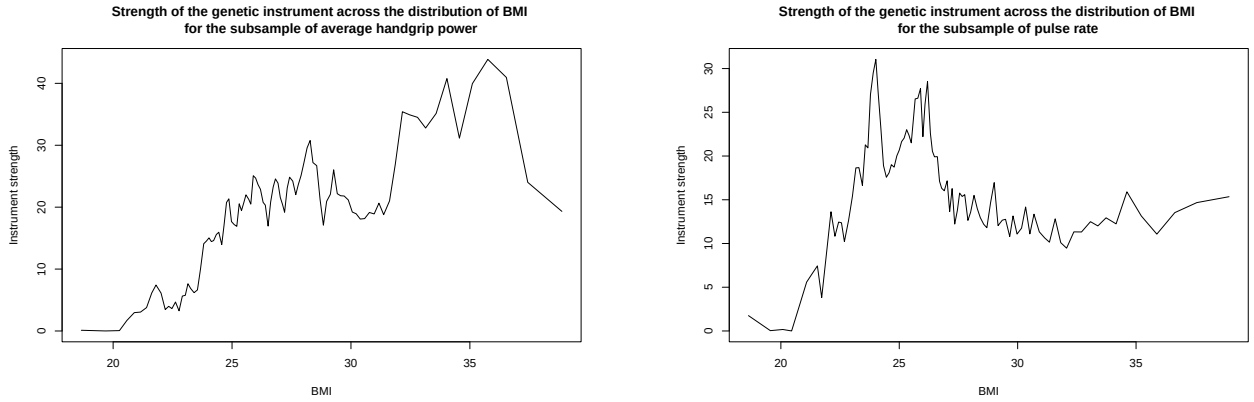


(c)



(f)

Figure 3.7: Nonparametric MR and stratified nonlinear MR estimates for the relationship between BMI and average handgrip power and between BMI and pulse rate. The black curve represents the estimates and the blue curves represent pointwise 95% confidence intervals. The red curve indicates a null effect.



(a) (b)

Figure 3.8: Strength of the genetic variant rs9939609 across the distribution of BMI for the subsamples of average handgrip power and pulse rate. The instrument strength is measured by the F-statistics of the instrument-exposure regressions within 100 strata defined by doubly-ranked stratification.

LACE estimates for the causal relationship between BMI and average handgrip power in the subsample

Stratum	Stratum-specific estimate	Standard error	95% confidence interval
1	−2.44	2.00	[−6.35, 1.48]
2	−0.131	2.08	[−4.20, 3.94]
3	−1.23	2.05	[−5.25, 2.80]
4	0.786	1.41	[−1.97, 3.55]
5	1.57	1.31	[−1.01, 4.14]
6	1.36	1.23	[−1.04, 3.77]
7	0.0978	0.974	[−1.81, 2.01]
8	−0.149	0.909	[−1.93, 1.63]
9	0.0493	0.712	[−1.35, 1.44]
10	−0.588	0.563	[−1.69, 0.516]

Table 3.15: Stratum-specific (LACE) estimates using doubly-ranked stratification for estimating the relationship between BMI and average handgrip power in the randomly drawn subsample.

LACE estimates for the causal relationship between BMI and pulse rate in the subsample

Stratum	Stratum-specific estimate	Standard error	95% confidence interval
1	3.64	2.69	[−1.63, 8.91]
2	0.0713	1.76	[−3.39, 3.53]
3	1.82	1.64	[−1.40, 5.03]
4	−2.49	1.79	[−6.00, 1.01]
5	−2.88	1.62	[−6.05, 0.288]
6	−0.858	1.46	[−3.71, 2.00]
7	−1.54	1.22	[−3.94, 0.862]
8	1.15	1.30	[−1.40, 3.70]
9	−2.16	1.12	[−4.35, 0.0398]
10	−0.737	0.719	[−2.15, 0.672]

Table 3.16: Stratum-specific (LACE) estimates using doubly-ranked stratification for estimating the relationship between BMI and pulse rate in the randomly drawn subsample.

3.6 Discussion and Concluding Remarks

In this chapter, we contribute to the MR literature by introducing nonparametric MR, a nonparametric extension to stratified nonlinear MR for characterising potentially nonlinear exposure-outcome relationships with discrete instrumental variables. Similar to the doubly-ranked stratification, nonparametric MR relaxes the constant genetic effect assumption and requires only the weaker rank-preserving assumption. Nonparametric MR improves the existing stratified nonlinear MR methods in the regard that it is able to flexibly characterise complex nonlinear relationships beyond fractional polynomials, and by using quantiles of the counterfactual exposure distributions in lieu of stratification, there is no need to pre-specify the number or the size of the strata.

In section 3.2, we firstly distinguish the concept of quantile average causal effect (QACE) that looks at the quantiles of the counterfactual exposure distributions from another causal concept called quantile treatment effect (QTE) that looks at the quantiles of the counterfactual

outcome distributions.

Then in section 3.3, we formally define the estimand of our nonparametric MR estimator as quantile average causal effect (QACE), a different version of localised average causal effect (LACE) that approximates the derivative of the structural function and has meaningful interpretations in practice. We then demonstrate through simulation studies in section 3.4 that nonparametric MR reliably estimates various structural functions under a range of scenarios with different instrument-outcome models and instrument strengths, and outperforms doubly-ranked nonlinear MR for estimating functions that do not belong to the class of fractional polynomials if either the rank-preserving assumption is satisfied or the instrument is not too weak. Among the two nonparametric regression methods we consider, nonparametric MR with kernel ridge regression provides a smoother estimate than local linear kernel estimation and is generally more robust to weak instruments when the rank-preserving assumption holds. Nevertheless, nonparametric MR can become problematic when the rank-preserving assumption is violated and the instrument only explains 0.3% of the exposure variance.

In section 3.4, we compare nonparametric MR against stratified nonlinear MR in real data examples from UK Biobank by considering the effect of BMI on average handgrip power and on pulse rate. Although nonparametric MR does not give a significant nonlinear causal effect potentially due to lack of power with relatively small sample size (we have only used a subset of all the available data because of computational constraint), the estimates are consistent with the standard MR estimates and the LACE estimates from doubly-ranked stratification in the same sample. On the other hand, we observe that stratified nonlinear MR estimates obtained using fractional polynomial regression might give misleading conclusions for the uncertainties of estimated local causal effects. Therefore, we suggest researchers examine LACE estimates from doubly-ranked stratification first before fitting the fractional polynomials regression. We leave the application of nonparametric MR to the full UK Biobank datasets for future research.

There are of course a number of limitations in this research. The most prominent limitation is that the proposed nonparametric MR method only works for a single instrument that takes a small number of discrete values, such as one genetic variant that takes only three values. This is not the most efficient use of genetic information. Indeed, in most MR investigations, researchers choose a large number of genetic variants and combine them into a continuous genetic risk score. We will overcome this disadvantage in the next chapter, where we extend nonparametric MR to be compatible with continuous instruments. A relevant limitation is that a single genetic variant is often very weak, so we might need massive sample size for nonparametric MR to render significant estimates as we have demonstrated with our UK Biobank examples. Nevertheless, we expect our method to be applicable to various situations outside of MR where the strength of the instrument could be much stronger, such as studies with an encouragement design.

Another important limitation is the choice of nonparametric regression methods. Theoretically, nonparametric MR estimators using any consistent nonparametric regression method will converge to the same QACE, but their empirical performances in finite samples depend crucially on the chosen nonparametric regression method. In this chapter, we have only considered two methods, namely kernel ridge regression and local linear kernel estimation, and have shown that kernel ridge regression generally performs better. We leave it for future research to investigate other, more sophisticated nonparametric regression methods. A very relevant issue with using nonparametric regressions is how to do hyperparameter tuning. Standard cross-validation on held-out datasets will result in overfitting for kernel ridge regression, and in order to avoid this, we impose an additional constraint that the hyperparameters need to be the same for all kernel ridge regressions involved with predicting the exposure and the outcome values using quantiles of the counterfactual exposure distributions. We justify for this constraint by arguing that we want the kernel ridge regressions to be similarly penalised, but it will be an interesting direction for future research to explore other hyperparameter tuning procedures for nonparametric MR. We will expand on this in

a bit more detail in Chapter 6 (section 6.2.2).

One further limitation is the computational time taken to implement nonparametric MR. Again, this depends crucially on the nonparametric regression method that we choose. Local linear kernel estimation runs relatively fast, while kernel ridge regression is much more computationally demanding as it involves inverting an $n \times n$ matrix. For both methods, the computational time increases exponentially as the sample size increases. This is why we restrict the analyses to only subsamples of UK Biobank in section 3.4. However, we anticipate that the major reason for the large computational time in our analyses is that R is a very slow language, and the algorithm should run much faster if we implement it in a different language such as Python or C++. Nevertheless, it will be an important direction for future research to speed up the computation, for example, by considering cross-validation with fewer folds.

In summary, in this chapter we present nonparametric MR with discrete instrumental variables that is able to flexibly estimate complex nonlinear exposure-outcome relationships and requires only the rank-preserving assumption. We recommend using nonparametric MR in addition to the standard MR and doubly-ranked nonlinear MR for more accurate understandings of nonlinearity.

Chapter 4

Nonparametric Mendelian Randomisation for Characterising Nonlinear Exposure-Outcome Relationships: An Extension with Continuous Instrumental Variables

4.1 Introduction

In Chapter 3, we have introduced nonparametric MR with discrete instrumental variables as a novel nonparametric extension to stratified nonlinear MR. We have shown both its theoretical advantages and empirical performances through simulation studies and UK Biobank examples.

In this chapter, we continue to develop the ideas in Chapter 3 and extend our nonparametric

MR approach to situations where the instrument is continuous. This extension is non-trivial because most MR applications involves combining multiple genetic variants into a continuous genetic risk score so that the instrument is much stronger and more explanatory than a single genetic variant. We obtain quantiles t of the counterfactual exposure distributions via Gaussian mixture modelling (e.g. Dempster et al., 1977; Bishop, 2006; Hastie et al., 2009), and estimate the quantile average causal effect (QACE) via separate two-dimensional nonparametric regressions that estimate the exposure and the outcome values from t. Our approach inherits the advantages of nonparametric MR with discrete IVs that it does not require the strong constant genetic effect assumption and that it is able to characterise complex nonlinear structural functions without the need to pre-specify the number or size of the strata.

Similar to Chapter 3, we demonstrate the advantages of nonparametric MR over stratified nonlinear MR through simulation studies in a wide range of scenarios that cover different instrument-exposure relationships and different instrument strengths. We also apply our method to real data examples from UK Biobank, using continuous genetic risk scores to investigate the causal effect of vitamin D and alcohol intake on systolic blood pressure (SBP).

4.2 Methodology

The first step of nonparametric MR is to obtain empirical quantiles t_i of the counterfactual exposure distributions. With dichotomous or discrete instrumental variables, it is easy to obtain the quantiles as we usually will have enough data on exposure values for each value of the instrument. With a continuous instrument, however, it becomes impossible to condition on every instrument value. Therefore, in order to get quantiles t_i, we need to give a model to the counterfactual exposure distributions.

We choose to model the counterfactual exposure distributions as Gaussian mixture distribu-

tions, a widely used model in statistics and machine learning that is easy to solve using standard numerical algorithms. Note that Deep IV (a deep-learning based nonlinear IV method reviewed in Chapter 2 (section 2.4.4)) adopts a similar approach that also uses Gaussian mixture distributions to model the counterfactual exposure distributions. Similar to the case with discrete IVs presented in Chapter 3, the rank-preserving assumption is enough to ensure that we are comparing valid counterfactuals so the constant genetic effect assumption is not needed. However, this approach does additionally require that the counterfactual exposure distributions cannot greatly deviate from a Gaussian mixture. We will discuss this limitation in later sections.

After having obtained quantiles t_i, we proceed to estimate the relationships between Y and T and X and T. However, unlike the method with discrete instruments where we do several one-dimensional nonparametric regressions using T, with a continuous instrument we need to do two separate two-dimensional nonparametric regressions using both T and Z for Y and X respectively. This is natural, as the quantiles t_i are obtained using all observations of Z so they must be considered jointly. A consequence of this change is that the computational time for fitting the nonparametric regressions will increase.

Finally, we combine the nonparametric regression estimates in a manner that resembles the usual Wald estimator to get the estimates of QACE for each value of $T = t$. These steps are summarised and formalised below.

With a continuous instrumental variable Z, nonparametric MR does the following:

STEP 1. Model the counterfactual exposure distribution, i.e. the distribution of $X_i(z_i)$ using a Gaussian mixture model with two clusters:

$$f(x_i(z_i)) = \alpha_i \phi(x_i(z_i); \mu_1, \sigma_1^2) + (1 - \alpha_i)\phi(x_i(z_i); \mu_2, \sigma_2^2) \tag{4.1}$$

where $f(\cdot)$ denotes the density function of the counterfactual exposure distribution $X_i(z_i)$, $\phi(\cdot)$ denotes the density of a Gaussian distribution and $\{\mu_1, \sigma_1^2, \mu_2, \sigma_2^2\}$ are Gaussian param-

eters. $\alpha_i = \Phi(z_i)$, where $\Phi(\cdot)$ denotes a standard Gaussian CDF. We use α_i to model the proportion coming from the first cluster and $(1 - \alpha_i)$ to model the proportion coming from the second cluster. We choose the Gaussian mixture model as it is a very flexible model for continuous distributions that can be easily solved using standard numerical algorithms.

Note that this model is a structural model (e.g. Robins et al., 2000) in the sense that it models the counterfactual variable $X(z)$. Because of the consistency assumption of counterfactuals, this is equivalent to write

$$f(x_i|z_i) = \alpha_i \phi(x_i; \mu_1, \sigma_1^2) + (1 - \alpha_i)\phi(x_i; \mu_2, \sigma_2^2) \tag{4.2}$$

which can be solved using Expectation-Maximisation (EM) algorithms (e.g. Bishop, 2006) with standard numerical methods. Denote the estimators by $(\hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2)$.

STEP 2. Obtain $t_i \in [0, 1]$ as estimated quantiles of the counterfactual exposure distributions by integrating (4.2):

$$t_i = \hat{F}(x_i) = \int_{-\infty}^{x_i} \hat{f}(x)dx = \alpha_i \Phi(x_i; \hat{\mu}_1, \hat{\sigma}_1^2) + (1 - \alpha_i)\Phi(x_i; \hat{\mu}_2, \hat{\sigma}_2^2) \tag{4.3}$$

where $\Phi(\cdot)$ denotes a general Gaussian CDF (with slight abuse of notation from above).

STEP 3. Nonparametrically estimate the relationships between Y and (T, Z) and X and (T, Z), denoted as $\hat{Y}(t, z)$ and $\hat{X}(t, z)$.

Similar to what we have mentioned in section 3.3.2, $\hat{Y}(t, z)$ is the estimated expectation of Y given $T = t$ and $Z = z$, i.e. $\hat{Y}(t, z) = \widehat{\mathbb{E}}\big[Y \mid T = t, Z = z\big]$. On the other hand, $\hat{X}(t, z)$ is the estimated deterministic function of X given $T = t$ and $Z = z$. We should note that this deterministic function is in fact already specified by equation (4.1), and we can invert (4.3) to obtain (estimated) values of X given $T = t$ and $Z = z$. A benefit of this approach is that the computational time of the algorithm can be reduced as we will only have to run one two-dimensional nonparametric regression. However, estimating values of X using nonparametric regression is more robust to model misspecification of the Gaussian mixture

model given in (4.1). In this chapter, we obtain estimated values of X using two-dimensional nonparametric regression on (T, Z).

STEP 4. Estimate an average causal effect based on t, which we call quantile average causal effect (QACE), by averaging across the distribution of Z:

$$\widehat{QACE}(t) = \frac{\frac{1}{n}\sum_{j=1}^{n}\hat{Y}(t, z_j)z_j - \frac{1}{n}\sum_{j=1}^{n}\hat{Y}(t, z_j)\frac{1}{n}\sum_{j=1}^{n} z_j}{\frac{1}{n}\sum_{j=1}^{n}\hat{X}(t, z_j)z_j - \frac{1}{n}\sum_{j=1}^{n}\hat{X}(t, z_j)\frac{1}{n}\sum_{j=1}^{n} z_j} \tag{4.4}$$

Note that (4.4) can be viewed as a quantile-specific sample analogue to the classic Wald estimand with continuous instruments:

$$\beta_{Wald} = \frac{Cov(Y, Z)}{Cov(X, Z)} = \frac{\mathbb{E}[YZ] - \mathbb{E}[Y]\mathbb{E}[Z]}{\mathbb{E}[XZ] - \mathbb{E}[X]\mathbb{E}[Z]} \tag{4.5}$$

$$= \frac{\mathbb{E}\big(\mathbb{E}[Y|Z]Z\big) - \mathbb{E}\big(\mathbb{E}[Y|Z]\big)\mathbb{E}[Z]}{\mathbb{E}\big(\mathbb{E}[X|Z]Z\big) - \mathbb{E}\big(\mathbb{E}[X|Z]\big)\mathbb{E}[Z]} \tag{4.6}$$

where we replace conditional expectations $\mathbb{E}[Y|Z]$ and $\mathbb{E}[X|Z]$ by nonparametric estimates of Y and X in **STEP 3** and replace the other expectation operators by sample averages to get our QACE estimator $\widehat{QACE}(t)$.

Similar to the nonparametric MR with discrete IVs, we define the theoretical QACE as

$$\frac{\mathrm{Cov}^{(t)}\big(h(X), Z\big)}{\mathrm{Cov}^{(t)}\left(X, Z\right)} \tag{4.7}$$

where $\mathrm{Cov}^{(t)}\left(X, Z\right)$ denotes the covariance between X and Z conditional on $T = t$, i.e. $\mathrm{Cov}^{(t)}\left(X, Z\right) = \mathrm{Cov}\left(X, Z \,|\, T = t\right)$, and similarly $\mathrm{Cov}^{(t)}\left(h(X), Z\right)$ denotes the covariance between $h(X)$ and Z conditional on $T = t$, i.e. $\mathrm{Cov}^{(t)}\left(h(X), Z\right) = \mathrm{Cov}\left(h(X), Z \,|\, T = t\right)$. This can be obtained by plugging the structural equation $Y = h(X) + \epsilon_Y$ into the right hand side of (4.5) and conditioning on t, which represents the average causal effect of X on Y at quantile t induced by the average change across the distribution of Z. Similar to the case the discrete IVs, (4.7) is an approximation of the derivative of the true structural function h.

We should note that nonparametric MR relies on the crucial condition that comparing individuals at the same values of T will not introduce collider bias. As we have argued in Chapter

3 (section 3.3.2), this is true because T is the quantile of the counterfactual exposure distribution, i.e. the distribution of $X(z)$, where z is fixed. With continuous instruments, t_i's are obtained by estimating the Gaussian mixture model in (4.1) and consequently, t_i will be empirically correlated with z_i due to estimation error. Nevertheless, this correlation is likely to be small when the sample size is large enough and our Gaussian mixture model well approximates the counterfactual exposure distribution. In practice, we can always examine the empirical correlation between t_i and z_i before running nonparametric MR.

4.3 Simulation Study

4.3.1 Simulation design

We now present a simulation study that demonstrates the empirical performances of nonparametric MR with continuous IVs and compare it against doubly-ranked nonlinear MR (Tian et al., 2023) using fractional polynomials. Same as before, we implement doubly-ranked nonlinear MR using the SUMnlmr package (Mason and Burgess, 2022) with the default choice of 10 strata.

We adopt a simulation design similar to that in Chapter 3 except the instrument is generated to be continuous. Specifically, we consider the following four structural functions h:

(1). Null function: $h(x) = 0$;

(2). Cubic function: $h(x) = 0.05x^3$;

(3). Exponential function: $h(x) = 0.8\exp(x/4)$;

(4). Threshold function: $h(x) = \begin{cases} 0.5(x-15), & \text{if } x > 15 \\ 0, & \text{if } x \leq 15 \end{cases}$

and generate Y according to the structural equation:

$$Y_i = h(X_i) + U_i + e_{Y,i}$$

where $U_i \overset{i.i.d}{\sim} \mathbb{N}(0,1)$, $e_{Y,i} \overset{i.i.d}{\sim} \mathbb{N}(0, 0.1^2)$ and $e_{Y,i} \perp\!\!\!\perp (U_i, X_i)$..

We generate X according to the following three different structural models:

(i). *Constant genetic effect*:

$$X_i = \alpha_0 + \alpha_Z Z_i + \alpha_U U_i + e_{X,i}$$

$e_{X,i} \overset{i.i.d}{\sim} \mathbb{N}(0,1)$ and $e_{X,i} \perp\!\!\!\perp (U_i, Z_i, e_{Y,i})$

(ii). *Non-constant but rank-preserving genetic effect*:

$$X_i = \alpha_0 + \alpha_Z Z_i + \delta_Z Z_i^3 + \alpha_U U_i + e_{X,i}$$

Here the genetic effect is non-constant because of the nonlinear component Z_i^3 in the structural equation. But since this nonlinear function is deterministic, this model preserves the ranks of the counterfactual exposure values.

(ii). *Rank-varying genetic effect*:

$$X_i = \alpha_0 + \alpha_Z Z_i + \delta_Z Z_i^3 + \gamma_Z Z_i \xi_i + \alpha_U U_i + e_{X,i}$$

where $\xi_i \overset{i.i.d}{\sim} \mathbb{N}(0,1)$ and $\xi_i \perp\!\!\!\perp (U_i, Z_i, e_{Y,i}, e_{X_i})$. The rank-preserving assumption is violated because of the random component $\xi_i Z_i$ in the structural equation.

For all three instrument-exposure models above, we generate $Z_i \overset{i.i.d}{\sim} \text{Unif}(-2, 2)$.

For each instrument-exposure model, we also consider three different levels of instrument strength $\in \{5\%, 2.5\%, 1\%\}$, characterised by the proportion of variance in the exposure explained by the instruments and measured by R^2 in the exposure-instrument regression. Note that the strengths of the instrument we consider with continuous IVs are stronger than those

we consider with discrete IVs in Chapter 3, since a genetic risk score combines multiple genetic variants together and therefore is typically much stronger than a single variant.

Table 4.1 summarises the parameterisation of the data-generating models for all simulated scenarios.

Same as in Chapter 3, we generate $D = 100$ simulated datasets and for each simulated dataset, we set the sample size to be $N = 10,000$. We report the bias between $\hat{h}'$ and h' as the magnitude of bias averaged across the exposure distribution, given by

$$\text{Bias}(\hat{h}', h') = \frac{1}{m}\sum_{j=1}^{m}\left|\text{bias}\Big(\hat{h}'(x_j), h'(x_j)\Big)\right| \tag{4.8}$$

where we set $m = 100$ for evaluation of the bias, x_j's are out-of-sample values of X across the distribution of X and

$$\text{bias}\Big(\hat{h}'(x_j), h'(x_j)\Big) = \frac{1}{D}\sum_{d=1}^{D}\big(\hat{h}'_d(x_j) - h'(x_j)\big)$$

is the bias of $\hat{h}'(x_j)$ across the 100 simulated datasets. $\hat{h}'_d$ denotes the estimator of h' for the d-th round of the simulation, We also report the standard deviation of $\hat{h}'$ as standard deviation of the estimates averaged across the exposure distribution, given by

$$\text{SD}(\hat{h}') = \frac{1}{m}\sum_{j=1}^{m}\text{s.d.}\big(\hat{h}'(x_j)\big) \tag{4.9}$$

where

$$\text{s.d.}\big(\hat{h}'(x_j)\big) = \frac{1}{D-1}\sum_{d=1}^{D}\left(\hat{h}'_d(x_j) - \frac{1}{D}\sum_{d=1}^{D}\hat{h}'_d(x_j)\right)^2$$

is the standard deviation of $\hat{h}'(x_j)$ across the 100 simulated datasets. Note that equations (4.8) and (4.9) and exactly the same as equations (3.9) and (3.10) in Chapter 3. These equations are duplicated for readers' convenience.

For nonparametric MR, we treat QACE estimates as $\hat{h}'$ and for stratified nonlinear MR, we differentiate the estimated h obtained from the fractional polynomial regression. Because

QACE is estimated based on quantiles t of the counterfactual exposure distributions, for nonparametric MR we get back to the exposure values x by averaging across $\hat{X}(t,z)$ over the observed values of the instrument. I.e. we let

$$x_j = \frac{1}{n}\sum_{i=1}^{n}\hat{X}(t_j, z_i)$$

where t_j is a given quantile of the counterfactual exposure distribution. We also trim the tails of QACE estimates at 5% level for consistent comparison of the bias and the standard deviation. In addition to bias between $\hat{h}'$ and h', we also present bias between estimated QACE and the theoretical QACE for nonparametric MR, where the theoretical QACE is given in equation (4.7). I.e. We replace h' by the theoretical QACE in equation (4.8).

Data generating model for simulated scenarios

Instrument-exposure model	IV strength	α_0	α_Z	γ_Z	δ_Z	α_U
constant genetic effect	5%	15	0.55	0	0	2.5
constant genetic effect	2.5%	15	0.4	0	0	2.5
constant genetic effect	1%	15	0.25	0	0	2.5
non-constant but rank-preserving genetic effect	5%	15	0.3	0.1	0	2.5
non-constant but rank-preserving genetic effect	2.5%	15	0.15	0.1	0	2.5
non-constant but rank-preserving genetic effect	1%	15	0	0.1	0	2.5
rank-varying genetic effect	5%	15	0.3	0.1	1	2.5
rank-varying genetic effect	2.5%	15	0.15	0.1	1	2.5
rank-varying genetic effect	1%	15	0	0.1	1	2.5

Table 4.1: Model parameters for the data generating mechanisms for instrument-exposure relationships. Instrument strength is the proportion of variance in the exposure explained by the instrument, measured by R^2 in the exposure-instrument regression. α_0 are chosen such that the mean of the exposure distribution is kept about the same in each scenario.

4.3.2 Optimisation of the Gaussian mixture model

We perform optimisation of the Gaussian mixture model in **STEP 1** of nonparametric MR with continuous IVs using the "optimx" package (Nash and Varadhan, 2011; Nash, 2014) in

R, which automatically implements a range of common numerical algorithms for optimisation, including the Nelder-Mead method (e.g. Press and Flannery, 2007; Gill et al., 2020), BFGS (e.g. Nocedal and Wright, 1999) and others. We select the algorithm that gives the maximum value of the objective function and whose KKT (Karush-Kuhn-Tucker) conditions are true.

4.3.3 Hyperparameter tuning for nonparametric regressions

Same as in Chapter 3, we perform the nonparametric regressions in **STEP 3** of nonparametric MR using both kernel ridge regression (e.g. Scholkopf et al., 2001) and local linear kernel estimation (Li and Racine, 2007), and adopt the same hyperparameter tuning procedure discussed in section 3.4.2. That is, we perform cross-validation on held-out datasets to avoid overfitting. For kernel ridge regressions, we choose the Gaussian kernel and impose the additional constraint that we choose the same hyperparameters for both regressions of Y on t, z and X on t, z. We keep the hyperparameters fixed when calculating bias and standard deviation and when performing nonparametric bootstrap for confidence intervals. Recall we have discussed in section 3.4.2 that fixing the hyperparameters will make the confidence intervals overly-optimistic, as the uncertainties in hyperparameter selection will not be taken into account.

Note that nonparametric MR with continuous IVs is more computationally demanding than nonparametric MR with discrete IVs. The first reason is that we need to perform two-dimensional instead of one-dimensional nonparametric regressions, for which the computational time will notably increase especially when sample size is large. The second reason is that for each value of t, we need to calculate $\hat{Y}(t, z_i)$ and $\hat{X}(t, z_i)$ for all observed values of z_i in order to get our QACE estimates in **STEP 4**. I.e. we need to calculate nonparametric estimates $2n$ times. In contrast, with discrete IVs we only have to calculate nonparametric estimates $2K$ times, where K is the number of discrete values the instrument can take and does not increase as the sample size increases.

Because of the increased computational complexity of nonparametric MR with continuous IVs, we reduce the number of bootstrapped samples to be $R = 100$ for calculating bootstrapped confidence intervals. The cost of this is that the confidence intervals could become narrower (or wider) than expected, as we have fewer bootstrapped samples to aggregate from.

4.3.4 Simulation results

Similar to what we have presented in Chapter 3, we firstly visualise nonparametric MR estimates using both kernel ridge regression and local linear kernel estimation and doubly-ranked nonlinear MR estimates using fractional polynomials in Figures 4.1 and 4.2. For illustration purpose, we present the ideal case where we have constant genetic effect for the instrument-exposure model with 5% instrument strength. The pointwise 95% confidence intervals (the blue curves) are constructed using nonparametric bootstrap for nonparametric MR and using asymptotic standard errors from the fractional polynomial regression for doubly-ranked nonlinear MR. The true derivatives of the structural functions (the red curves) are also plotted as comparisons. Each figure is generated using one single dataset.

We note that for the null and the cubic functions (Figure 4.1), all three methods correctly estimate the true structural functions. For the exponential and the threshold functions (Figure 4.2), both nonparametric MR methods continue to give correct estimates while doubly-ranked nonlinear MR is no longer able to estimate the true shapes. These observations are in line with what we have observed in the case of discrete IVs in Chapter 3 (section 3.4.3) and demonstrate that nonparametric MR with continuous IVs also exhibits the flexibility of estimating complex exposure-outcome relationships beyond fractional polynomials. The discussion about the smoothness of these three approaches in section 3.4.3 also applies here with continuous IVs.

In addition, we observe that the confidence intervals for nonparametric MR using local linear kernel estimation are wider than those for nonparametric MR using kernel ridge regression.

This is particularly obvious with the null and the threshold functions, and likely shows the difficulties of using local linear kernel estimation for two-dimensional nonparametric regressions. However, we should also note that in subfigure (a) of Figure 4.2, the confidence intervals for nonparametric MR using kernel ridge regression slightly exclude the truth for a small part of the structural function. In contrast, with discrete IVs the confidence intervals for nonparametric MR fully cover the truth for all scenarios (except around the change point of the threshold function) as shown in section 3.4.3. This is likely due to the overly-optimistic confidence intervals by using only 100 bootstrapped samples as discussed in section 4.3.3.

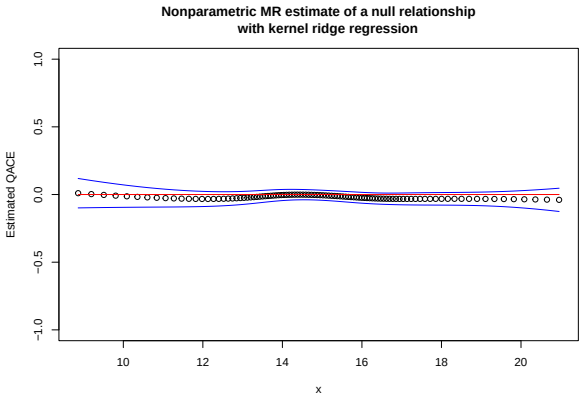
Tables 4.2 - 4.4 show average magnitude of bias and standard deviation of the estimates for all simulated scenarios. We firstly observe that doubly-ranked nonlinear MR outperforms nonparametric MR in estimating the null function. This is expected and is in line with what we observe in Chapter 3 (section 3.4.3), as the null is the simplest objective to estimate so a parametric approach is understandably more advantageous. For estimating the cubic, the exponential and the threshold functions, on the other hand, nonparametric MR generally outperforms doubly-ranked nonlinear MR. In particular, we note that nonparametric MR with kernel ridge regression consistently outperforms doubly-ranked nonlinear MR for estimating the exponential and the threshold functions in all simulated scenarios both in terms of bias and standard deviation. This confirms the advantages of using nonparametric MR for estimating complex structural functions, even when the rank-preserving assumption is violated. For estimating the cubic function, nonparametric MR with kernel ridge regression outperforms doubly-ranked nonlinear MR when the rank-preserving assumption holds, but admits bigger bias than doubly-ranked nonlinear MR when the genetic effect is rank-varying. This is understandable, as the cubic function belongs to the class of fractional polynomials. Nevertheless, nonparametric MR with kernel ridge regression renders smaller standard deviation of estimates than doubly-ranked nonlinear MR in all simulated scenarios for the cubic function, even when the rank-preserving assumption is violated.

Nonparametric MR with local linear kernel estimation also outperforms doubly-ranked non-

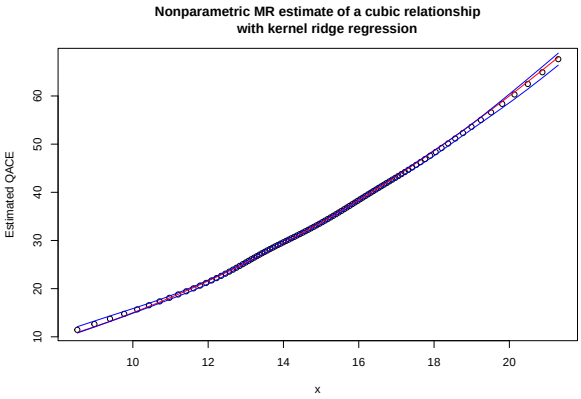
linear MR for estimating the exponential and the threshold functions in terms of bias, but it can result in larger standard deviations, especially when the instrument is weak. For estimating the cubic function, nonparametric MR with local linear kernel estimation outperforms doubly-ranked nonlinear MR in general, but can give bigger bias when the instrument strength is only 1%. Among the two nonparametric regression methods, local linear kernel estimation can render smaller bias than kernel ridge regression, especially for estimating the cubic and the exponential functions, but local linear kernel estimation results in estimates with bigger standard deviation than kernel ridge regression, particularly when the instrument is weak. As we discussed in Chapter 3 (section 3.4.3), this is likely due to the different degree of smoothing imposed by these two methods.

A relevant note is that when the rank-preserving assumption is violated, doubly-ranked nonlinear MR gives much smaller bias for the null and the cubic functions (i.e. when the structural function belongs to the class of fractional polynomials) with continuous IVs than that with discrete IVs if we compare Table 4.4 against Tables 3.5 and 3.8. While there might be other factors (such as difference in the exposure distributions due to different instrument distributions in the data generating mechanisms) involved, this more likely suggests that the doubly-ranked stratification method works better with continuous IVs than discrete IVs and is therefore more robust to the violation of the rank-preserving assumption.

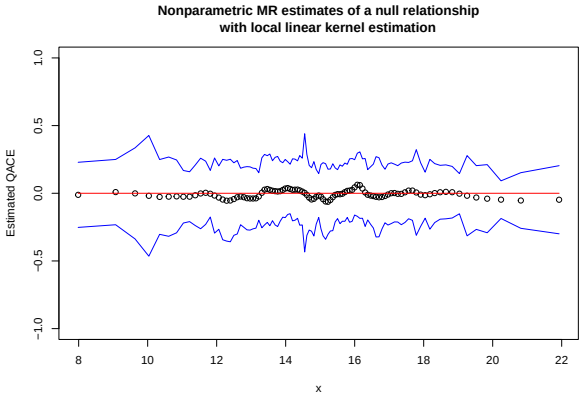
Tables 4.5 - 4.7 compare average magnitude of bias with respect to h' and average magnitude of bias with respect to theoretical QACE for nonparametric MR. Results for the null function are omitted as the derivative of the null function coincides with its theoretical QACE (both are null). As expected, the bias with respect to theoretical QACE are smaller than the bias with respect to h' for most of the simulated scenarios, especially for estimating the threshold function.



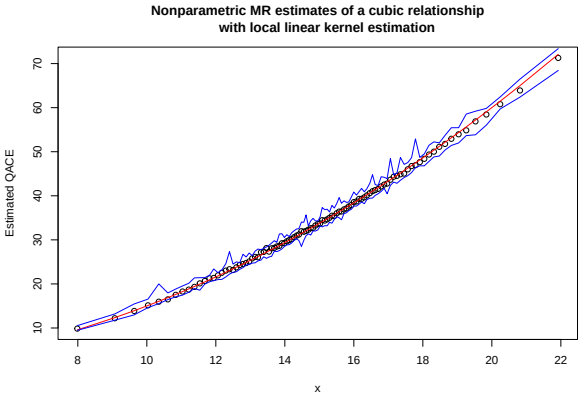
(a)



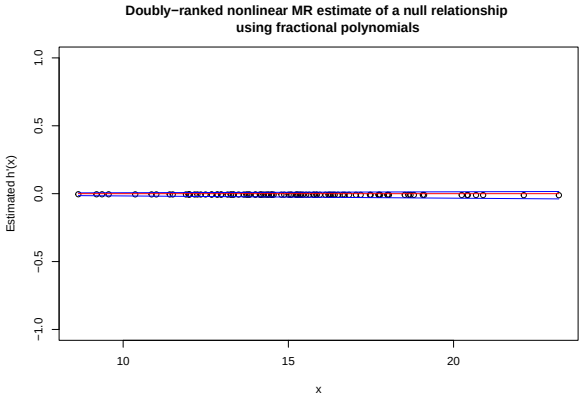
(d)



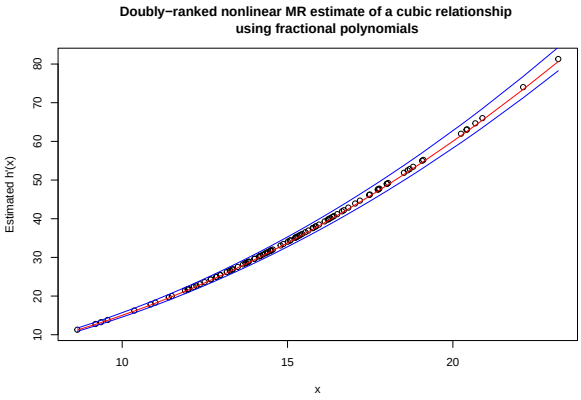
(b)



(e)

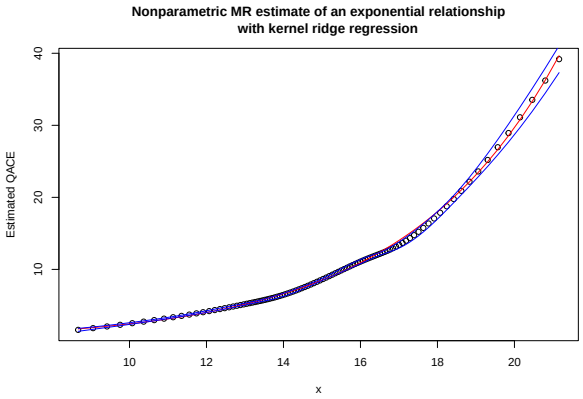


(c)

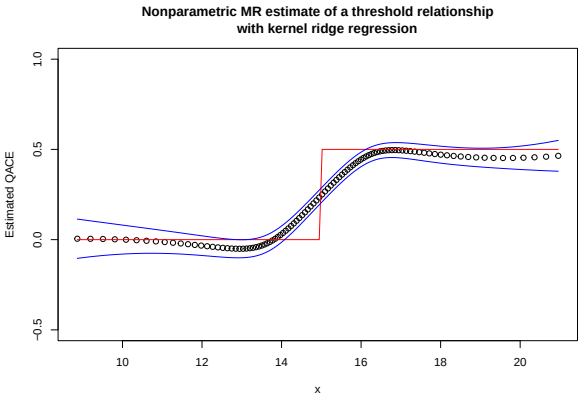


(f)

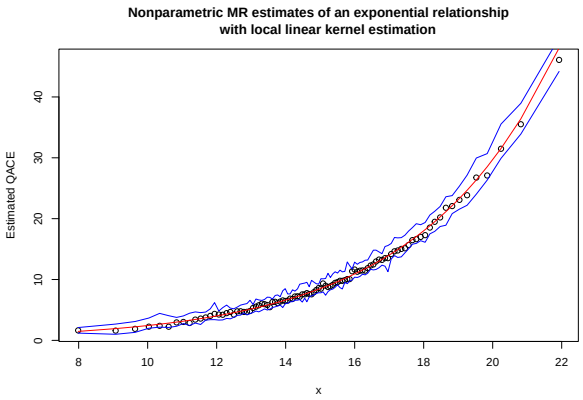
Figure 4.1: Nonparametric MR and stratified nonlinear MR estimates for null and cubic functions with a continuous instrument, constant genetic effect and 5% instrument strength. The black dots represent the estimates, the blue curves represent pointwise 95% confidence intervals and the red curve represents the true derivative of h.



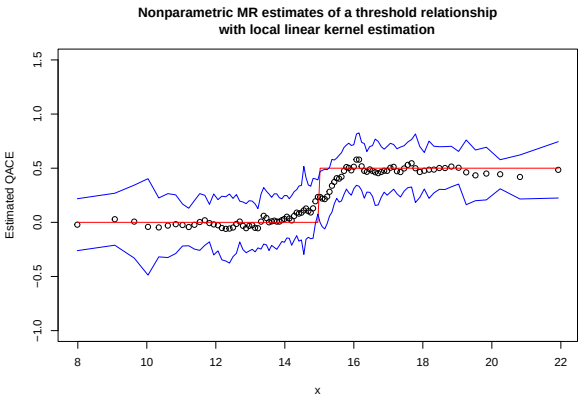
(a)



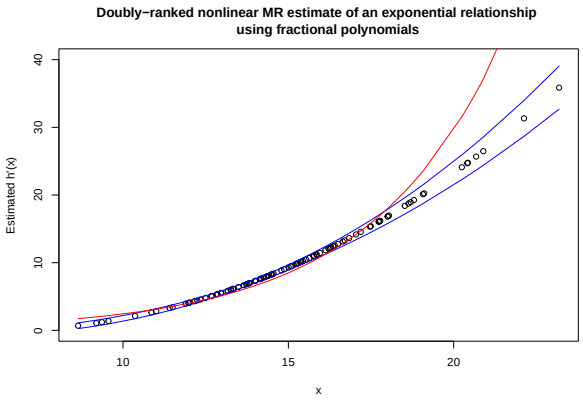
(d)



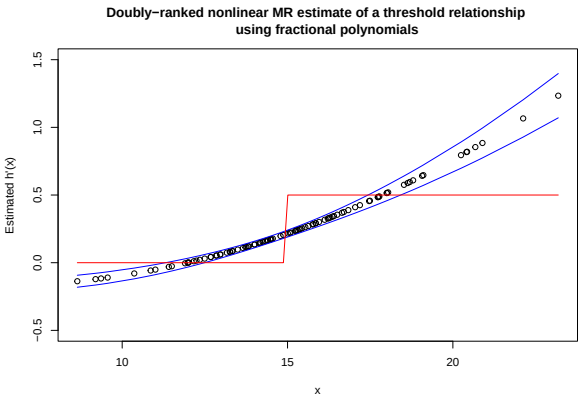
(b)



(e)



(c)



(f)

Figure 4.2: Nonparametric MR and stratified nonlinear MR estimates for exponential and threshold functions with a continuous instrument, constant genetic effect and 5% instrument strength. The black dots represent the estimates, the blue curves represent pointwise 95% confidence intervals and the red curve represents the true derivative of h.

Bias and standard deviation with constant genetic effect

		Nonparametric MR with kernel ridge regression		Nonparametric MR with local linear kernel estimation		Stratified nonlinear MR with doubly-ranked stratification	
Structural function h	IV strength	Bias	SD	Bias	SD	Bias	SD
null	5%	0.020	0.024	0.018	0.035	0.003	0.018
null	2.5%	0.014	0.034	0.022	0.053	0.004	0.025
null	1%	0.018	0.050	0.012	0.081	0.007	0.042
cubic	5%	0.169	0.220	0.192	0.451	0.228	2.009
cubic	2.5%	0.113	0.265	0.152	0.681	0.198	1.993
cubic	1%	0.103	0.354	0.175	1.121	0.160	2.002
exponential	5%	0.253	0.188	0.098	0.493	0.932	0.939
exponential	2.5%	0.269	0.288	0.106	0.677	0.906	0.931
exponential	1%	0.290	0.378	0.179	1.034	0.887	0.943
threshold	5%	0.056	0.025	0.048	0.043	0.101	0.043
threshold	2.5%	0.052	0.035	0.047	0.064	0.101	0.048
threshold	1%	0.054	0.050	0.027	0.104	0.101	0.063

Table 4.2: Magnitude of the bias and standard deviation of the estimated derivatives with a continuous instrument, constant genetic effect and instrument strength $\in \{5\%, 2.5\%, 1\%\}$.

Bias and standard deviation with nonconstant but rank-preserving genetic effect

		Nonparametric MR with kernel ridge regression		Nonparametric MR with local linear kernel estimation		Stratified nonlinear MR with doubly-ranked stratification	
Structural function h	IV strength	Bias	SD	Bias	SD	Bias	SD
null	5%	0.029	0.027	0.046	0.036	0.003	0.019
null	2.5%	0.029	0.034	0.052	0.053	0.004	0.026
null	1%	0.035	0.056	0.079	0.094	0.008	0.044
cubic	5%	0.186	0.227	0.160	0.436	0.235	2.022
cubic	2.5%	0.122	0.266	0.115	0.698	0.202	2.008
cubic	1%	0.207	0.366	0.147	0.967	0.156	2.023
exponential	5%	0.375	0.219	0.084	0.488	0.936	0.949
exponential	2.5%	0.303	0.270	0.089	0.581	0.909	0.942
exponential	1%	0.434	0.377	0.201	1.106	0.890	0.961
threshold	5%	0.060	0.027	0.067	0.045	0.101	0.044
threshold	2.5%	0.058	0.034	0.068	0.065	0.101	0.049
threshold	1%	0.063	0.056	0.090	0.115	0.102	0.066

Table 4.3: Magnitude of the bias and standard deviation of the estimated derivatives with a continuous instrument, nonconstant but rank-preserving genetic effect and instrument strength $\in \{5\%, 2.5\%, 1\%\}$.

Bias and standard deviation with rank-varying genetic effect

		Nonparametric MR with kernel ridge regression		Nonparametric MR with local linear kernel estimation		Stratified nonlinear MR with doubly-ranked stratification	
Structural function h	IV strength	Bias	SD	Bias	SD	Bias	SD
null	5%	0.031	0.030	0.047	0.045	0.002	0.020
null	2.5%	0.036	0.041	0.050	0.065	0.002	0.028
null	1%	0.045	0.072	0.065	0.125	0.005	0.047
cubic	5%	0.312	0.281	0.134	0.583	0.241	2.055
cubic	2.5%	0.229	0.332	0.154	1.027	0.220	2.042
cubic	1%	0.307	0.688	0.275	1.868	0.207	2.075
exponential	5%	0.779	0.325	0.131	0.832	1.031	1.033
exponential	2.5%	0.642	0.346	0.204	1.091	1.001	1.039
exponential	1%	0.787	0.975	0.664	3.630	0.978	1.097
threshold	5%	0.061	0.030	0.064	0.062	0.101	0.045
threshold	2.5%	0.062	0.041	0.064	0.090	0.102	0.051
threshold	1%	0.070	0.073	0.087	0.233	0.102	0.076

Table 4.4: Magnitude of the bias and standard deviation of the estimated derivatives with a continuous instrument, rank-varying genetic effect and instrument strength $\in \{5\%, 2.5\%, 1\%\}$.

Bias w.r.t QACE with constant genetic effect

		Nonparametric MR with kernel ridge regression		Nonparametric MR with local linear kernel estimation	
Structural function h	IV strength	w.r.t. h'	w.r.t. QACE	w.r.t. h'	w.r.t. QACE
cubic	5%	0.169	0.118	0.192	0.117
cubic	2.5%	0.113	0.087	0.152	0.127
cubic	1%	0.103	0.099	0.175	0.178
exponential	5%	0.253	0.251	0.098	0.062
exponential	2.5%	0.269	0.270	0.106	0.103
exponential	1%	0.290	0.290	0.685	0.683
threshold	5%	0.056	0.029	0.048	0.025
threshold	2.5%	0.052	0.032	0.047	0.031
threshold	1%	0.054	0.042	0.027	0.015

Table 4.5: Comparison of bias of nonparametric MR estimates with respect to the derivative of the structural function h and with respect to the theoretical QACE, with a continuous instrument, constant genetic effect and instrument strength $\in \{5\%, 2.5\%, 1\%\}$. Results for the null function are omitted.

Bias w.r.t QACE with nonconstant but rank-preserving genetic effect

		Nonparametric MR with kernel ridge regression		Nonparametric MR with local linear kernel estimation	
Structural function h	IV strength	w.r.t. h'	w.r.t. QACE	w.r.t. h'	w.r.t. QACE
cubic	5%	0.186	0.158	0.160	0.122
cubic	2.5%	0.122	0.104	0.115	0.113
cubic	1%	0.207	0.204	0.147	0.179
exponential	5%	0.375	0.375	0.084	0.070
exponential	2.5%	0.303	0.303	0.089	0.091
exponential	1%	0.434	0.434	0.201	0.201
threshold	5%	0.060	0.035	0.067	0.049
threshold	2.5%	0.058	0.040	0.068	0.056
threshold	1%	0.063	0.052	0.090	0.084

Table 4.6: Comparison of bias of nonparametric MR estimates with respect to the derivative of the structural function h and with respect to the theoretical QACE, with a continuous instrument, nonconstant but rank-preserving genetic effect and instrument strength $\in \{5\%, 2.5\%, 1\%\}$. Results for the null function are omitted.

Bias w.r.t. QACE with rank-varying genetic effect

		Nonparametric MR with kernel ridge regression		Nonparametric MR with local linear kernel estimation	
Structural function h	IV strength	w.r.t. h'	w.r.t. QACE	w.r.t. h'	w.r.t. QACE
cubic	5%	0.312	0.280	0.134	0.107
cubic	2.5%	0.229	0.209	0.154	0.155
cubic	1%	0.307	0.278	0.275	0.274
exponential	5%	0.779	0.731	0.130	0.116
exponential	2.5%	0.642	0.600	0.204	0.201
exponential	1%	0.787	0.740	0.664	0.652
threshold	5%	0.061	0.039	0.064	0.051
threshold	2.5%	0.062	0.047	0.064	0.055
threshold	1%	0.070	0.061	0.087	0.080

Table 4.7: Comparison of bias of nonparametric MR estimates with respect to the derivative of the structural function h and with respect to the theoretical QACE, with a continuous instrument, rank-varying genetic effect and instrument strength $\in \{5\%, 2.5\%, 1\%\}$. Results for the null function are omitted.

4.4 Applied Examples from UK Biobank

4.4.1 Description of the examples

We now illustrate nonparametric MR with continuous IVs using real data examples from UK Biobank. We consider the causal effect of two risk factors, including vitamin D concentration and alcohol intake, on systolic blood pressure (SBP) as the outcome of interest.

Vitamin D concentration is measured by the concentration of 25-hydroxy-vitamin D (25[OH]D), a commonly used clinical indicator of vitamin D status that is observationally associated with risks of cardiovascular diseases and all-cause mortality (Wang et al., 2008; Heath et al., 2019; Sofianopoulou et al., 2024). Because 25[OH]D concentration varies naturally across seasons, we remove the seasonal effect by regressing 25[OH]D concentrations on months. The de-seasonalised 25[OH]D concentration is then obtained by adding the residuals of this regression to a pre-chosen reference level of 25[OH]D concentration. This is to ensure that the rank-preserving assumption is not violated due to seasonal variability of vitamin D level. We take the genetic risk score created by Sofianopoulou et al. (2024) from 21 genetic variants that have been shown to be associated with 25[OH]D concentration at genome-wide significance level as the instrument.

Alcohol intake, measured by units per day (where 1 unit equals 8 grams of alcohol), is another important risk factor that has been shown to have positive effect on cardiovascular risks by various Mendelian randomisation analyses (Chen et al., 2008; Cho et al., 2015; Larsson et al., 2020). Tian et al. (2023) investigate the nonlinear causal effect of alcohol intake on SBP using the doubly-ranked stratification method, and discover that the positive effect exists only for individuals with high alcohol consumption. We take the genetic risk score constructed by Tian et al. (2023) from 93 genetic variants that have been shown to be associated with alcohol intake at genome-wide significance level (Liu et al., 2019) as the instrument.

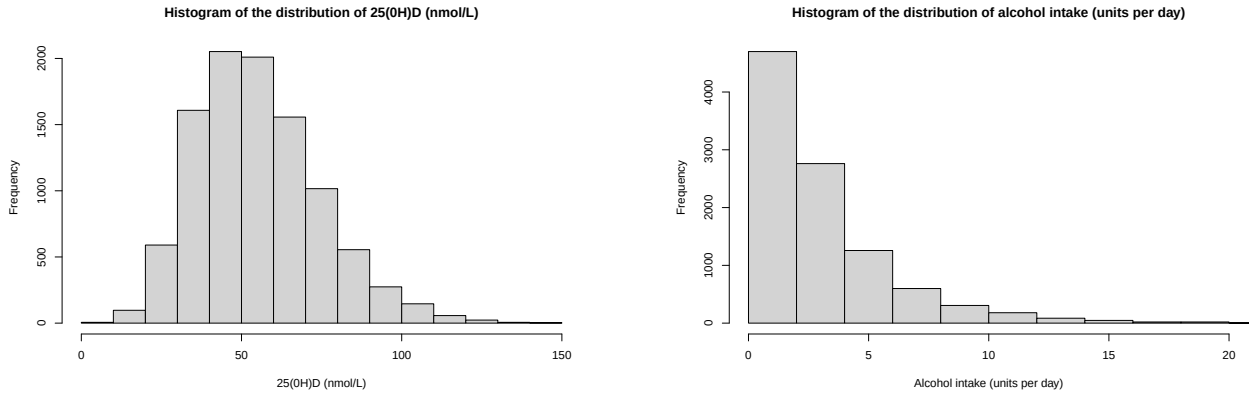
We drop the small number of missing individuals for complete-case analyses. Same as in Chapter 3 (section 3.5), for simplicity we do not consider any covariates. For the full UK Biobank sample, the standard MR estimate for the causal effect of 25[OH]D on SBP is null (the estimate is 0.00108 with standard error 0.00769); the standard MR estimate for the causal effect of alcohol intake on SBP is 0.740 (with standard error 0.126).

Also same as in section 3.5, we restrict our analyses to a randomly drawn subsample of size 10,000 for each of the exposure due to computational complexity of bootstrapping the nonparametric MR estimator. The strength of the instrument is about 4.56%, measured by R^2 of the exposure-instrument regression, for the subsample of 25[OH]D concentration and about 0.711% for the subsample of alcohol intake. The standard MR estimate for the subsample of 25[OH]D concentration is null (the estimate is 0.0398 with standard error 0.0455) and for the subsample of alcohol intake is 1.819 (with standard error 0.713). These estimates are consistent with the respective standard MR estimates on the full UK Biobank sample.

Figure 4.3 gives histograms that show the distribution of 25[OH]D concentration and alcohol intake in the subsamples. The distribution of 25[OH]D concentration is close to a normal distribution and therefore the Gaussian mixture model in **STEP 1** of nonparametric MR can be fitted reasonably well.

However, recall we have mentioned in Chapter 2 (section 2.6.2) that alcohol intake is an example of the so-called coarsened exposure, which essentially takes a number of discrete values (though this number is large) rather than being absolutely continuous. In addition, alcohol intake is truncated below at zero (see subfigure (b) in Figure 4.3), as we cannot have individuals taking negative alcohol consumption. This effectively means our assumption that the counterfactual exposure distributions can be modelled using Gaussian mixture distributions is violated. In particular, the nonparametric MR estimates at the lower tails of the exposure distributions are unreliable, as those estimates will correspond to negative

exposure values (as predicted from the lower quantiles of the counterfactual distributions of alcohol intake that are estimated from the Gaussian mixture model). Nevertheless, the nonparametric MR estimates will still be valid for the rest of the exposure distribution, as the large number of discrete values of alcohol measurements means that the distribution of the values of alcohol intake that are bounded away from zero can be well approximated by a Gaussian mixture distribution. We leave it for future research to consider more appropriate models for coarsened or truncated exposures (e.g. a truncated normal model might be more appropriate for truncated exposures).



(a) (b)

Figure 4.3: Histograms showing the distributions of 25[OH]D concentration (nmol/L) and alcohol intake (units per day) in the respective subsamples.

4.4.2 Results of the examples

Figure 4.4 presents nonparametric MR estimates with both kernel ridge regression and local linear kernel estimation and doubly-ranked nonlinear MR estimates. Same as before, for nonparametric MR, we present QACE estimates and for doubly-ranked nonlinear MR, we present the derivative of the estimated structural function from fractional polynomial regression. As mentioned above, for alcohol intake we trim QACE estimates at the lower tails of

the counterfactual exposure distributions. The pointwise 95% confidence intervals (the blue curves) are constructed using nonparametric bootstrap with $R = 100$ bootstrapped samples for nonparametric MR and using asymptotic standard errors of fractional polynomial regression for doubly-ranked nonlinear MR. The null effects (the red curves) are also plotted for reference.

For the causal effect of vitamin D level, all three methods produce clear null estimates (subfigures (a), (b) and (c) in Figure 4.4). This is consistent with the standard MR estimates for both the subsample and the full UK Biobank sample, and is also consistent with the stratum-specific estimates obtained from the doubly ranked stratification which, as shown in Table 4.8, are all null.

The confidence intervals for nonparametric MR estimates with kernel ridge regression (subfigure (a) in Figure 4.4) are wide for small values of 25[OH]D concentration, become much tighter for larger values of 25[OH]D concentration, and get slightly wide again for largest values of 25[OH]D concentration. For nonparametric MR with local linear kernel estimation (subfigure (b) in Figure 4.4), the confidence intervals are generally much wider but also follow a similar pattern that they are relatively tight in the middle and are wide at the two tails of the distribution of 25[OH]D concentration. This is most likely due to the fact that the strength of the genetic risk score varies across the distribution of 25[OH]D concentration. Similar to Chapter 3, we use F-statistics to measure the instrument strength within 100 strata defined by doubly-ranked stratification. As shown by subfigure (a) in Figure 4.5, the genetic instrument is the weakest at the lower tail and is the strongest in the middle of the distribution of 25[OH]D concentration.

For doubly-ranked nonlinear MR, the confidence intervals are wide for small values of 25[OH]D concentration and becomes increasingly narrow as the values of 25[OH]D concentration increases. This is observed for both fractional polynomial estimates (subfigure (c) in Figure 4.4) and LACE estimates (Table 4.8), and is similar to what we observe with nonparametric

MR with kernel ridge regression except the confidence intervals of doubly-ranked nonlinear MR estimates do not increase at the upper tail of the exposure distribution. This is possibly because we only choose 10 strata for performing doubly-ranked stratification, so that the drop in instrument strength for the extreme values of 25[OH]D concentration does not affect the strata-specific estimates.

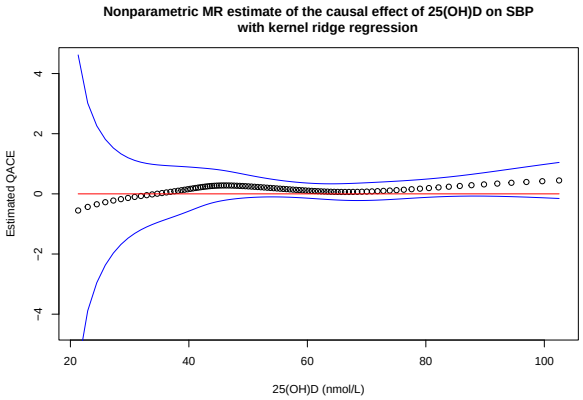
For the causal effect of alcohol intake, we see that nonparametric MR with kernel ridge regression (subfigure (d) in Figure 4.4) suggests an increasing causal effect that is significant for moderate values of alcohol intake (about 4 to 7 units per day). For small and large values of alcohol intake, the confidence intervals become exceedingly wide that the estimates are nonsignificant. This result is consistent with the positive causal effect estimated by standard MR in both the subsample and the full UK Biobank sample, and also in general agrees with the finding in Tian et al. (2023) that there is no significant causal effect for light drinkers but there is generally positive effect for moderate and heavy drinkers. The change in the width of the confidence intervals for nonparametric MR estimates with kernel ridge regression is again likely due to the varying instrument strength across the exposure distribution. As shown by subfigure (b) in Figure 4.5, the genetic instrument is very weak at the two tails and is much stronger towards the middle of the distribution of alcohol intake. This suggests that the nonsignificance of nonparametric MR estimates with kernel ridge regression for large values of alcohol intake is most likely due to lack of power in the subsample. We should also note that the scales of y-axis of subfigures (a) and (b) in Figure 4.5 are different: the genetic risk score for 25[OH]D concentration is much stronger than the genetic risk score for alcohol intake. This is consistent with the R^2 measurements on the overall subsamples as mentioned in section 4.4.1 (4.56% for 25[OH]D concentration and 0.711% for alcohol intake). This explains the difference in the widths of confidence intervals for nonparametric MR estimates for the causal effect of vitamin D level and that of alcohol intake.

The estimates from nonparametric MR with local linear kernel estimation (subfigure (e) in Figure 4.4) follow a similar increasing pattern, but they are all nonsignificant because of the

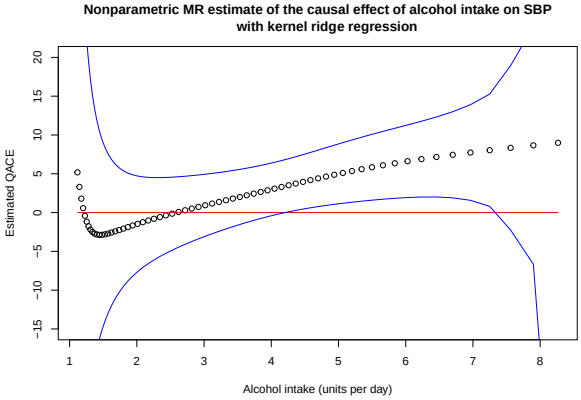
wide confidence intervals. Although the confidence intervals also follow a similar pattern as those for nonparametric MR with kernel ridge regression that they are very wide at the tails and are tighter at the middle of the distribution of alcohol intake, the confidence intervals are not strict enough to make the estimates informative. This is consistent with what we have observed in the simulation studies (and in Chapter 3 as well), and highlight that local linear kernel estimation is less efficient for implementing nonparametric MR.

The estimates from doubly-ranked nonlinear MR (subfigure (f) in Figure 4.4) also suggest an increasing and positive causal effect, and the confidence intervals excludes the null for the whole distribution of alcohol intake. However, if we examine the LACE estimates obtained from doubly-ranked stratification, as shown in Table 4.9, we see that the stratum-specific estimate is only significant in stratum 10. The standard errors are decreasing from stratum 1 to stratum 10, which contradicts to the confidence interval obtained from fractional polynomial regression whose width is the smallest for small values of alcohol intake and increases as the value of alcohol intake increases. This is another example, as we have discussed in Chapter 2 and 3, where fractional polynomial regression can result in misleading causal conclusions if we interpret the estimates locally.

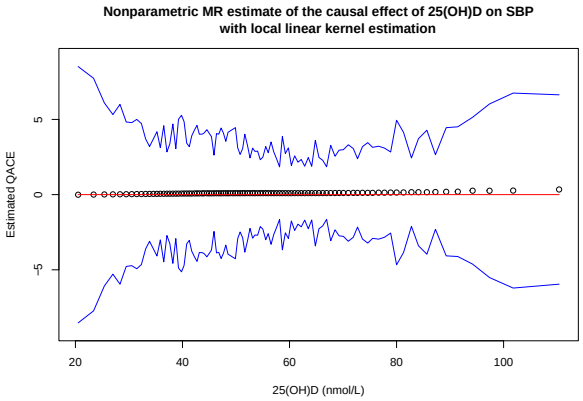
The LACE estimates from the doubly-ranked stratification in general agree with the nonparametric MR estimates with kernel ridge regression, but there are two discrepancies. The first one is that the LACE estimates do not show increasing standard errors towards the upper tails of the distribution of alcohol intake. Again, this is likely due to the fact that we only choose 10 strata, so the weak instrument for large values of alcohol intake does not kick in. The second difference is that the LACE estimates are not significant in the middle strata, which is most likely due to lack of power in the subsample (as we do see significant effects in the middle strata in the full UK Biobank sample, shown by Tian et al. (2023)).



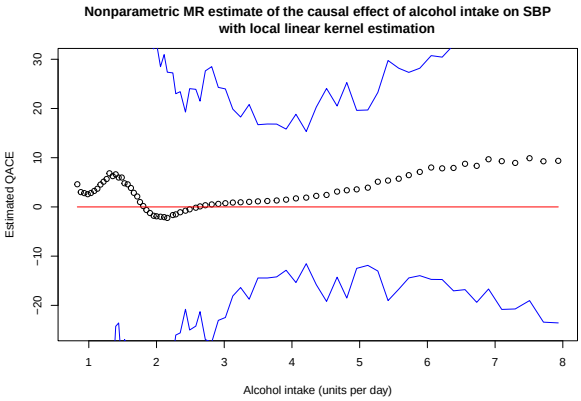
(a)



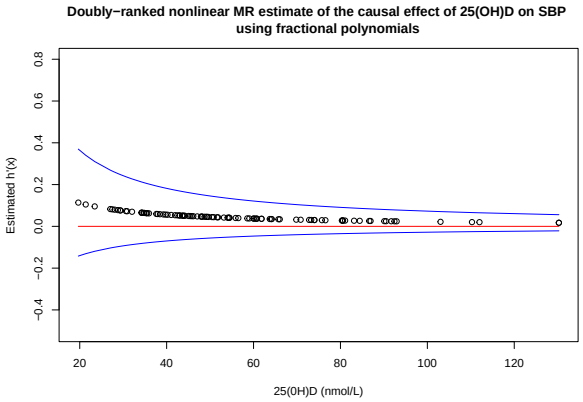
(d)



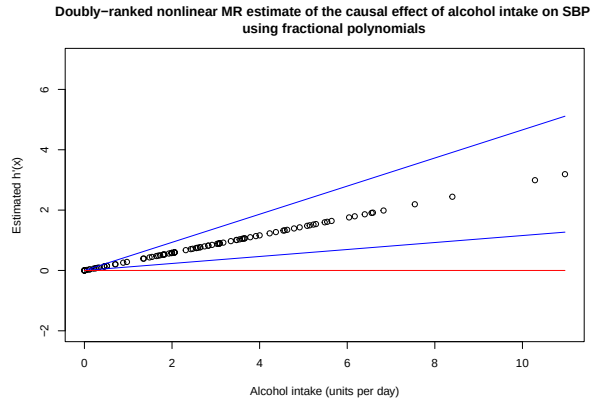
(b)



(e)

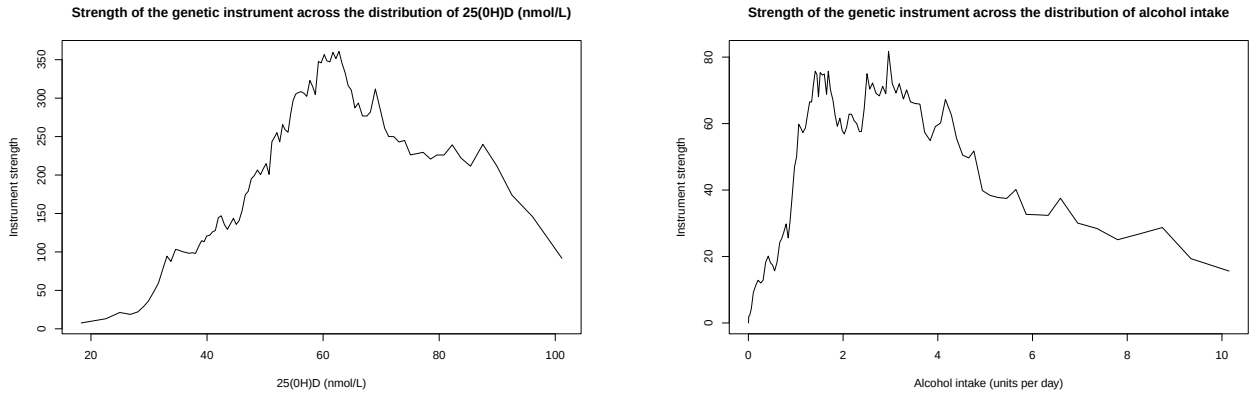


(c)



(f)

Figure 4.4: Nonparametric MR and stratified nonlinear MR estimates for the relationship between 25[OH]D and SBP and between alcohol intake and SBP. The black curve represents the estimates and the blue curves represent pointwise 95% confidence intervals. The red curve indicates a null effect.



(a) (b)

Figure 4.5: Strength of the genetic risk scores across the distribution of 25[OH]D concentration and alcohol intake in their respective subsamples. The instrument strength is measured by the F-statistics of the instrument-exposure regressions within 100 strata defined by doubly-ranked stratification.

LACE estimates for the causal relationship between 25(OH)D concentration and systolic blood pressure (SBP) in the subsample

Stratum	Stratum-specific estimate	Standard error	95% confidence interval
1	0.117	0.355	[−0.578, 0.812]
2	−9.99E-04	0.255	[−0.500, 0.498]
3	0.0603	0.212	[−0.356, 0.476]
4	0.0801	0.193	[−0.297, 0.458]
5	0.0703	0.159	[−0.242, 0.383]
6	0.0132	0.134	[−0.250, 0.276]
7	0.127	0.122	[−0.112, 0.366]
8	−0.0745	0.112	[−0.294, 0.145]
9	0.0320	0.102	[−0.167, 0.231]
10	0.0441	0.0813	[−0.115, 0.204]

Table 4.8: Stratum-specific (LACE) estimates using the doubly-ranked stratification for estimating the relationship between 25(OH)D concentration (nmol/L) and systolic blood pressure (SBP) in the randomly drawn subsample.

LACE estimates for the causal relationship between alcohol intake and systolic blood pressure (SBP) in the subsample

Stratum	Stratum-specific estimate	Standard error	95% confidence interval
1	11.0	11.24	[−11.0, 33.1]
2	5.16	6.44	[−7.46, 17.8]
3	−3.91	4.29	[−12.3, 4.50]
4	0.710	2.94	[−5.05, 6.47]
5	0.598	2.83	[−4.95, 6.14]
6	2.00	2.23	[−2.36, 6.36]
7	0.965	2.13	[−3.21, 5.14]
8	2.29	1.51	[−0.683, 5.25]
9	2.61	1.45	[−0.237, 5.47]
10	2.09	0.911	[0.305, 3.88]

Table 4.9: Stratum-specific (LACE) estimates using doubly-ranked stratification for estimating the relationship between alcohol intake (units per day) and systolic blood pressure (SBP) in the randomly drawn subsample.

4.5 Discussion and Concluding Remarks

In this chapter, we develop an important extension to the nonparametric MR framework introduced in Chapter 3 to situations where instrument is continuous. The quantiles t of the counterfactual exposure distributions are obtained using Gaussian mixture model that can be solved via standard numerical algorithms. The quantile average causal effect (QACE) is then estimated by combining two-dimensional nonparametric regression estimates in a way that can be viewed as quantile-specific analogue of the usual Wald estimator.

By comparing against doubly-ranked nonlinear MR in a wide range of simulated scenarios, we demonstrate in section 4.3 that our nonparametric MR method with continuous IVs inherits the advantages of nonparametric MR with discrete IVs in the sense that it is able to flexibly characterise complex nonlinear structural functions beyond fractional polynomials without

the need to pre-specify the number or size of the strata, and the constant genetic effect assumption is not required for our method to work. In particular, both nonparametric MR methods consistently outperform doubly-ranked nonlinear MR for estimating functions that are not fractional polynomials in terms of bias, and nonparametric MR with kernel ridge regression wins additionally in terms of standard deviation of the estimates.

In section 4.4, we apply our method to real datasets from UK Biobank to investigate the causal effects of vitamin D level and alcohol intake on systolic blood pressure. The results from nonparametric MR with kernel ridge regression are consistent with standard MR estimates and in general agree with LACE estimates obtained from doubly-ranked stratification. Even for alcohol intake, a coarsened exposure truncated at zero so that the Gaussian mixture model is misspecified for exposure values near zero, nonparametric MR with kernel ridge regression still provides informative estimates for values of alcohol intake around 4 to 7 units per day. Nonparametric MR with local linear kernel estimation, however, results in wide confidence intervals and nonsignificant causal effect estimates for the two example considered. Doubly-ranked nonlinear MR with fractional polynomial regression, on the other hand, gives confidence intervals that are different from the confidence intervals of the corresponding LACE estimates for alcohol intake. This is a general disadvantage of using fractional polynomials regression to estimate the structural function that we have mentioned in previous chapters, and shows the advantage of using nonparametric MR with kernel ridge regression in practice. We leave it for future research to apply nonparametric MR to full UK Biobank datasets.

There are also a number of limitations with our method. The obvious one is that Gaussian mixture model might not be suitable for discrete or truncated exposures, such as alcohol intake considered in this chapter. Although nonparametric MR is still valid for exposure values that are bounded away from truncation, it is a disadvantage that we cannot reliably estimate causal effects of exposure values near truncation. It will be an important direction for future research to develop other models that are able to estimate quantiles over the whole

counterfactual distributions of the truncated exposure.

Another disadvantage, as mentioned in Chapter 3, is that we do not consider the adjustment for covariates for nonparametric MR. While the estimates will remain valid as long as we have valid instruments, not adjusting for baseline covariates can lead to inefficient estimates. One option is to pre-adjust the outcome by regressing the outcome on the covariates and choosing the residuals of this regression as the adjusted outcome. This idea is similar to the famous partitioned regression method, but it apparently assumes that the covariates affects the outcome linearly and cannot handle complex models such as interactions between the covariates and the exposure. How to handle covariates will be an interesting direction for future research.

One further limitation is the choice of nonparametric regression methods and their hyperparameter tuning procedures. In this chapter, we adopt the same nonparametric regression methods and hyperparameter tuning procedures that we use for nonparametric MR with discrete IVs, and we have discussed the potential limitations in Chapter 3. We already observe that local linear kernel estimation is not the best choice, and we leave it for future research to consider other nonparametric regression and relevant hyperparameter tuning methods.

Another relevant limitation is the computational complexity for nonparametric MR. Again, we have discussed this limitation in Chapter 3, but we have noted that nonparametric MR with continuous IVs is more computationally demanding than nonparametric MR with discrete IVs because of the need of two-dimensional nonparametric regressions and the necessity of calculating $2n$ nonparametric estimates where n is the sample size. Implementing the algorithms using languages much faster than R seems to be the most straightforward solution, but it will also be an important direction for future research to speed up the computation algorithmically.

Chapter 5

Investigating Selection Bias in Linear and Stratified Non-linear Mendelian Randomisation Analyses

5.1 Introduction

Selection bias is one of the major sources of bias in a statistical analysis. It can arise due to many different reasons, such as missing data, loss to follow-up, participants self-selection (also known as volunteering bias or participation bias) and the nature of certain statistical design like a case-control study (Hernán et al., 2004; Evans and Didelez, 2015). It is common to illustrate selection bias using DAGs, where selection can be conveniently represented as conditioning on a collider, an example of the so-called "collider-stratification bias" (Greenland, 2003; Hernán et al., 2004).

In recent years, selection bias has been increasingly discussed in epidemiological, medical and genetic studies using large-scale biobanks, which are powerful and easily accessible data

source but have been demonstrated to suffer extensively from volunteering bias (e.g. Banks et al., 2012; Keyes and Westreich, 2019; van Alten et al., 2024). For example, studies have shown that UK biobank is not representative of the UK population: among the 9.2 million individuals who were invited to enter the UK Biobank cohort, only about 5.5% participated in the study, and the participation rate was higher among women, older age groups and people with higher wealth and higher level of education (Fry et al., 2017; Munafo et al., 2018; Schoeler et al., 2023).

Many studies using these large-scale biobanks involve MR analyses. When a MR study is subject to selection, typically either no unmeasured confounding or exclusion restriction (or both) assumption is violated, and therefore the analysis is biased. Note, however, that selection bias is different from non-representativeness. Non-representativeness occurs when we focus our analyses on a specific target population instead of the full population. For example, restricting the analyses to women will only give rise to non-representativeness but not selection bias. A similar but different example that could give rise to selection bias is when women are more likely to enter the study than men, often in a way that the researchers cannot control.

An abundance of literatures have examined selection bias in the context of a standard IV or MR analysis, and many of them propose adjusting for covariates or inverse probability weighting (IPW) as solutions to reduce or eliminate selection bias. On the one hand, while covariate adjustment may help reduce bias in some situations (Burgess et al., 2024), the usefulness of this approach is doubtful in the presence of unmeasured confounders (which is exactly the reason why we use MR). On the other hand, IPW, which aims to inversely weight the selected sample by the probability of selection to get a pseudo-population that is representative of the true population (e.g. Hernán and Robins, 2006; Seaman and White, 2013), seems to be a more promising approach. The usefulness of IPW depends on whether our data is missing at random or not. The missing data (i.e. the unselected data) is said to be missing at random (MAR) if the probabilities of missingness depend only on observed

data (e.g. Rubin, 1976; Schafer and Graham, 2002). Otherwise, the data is said to be missing not at random (MNAR). While it is well-known that IPW will be biased when the selection mechanism depends on unmeasured variables, i.e. when the data is MNAR (e.g. Robins et al., 2000; Gkatzionis and Burgess, 2019; Metten et al., 2022), many literatures claim that IPW can eliminate selection bias when the data is MAR and the selection model is correctly specified (e.g. Canan et al., 2017; Gkatzionis and Burgess, 2019; Hughes et al., 2019). There are also lots of efforts in practice that aims to create participation weights for UK Biobank and then use IPW to estimate causal effects (e.g. Schoeler et al., 2023; van Alten et al., 2024). However, exactly how to perform IPW in the context of an IV or a MR analysis is often unclear in existing literatures.

In this chapter, we make an explicit account of inverse probability weighted MR estimator and its asymptotic distribution. We note that the standard errors used in many literatures and existing softwares are not suitable for inverse probability weighted MR, and we provide instead a robust variance estimator that delivers correct coverage under regularity conditions. More importantly, we observe that the performance of IPW depends crucially on the degree of selection. Through simulation studies and real data examples from UK Biobank, we demonstrate that even in the ideal situation where the data is MAR and the selection model is correctly specified, inverse probability weighted MR is not guaranteed to reduce bias when the selection mechanism depends strongly on observed data and when the selection probability is small. Our findings have meaningful implications because the participation probabilities of UK Biobank and many other large-scale biobanks are usually very low.

In MR literatures, there are also many efforts that attempt to go beyond standard linear MR to estimate more complicated versions of causal effect. As we have discussed in Chapter 2, stratified nonlinear MR is one of the widely used MR approaches to estimate nonlinear causal effects. However, we have also discussed some recent findings suggesting that some datasets analysed by either residual or doubly-ranked stratification are likely to suffer from selection bias (Wade et al., 2023; Hamilton et al., 2024). Nevertheless, there has been no

systematic investigations on selection bias in the context of stratified nonlinear MR (or any other nonlinear IV/MR methods).

In this chapter, we also explore how selection bias affects stratum-specific estimates obtained by stratified nonlinear MR. In particular, we observe that, unlike in standard MR case, IPW is unable to eliminate selection bias for stratum-specific estimates even if the selection mechanism on the overall population is fully known, the selection model is correctly specified and the selection effect is only moderate. We argue that this is because while IPW is able to create representative pseudo-population for the overall population, it fails to do so within each stratum. In order to eliminate selection bias, we need to put selected individuals in strata that would have been formed without selection and then perform inverse-probability-weighted MR in each stratum. We also demonstrate our arguments through simulation studies.

Note that both inverse probability weighted MR and stratified nonlinear MR require individual-level data (recall we assume throughout this thesis that we have access to individual-level data). We have also mentioned in Chapter 1 (section 1.3.3) that, with individual-level data, standard MR is equivalent to two-stage-least-square (2SLS). Therefore, in the rest of this chapter we shall use MR and 2SLS interchangeably.

5.2 Related Works

Before we move on, it is worth mentioning a few other related works that explore selection bias, or more generally collider bias in the context of MR. Gkatzionis et al. (2024) propose an "instrument for selection" model that can be used to mitigate selection bias when the data is MNAR. This approach is motivated by the Heckman's selection model, in which a variable is called an instrument for selection if it is associated with the selection indicator and is independent of the outcome (Heckman, 1979; Gkatzionis et al., 2024). If such a

variable exists and is observed for the whole population, the instrument for selection model can be applied to reduce selection bias when the selection model cannot be fully specified or estimated. See Figure 5.1 for a graphical illustration of this approach.

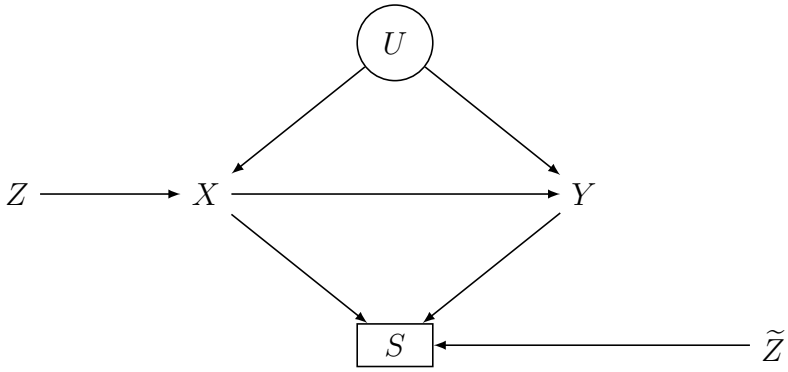


Figure 5.1: DAG for the instrument for selection model (Gkatzionis et al., 2024). Z denotes the instruments for the exposures, X denotes the exposures, Y denotes the outcome, U denotes unmeasured confounders, S denotes the selection indicator and $\widetilde{Z}$ denotes the instrument for selection.

As we mentioned previously, selection bias is only one example of collider bias. There are other types of collider bias in the context of GWAS and MR. One interesting example occurs in mediation analysis, where researchers are typically interested in the so-called natural direct effect or controlled direct effect (e.g. VanderWeele, 2013) that requires conditioning or stratifying on a mediating variable that happens to be a collider. Figure 5.2 is a graphical illustration of this type of problems (Cai et al., 2022). A practical example is the studies of disease progression, where X is an indicator of having the disease, Y is disease progression (such as recovery from COVID) and G is the exposure or SNP of interest. People are interested in the direct causal effect of G on Y, but this effect can only be estimated among those who have the disease, i.e., conditioning on X. This opens up the path $G \longrightarrow X \longleftarrow U \longrightarrow Y$, which is the source of collider bias.

Dudbridge et al. (2019) observe that the collider bias is approximately linear on top of the effect of G on X, and provide a bias-correction formula assuming linear structural equation

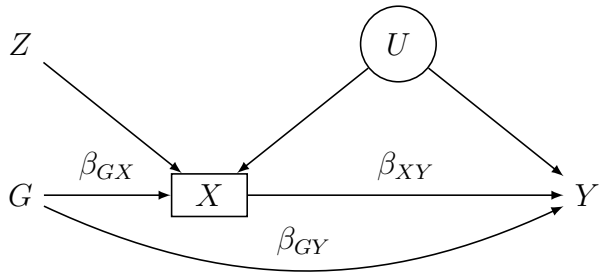


Figure 5.2: DAG for the mediation model (Cai et al., 2022). Z denotes the instrument, G denotes the exposure or SNP of interest, X denotes the mediator, Y denotes the outcome and U denotes unmeasured confounders. β_{GX}, β_{GY}, β_{XY} are coefficients assuming linear SEMs.

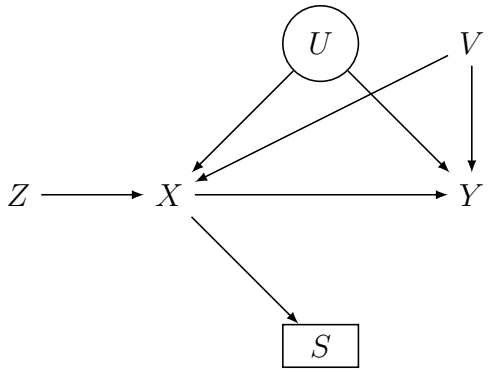
models:

$$\beta_{GY} = \tilde{\beta}_{GY} - b\beta_{GX}$$

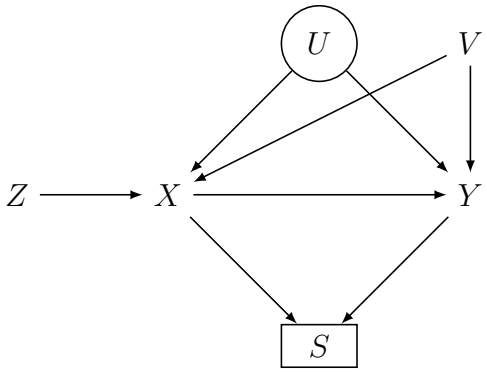
where $\tilde{\beta}_{GY}$ is the total effect of G on Y conditioning on X. b can be estimated using standard MR provided we have access to a valid set of instruments Z. Mahmoud et al. (2022) give an algorithm called "slope-hunter" that identifies valid SNPs as IVs.

5.3 Selection Mechanisms

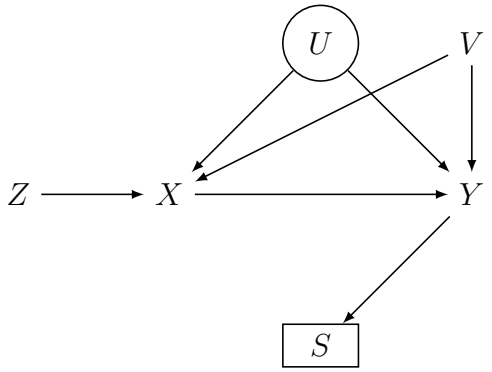
There are many ways one can think of in which selection affects an IV or a MR analysis. For example, Hughes et al. (2019) list nine common selection mechanisms (and seven more in their appendix) for an IV model. Figure 5.3 illustrates four different mechanisms of selection that will lead to selection bias in general situations, and these will be the selection mechanisms that we consider in this chapter. Note that we assume the data is MAR (i.e. the selection indicator S does not depend on the unmeasured confounders U except through observed variables X, Y or V). Although these are by no means exhaustive, our arguments are easily generalisable to many other situations as long as the data is also MAR (e.g. four other different but similar selection mechanisms are listed in Figure A.1 in the Appendix).



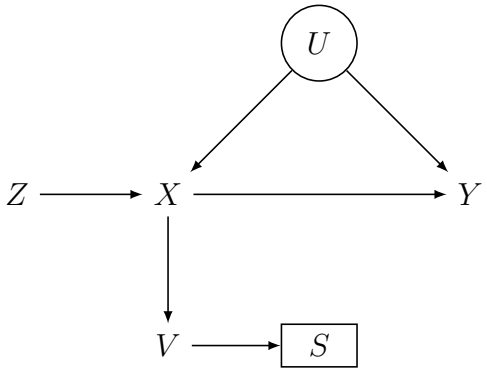
(A) selection according to the exposure X



(B) selection according to both the exposure X and the outcome Y



(C) selection according to the outcome Y



(D) selection according to a downstream covariate V of the exposure X

Figure 5.3: DAGs illustrating four different selection mechanisms. Z denotes the instruments, X denotes the exposures, Y denotes the outcome, V denotes measured covariates, U denotes unmeasured confounders and S denotes the selection indicator.

Basic graph theories can help us understand where selection bias comes from in each of the selection mechanisms. In lines with the observations made by Hughes et al. (2019), in scenario A, selection bias comes from the fact that conditioning on S opens up the path $Z \longrightarrow X \longleftarrow U \longrightarrow Y$, violating the no unmeasured confounding for the instrument assumption. In scenario B, conditioning on S opens up path $Z \longrightarrow X \longleftarrow U \longrightarrow Y$, path $Z \longrightarrow X \longrightarrow Y \longleftarrow U$ and path $Z \longrightarrow X \longrightarrow S \longleftarrow Y$, thus violating both the no unmeasured confounding for the instrument and the exclusion restriction assumptions. In scenario C,

conditioning on S opens up path $Z \longrightarrow X \longleftarrow U \longrightarrow Y$ and path $Z \longrightarrow X \longrightarrow Y \longleftarrow U$, also violating the no unmeasured confounding for the instrument assumption. However, note that in this scenario, selection bias only exists if X actually causes Y. If the causal effect is null, Y will not be a collider so no collider bias will be induced. Scenario D is similar to A, where conditioning on S induces correlation between Z and U.

It is straightforward to note that adjusting for measured covariates V will reduce selection bias in scenarios A, B and C, because V is similar to U and adjusting for V will block paths containing V. Nevertheless, bias cannot be eliminated because of the existence of unmeasured confounders U. Moreover, adjusting for V will not help in scenario D because V itself is a (child of) collider. In fact, Burgess et al. (2024) show through their simulation studies (in their supplementary materials) that adjusting for downstream covariates of the exposure will make the selection bias worse. This is in line with the general observation that colliders and children of colliders should not be adjusted for.

5.4 Inverse Probability Weighted MR

5.4.1 Methodology

In this section, we will continue to assume linear structural equation models as in section 5.3, except the account will be general for data with higher dimensions. Therefore, according to the DAGs in Figure 5.3, the structural model for scenario A, B and C is given by

$$\begin{aligned} X_i &= \alpha_0 + \alpha_Z^T Z_i + \alpha_U^T U_i + \alpha_V^T V_i + e_{X,i} \\ &= \alpha^T Z_i + \epsilon_{X,i} \end{aligned}$$

$$\begin{aligned} Y_i &= \beta_0 + \beta_X^T X_i + \beta_U^T U_i + \beta_V^T V_i + e_{Y,i} \\ &= \beta^T X_i + \epsilon_{Y,i} \end{aligned}$$

where $\alpha = (\alpha_0, \alpha_Z, \alpha_V)^T$, $\beta = (\beta_0, \beta_X, \beta_V)^T$, $\epsilon_{X,i} = \alpha_U^T U_i + e_{X,i}$, $\epsilon_{Y,i} = \beta_U^T U_i + e_{Y,i}$, $e_{X,i}$, $e_{Y,i}$ are random noises that are mutually independent with zero mean and finite variance, $e_{X,i} \perp\!\!\!\perp (Z_i, U_i, V_i)$ and $e_{Y,i} \perp\!\!\!\perp (X_i, Z_i, U_i, V_i, e_{X,i})$. With slight abuse of notations, we use X_i (and similarly Z_i) to denote both the exposure and the design vector that includes the exposure, the measured covariates V_i and the intercept.

For scenario D, the structural model is given by

$$\begin{aligned} X_i &= \alpha_0 + \alpha_Z^T Z_i + \alpha_U^T U_i + e_{X,i} \\ &= \alpha^T Z_i + \epsilon_{X,i} \end{aligned}$$

$$\begin{aligned} Y_i &= \beta_0 + \beta_X^T X_i + \beta_U^T U_i + e_{Y,i} \\ &= \beta^T X_i + \epsilon_{Y,i} \end{aligned}$$

where $\alpha = (\alpha_0, \alpha_Z)^T$, $\beta = (\beta_0, \beta_X)^T$, $\epsilon_{X,i} = \alpha_U^T U_i + e_{X,i}$, $\epsilon_{Y,i} = \beta_U^T U_i + e_{Y,i}$, $e_{X,i}$, $e_{Y,i}$ are random noises that are mutually independent with zero mean and finite variance, $e_{X,i} \perp\!\!\!\perp (Z_i, U_i)$ and $e_{Y,i} \perp\!\!\!\perp (X_i, Z_i, U_i, e_{X,i})$. Similar to above, with slight abuse of notations, we use X_i (and similarly Z_i) to denote both the exposure and the design vector that includes the exposure and the intercept. In addition, we also have

$$V_i = \gamma_0 + \gamma_X^T X_i + e_{V_i}$$

where $e_{V,i}$ are random noises that are mutually independent with zero mean and finite variance and $e_{V,i} \perp\!\!\!\perp (X_i, Z_i, U_i, e_{X_i}, e_{Y,i})$.

We make the usual rank and IV exogeneity assumptions (Assumptions 3 and 4 as defined in Chapter 1 (section 1.2.2)) for instruments Z_i so that we have valid instruments. These assumptions are also evident from the DAGs in Figure 5.3.

For the ease of notation, we will also use Ψ to denote the set of variables that determines selection. Specifically, in scenario A, $\Psi = X$; in scenario B. $\Psi = (Y, X^T)^T$; in scenario C, $\Psi = Y$; and in scenario D, $\Psi = V$.

The inverse probability weighted two-stage-least-square (IPW-2SLS) involves weighting both stages by the inverse of the probability of getting selected into the sample. When the model is just-identified (i.e. the number of instruments equals to the number of exposures), the inverse probability weighted 2SLS estimator is given by

$$\hat{\beta}_{IPW-2SLS} = \left(\sum_{s_i=1} z_i w_i x_i^T \right)^{-1} \sum_{s_i=1} z_i w_i y_i \tag{5.1}$$

where $W_i = 1/\mathbb{P}\big[S_i = 1 \,|\Psi_i\big]$ is the inverse probability weight for individual i.

The asymptotics of $\hat{\beta}_{IPW-2SLS}$ are easily generalisable to cases where the model is over-identified (i.e. the number of instruments exceeds the number of exposures). Therefore, for the ease of notation we focus on the just-identified model. We propose the following theorem that provides theoretical guarantees on when we can expect IPW to eliminate selection bias in the context of IV analyses.

Theorem 8 (Asymptotics of inverse probability weighted 2SLS estimator)**.** *Assume the data is missing at random and the inverse probability weights are known variables. Under regularity conditions and the positivity assumption that*

$$\exists\, c > 0 \;\; \textit{such that} \;\; \forall \psi \;\; \textit{where} \;\; \mathbb{P}(\Psi = \psi) > 0,\; \mathbb{P}\big[S = 1 \,|\Psi = \psi\big] > c$$

where S denotes the selection indicator and Ψ denotes the set of variables that determine selection, we have

$$\sqrt{n}(\hat{\beta}_{IPW-2SLS} - \beta) \xrightarrow{d} \mathbb{N}\big(0,\; \Sigma_{ZWX}\, \Omega\, \Sigma_{ZWX}^T\big) \tag{5.2}$$

where $\Sigma_{ZWX} = \mathbb{E}\left[ZWX^T\, \mathbb{I}(S=1)\right]^{-1}$ *and* $\Omega = \mathbb{E}\left[\epsilon_Y^2 W^2 Z Z^T\, \mathbb{I}(S=1)\right]$.

Proof. We first show how we can derive the formula given in equation (5.1). Using standard weighted regression formula, we have the first stage estimator given by

$$\hat{\alpha} = \left(\sum_{s_i=1} z_i w_i z_i^T \right)^{-1} \sum_{s_i=1} z_i w_i x_i^T$$

The weighted 2SLS estimator then involves regressing (with weights) the outcome on the instrument-predicted exposure in the second stage:

$$
\begin{aligned}
\hat{\beta}_{IPW-2SLS} &= \Big(\sum_{s_i=1}(\hat{\alpha}^T z_i) w_i (\hat{\alpha}^T z_i)^T\Big)^{-1} \sum_{s_i=1}(\hat{\alpha}^T z_i) w_i y_i \\
&= \Big(\sum_{s_i=1}\big(\sum_{s_j=1} x_j w_j z_j^T\big)\big(\sum_{s_j=1} z_j w_j z_j^T\big)^{-1} z_i w_i z_i^T \big(\sum_{s_j=1} z_j w_j z_j^T\big)^{-1}\big(\sum_{s_j=1} z_j w_j x_j^T\big)\Big)^{-1} \cdot \\
&\quad \sum_{s_i=1}\big(\sum_{s_j=1} x_j w_j z_j^T\big)\big(\sum_{s_j=1} z_j w_j z_j^T\big)^{-1} z_i w_i y_i \\
&= \Big(\big(\sum_{s_j=1} x_j w_j z_j^T\big)\big(\sum_{s_j=1} z_j w_j z_j^T\big)^{-1}\big(\sum_{s_i=1} z_i w_i z_i^T\big)\big(\sum_{s_j=1} z_j w_j z_j^T\big)^{-1}\big(\sum_{s_j=1} z_j w_j x_j^T\big)\Big)^{-1} \cdot \\
&\quad \sum_{s_j=1} x_j w_j z_j^T \big(\sum_{s_j=1} z_j w_j z_j^T\big)^{-1} \sum_{s_i=1} z_i w_i y_i \\
&= \Big(\big(\sum_{s_j=1} x_j w_j z_j^T\big)\big(\sum_{s_j=1} z_j w_j z_j^T\big)^{-1}\big(\sum_{s_j=1} z_j w_j x_j^T\big)\Big)^{-1} \cdot \\
&\quad \sum_{s_j=1} x_j w_j z_j^T \big(\sum_{s_j=1} z_j w_j z_j^T\big)^{-1} \sum_{s_i=1} z_i w_i y_i
\end{aligned}
$$

When the model is just-identified, matrix $\sum_{s_j=1} x_j w_j z_j^T$ becomes invertible and the above equation reduces to

$$\hat{\beta}_{IPW-2SLS} = \Big(\sum_{s_j=1} z_j w_j x_j^T\Big)^{-1} \sum_{s_i=1} z_i w_i y_i$$

which is exactly what we have in equation (5.1).

Next, we show that $\hat{\beta}_{IPW-2SLS}$ is a consistent estimator of the population parameter β. Plug the structural model $Y_i = \beta^T X_i + \epsilon_{Y,i}$ into equation (5.1), we have

$$
\begin{aligned}
\hat{\beta}_{IPW-2SLS} &= \beta + \Big(\sum_{s_i=1} z_i w_i x_i^T\Big)^{-1} \sum_{s_i=1} z_i w_i \epsilon_{Y,i} \\
&= \beta + \Big(\frac{1}{n}\sum_{i=1}^{n} z_i w_i x_i^T \mathbb{I}(s_i=1)\Big)^{-1} \frac{1}{n}\sum_{i=1}^{n} z_i w_i \epsilon_{Y,i} \mathbb{I}(s_i=1) \qquad (5.3) \\
&\longrightarrow \beta + \mathbb{E}\Big[ZWX^T\,\mathbb{I}(S=1)\Big]^{-1} \mathbb{E}\Big[ZW\epsilon_Y\,\mathbb{I}(S=1)\Big]
\end{aligned}
$$

Therefore, we want to show that $\mathbb{E}\Big[ZW\epsilon_Y\,\mathbb{I}(S=1)\Big]=0$. Recalling from the DAGs in Figure 5.3 that the inverse probability weights depends on Ψ only, we have

$$\begin{aligned}\mathbb{E}\Big[ZW\epsilon_Y\,\mathbb{I}(S=1)\Big] &= \mathbb{E}\Big[ZW\epsilon_Y S\Big]\\ &= \int \mathbb{E}\Big[ZW\epsilon_Y S\,|\Psi=\psi\Big]p(\psi)d\psi\\ &= \int W(\psi)\,\mathbb{E}\Big[Z\epsilon_Y S\,|\Psi=\psi\Big]p(\psi)d\psi\\ &= \int W(\psi)\,\mathbb{E}\Big[Z\epsilon_Y\,|S=1,\Psi=\psi\Big]\mathbb{P}\big[S=1\,|\Psi=\psi\big]p(\psi)d\psi \qquad (5.4)\end{aligned}$$

By definition, $W(\psi)=1/\mathbb{P}\big[S=1\,|\Psi=\psi\big]$ and using the DAGs in Figure 5.3, we see that conditioning on Ψ, the selection indicator S is independent of Z and U (and hence ϵ_Y), so equation (5.4) reduces to

$$\mathbb{E}\Big[ZW\epsilon_Y\,\mathbb{I}(S=1)\Big]=\int \mathbb{E}\Big[Z\epsilon_Y\,|\Psi=\psi\Big]p(\psi)d\psi=\mathbb{E}\big[Z\epsilon_Y\big]=0$$

This proves the consistency of $\hat{\beta}_{IPW-2SLS}$. Note that the positivity assumption is required so that the above integrals are well-defined.

Once we've established consistency, using equation (5.3), we have

$$\sqrt{n}(\hat{\beta}_{IPW-2SLS}-\beta)=\left(\frac{1}{n}\sum_{i=1}^{n}z_i w_i x_i^T\mathbb{I}(s_i=1)\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}z_i w_i\epsilon_{Y,i}\mathbb{I}(s_i=1) \qquad (5.5)$$

Classic convergence theorem tells us that, under regularity conditions,

$$\sqrt{n}(\hat{\beta}_{IPW-2SLS}-\beta)\xrightarrow{d}\mathbb{N}\big(0,\ \Sigma_{ZWX}\,\Omega\,\Sigma_{ZWX}^T\big)$$

where $\Sigma_{ZWX}=\mathbb{E}\Big[ZWX^T\,\mathbb{I}(S=1)\Big]^{-1}$ and $\Omega=\mathbb{E}\Big[\epsilon_Y^2W^2ZZ^T\,\mathbb{I}(S=1)\Big]$.

This completes the proof. □

In many MR applications, X and Z are both simple one-dimensional variables. For example, X can be a certain biological risk factor of interest, such as BMI or blood pressure, and Z

is the associated genetic risk score. In this case, we will call the inverse probability weighted 2SLS estimator for the causal parameter β_X as the inverse probability weighted MR estimator, which, after some algebra, is given by

$$\hat{\beta}_{IPW-MR} = \frac{\sum_{s_i=1} w_i\big(z_i - \frac{\sum_{s_i=1} w_i z_i}{\sum_{s_i=1} w_i}\big)\big(y_i - \frac{\sum_{s_i=1} w_i y_i}{\sum_{s_i=1} w_i}\big)}{\sum_{s_i=1} w_i\big(z_i - \frac{\sum_{s_i=1} w_i z_i}{\sum_{s_i=1} w_i}\big)\big(x_i - \frac{\sum_{s_i=1} w_i x_i}{\sum_{s_i=1} w_i}\big)}$$

Let

$$\tilde{z}_i = z_i - \frac{\sum_{s_i=1} w_i z_i}{\sum_{s_i=1} w_i}, \quad \tilde{x}_i = x_i - \frac{\sum_{s_i=1} w_i x_i}{\sum_{s_i=1} w_i}, \quad \tilde{y}_i = y_i - \frac{\sum_{s_i=1} w_i y_i}{\sum_{s_i=1} w_i}, \quad \tilde{\epsilon}_{y,i} = \epsilon_i - \frac{\sum_{s_i=1} w_i \epsilon_i}{\sum_{s_i=1} w_i}$$

be variables centred by their weighted mean, we have

$$\hat{\beta}_{IPW-MR} = \frac{\sum_{s_i=1} w_i \tilde{z}_i \tilde{y}_i}{\sum_{s_i=1} w_i \tilde{z}_i \tilde{x}_i} = \beta_X + \frac{\sum_{s_i=1} w_i \tilde{z}_i \tilde{\epsilon}_{y,i}}{\sum_{s_i=1} w_i \tilde{z}_i \tilde{x}_i} \tag{5.6}$$

Theorem 8 tells us that under regularity conditions,

$$\sqrt{n}\big(\hat{\beta}_{IPW-MR} - \beta_X\big) \xrightarrow{d} \mathbb{N}\left(0, \quad \frac{\mathbb{E}\left[\widetilde{\epsilon}_Y^2 W^2 \widetilde{Z}^2 \, \mathbb{I}(S=1)\right]}{\mathbb{E}\left[\widetilde{Z} W \widetilde{X} \, \mathbb{I}(S=1)\right]^2}\right)$$

A natural consistent variance estimator is therefore given by

$$\widehat{\Sigma}_{IPW-MR} = \frac{\frac{1}{n_s}\sum_{s_i=1}\left(\tilde{y}_i - \hat{\beta}_{IPW-MR}\tilde{x}_i\right)^2 w_i^2 \tilde{z}_i^2}{\left(\frac{1}{n_s}\sum_{s_i=1} \tilde{z}_i w_i \tilde{x}_i\right)^2} \tag{5.7}$$

where n_s is the size of the selected sample.

Although the inverse probability weighted MR estimator is asymptotically unbiased, it can be biased in finite samples. This is because when selection is severe, selection probability for certain parts of the data will be very small that those data will almost never be selected. Effectively, this means that the positivity assumption fails in finite sample.

Figure 5.4 provides an illustrating example, where the selection model is a simple logistic function of the exposure (as in scenario A): $S \,|X \sim \text{Bernoulli}(\pi)$ with

$$\text{logit}(\pi) = \delta_0 + \delta_X X$$

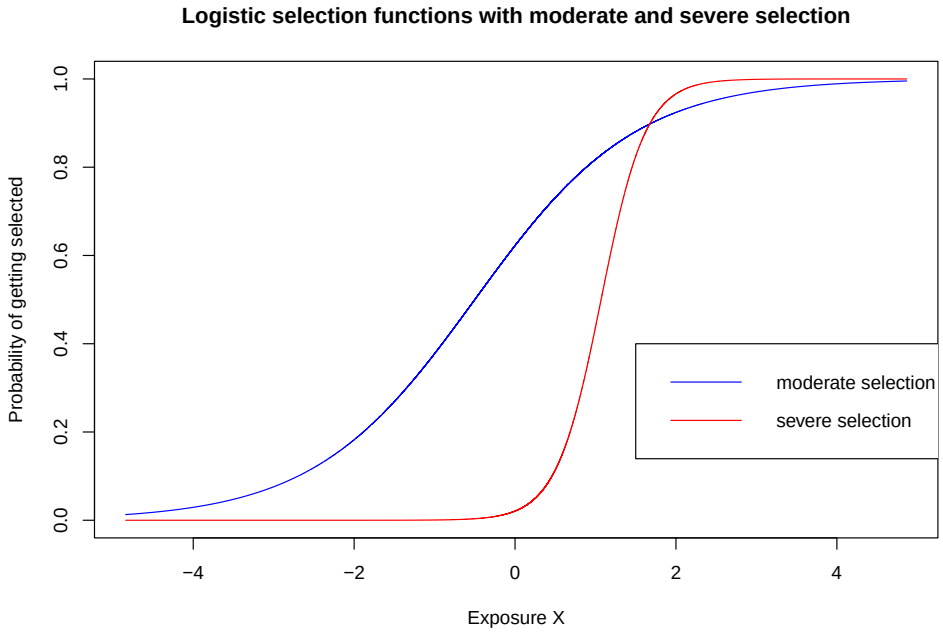


Figure 5.4: Moderate and severe selection when the selection model is a logistic function of the exposure. For moderate selection, $\delta_0 = 0.5$ and $\delta_X = 1$; for severe selection, $\delta_0 = -3.85$ and $\delta_X = 3.6$.

In this example, under severe selection, the selection probabilities for lower tails of the exposure distribution are so small that a massive sample size is required for the convergence in Theorem 8. More importantly, for any given sample size, it is always possible to find a scenario with small selection probability that the asymptotic convergence fails. I.e. we can always specify a selection model that pushes the red curve further towards zero for which the given sample size is not large enough. In fact, we will see later in the simulation studies and UK Biobank examples that under severe selection, considerable bias exists even for very large sample size.

Recall that we have been assuming the weights W_i are known. In practice, the weights will need to be estimated from external data (e.g. Hughes et al., 2019; Schoeler et al., 2023; van Alten et al., 2024), and therefore, the inverse probability weighted MR estimator depends crucially on how well the weights (i.e. the propensity scores) are estimated. In our

following simulation studies and UK Biobank examples, we will assume that we are in the ideal situation where the propensity scores are either directly known or are known functions of observed data that can be very precisely estimated (e.g. via simple logistic regressions). This is deliberate, because we want to stress-test, in the most ideal cases possible, how inverse probability weighted MR performs under severe selection. However, we need to bear in mind that the weights we are given in practice, such as participation weights for UK Biobank, can potentially be very imprecise.

5.4.2 Discussion on large weights

Another issue comes with severe selection is that small selection probabilities can render extremely large weights, resulting in inefficient estimates with very large variance. There are two common methods proposed in the literature that attempt to mitigate this issue. Robins et al. (2000) suggest replacing the weights with the following so-called stabilised weights $\widetilde{W}_i$:

$$\widetilde{W}_i = \frac{\mathbb{P}\big[S_i = 1\big]}{\mathbb{P}\big[S_i = 1 \,|\Psi_i\big]} \tag{5.8}$$

which are used in various applied investigations (e.g. van Alten et al., 2024). While Robins (1999) shows that stabilised weights can reduce variability of IPW estimates for unsaturated marginal structural models with no unmeasured confounders, this version of stabilised weights will not help in the context of using inverse probability weighted MR to correct for selection bias. This is because $\mathbb{P}\big[S_i = 1\big] = \mathbb{E}\left[S_i\right]$ is the average probability of getting selected into the study and is therefore a constant, which is cancelled as we divide the coefficient of $Y - Z$ regression by the coefficient of $X - Z$ regression in two-stage-least-square.

A potential solution to this is to replace $\mathbb{P}\big[S_i = 1\big]$ by $\mathbb{P}\big[S_i = 1 \,|Z\big]$ as the numerator for the stabilised weights. Because $\mathbb{P}\big[S_i = 1 \,|Z\big]$ is a function of Z, it will not get cancelled in two-stage-least-square and can therefore serve to stabilise the weights. By slightly modifying

the proof of Theorem 8 (i.e. replace W by $\widetilde{W}$), it is easy to show that with

$$\widetilde{W}_i = \frac{\mathbb{P}\big[S_i = 1 \,|Z\big]}{\mathbb{P}\big[S_i = 1 \,|\Psi_i\big]} \tag{5.9}$$

equation (5.4) reduces to

$$\mathbb{E}\left[Z\widetilde{W}\epsilon_Y \,\mathbb{I}(S=1)\right] = \mathbb{E}\big[Z\,\mathbb{P}(S_i = 1\,|Z)\,\epsilon_Y\big] = 0 \tag{5.10}$$

because ϵ_Y is independent of Z according to the DAGs in Figure 5.3. Therefore, the inverse probability weighted 2SLS estimator with these stabilised weights is consistent.

Alternatively, we might also use $\mathbb{P}\big[S_i = 1 \,|Z, V\big]$ as the numerator for the stabilised weights. With this version of stabilised weights, equation (5.4) reduces to

$$\mathbb{E}\left[Z\widetilde{W}\epsilon_Y \,\mathbb{I}(S=1)\right] = \mathbb{E}\big[Z\,\mathbb{P}(S_i = 1\,|Z, V)\,\epsilon_Y\big] = 0 \tag{5.11}$$

which is true for scenarios A, B and C in Figure 5.3 because U (and hence ϵ_Y) is independent of V. However, note that equation (5.11) does not hold for scenario D, as U is correlated with V in scenario D.

In practice, $\mathbb{P}\big[S_i = 1 \,|Z\big]$ or $\mathbb{P}\big[S_i = 1 \,|Z, V\big]$ will need to be estimated. The good news is that the estimators for these conditional probabilities need not to be consistent, because the inconsistency of these estimators will not affect the validity of equations (5.10) and (5.11). Likewise, the first-stage model needs not to be correctly specified for this approach to work.

Another method proposed is to trim or to truncate large weights by setting weights larger than its p_{th} percentile equal to the value at the p_{th} percentile (e.g. Lee and Stuart, 2011; van Alten et al., 2024). However, weight trimming reduces the variance at the expense of increasing bias. Cole and Hernán (2008) demonstrate growing bias as the weights are trimmed progressively for estimating marginal structural models, and Gkatzionis and Burgess (2019) note the same observation for inverse probability weighted MR estimator. This is not beneficial for causal inference, as large bias with small variance can easily result in misleading causal conclusions. Therefore, we do not recommend weight trimming in general.

5.4.3 Calculation of standard errors

As mentioned in Chapter 1 (section 1.3.3), in many MR applications, a simple method of calculating standard errors that is based on the first-order delta method expansion of the variance of the ratio estimator is often applied (Burgess and Thompson, 2021):

$$\text{s.e.}\left(\hat{\beta}_{MR}\right) = \text{s.e.}\left(\frac{\hat{\beta}_{YZ}}{\hat{\beta}_{XZ}}\right) \approx \left|\frac{\text{s.e.}(\hat{\beta}_{YZ})}{\hat{\beta}_{XZ}}\right| \tag{5.12}$$

where $\hat{\beta}_{MR}$ is the standard MR estimator, $\hat{\beta}_{YZ}$ is the estimated genetic association with the outcome and $\hat{\beta}_{XZ}$ is the estimated genetic association with the exposure.

Some literatures also apply this method to calculate standard errors of the inverse probability weighted MR estimator (e.g. Gkatzionis and Burgess, 2019), replacing $\hat{\beta}_{YZ}$ with the weighted genetic association with the outcome and $\hat{\beta}_{XZ}$ with the weighted genetic association with the exposure respectively. It can be shown that this way of calculating standard errors is equivalent to calculating the variance by

$$\widetilde{\Sigma}_{IPW-MR} = \frac{\frac{1}{n_s}\sum_{s_i=1}\left(\tilde{y}_i - \hat{\beta}_{YZ}\tilde{z}_i\right)^2 w_i^2\tilde{z}_i^2}{\left(\frac{1}{n_s}\sum_{s_i=1}\tilde{z}_i w_i\tilde{x}_i\right)^2} \tag{5.13}$$

If we assume in addition conditional homoskedasticity (for example, if we use standard statistical softwares), (5.13) will reduce to

$$\widetilde{\Sigma}_{IPW-MR} = \frac{\frac{1}{n_s}\sum_{s_i=1}(\tilde{y}_i - \hat{\beta}_{YZ}\tilde{z}_i)^2 w_i \sum_{s_i=1} w_i\tilde{z}_i^2}{\left(\frac{1}{n_s}\sum_{s_i=1}\tilde{z}_i w_i\tilde{x}_i\right)^2} \tag{5.14}$$

which effectively uses $\frac{1}{n_s}\sum_{s_i=1}(\tilde{y}_i - \hat{\beta}_{YZ}\tilde{z}_i)^2 w_i$ to estimate the residual standard error in the weighted instrument-outcome regression.

Apparently, neither (5.13) nor (5.14) can consistently estimate the variance of the inverse probability weighted MR estimator. While this simple method (5.12) may be innocuous in certain applications in which the true causal effect is small (so that the residuals of the

instrument-outcome regression are good approximates for the true error ϵ_Y), it will provide poor standard error approximation when the true effect is substantial. Moreover, conditional homoskedasticity will not hold in the selected sample even if it holds in the true population, as selection breaks the core IV assumptions. We will see later in the simulation studies that using (5.12) delivers considerably invalid confidence intervals for inverse probability weighted MR.

5.5 Simulation Study

5.5.1 Simulation design

We design our simulation studies to investigate the performance of covariate adjustment and IPW to reduce selection bias in MR analyses under a range of selection scenarios for each of the selection mechanisms listed in Figure 5.3. We would like to consider the ideal situation where the true data generating models are linear SEMs and the selection model is fully specified and correctly estimated, and do stress-testing based on different degrees of selection.

To this purpose, we simulate a simple dataset where Y, X, Z, U, V are just one-dimensional variables. We assume the exposure X has a constant causal effect on the outcome Y. The data are generated according to linear SEMs:

$$\begin{aligned} X_i &= \alpha_0 + \alpha_Z Z_i + \alpha_U U_i + \alpha_V V_i + e_{X,i} \\ Y_i &= \beta_0 + \beta_X X_i + \beta_U U_i + \beta_V V_i + e_{Y,i} \\ V_i &= \gamma_0 + \gamma_X X_i + e_{V,i} \end{aligned}$$

where $Z_i, U_i, e_{Y,i}, e_{X,i}, e_{V,i} \overset{i.i.d}{\sim} \mathbb{N}(0,1)$ and are mutually independent.

In scenario A, B and C, $\gamma_0 = \gamma_X = 0$ so V is a simple standard normal variable like Z and

U. In scenario D, $\alpha_V = \beta_V = 0$.

The selection indicator follows a classic logistic model: $S_i \,|\Psi_i \sim \text{Bernoulli}(\pi_i)$ with

$$\text{logit}(\pi_i) = \delta_0 + \delta_X \widetilde{X}_i + \delta_Y \widetilde{Y}_i + \delta_V \widetilde{V}_i$$

where $\widetilde{X}_i = \big(X_i - \text{mean}(X_i)\big)/\text{s.d}(X_i)$, $\widetilde{Y}_i = \big(Y_i - \text{mean}(Y_i)\big)/\text{s.d}(Y_i)$ and $\widetilde{V}_i = \big(V_i - \text{mean}(V_i)\big)/\text{s.d}(V_i)$ are standardised exposure, outcome and covariate with zero mean and unit variance. Standardisation will make it easier to parameterise the selection models.

Because a null causal effect will not induce bias for scenario C, we choose a non-null causal effect of interest. For simplicity, we set $\beta_X = 1$. Recall we have discussed in section 5.3 that selection bias depends on the instrument strength and confounding level. Therefore, we fix the strength of the instrument to be about 4% (the proportion of variance in the exposure explained by the instrument, as measured by R^2 in the instrument-exposure regression) and confounding level to be about 40% (the proportion of variance in the outcome not explained by the exposure, as measured by $1-R^2$ in the exposure-outcome regression), which is common for MR applications. Table 5.1 summarises the parameterisation of the data generating models.

We alter the selection model parameters $\delta_0, \delta_X, \delta_Y$ and δ_V to generate scenarios with different degrees of selection, which are measured by selection correlation and selection probability. We define selection correlation as the correlation between the selection indicator S and selection-determining variables Ψ, which will reflect the level of collider bias. We define selection probability as the mean probability of getting selected into the study, measured by the proportion of individuals with $S = 1$. Intuitively, selection is severe when the selection correlation is high and the selection probability is low, and vice versa.

We consider the combinations of selection correlation $\in \{0.25, 0.45, 0.65\}$ and selection probability $\in \{60\%, 40\%, 20\%, 5\%\}$. These scenarios are of course not exhaustive, but hopefully serve as representative examples. Selection probability of 5% can be particularly interest-

ing, as it is the average participation rate of UK Biobank. However, note that there is no scenario with selection correlation being 0.65 and selection probability being 5%. This is because for many datasets (including our simulated datasets), when selection probability is only 5%, selection correlation cannot be as high as 0.65. Tables 5.2 - 5.5 summarise the parameterisations of the selection models.

In scenarios where the selection probability $\in \{60\%, 40\%, 20\%\}$, we set the full sample size (i.e. population without selection) to be $n = 100,000$; in scenarios where the selection probability is 5%, we set the full sample size to be $n = 250,000$. Among all individuals who are selected into the study (i.e. individuals with $S = 1$), we draw a random sample of size $N = 10,000$. This is to say that we keep the selected sample size at $N = 10,000$. The main metrics for comparison include bias, estimated standard error and coverage rate. For all scenarios, results are aggregated and averaged across $D = 1,000$ simulated datasets.

Data generating models

Scenario	IV strength	Confounding level	β_X	α_Z	α_0	β_0	$\alpha_U = \beta_U$	$\alpha_V = \beta_V$	γ_0	γ_X
A, B, C	4%	40%	1	0.25	0	15	0.5	0.3	0	0
D	4%	40%	1	0.25	0	15	0.6	0	0	0.8

Table 5.1: Model parameters for the data generating mechanisms. Instrument strength is the proportion of variance in the exposure explained by the instrument, measured by R^2 in the instrument-exposure regression; confounding level is the proportion of variance in the outcome not explained by the exposure, measured by $1 - R^2$ in the exposure-outcome regression.

Selection according to the exposure X

Scenario	Selection correlation	Selection probability	δ_0	δ_X	δ_Y	δ_V
A1	0.25	60%	0.45	0.45	0	0
A2	0.25	40%	−0.45	0.45	0	0
A3	0.25	20%	−1.5	0.56	0	0
A4	0.25	5%	−3.6	1.07	0	0
A5	0.45	60%	0.5	1	0	0
A6	0.45	40%	−0.5	1	0	0
A7	0.45	20%	−1.8	1.2	0	0
A8	0.45	5%	−9.05	4.4	0	0
A9	0.65	60%	0.8	2.15	0	0
A10	0.65	40%	−0.8	2.15	0	0
A11	0.65	20%	−3.85	3.6	0	0

Table 5.2: Model parameters for selection scenarios where the selection depends on the exposure X only. Selection correlation is the correlation between S and X, and selection probability is the mean probability of getting selected into the sample.

Selection according to the exposure X and the outcome Y

Scenario	Selection correlation	Selection probability	δ_0	δ_X	δ_Y	δ_V
B1	0.25	60%	0.45	0.28	0.28	0
B2	0.25	40%	−0.43	0.27	0.29	0
B3	0.25	20%	−1.53	0.3	0.4	0
B4	0.25	5%	−3.7	0.58	0.75	0
B5	0.45	60%	0.53	0.56	0.68	0
B6	0.45	40%	−0.54	0.56	0.68	0
B7	0.45	20%	−2	0.715	0.88	0
B8	0.45	5%	−17.2	4.6	5.5	0
B9	0.65	60%	0.9	1.38	1.65	0
B10	0.65	40%	−0.9	1.38	1.65	0
B11	0.65	20%	−7.45	3.85	4.55	0

Table 5.3: Model parameters for selection scenarios where the selection depends on both the exposure X and the outcome Y. Selection correlation refers to both the correlation between S and X and the correlation between S and Y (e.g. in scenario B1, both corr(S, X) and corr(S, Y) are kept at 0.25), and selection probability is the mean probability of getting selected into the sample.

Selection according to the outcome Y

Scenario	Selection correlation	Selection probability	δ_0	δ_X	δ_Y	δ_V
C1	0.25	60%	0.435	0	0.545	0
C2	0.25	40%	−0.432	0	0.55	0
C3	0.25	20%	−1.51	0	0.67	0
C4	0.25	5%	−3.62	0	1.28	0
C5	0.45	60%	0.52	0	1.17	0
C6	0.45	40%	−0.52	0	1.17	0
C7	0.45	20%	−1.9	0	1.5	0
C8	0.45	5%	−9.9	0	5.8	0
C9	0.65	60%	0.79	0	2.54	0
C10	0.65	40%	−0.775	0	2.53	0
C11	0.65	20%	−4.05	0	4.52	0

Table 5.4: Model parameters for selection scenarios where the selection depends on the outcome Y only. Selection correlation is the correlation between S and Y, and selection probability is the mean probability of getting selected into the sample.

Selection according to the downstream covariate V of the exposure X

Scenario	Selection correlation	Selection probability	δ_0	δ_X	δ_Y	δ_V
D1	0.25	60%	0.43	0	0	0.55
D2	0.25	40%	−0.42	0	0	0.55
D3	0.25	20%	−1.5	0	0	0.66
D4	0.25	5%	−3.6	0	0	1.27
D5	0.45	60%	0.51	0	0	1.16
D6	0.45	40%	−0.51	0	0	1.16
D7	0.45	20%	−1.9	0	0	1.48
D8	0.45	5%	−10.3	0	0	6
D9	0.65	60%	0.78	0	0	2.54
D10	0.65	40%	−0.78	0	0	2.54
D11	0.65	20%	−4.1	0	0	4.52

Table 5.5: Model parameters for selection scenarios where the selection depends on a downstream covariate V of the exposure X. Selection correlation is the correlation between S and V, and selection probability is the mean probability of getting selected into the sample.

5.5.2 Simulation results

In Tables 5.6 - 5.13, we present and compare bias, mean estimated standard error and coverage of the 95% confidence interval for the inverse probability weighted MR estimator, the covariate adjusted MR estimator (in complete cases) and the simple MR estimator (simple complete-case analysis without adjusting for covariates) under different degrees of selection for each of the selection mechanisms we consider.

For results of inverse probability weighted MR, we present standard errors and their respective coverages using both the consistent variance estimator in equation (5.7), which we call robust standard error (robust SE), and the first-order approximation in equation (5.12), which we call naive standard error (naive SE). For covariate adjusted MR and standard MR, we present standard errors using their asymptotic variances without assuming conditional homoskedasticity so that the results are comparable to those of inverse probability weighted MR using robust standard errors.

We firstly note that the naive standard error provides considerably incorrect confidence intervals for the inverse probability weighted MR estimator in all simulated scenarios. On the one hand, when selection is mild (e.g. when selection correlation is 0.25 and selection probability is 20% or higher), the naive standard error is overly conservative and not efficient. On the other hand, when the selection is severe (e.g. when selection correlation is 0.65 or when selection probability is 5%), the naive standard error is overly optimistic and tremendously under-covers the truth. We also note that the naive standard error stays largely constant in most scenarios, and only increases slightly when selection becomes very severe. This demonstrates that this simple method of calculating standard error fails to take into account the dependencies in the selected sample and the uncertainties of weighting rarely selected cases, and is therefore inconsistent. In contrast, the robust standard error provides correct or close-to-correct coverage in most scenarios for each of the selection mechanisms considered.

Comparing the magnitude of biases of inverse probability weighted MR, covariate adjusted MR and standard MR in those scenarios, we observe that IPW indeed eliminates selection bias when the selection is moderate. However, in scenarios where the selection is more severe (e.g. when selection correlation is 0.45 and selection probability is 5% or when selection correlation is 0.65), there is still noticeable bias from IPW. In some of the most extreme cases (scenarios A10, A11, D10 and D11), the bias from inverse probability weighted MR exceeds the bias from simple MR without any adjustment for selection. This confirms our argument in section 5.4.1 that the asymptotic convergence of the inverse probability weighted MR estimator holds under moderate selection, but fails under severe selection. Moreover, the bias from IPW can be larger than the bias from selection when selection is severe. We also note that the robust standard error quickly gets large when selection becomes more severe. This corroborates the arguments from existing literatures that IPW can be very inefficient when selection correlation is high and selection probability is small (Robins et al., 2000). Nevertheless, the issue with large standard error could be mitigated by using stabilised weights discussed in section 5.4.2, which we did not implement in the simulation. But again, even with very large standard errors, the confidence intervals still exhibit under-coverage under severe selection.

Finally, we observe that covariate adjustment reduces selection bias and improves coverage compared to simple MR, but the extent to which the bias is reduced is quite limited. Noticeable bias and considerable under-covered confidence interval exist for almost all simulated scenarios. The exception is scenario D, in which adjusting for covariate V largely increases bias because V is a child of collider. This is consistent with our path analysis argument in section 5.3 and with previous simulation studies by Burgess et al. (2024).

Selection according to the exposure X							
Selection	Selection	Selection	Inverse probability weighted MR				
scenario	correlation	probability	Bias	Robust SE	Coverage	Naive SE	Coverage
A1	0.25	60%	−0.00005	0.048	94.8%	0.074	99.6%
A2	0.25	40%	−0.001	0.050	95.3%	0.074	99.8%
A3	0.25	20%	−0.002	0.055	94.3%	0.074	99.1%
A4	0.25	5%	−0.010	0.100	93.7%	0.075	85.7%
A5	0.45	60%	−0.001	0.056	94.8%	0.074	98.8%
A6	0.45	40%	−0.002	0.065	95.9%	0.074	98.1%
A7	0.45	20%	−0.005	0.095	95.6%	0.075	88.0%
A8	0.45	5%	−0.191	234.386	84.3%	0.345	13.3%
A9	0.65	60%	−0.022	0.194	93.6%	0.081	80.8%
A10	0.65	40%	−0.234	65.844	92.1%	0.120	63.6%
A11	0.65	20%	−0.910	1325.848	88.1%	0.480	22.9%

Table 5.6: Bias, standard error and coverage for inverse probability weighted MR estimator, scenarios A1 - A11.

Selection according to the exposure X								
Selection	Selection	Selection	Covariate adjusted MR			Simple MR		
scenario	correlation	probability	Bias	SE	Coverage	Bias	SE	Coverage
A1	0.25	60%	−0.011	0.048	93.4%	−0.016	0.049	92.7%
A2	0.25	40%	−0.013	0.048	94.5%	−0.018	0.050	94.1%
A3	0.25	20%	−0.016	0.048	93.0%	−0.021	0.050	92.9%
A4	0.25	5%	−0.029	0.052	92.4%	−0.040	0.054	90.3%
A5	0.45	60%	−0.044	0.055	89.7%	−0.061	0.058	83.5%
A6	0.45	40%	−0.051	0.057	86.1%	−0.069	0.060	80.0%
A7	0.45	20%	−0.068	0.060	82.4%	−0.094	0.064	71.7%
A8	0.45	5%	−0.554	0.182	10.8%	−0.754	0.205	1.6%
A9	0.65	60%	−0.118	0.073	64.1%	−0.161	0.078	46.7%
A10	0.65	40%	−0.157	0.082	51.9%	−0.213	0.089	30.5%
A11	0.65	20%	−0.343	0.128	22.2%	−0.469	0.142	6.6%

Table 5.7: Bias, standard error and coverage for standard MR estimator with and without covariate adjustment, scenarios A1 - A11.

Selection according to the exposure X and the outcome Y

Selection scenario	Selection correlation	Selection probability	Inverse probability weighted MR				
			Bias	Robust SE	Coverage	Naive SE	Coverage
B1	0.25	60%	0.0004	0.049	94.8%	0.074	99.7%
B2	0.25	40%	0.001	0.051	96.1%	0.074	99.8%
B3	0.25	20%	−0.005	0.059	95.9%	0.074	98.7%
B4	0.25	5%	−0.018	0.133	95.5%	0.076	76.7%
B5	0.45	60%	−0.002	0.062	95.4%	0.074	98.3%
B6	0.45	40%	0.002	0.076	95.1%	0.074	94.6%
B7	0.45	20%	−0.032	0.214	96.2%	0.079	74.4%
B8	0.45	5%	−0.812	534.672	80.2%	0.332	7.9%
B9	0.65	60%	−0.068	2.170	93.6%	0.085	60.5%
B10	0.65	40%	−0.062	6.078	91.7%	0.115	43.8%
B11	0.65	20%	0.469	136.853	85.4%	0.284	13.0%

Table 5.8: Bias, standard error and coverage for inverse probability weighted MR estimator, scenarios B1 - B11.

Selection according to the exposure X and the outcome Y

Selection scenario	Selection correlation	Selection probability	Covariate adjusted MR			Simple MR		
			Bias	SE	Coverage	Bias	SE	Coverage
B1	0.25	60%	−0.026	0.047	91.3%	−0.032	0.049	89.9%
B2	0.25	40%	−0.028	0.047	91.1%	−0.033	0.049	91.0%
B3	0.25	20%	−0.040	0.048	88.1%	−0.046	0.050	85.6%
B4	0.25	5%	−0.074	0.051	70.3%	−0.087	0.053	63.3%
B5	0.45	60%	−0.106	0.053	46.4%	−0.123	0.056	39.5%
B6	0.45	40%	−0.122	0.054	37.8%	−0.141	0.057	29.6%
B7	0.45	20%	−0.180	0.059	12.1%	−0.209	0.062	6.3%
B8	0.45	5%	−1.122	0.115	0%	−1.220	0.120	0%
B9	0.65	60%	−0.273	0.066	0.7%	−0.315	0.070	0.3%
B10	0.65	40%	−0.369	0.073	0%	−0.421	0.078	0%
B11	0.65	20%	−0.812	0.101	0%	−0.902	0.108	0%

Table 5.9: Bias, standard error and coverage for standard MR estimator with and without covariate adjustment, scenarios B1 - B11.

Selection according to the outcome Y

Selection scenario	Selection correlation	Selection probability	Inverse probability weighted MR				
			Bias	Robust SE	Coverage	Naive SE	Coverage
C1	0.25	60%	0.0003	0.049	96.8%	0.073	99.7%
C2	0.25	40%	−0.0002	0.052	95.7%	0.074	99.4%
C3	0.25	20%	−0.0006	0.060	95.5%	0.074	99.1%
C4	0.25	5%	−0.015	0.131	95.3%	0.075	74.0%
C5	0.45	60%	−0.007	0.064	95.6%	0.073	97.9%
C6	0.45	40%	−0.006	0.078	95.4%	0.074	94.1%
C7	0.45	20%	−0.017	0.136	95.4%	0.075	70.6%
C8	0.45	5%	−0.135	390.592	85.5%	0.229	9.1%
C9	0.65	60%	−0.032	0.162	94.6%	0.076	65.3%
C10	0.65	40%	−0.083	0.350	93.8%	0.082	48.9%
C11	0.65	20%	−0.357	21.517	88.7%	0.164	15.9%

Table 5.10: Bias, standard error and coverage for inverse probability weighted MR estimator, scenarios C1 - C11.

Selection according to the outcome Y

Selection scenario	Selection correlation	Selection probability	Covariate adjusted MR			Simple MR		
			Bias	SE	Coverage	Bias	SE	Coverage
C1	0.25	60%	−0.028	0.046	92.1%	−0.030	0.047	90.9%
C2	0.25	40%	−0.030	0.046	91.0%	−0.034	0.047	90.9%
C3	0.25	20%	−0.034	0.046	90.5%	−0.038	0.048	89.6%
C4	0.25	5%	−0.066	0.047	72.0%	−0.073	0.049	69.1%
C5	0.45	60%	−0.097	0.047	46.1%	−0.107	0.049	40.6%
C6	0.45	40%	−0.111	0.048	34.2%	−0.122	0.050	29.2%
C7	0.45	20%	−0.153	0.049	10.8%	−0.169	0.050	6.7%
C8	0.45	5%	−0.600	0.049	0%	−0.627	0.050	0%
C9	0.65	60%	−0.224	0.050	0.6%	−0.245	0.052	0.2%
C10	0.65	40%	−0.275	0.050	0%	−0.298	0.053	0%
C11	0.65	20%	−0.470	0.051	0%	−0.498	0.053	0%

Table 5.11: Bias, standard error and coverage for standard MR estimator with and without covariate adjustment, scenarios C1 - C11.

Selection according to the downstream covariate V of the exposure X

Selection scenario	Selection correlation	Selection probability	Inverse probability weighted MR				
			Bias	Robust SE	Coverage	Naive SE	Coverage
D1	0.25	60%	−0.0002	0.048	96.3%	0.074	99.9%
D2	0.25	40%	0.0003	0.050	95.3%	0.075	99.8%
D3	0.25	20%	0.0004	0.054	95.7%	0.074	99.6%
D4	0.25	5%	−0.008	0.098	94.6%	0.075	84.3%
D5	0.45	60%	−0.0008	0.054	95.4%	0.075	99.4%
D6	0.45	40%	0.0004	0.063	95.4%	0.075	98.7%
D7	0.45	20%	0.001	0.097	95.0%	0.075	86.3%
D8	0.45	5%	−0.890	341.204	86.4%	0.413	14.3%
D9	0.65	60%	0.002	0.102	93.3%	0.076	83.4%
D10	0.65	40%	−0.056	1.194	92.2%	0.087	65.6%
D11	0.65	20%	−0.491	147.381	88.2%	0.246	24.8%

Table 5.12: Bias, standard error and coverage for inverse probability weighted MR estimator, scenarios D1 - D11.

Selection according to the downstream covariate V of the exposure X

Selection scenario	Selection correlation	Selection probability	Covariate adjusted 2SLS			Simple MR		
			Bias	SE	Coverage	Bias	SE	Coverage
D1	0.25	60%	−0.233	0.092	26.1%	−0.008	0.048	95.9%
D2	0.25	40%	−0.231	0.092	27.0%	−0.007	0.048	95.4%
D3	0.25	20%	−0.232	0.091	26.1%	−0.009	0.048	95.9%
D4	0.25	5%	−0.233	0.092	28.3%	−0.019	0.050	93.9%
D5	0.45	60%	−0.232	0.092	26.3%	−0.027	0.052	92.9%
D6	0.45	40%	−0.234	0.092	26.9%	−0.030	0.052	91.2%
D7	0.45	20%	−0.229	0.091	27.0%	−0.039	0.054	90.4%
D8	0.45	5%	−0.229	0.091	26.7%	−0.148	0.075	51.4%
D9	0.65	60%	−0.225	0.091	28.6%	−0.055	0.058	84.9%
D10	0.65	40%	−0.233	0.092	27.2%	−0.073	0.061	78.9%
D11	0.65	20%	−0.236	0.092	24.4%	−0.123	0.070	57.1%

Table 5.13: Bias, standard error and coverage for standard MR estimator with and without covariate adjustment, scenarios D1 - D11.

5.6 Applied Examples from UK Biobank

5.6.1 Description of the examples

We further illustrate the performances of the inverse probability weighted MR estimator using real data examples from UK Biobank. Because of the limited benefits in reducing selection bias by covariate adjustment as demonstrated in our simulation studies, we do not perform covariate adjusted MR for real data examples.

We consider the causal effect of systolic blood pressure (SBP) on four different outcomes, including concentration of sodium in the blood, concentration of creatinine in the blood, coronary heart disease and stroke. The first two outcomes are continuous while the last two are binary indicators of getting the disease or not. The instrument is a weighted genetic risk score, constructed by Higgins et al. (2021), from 258 genetic variants that have previously been shown to be associated with blood pressure at genome-wide significance level. The strength of the score for systolic blood pressure is about 1.82%. The full sample size of our example is 367,703 individuals. A very small number of missing individuals for each of the outcomes are dropped for complete-case analyses. Because covariate adjustment is not performed, for simplicity we do not consider any covariate.

We should note that for binary outcomes, the usual measure of effect is odds ratio, often estimated using logistic regressions. However, logistic regressions with instrumental variables are generally considered as "forbidden regressions" due to non-collapsibility of the logistic model (e.g. Angrist and Pischke, 2009; Burgess, 2017; Burgess and Thompson, 2021). Indeed, consistency of IV estimators relies on the linearity assumption of the exposure-outcome model. The linearity assumption is also required for proving the asymptotic convergence of inverse probability weighted 2SLS estimator in Theorem 8. Moreover, naively applying IPW to logistic regressions will result in non-integer number of success in the logistic model, so the objective function needs to be defined carefully. Because of these limitations, we perform the

usual (weighted) 2SLS estimation instead of logistic regression for the two binary outcomes in our example. I.e. we assume a linear probability model and estimate the absolute risk difference as the exposure effect. This is in line with various applied IV or MR analyses conducted in the literatures (e.g. Hogan and Lancaster, 2004; Hughes et al., 2019; Schoeler et al., 2023). We leave it for future work to investigate using IPW for binary outcome models in the context of IV and MR studies.

5.6.2 Construction of the examples

In order to calculate bias and coverage rates, we manually create selection in the same way as we create selection in the simulation studies. Because we do not consider any covariates, we only construct selection models according to scenarios A, B and C and omit scenario D. Since in scenario D the selection depends on the downstream covariate of the exposure, the performances of inverse probability weighted MR and simple MR will be similar to those in scenario A. Therefore, the examples we constructed will still be representative.

Specifically, for the continuous outcomes, we define the selection indicator to follow a simple logistic model: $S_i \,|\Psi_i \sim \text{Bernoulli}(\pi_i)$ with

$$\text{logit}(\pi_i) = \delta_0 + \delta_X \widetilde{X}_i + \delta_Y \widetilde{Y}_i$$

where $\widetilde{X}_i$ and $\widetilde{Y}_i$ are standardised exposures and outcomes for all the scenarios A, B and C.

For the binary outcomes, we define $S_i \,|\Psi_i \sim \text{Bernoulli}(\pi_i)$ with separate selection functions for each of the selection mechanisms. For scenario A, we define the selection model to be a logistic function in the same way as above:

$$\text{logit}(\pi_i) = \delta_0 + \delta_X \widetilde{X}_i$$

For scenario B, we define differential selection functions for cases and controls such that the

selection probability depends on the exposure only for controls:

$$\pi_i = \begin{cases} \text{expit}\big(\delta_0 + \delta_X \widetilde{X}_i\big), & \text{if} \quad Y_i = 0 \\ \pi_1, & \text{if} \quad Y_i = 1 \end{cases}$$

where the expit function is the inverse of the logistic function.

For scenario C, we define different but constant selection probabilities for cases and controls:

$$\pi_i = \begin{cases} \pi_0, & \text{if} \quad Y_i = 0 \\ \pi_1, & \text{if} \quad Y_i = 1 \end{cases}$$

This is in fact equivalent to defining a logistic selection function on Y: $\text{logit}(\pi_i) = \delta_0 + \delta_Y Y_i$.

We firstly derive the full sample MR estimates (given in Table 5.14), and calculate bias and coverage of the inverse probability weighted MR estimator and the simple MR estimator from the selected samples by regarding the full sample estimates as the "truth". The examples created are of course artificial, but hopefully they will serve as illustrating examples as how selection bias might occur in real datasets.

We alter the selection model parameters δ_0, δ_X, δ_Y, π_0 and π_1 to generate different degrees of selection. For scenario A, similar to the simulation studies before, we consider combinations of selection correlation $\in \{0.25, 0.45, 0.65\}$ and selection probability $\in \{60\%, 40\%, 20\%, 5\%\}$. However, for scenario B and C, we keep selection probability $\in \{60\%, 40\%, 20\%, 5\%\}$ but choose different sets of selection correlations as examples to present. This is deliberate, as different outcome distributions require different scenarios for the selection effect to be observable. Specifically, for the two continuous outcomes, we choose negative correlations between the outcome and the selection indicator. As for the two binary outcomes, due to their binary nature it is difficult to control the correlation between the outcome and the selection indicator. Therefore, for the ease of constructing examples, in scenario B we only consider correlation between the exposure and the selection indicator in the controls; in scenario C, we do not specifically choose selection correlation but choose the probability of selection for the

cases (i.e. π_1) instead. Tables 5.15 - 5.23 summarise the parameterisations of the selection models

Similar to the simulation studies in section 5.5, among all individuals who are selected, we draw a random sample of size $N = 10,000$. I.e., we keep the selected sample size at $N = 10,000$. All results are aggregated and averaged across $D = 1,000$ randomly drawn datasets from the selected sample.

Full UK Biobank sample estimates

Outcomes	Sodium	Creatinine	Coronary heart disease	Stroke
Full sample estimates	−0.081	0.0054	0.0045	0.0015

Table 5.14: Standard MR estimates in the full UK Biobank datasets for the causal effect of systolic blood pressure on concentration of sodium in the blood (sodium), concentration creatinine in the blood (creatinine), coronary heart disease and stroke.

Selection according to the exposure (SBP)

Scenario	Selection correlation	Selection probability	δ_0	δ_X	δ_Y
A1	0.25	60%	0.45	0.56	0
A2	0.25	40%	−0.42	0.53	0
A3	0.25	20%	−1.5	0.62	0
A4	0.25	5%	−3.4	1	0
A5	0.45	60%	0.56	1.25	0
A6	0.45	40%	−0.48	1.13	0
A7	0.45	20%	−1.8	1.3	0
A8	0.45	5%	−5.35	2.55	0
A9	0.65	60%	1.04	2.96	0
A10	0.65	40%	−0.6	2.42	0
A11	0.65	20%	−2.86	3.02	0

Table 5.15: Model parameters for selection scenarios where the selection depends on the exposure X only. The exposure is systolic blood pressure (SBP). Selection correlation is the correlation between S and X, and selection probability is the mean probability of getting selected into the sample. The same parameterisations for scenarios A1 - A11 are used for all four outcomes.

Selection according to the exposure (SBP) and the outcome (sodium)

Scenario	Selection correlation	Selection probability	δ_0	δ_X	δ_Y
B1	0.25; −0.1	60%	0.45	0.6	−0.26
B2	0.25; −0.1	40%	−0.43	0.56	−0.27
B3	0.25; −0.1	20%	−1.51	0.64	−0.34
B4	0.25; −0.1	5%	−3.6	1.05	−0.73
B5	0.45; −0.15	60%	0.6	1.36	−0.5
B6	0.45; −0.15	40%	−0.5	1.22	−0.5
B7	0.45; −0.15	20%	−1.88	1.39	−0.67
B8	0.45; −0.15	5%	−7.98	3.65	−2.75
B9	0.65; −0.2	60%	1.35	3.9	−1.33
B10	0.65; −0.2	40%	−0.72	2.98	−1.17
B11	0.65; −0.2	20%	−4.08	4.34	−2.11

Table 5.16: Model parameters for selection scenarios where the selection depends on both the exposure X and the outcome Y. The exposure is systolic blood pressure (SBP) and the outcome is concentration of sodium in the blood (sodium). Selection correlation refers to both the correlation between S and X and the correlation between S and Y (e.g. in scenario B1, $\text{corr}(S, X) = 0.25$ and $\text{corr}(S, X) = -0.1$), and selection probability is the mean probability of getting selected into the sample.

Selection according to the outcome (sodium)					
Scenario	Selection correlation	Selection probability	δ_0	δ_X	δ_Y
C1	−0.1	60%	0.41	0	−0.2
C2	−0.1	40%	−0.4	0	−0.21
C3	−0.1	20%	−1.4	0	−0.27
C4	−0.1	5%	−3.08	0	−0.59
C5	−0.25	60%	0.42	0	−0.53
C6	−0.25	40%	−0.45	0	−0.59
C7	−0.25	20%	−1.58	0	−0.85
C8	−0.25	5%	−5.54	0	−3.57
C9	−0.5	60%	0.46	0	−1.3
C10	−0.5	40%	−0.72	0	−1.66
C11	−0.5	20%	−3.92	0	−4.25

Table 5.17: Model parameters for selection scenarios where the selection depends on the outcome Y only. The outcome is concentration of sodium in the blood (sodium). Selection correlation is the correlation between S and Y, and selection probability is the mean probability of getting selected into the sample.

Selection according to the exposure (SBP) and the outcome (creatinine)

Scenario	Selection Correlation	Selection Probability	δ_0	δ_X	δ_Y
B1	0.25; −0.1	60%	0.45	0.57	−0.21
B2	0.25; −0.1	40%	−0.43	0.54	−0.22
B3	0.25; −0.1	20%	−1.51	0.62	−0.28
B4	0.25; −0.1	5%	−3.58	1	−0.68
B5	0.45; −0.15	60%	0.6	1.3	−0.38
B6	0.45; −0.15	40%	−0.48	1.16	−0.4
B7	0.45; −0.15	20%	−1.85	1.31	−0.54
B8	0.45; −0.15	5%	−7.15	3.15	−2.28
B9	0.65; −0.2	60%	1.2	3.35	−0.92
B10	0.65; −0.2	40%	−0.68	2.68	−0.88
B11	0.65; −0.2	20%	−3.48	3.58	−1.47

Table 5.18: Model parameters for selection scenarios where the selection depends on both the exposure X and the outcome Y. The exposure is systolic blood pressure (SBP) and the outcome is concentration of creatinine in the blood (creatinine). Selection correlation refers to both the correlation between S and X and the correlation between S and Y (e.g. in scenario B1, $\text{corr}(S, X) = 0.25$ and $\text{corr}(S, X) = -0.1$), and selection probability is the mean probability of getting selected into the sample.

Selection according to the outcome (creatinine)

Scenario	Selection correlation	Selection probability	δ_0	δ_X	δ_Y
C1	-0.1	60%	0.41	0	-0.205
C2	-0.1	40%	-0.405	0	-0.215
C3	-0.1	20%	-1.41	0	-0.29
C4	-0.1	5%	-3.09	0	-0.64
C5	-0.25	60%	0.42	0	-0.54
C6	-0.25	40%	-0.45	0	-0.61
C7	-0.25	20%	-1.6	0	-0.93
C8	-0.25	5%	-7.15	0	-5.45
C9	-0.5	60%	0.43	0	-1.38
C10	-0.5	40%	-0.81	0	-1.87
C11	-0.5	20%	-5.2	0	-6.1

Table 5.19: Model parameters for selection scenarios where the selection depends on the outcome Y only. The outcome is concentration of creatinine in the blood (creatinine). Selection correlation is the correlation between S and Y, and selection probability is the mean probability of getting selected into the sample.

Selection according to the exposure (SBP) and the outcome (coronary heart disease)

Scenario	Selection correlation	Selection probability	δ_0	δ_X	π_1
B1	0.25	60%	0.245	0.56	1
B2	0.25	40%	−0.72	0.57	0.95
B3	0.25	20%	−2.13	0.81	0.8
B4	0.25	5%	−10.42	4.3	0.3
B5	0.45	60%	0.36	1.32	1
B6	0.45	40%	−0.8	1.33	0.9
B7	0.45	20%	−2.37	1.8	0.5
B8	0.45	5%	−16.2	8.95	0.08
B9	0.65	60%	0.9	3.7	1
B10	0.65	40%	−1.24	3.9	0.8
B11	0.65	20%	−9	11.3	0.25

Table 5.20: Model parameters for selection scenarios where the selection depends on both the exposure X and the outcome Y. The exposure is systolic blood pressure (SBP) and the outcome is coronary heart disease. Selection correlation is the correlation between S and X in the controls (i.e. individuals with $Y = 0$), and selection probability is the mean probability of getting selected into the sample.

Selection according to the outcome (coronary heart disease)			
Scenario	Selection probability	π_0	π_1
C1	60%	0.576	0.8
C2	40%	0.352	0.8
C3	20%	0.128	0.8
C4	5%	0.026	0.25
C5	60%	0.564	0.9
C6	40%	0.34	0.9
C7	20%	0.116	0.9
C8	5%	0.014	0.35
C9	60%	0.553	1
C10	40%	0.328	1
C11	20%	0.105	1
C12	5%	0.002	0.45

Table 5.21: Model parameters for selection scenarios where the selection depends on the outcome Y only. The outcome is coronary heart disease. Selection probability is the mean probability of getting selected into the sample.

Selection according to the exposure (SBP) and the outcome (stroke)

Scenario	Selection correlation	Selection probability	δ_0	δ_X	π_1
B1	0.25	60%	0.38	0.56	1
B2	0.25	40%	−0.5	0.535	1
B3	0.25	20%	−1.72	0.66	0.95
B4	0.25	5%	−8.84	3.42	0.9
B5	0.45	60%	0.5	1.27	1
B6	0.45	40%	−0.58	1.18	0.95
B7	0.45	20%	−2.1	1.48	0.9
B8	0.45	5%	−18	9.45	0.3
B9	0.65	60%	1	3.15	1
B10	0.65	40%	−0.79	2.7	0.95
B11	0.65	20%	−5.36	5.6	0.9

Table 5.22: Model parameters for selection scenarios where the selection depends on both the exposure X and the outcome Y. The exposure is systolic blood pressure (SBP) and the outcome is stroke. Selection correlation is the correlation between S and X in the controls (i.e. individuals with $Y = 0$), and selection probability is the mean probability of getting selected into the sample.

Selection according to the outcome (stroke)			
Scenario	Selection probability	π_0	π_1
C1	60%	0.592	0.8
C2	40%	0.385	0.8
C3	20%	0.178	0.8
C4	5%	0.022	0.8
C5	60%	0.588	0.9
C6	40%	0.38	0.9
C7	20%	0.174	0.9
C8	5%	0.018	0.9
C9	60%	0.585	1
C10	40%	0.377	1
C11	20%	0.171	1
C12	5%	0.014	1

Table 5.23: Model parameters for selection scenarios where the selection depends on the outcome Y only. The outcome is stroke. Selection probability is the mean probability of getting selected into the sample.

5.6.3 Results of the examples

Tables 5.24 - 5.35 present and compare bias, mean estimated standard error and coverage of the 95% confidence intervals for the inverse probability weighted MR estimator and the standard MR estimator under different degrees of selection for scenarios A, B and C. Since in section 5.4.3 and 5.5.2 we have demonstrated that the naive standard error provides invalid confidence intervals, we use the robust standard error in all scenarios considered.

For the two continuous outcomes, we firstly note that the selection bias with concentration of sodium in the blood is in general greater than that with concentration of creatinine in the blood under same selection scenarios. This is because SBP has greater effect on the concentration of sodium than the concentration of creatinine in the overall UK Biobank population (see Table 5.14), resulting in more substantial selection effect. Despite the differences in

the magnitude of causal effect, the results with both outcomes are similar to what we have observed in the simulation studies. The inverse probability weighted MR largely reduces selection bias when the selection is moderate, but fails to do so when the selection is severe. In particular, in scenarios A8, B8, B11, C8 and C11 with both outcomes, the bias from inverse probability weighted MR greatly exceeds the bias from simple MR without accounting for selection. The standard error gets exceedingly large in these scenarios, but still fails to provide correct coverage.

For the two binary outcomes, IPW seems to work better. We note that the inverse probability weighted MR estimator eliminates selection bias in almost all scenarios A and scenarios C. The only exception is scenario C12 where the outcome is coronary heart disease, in which the bias from inverse probability weighted MR exceeds the bias from simple MR, though the robust standard error still gives close-to-correct coverage. This is likely due to the fact that the effects of SBP on the absolute risk differences of coronary heart disease and stroke are quite small, resulting in relatively small selection bias from which IPW can easily recover. However, when the selection function becomes more complex as in scenarios B, inverse probability weighted MR starts failing to eliminate bias under severe selection. In particular, in scenarios B4, B8 and B11 with both coronary heart disease and stroke, nonignorable bias exists with large standard error that delivers under-covered confidence interval, similar to what we have observed with the continuous outcomes.

We also note that, for the two binary outcomes, the robust standard error can give rise to an overly wide confidence interval for the inverse probability weighted MR estimator. For example, in scenario A8 for coronary heart disease and scenarios A8, A10, A11 for stroke, the coverages of the robust standard error are over 98%. This is likely because the robust standard error is derived with the linearity assumption, and therefore it can be overly conservative for binary outcomes. This highlights the disadvantages of using linear structural equations to model binary outcomes.

Effect of SBP on sodium with selection according to the exposure

Selection	Selection	Selection	Inverse probability weighted MR			Simple MR		
scenario	correlation	probability	Bias	SE	Coverage	Bias	SE	Coverage
A1	0.25	60%	0.004	0.179	96.7%	−0.030	0.178	96.7%
A2	0.25	40%	0.003	0.183	94.6%	−0.037	0.176	95.0%
A3	0.25	20%	0.008	0.195	94.6%	−0.053	0.171	94.9%
A4	0.25	5%	−0.008	0.274	94.6%	−0.107	0.170	92.1%
A5	0.45	60%	−0.004	0.196	96.0%	−0.113	0.207	92.5%
A6	0.45	40%	−0.0004	0.216	95.7%	−0.115	0.206	91.8%
A7	0.45	20%	−0.006	0.285	95.5%	−0.171	0.216	89.1%
A8	0.45	5%	3.830	9729.985	91.7%	−0.486	0.409	83.5%
A9	0.65	60%	−0.035	0.439	95.0%	−0.259	0.288	87.9%
A10	0.65	40%	−0.047	0.733	95.1%	−0.306	0.305	85.3%
A11	0.65	20%	0.398	222.074	93.8%	−0.588	0.436	74.1%

Table 5.24: Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is concentration of sodium, scenarios A1 - A11.

Effect of SBP on sodium with selection according to the exposure and the outcome

Selection	Selection	Selection	Inverse probability weighted MR			Simple MR		
scenario	correlation	probability	Bias	SE	Coverage	Bias	SE	Coverage
B1	0.25; −0.1	60%	0.001	0.192	96.6%	0.040	0.169	94.9%
B2	0.25; −0.1	40%	−0.004	0.207	96.0%	0.027	0.162	95.9%
B3	0.25; −0.1	20%	0.006	0.240	94.6%	0.029	0.150	94.1%
B4	0.25; −0.1	5%	0.052	0.568	96.0%	0.057	0.131	92.5%
B5	0.45; −0.15	60%	0.0003	0.243	95.5%	0.127	0.188	89.6%
B6	0.45; −0.15	40%	−0.0004	0.290	96.3%	0.108	0.181	89.5%
B7	0.45; −0.15	20%	0.023	0.475	95.5%	0.118	0.175	89.5%
B8	0.45; −0.15	5%	−2.521	2835.575	85.8%	0.478	0.302	61.8%
B9	0.65; −0.2	60%	0.160	10.717	93.2%	0.337	0.257	73.9%
B10	0.65; −0.2	40%	0.463	17.321	93.4%	0.313	0.259	77.1%
B11	0.65; −0.2	20%	−3.048	602.223	89.2%	0.445	0.345	70.0%

Table 5.25: Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is concentration of sodium, scenarios B1 - B11.

Effect of SBP on sodium with selection according to the outcome

Selection	Selection	Selection	Inverse probability weighted MR			Simple MR		
scenario	correlation	probability	Bias	SE	Coverage	Bias	SE	Coverage
C1	−0.1	60%	−0.003	0.185	95.9%	0.002	0.168	95.7%
C2	−0.1	40%	0.009	0.190	95.5%	0.013	0.163	95.1%
C3	−0.1	20%	−0.002	0.205	95.9%	0.004	0.156	95.6%
C4	−0.1	5%	−0.001	0.307	95.7%	0.017	0.131	95.9%
C5	−0.25	60%	0.001	0.217	94.4%	0.011	0.151	95.4%
C6	−0.25	40%	0.006	0.243	95.4%	0.015	0.145	95.8%
C7	−0.25	20%	0.021	0.400	95.1%	0.024	0.121	94.6%
C8	−0.25	5%	−3.815	1356.891	87.6%	0.028	0.051	92.5%
C9	−0.5	60%	−0.006	0.462	95.4%	0.022	0.122	94.6%
C10	−0.5	40%	0.010	0.946	95.2%	0.025	0.099	93.8%
C11	−0.5	20%	−0.490	26.227	87.9%	0.034	0.053	89.5%

Table 5.26: Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is concentration of sodium, scenarios C1 - C11.

Effect of SBP on creatinine with selection according to the exposure

Selection	Selection	Selection	Inverse probability weighted MR			Simple MR		
scenario	correlation	probability	Bias	SE	Coverage	Bias	SE	Coverage
A1	0.25	60%	0.0003	0.024	95.8%	−0.005	0.023	94.3%
A2	0.25	40%	−0.0003	0.024	95.7%	−0.008	0.023	94.0%
A3	0.25	20%	0.0002	0.026	94.9%	−0.013	0.022	90.7%
A4	0.25	5%	0.001	0.037	95.6%	−0.022	0.021	83.9%
A5	0.45	60%	0.001	0.026	94.6%	−0.010	0.027	93.8%
A6	0.45	40%	0.001	0.029	95.1%	−0.014	0.026	92.1%
A7	0.45	20%	0.002	0.039	96.4%	−0.020	0.027	89.2%
A8	0.45	5%	0.123	122.685	93.8%	−0.035	0.049	92.3%
A9	0.65	60%	0.001	0.097	95.4%	−0.015	0.036	93.4%
A10	0.65	40%	0.002	0.067	96.7%	−0.022	0.038	92.8%
A11	0.65	20%	−0.026	12.743	93.9%	−0.036	0.053	90.2%

Table 5.27: Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is concentration of creatinine, scenarios A1 - A11.

Effect of SBP on creatinine with selection according to the exposure and the outcome

Selection	Selection	Selection	Inverse probability weighted MR			Simple MR		
scenario	correlation	probability	Bias	SE	Coverage	Bias	SE	Coverage
B1	0.25; −0.1	60%	−0.001	0.026	95.9%	−0.001	0.022	95.7%
B2	0.25; −0.1	40%	0.001	0.028	95.0%	−0.002	0.021	95.8%
B3	0.25; −0.1	20%	−0.0004	0.034	94.7%	−0.007	0.019	93.9%
B4	0.25; −0.1	5%	0.009	0.090	96.1%	−0.009	0.015	91.3%
B5	0.45; −0.15	60%	0.001	0.034	95.9%	0.009	0.024	94.1%
B6	0.45; −0.15	40%	0.002	0.041	96.0%	0.006	0.023	95.7%
B7	0.45; −0.15	20%	0.002	0.069	96.2%	0.002	0.021	95.5%
B8	0.45; −0.15	5%	1.176	1576.086	89.2%	0.043	0.031	72.7%
B9	0.65; −0.2	60%	0.048	1.955	94.6%	0.034	0.032	81.5%
B10	0.65; −0.2	40%	0.021	0.682	94.5%	0.028	0.031	85.7%
B11	0.65; −0.2	20%	0.035	6.032	92.4%	0.045	0.039	80.1%

Table 5.28: Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is concentration of creatinine, scenarios B1 - B11.

Effect of SBP on creatinine with selection according to the outcome

Selection	Selection	Selection	Inverse probability weighted MR			Simple MR		
scenario	correlation	probability	Bias	SE	Coverage	Bias	SE	Coverage
C1	−0.1	60%	−0.0004	0.025	97.3%	−0.001	0.021	97.1%
C2	−0.1	40%	−0.001	0.026	95.6%	−0.001	0.021	95.3%
C3	−0.1	20%	0.001	0.029	95.6%	−0.001	0.019	95.7%
C4	−0.1	5%	0.001	0.054	95.6%	−0.004	0.016	96.1%
C5	−0.25	60%	−0.0001	0.034	95.2%	−0.003	0.019	94.2%
C6	−0.25	40%	0.002	0.043	96.0%	−0.005	0.017	95.5%
C7	−0.25	20%	0.008	0.078	94.8%	−0.004	0.014	94.4%
C8	−0.25	5%	0.165	21.081	86.5%	−0.004	0.004	83.3%
C9	−0.5	60%	0.008	0.105	96.6%	−0.005	0.015	93.9%
C10	−0.5	40%	0.010	0.351	94.4%	−0.004	0.011	93.2%
C11	−0.5	20%	−0.110	33.069	86.9%	−0.004	0.005	84.5%

Table 5.29: Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is concentration of creatinine, scenarios C1 - C11.

Effect of SBP on CAD with selection according to the exposure

Selection scenario	Selection correlation	Selection probability	Inverse probability weighted MR			Simple MR		
			Bias	SE	Coverage	Bias	SE	Coverage
A1	0.25	60%	−0.00001	0.001	96.9%	−0.0001	0.001	96.8%
A2	0.25	40%	−0.00001	0.001	96.4%	−0.0001	0.001	95.6%
A3	0.25	20%	0.00003	0.001	95.1%	−0.0004	0.001	94.3%
A4	0.25	5%	0.0002	0.002	96.2%	−0.001	0.001	90.8%
A5	0.45	60%	−0.00006	0.001	95.3%	0.0002	0.002	95.0%
A6	0.45	40%	0.0001	0.001	95.8%	0.0001	0.002	96.0%
A7	0.45	20%	0.0001	0.002	96.5%	0.0001	0.002	96.5%
A8	0.45	5%	0.0004	0.620	98.6%	0.002	0.003	97.1%
A9	0.65	60%	0.0003	0.002	97.7%	0.001	0.002	92.4%
A10	0.65	40%	0.0004	0.005	96.0%	0.001	0.002	92.4%
A11	0.65	20%	0.008	36.778	97.4%	0.003	0.003	87.9%

Table 5.30: Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is coronary heart disease (CAD), scenarios A1 - A11.

Effect of SBP on CAD with selection according to the exposure and the outcome

Selection scenario	Selection correlation	Selection probability	Inverse probability weighted MR			Simple MR		
			Bias	SE	Coverage	Bias	SE	Coverage
B1	0.25	60%	0.0001	0.001	95.4%	0.001	0.002	94.4%
B2	0.25	40%	0.00003	0.001	95.2%	0.001	0.002	92.5%
B3	0.25	20%	−0.00003	0.001	96.0%	0.002	0.002	86.2%
B4	0.25	5%	0.006	1.618	75.7%	−0.016	0.001	0%
B5	0.45	60%	0.00001	0.010	96.0%	−0.0003	0.002	95.2%
B6	0.45	40%	−0.00002	0.001	97.1%	−0.002	0.002	83.1%
B7	0.45	20%	0.0001	0.001	97.3%	−0.008	0.002	11.2%
B8	0.45	5%	0.003	0.964	80.0%	−0.018	0.002	0%
B9	0.65	60%	0.001	0.901	96.1%	−0.002	0.003	89.1%
B10	0.65	40%	0.0002	0.149	94.7%	−0.007	0.003	43.0%
B11	0.65	20%	0.005	7.104	80.2%	−0.013	0.003	6.4%

Table 5.31: Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is coronary heart disease (CAD), scenarios B1 - B11.

Effect of SBP on CAD with selection according to the outcome

Selection	Selection	Inverse probability weighted MR			Simple MR		
scenario	probability	Bias	SE	Coverage	Bias	SE	Coverage
C1	60%	0.00004	0.001	95.6%	0.001	0.001	86.5%
C2	40%	0.00001	0.001	96.3%	0.004	0.002	45.5%
C3	20%	0.00001	0.001	95.1%	0.008	0.002	4.7%
C4	5%	0.0001	0.001	96.2%	0.008	0.002	3.0%
C5	60%	0.00002	0.001	95.3%	0.002	0.002	77.6%
C6	40%	0.00003	0.001	95.8%	0.004	0.002	32.2%
C7	20%	0.00004	0.001	97.0%	0.008	0.002	1.9%
C8	5%	0.00006	0.001	95.5%	0.005	0.002	26.4%
C9	60%	0.00007	0.001	95.4%	0.003	0.002	65.2%
C10	40%	−0.00003	0.001	96.2%	0.005	0.002	23.9%
C11	20%	0.00004	0.001	95.0%	0.008	0.003	3.5%
C12	5%	−0.013	18.275	94.0%	−0.003	0.001	33.8%

Table 5.32: Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is coronary heart disease (CAD), scenarios C1 - C12.

Effect of SBP on stroke with selection according to the exposure

Selection	Selection	Selection	Inverse probability weighted MR			Simple MR		
scenario	correlation	probability	Bias	SE	Coverage	Bias	SE	Coverage
A1	0.25	60%	0.000001	0.001	95.3%	0.00002	0.001	95.9%
A2	0.25	40%	−0.00001	0.001	96.1%	0.00002	0.001	96.3%
A3	0.25	20%	−0.000002	0.001	94.7%	0.00002	0.001	95.3%
A4	0.25	5%	−0.00001	0.001	96.6%	0.0001	0.001	96.1%
A5	0.45	60%	0.00001	0.001	96.3%	0.0001	0.001	95.4%
A6	0.45	40%	0.000005	0.0001	96.1%	0.0002	0.001	96.3%
A7	0.45	20%	0.00005	0.001	96.9%	0.0003	0.001	95.0%
A8	0.45	5%	0.001	0.428	99.2%	0.002	0.002	87.1%
A9	0.65	60%	0.0001	0.001	97.6%	0.001	0.001	93.7%
A10	0.65	40%	0.0002	0.008	98.5%	60.001	0.001	93.9%
A11	0.65	20%	−0.001	5.205	98.1%	0.002	0.002	91.3%

Table 5.33: Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is stroke, scenarios A1 - A11.

Effect of SBP on stroke with selection according to the exposure and the outcome

Selection	Selection	Selection	Inverse probability weighted MR			Simple MR		
scenario	correlation	probability	Bias	SE	Coverage	Bias	SE	Coverage
B1	0.25	60%	−0.00002	0.001	96.2%	0.0003	0.001	95.9%
B2	0.25	40%	−0.00002	0.001	95.4%	0.001	0.002	93.2%
B3	0.25	20%	0.00002	0.0004	96.8%	0.001	0.002	94.8%
B4	0.25	5%	−0.005	18.915	79.0%	−0.012	0.001	0%
B5	0.45	60%	0.00002	0.001	96.3%	−0.00002	0.001	96.2%
B6	0.45	40%	0.00005	0.001	96.0%	−0.0003	0.001	94.9%
B7	0.45	20%	0.00002	0.0004	97.9%	−0.003	0.002	62.5%
B8	0.45	5%	−0.032	7.205	77.4%	−0.016	0.001	0%
B9	0.65	60%	0.0001	0.002	96.7%	−0.0005	0.002	95.3%
B10	0.65	40%	0.0001	0.002	97.7%	−0.002	0.002	85.3%
B11	0.65	20%	−0.003	5.212	76.4%	−0.011	0.003	5.9%

Table 5.34: Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is stroke, scenarios B1 - B11.

Effect of SBP on stroke with selection according to the outcome

Selection	Selection	Inverse probability weighted MR			Simple MR		
scenario	probability	Bias	SE	Coverage	Bias	SE	Coverage
C1	60%	0.00002	0.001	95.5%	0.001	0.001	91.6%
C2	40%	0.00001	0.001	96.3%	0.001	0.001	74.6%
C3	20%	0.00001	0.0004	96.2%	0.004	0.001	21.7%
C4	5%	0.00003	0.0003	97.2%	0.009	0.002	0.1%
C5	60%	−0.00001	0.001	95.8%	0.001	0.001	90.6%
C6	40%	0.00002	0.001	97.0%	0.002	0.001	64.7%
C7	20%	0.00002	0.0004	95.4%	0.004	0.002	14.1%
C8	5%	0.00001	0.0004	96.6%	0.009	0.002	0.7%
C9	60%	0.00003	0.001	96.3%	0.001	0.001	83.8%
C10	40%	0.00003	0.001	96.4%	0.002	0.001	55.4%
C11	20%	0.00001	0.0004	97.1%	0.005	0.002	9.9%
C12	5%	0.00003	0.0004	96.5%	0.007	0.002	1.4%

Table 5.35: Bias, standard error and coverage for inverse probability weighted MR estimator and standard MR estimator where the outcome is stroke, scenarios C1 - C12.

5.7 Stratified Nonlinear MR under Selection

5.7.1 Inverse probability weighting for stratified nonlinear MR

We take one further step to consider selection bias when estimating non-linear causal effects using the widely applied stratified nonlinear MR approach, motivated by recent findings that some applied results obtained from stratified nonlinear MR might suffer from selection bias as discussed in section 5.1 and Chapter 2 (section 2.9).

Because stratum-specific estimates are essentially standard MR estimates (though performed on strata of the study sample), it is intuitive that selection will bias stratum-specific estimator in the same way as it biases the standard MR estimator. Likewise, adjusting for measured covariates that are not colliders can reduce but cannot eliminate selection bias. All of our arguments in section 5.3 equally apply for stratum-specific MR estimators.

But how does IPW change stratum-specific estimates? One might think that Theorem 8 also holds for applying inverse probability weighted MR within each stratum. Unfortunately, this is not true if the stratification is performed on the selected sample instead of the overall population. Weighting the data within each stratum cannot recover the true stratification that would have been performed on the overall population had there been no selection. This is equivalent to say that IPW cannot create pseudo-population that is representative of the true population within each stratum. Specifically, consider an example where an individual will be stratified into stratum 2 if the stratification is performed on the overall population. When the stratification is performed on the selected sample, however, this individual who should be in stratum 2 could have been stratified into stratum 1 because some other neighbouring individuals are not selected. IPW can only weight this individual within stratum 1 but cannot move him/her to stratum 2.

More concretely, consider the following toy example where the true population consists of six

individuals:

$$\{Z_1 = 0.9, X_1 = 3.14; Z_2 = 0.95, X_2 = 3.33; Z_3 = 1.01, X_3 = 3.82; \\ Z_4 = 1.05, X_4 = 4.02; Z_5 = 1.5, X_5 = 4.9; Z_6 = 1.6, X_6 = 5.05\}$$

and we want to divide this population into two strata. We will use the doubly-ranked stratification method for illustration but our conclusion also holds for the residual stratification method.

Doubly-ranked stratification will first construct three pre-strata based on the instrument values:

$$\text{pre-stratum 1 :}\{Z_1 = 0.9, X_1 = 3.14; Z_2 = 0.95, X_2 = 3.33\}$$
$$\text{pre-stratum 2 :}\{Z_3 = 1.01, X_3 = 3.82; Z_4 = 1.05, X_4 = 4.02\}$$
$$\text{pre-stratum 3 :}\{Z_5 = 1.5, X_5 = 4.9; Z_6 = 1.6, X_6 = 5.05\}$$

It then groups individuals with the smaller exposure value in each pre-stratum into stratum 1, and group individuals with the larger exposure value in each pre-stratum into stratum 2:

$$\text{stratum 1 :}\{Z_1 = 0.9, X_1 = 3.14; Z_3 = 1.01, X_3 = 3.82; Z_5 = 1.5, X_5 = 4.9\}$$
$$\text{stratum 2 :}\{Z_2 = 0.95, X_2 = 3.33; Z_4 = 1.05, X_4 = 4.02; Z_6 = 1.6, X_6 = 5.05\}$$

Now consider a selection mechanism that selects individuals 2, 3, 5 and 6 but misses individuals 1 and 4. The selected sample is

$$\{Z_2 = 0.95, X_2 = 3.33; Z_3 = 1.01, X_3 = 3.82; Z_5 = 1.5, X_5 = 4.9; Z_6 = 1.6, X_6 = 5.05\}$$

Doubly-ranked stratification performed on the selected sample will first construct two pre-strata based on the instrument values:

$$\text{pre-stratum 1 :}\{Z_2 = 0.95, X_2 = 3.33; Z_3 = 1.01, X_3 = 3.82\}$$
$$\text{pre-stratum 2 :}\{Z_5 = 1.5, X_5 = 4.9; Z_6 = 1.6, X_6 = 5.05\}$$

and gives the following stratification:

$$\text{stratum 1 :}\{Z_2 = 0.9, X_1 = 3.14; Z_5 = 1.5, X_5 = 4.9\}$$
$$\text{stratum 2 :}\{Z_3 = 1.01, X_3 = 3.82; Z_6 = 1.6, X_6 = 5.05\}$$

Comparing the stratifications on the overall population and on the selected sample, we see that two individuals end up in the wrong stratum. Individual 2 should be in stratum 2 in the true population but is stratified into stratum 1 in the selected sample. Oppositely, individual 3 should be in stratum 1 in the true population but is stratified into stratum 2 in the selected sample. This incorrect stratum membership cannot be dealt with using IPW, even if the selection model is fully known.

However, if for some reason we know the true stratification that would have been performed on the overall population without selection, we can then put selected individuals in correct strata. I.e. we form:

$$\text{correct stratum 1 :}\{Z_3 = 1.01, X_3 = 3.82; Z_5 = 1.5, X_5 = 4.9\}$$
$$\text{correct stratum 2 :}\{Z_2 = 0.95, X_2 = 3.33; Z_6 = 1.6, X_6 = 5.05\}$$

and performing inverse probability weighted MR within above correct strata will eliminate selection bias for stratum-specific estimates under moderate selection. This is because with correct stratification, within each stratum we have the standard selection problem so Theorem 8 applies.

It is possible to know the true stratification if, for example, the data is only missing for the outcome but not for the exposure and the instrument. This can happen, for example, with scenarios A and D in which selection does not depend on the outcome (so the data is missing at random). Because stratification does not require outcome data, we can stratify the population using the full exposure and instrument data and then perform IPW with the selected sample in correct strata. On the contrary, it will not be possible to know the correct

stratum memberships if either the exposure or the instrument (or both) contains missing values.

We should also mention that we might be able to weight the data directly before forming the pre-strata. It is then possible that doubly-ranked stratification performed on the weighted dataset might reduce selection bias for stratified nonlinear MR. The problem is that it is not clear how we can directly weight the data using inverse probability weights. We leave it for future research to investigate the feasibility of this approach.

5.7.2 Simulation study

We further illustrate our arguments through a simulation study. The simulation follows exactly the same design and parameterisation as in section 5.5.1, except we use stratified nonlinear MR instead of standard MR. Similar to previous chapters, we adopt the doubly-ranked stratification method (implemented by the SUMnlmr package in R (Mason and Burgess, 2022)), though the results will be similar if we choose to use residual stratification as the constant genetic effect assumption is satisfied in our simulation design. Same as in previous chapters, we choose the number of strata to be 10. Because we let the exposure have a constant causal effect with magnitude of 1, the true causal effect in each stratum is 1. For simplicity, we only present results in scenario A as results in other scenarios will be similar.

In scenarios where the selection probability $\in \{60\%, 40\%, 20\%\}$, we set the full sample size to be $n = 1000,000$; in scenarios where the selection probability is 5%, we set the full sample size to be $n = 2500,000$. Among all individuals who are selected, we draw a random sample of size $N = 100,000$. In this way, we keep the sample size in each stratum of the selected sample to be $10,000$ so that the results are comparable to the results of simulation studies for the standard linear MR in section 5.5.2. For all scenarios, results are aggregated and averaged across $D = 1,000$ simulated datasets.

Tables 5.36 - 5.57 present and compare bias, mean estimated standard error, coverage of the 95% confidence interval and size of each stratum for inverse probability weighted stratum-specific estimates and standard doubly-ranked stratum-specific estimates with and without covariate adjustment under different degrees of selection in scenario A. We show inverse probability weighted stratum-specific estimates with both correct strata formed using the full exposure and instrument data and incorrect strata formed using the selected sample. For the covariate adjusted stratum-specific estimates and simple stratum-specific estimates, we show results with incorrect strata formed using the selected sample only, as this is what happens if we directly apply stratified nonlinear MR to datasets that suffer from selection bias. We use robust standard errors for all scenarios.

We firstly note that inverse probability weighted MR performed on correct strata eliminates selection bias in most strata when the selection effect is mild or moderate as expected. The size of the correct stratum increases from stratum 1 to stratum 10 because of the way we define the selection function. Recall in section 5.5.1 we have defined the selection function for scenario A to be

$$\text{logit}(\pi_i) = \delta_0 + \delta_X \widetilde{X}_i$$

where δ_0 and δ_X are given in Table 5.2. Since δ_X are all positive, the selection probability increases as the exposure value increases, i.e. individuals with higher exposure values are more likely to be selected. For scenarios where selection is moderate, such as scenarios A6, A7, A9 and A10, there is still some bias in the first (and perhaps also the second) stratum for correctly stratified inverse probability weighted MR since selection is more severe for lower tails of the exposure distribution, but biases are largely eliminated in other strata where the stratum sizes are bigger.

However, considerable bias exists for inverse probability weighted MR performed on correct strata in scenarios where the selection is severe, such as scenarios A8 and A10. For strata at the lower tails of the exposure distribution, stratum sizes are very small and therefore biases for inverse probability weighted stratum-specific estimates are huge, exceeding the biases for

inverse probability weighted overall MR estimates under the same degrees of selection as presented in Table 5.6. This also corroborates our argument in section 5.4.1 that bias exists under severe selection when we use IPW for standard linear MR, as individuals with small exposure values are rarely selected. Standard errors in lower strata are also very large but the confidence intervals still under-cover the truth. In addition, we note that biases also exist in strata at the upper tails of the exposure distribution where the stratum sizes are much bigger, though the magnitudes of the biases are much smaller than those in lower strata. This demonstrates that severe selection induces bias for all strata when we perform inverse probability weighted MR with true stratification, not just for strata where the selection probabilities are small.

On the other hand, IPW does not help to mitigate selection bias for all selection scenarios when the stratification is performed using the selected sample. The stratum size is same for all strata as doubly-ranked stratification imposes equal stratum size. With incorrect strata, considerable bias exists for inverse probability weighted stratum-specific estimates in almost all 10 strata for all scenarios A1 - A11, similar to simple stratum-specific estimates without accounting for selection. When selection becomes more severe, the inverse probability weighted stratum-specific estimator exhibits greater bias and hugely under-covered confidence intervals. This confirms our arguments in section 5.7.1 that IPW cannot reduce selection bias for stratum-specific MR unless the stratification is performed on the overall population. We also note that the coverages are better in stratum 1 for inverse probability weighted stratum-specific estimates because of relatively large standard errors. This is again due to the way we specify the selection function, which imposes small selection probabilities for lower tails of the exposure distribution. However, the confidence intervals in stratum 1 still under-cover the truth except in scenarios A1 - A4 where selection is mild, because the stratum membership still differs from the true stratum membership in the overall population.

As for simple stratum-specific MR estimates, we note that selection bias exists for all 10 strata, and the magnitude of the bias can differ across strata. For example, in scenario A4,

the bias increases from stratum 1 to stratum 10 while in scenario A9, the bias decreases from stratum 1 to stratum 10. In other scenarios, the bias can increase first and then decrease, or varies up and down across strata. This shows that different selection model will affect stratum-specific estimates in different ways, likely due to the fact that different selection mechanisms will result in different stratum memberships as discussed in section 5.7.1. The average bias across strata increases as the selection becomes more severe, and is roughly of the same magnitude as the bias for standard linear MR presented in Table 5.7 under the same degree of selection. Covariate adjustment reduces bias and improves coverage in all 10 strata for all selection scenarios, but the improvement is fairly limited as expected.

Scenario A1								
	Correctly stratified IPW-MR				IPW-MR			
Stratum	Bias	SE	Coverage	Stratum size	Bias	SE	Coverage	Stratum size
1	−0.001	0.056	95.1%	6877	−0.008	0.048	96.4%	10000
2	0.001	0.051	94.7%	7988	−0.012	0.048	95.0%	10000
3	0.004	0.049	94.4%	8721	−0.013	0.048	94.4%	10000
4	0.001	0.047	94.6%	9308	−0.013	0.048	94.2%	10000
5	−0.001	0.046	93.5%	9840	−0.014	0.048	93.4%	10000
6	−0.001	0.045	95.4%	10331	−0.011	0.048	94.0%	10000
7	−0.0004	0.044	95.0%	10839	−0.013	0.048	93.2%	10000
8	−0.001	0.043	95.6%	11356	−0.013	0.048	93.1%	10000
9	0.003	0.041	93.8%	11952	−0.010	0.047	96.1%	10000
10	−0.003	0.040	95.1%	12788	−0.010	0.047	94.4%	10000

Table 5.36: Bias, standard error and coverage for inverse probability weighted stratum-specific MR estimates with correct strata formed using the full exposure and instrument data and incorrect strata formed using the selected sample, in scenario A1 where selection correlation = 0.25 and selection probability = 60%.

Scenario A1								
	Covariate adjusted MR				Simple MR			
Stratum	Bias	SE	Coverage	Stratum size	Bias	SE	Coverage	Stratum size
1	−0.010	0.047	95.8%	10000	−0.014	0.048	96.2%	10000
2	−0.011	0.047	94.3%	10000	−0.015	0.048	93.9%	10000
3	−0.013	0.047	94.0%	10000	−0.016	0.048	93.7%	10000
4	−0.013	0.047	94.2%	10000	−0.016	0.048	93.4%	10000
5	−0.013	0.047	93.7%	10000	−0.016	0.048	93.4%	10000
6	−0.011	0.047	94.3%	10000	−0.013	0.048	94.0%	10000
7	−0.012	0.047	94.0%	10000	−0.015	0.048	92.8%	10000
8	−0.012	0.047	93.1%	10000	−0.015	0.048	93.0%	10000
9	−0.010	0.047	96.2%	10000	−0.013	0.048	95.4%	10000
10	−0.011	0.047	94.7%	10000	−0.014	0.048	94.4%	10000

Table 5.37: Bias, standard error and coverage for stratum-specific MR estimates with and without covariate adjustment with incorrect strata formed using the selected sample, in scenario A1 where selection correlation = 0.25 and selection probability = 60%.

Scenario A2								
	Correctly stratified IPW-MR				IPW-MR			
Stratum	Bias	SE	Coverage	Stratum size	Bias	SE	Coverage	Stratum size
1	−0.002	0.063	95.5%	5739	−0.007	0.049	96.0%	10000
2	0.001	0.055	95.6%	7013	−0.011	0.048	95.2%	10000
3	0.002	0.052	94.8%	7932	−0.012	0.048	93.4%	10000
4	−0.001	0.049	95.1%	8736	−0.015	0.048	94.2%	10000
5	0.003	0.047	94.8%	9490	−0.013	0.048	94.4%	10000
6	0.002	0.045	94.7%	10249	−0.014	0.048	93.3%	10000
7	−0.001	0.043	94.9%	11051	−0.013	0.048	94.3%	10000
8	−0.003	0.042	95.0%	11949	−0.015	0.048	94.9%	10000
9	0.002	0.040	94.0%	13072	−0.014	0.048	93.8%	10000
10	0.0005	0.038	96.3%	14769	−0.011	0.048	95.4%	10000

Table 5.38: Bias, standard error and coverage for inverse probability weighted stratum-specific MR estimates with correct strata formed using the full exposure and instrument data and incorrect strata formed using the selected sample, in scenario A2 where selection correlation = 0.25 and selection probability = 40%.

Scenario A2								
	Covariate adjusted MR				Simple MR			
Stratum	Bias	SE	Coverage	Stratum size	Bias	SE	Coverage	Stratum size
1	−0.009	0.047	95.8%	10000	−0.013	0.048	94.9%	10000
2	−0.011	0.047	95.2%	10000	−0.014	0.048	94.2%	10000
3	−0.012	0.047	93.7%	10000	−0.015	0.048	92.7%	10000
4	−0.014	0.047	94.0%	10000	−0.018	0.048	94.0%	10000
5	−0.013	0.047	94.5%	10000	−0.016	0.048	94.1%	10000
6	−0.013	0.047	94.1%	10000	−0.017	0.048	92.9%	10000
7	−0.012	0.047	94.6%	10000	−0.015	0.048	93.9%	10000
8	−0.015	0.047	94.2%	10000	−0.018	0.048	94.6%	10000
9	−0.014	0.047	94.3%	10000	−0.017	0.048	93.9%	10000
10	−0.013	0.047	94.4%	10000	−0.016	0.048	94.1%	10000

Table 5.39: Bias, standard error and coverage for stratum-specific MR estimates with and without covariate adjustment with incorrect strata formed using the selected sample, in scenario A2 where selection correlation = 0.25 and selection probability = 40%.

Scenario A3								
	Correctly stratified IPW-MR				IPW-MR			
Stratum	Bias	SE	Coverage	Stratum size	Bias	SE	Coverage	Stratum size
1	−0.002	0.078	94.6%	4014	−0.012	0.052	95.2%	10000
2	0.001	0.065	94.0%	5378	−0.012	0.050	94.1%	10000
3	−0.001	0.058	95.2%	6503	−0.016	0.050	94.7%	10000
4	0.003	0.054	93.9%	7581	−0.016	0.049	94.6%	10000
5	0.002	0.050	95.0%	8666	−0.018	0.049	94.8%	10000
6	0.0001	0.047	94.0%	9853	−0.018	0.049	93.5%	10000
7	−0.00003	0.044	94.9%	11197	−0.015	0.049	94.1%	10000
8	0.0005	0.041	96.1%	12844	−0.015	0.049	93.3%	10000
9	0.002	0.038	95.2%	15059	−0.016	0.049	93.4%	10000
10	−0.001	0.034	94.9%	18903	−0.012	0.049	93.9%	10000

Table 5.40: Bias, standard error and coverage for inverse probability weighted stratum-specific MR estimates with correct strata formed using the full exposure and instrument data and incorrect strata formed using the selected sample, in scenario A3 where selection correlation = 0.25 and selection probability = 20%.

Scenario A3								
	Covariate adjusted MR				Simple MR			
Stratum	Bias	SE	Coverage	Stratum size	Bias	SE	Coverage	Stratum size
1	−0.012	0.047	93.9%	10000	−0.015	0.048	94.4%	10000
2	−0.012	0.047	94.1%	10000	−0.016	0.048	94.0%	10000
3	−0.015	0.047	94.4%	10000	−0.019	0.048	93.7%	10000
4	−0.015	0.047	94.2%	10000	−0.019	0.048	93.8%	10000
5	−0.017	0.047	95.0%	10000	−0.021	0.048	94.4%	10000
6	−0.017	0.047	93.6%	10000	−0.021	0.049	93.3%	10000
7	−0.014	0.048	94.2%	10000	−0.019	0.049	92.7%	10000
8	−0.015	0.048	94.4%	10000	−0.019	0.049	93.0%	10000
9	−0.017	0.048	92.4%	10000	−0.022	0.049	92.7%	10000
10	−0.016	0.048	93.6%	10000	−0.021	0.049	92.7%	10000

Table 5.41: Bias, standard error and coverage for stratum-specific MR estimates with and without covariate adjustment with incorrect strata formed using the selected sample, in scenario A3 where selection correlation = 0.25 and selection probability = 20%.

Scenario A4								
	Correctly stratified IPW-MR				IPW-MR			
Stratum	Bias	SE	Coverage	Stratum size	Bias	SE	Coverage	Stratum size
1	−0.011	0.200	94.9%	1055	−0.014	0.074	94.6%	10000
2	0.002	0.131	94.8%	1880	−0.018	0.061	93.6%	10000
3	0.004	0.104	95.1%	2791	−0.021	0.059	93.9%	10000
4	−0.001	0.087	95.1%	3875	−0.032	0.058	91.9%	10000
5	−0.002	0.074	95.6%	5233	−0.029	0.058	92.3%	10000
6	0.001	0.064	95.5%	7007	−0.033	0.058	92.1%	10000
7	−0.002	0.055	94.6%	9471	−0.034	0.058	91.3%	10000
8	0.001	0.047	94.4%	13211	−0.036	0.058	90.2%	10000
9	−0.001	0.039	95.7%	19763	−0.037	0.058	89.3%	10000
10	0.002	0.030	96.0%	35714	−0.033	0.057	89.7%	10000

Table 5.42: Bias, standard error and coverage for inverse probability weighted stratum-specific MR estimates with correct strata formed using the full exposure and instrument data and incorrect strata formed using the selected sample, in scenario A4 where selection correlation = 0.25 and selection probability = 5%.

Scenario A4								
	Covariate adjusted MR				Simple MR			
Stratum	Bias	SE	Coverage	Stratum size	Bias	SE	Coverage	Stratum size
1	−0.014	0.048	95.0%	10000	−0.020	0.050	93.8%	10000
2	−0.020	0.049	94.0%	10000	−0.026	0.050	92.5%	10000
3	−0.023	0.049	92.7%	10000	−0.030	0.051	91.5%	10000
4	−0.030	0.050	90.6%	10000	−0.038	0.051	87.8%	10000
5	−0.028	0.051	91.8%	10000	−0.036	0.052	88.9%	10000
6	−0.034	0.051	90.6%	10000	−0.043	0.053	89.4%	10000
7	−0.036	0.052	89.6%	10000	−0.046	0.053	86.3%	10000
8	−0.041	0.053	87.3%	10000	−0.052	0.054	83.7%	10000
9	−0.044	0.054	87.3%	10000	−0.056	0.055	82.9%	10000
10	−0.049	0.055	85.5%	10000	−0.063	0.057	78.2%	10000

Table 5.43: Bias, standard error and coverage for stratum-specific MR estimates with and without covariate adjustment with incorrect strata formed using the selected sample, in scenario A4 where selection correlation = 0.25 and selection probability = 5%.

Scenario A5								
	Correctly stratified IPW-MR				IPW-MR			
Stratum	Bias	SE	Coverage	Stratum size	Bias	SE	Coverage	Stratum size
1	−0.002	0.087	94.6%	4016	−0.040	0.063	90.4%	10000
2	−0.002	0.065	94.3%	5907	−0.053	0.058	86.6%	10000
3	−0.003	0.056	93.9%	7373	−0.057	0.057	83.7%	10000
4	0.002	0.051	95.3%	8638	−0.056	0.056	84.3%	10000
5	0.001	0.047	95.6%	9780	−0.053	0.056	85.5%	10000
6	0.001	0.045	94.3%	10846	−0.052	0.055	86.1%	10000
7	−0.0004	0.042	95.4%	11851	−0.051	0.055	85.1%	10000
8	−0.001	0.041	94.0%	12840	−0.046	0.054	87.3%	10000
9	−0.0003	0.039	95.5%	13830	−0.042	0.053	86.8%	10000
10	−0.0004	0.037	95.0%	14919	−0.035	0.052	90.6%	10000

Table 5.44: Bias, standard error and coverage for inverse probability weighted stratum-specific MR estimates with correct strata formed using the full exposure and instrument data and incorrect strata formed using the selected sample, in scenario A5 where selection correlation = 0.45 and selection probability = 60%.

Scenario A5								
	Covariate adjusted MR				Simple MR			
Stratum	Bias	SE	Coverage	Stratum size	Bias	SE	Coverage	Stratum size
1	−0.054	0.056	84.7%	10000	−0.070	0.058	78.5%	10000
2	−0.055	0.056	84.8%	10000	−0.070	0.058	79.0%	10000
3	−0.055	0.056	84.3%	10000	−0.070	0.058	77.8%	10000
4	−0.052	0.056	85.4%	10000	−0.067	0.057	78.7%	10000
5	−0.050	0.055	85.5%	10000	−0.063	0.057	80.6%	10000
6	−0.047	0.055	87.0%	10000	−0.061	0.056	81.6%	10000
7	−0.046	0.054	87.3%	10000	−0.059	0.056	82.9%	10000
8	−0.041	0.054	88.1%	10000	−0.053	0.055	83.5%	10000
9	−0.038	0.053	88.3%	10000	−0.049	0.054	84.0%	10000
10	−0.032	0.052	90.8%	10000	−0.043	0.054	89.6%	10000

Table 5.45: Bias, standard error and coverage for stratum-specific MR estimates with and without covariate adjustment with incorrect strata formed using the selected sample, in scenario A5 where selection correlation = 0.45 and selection probability = 60%.

Scenario A6

	Correctly stratified IPW-MR				IPW-MR			
Stratum	Bias	SE	Coverage	Stratum size	Bias	SE	Coverage	Stratum size
1	−0.003	0.112	93.6%	2718	−0.043	0.068	93.0%	10000
2	0.005	0.080	94.0%	4334	−0.054	0.061	85.2%	10000
3	−0.003	0.066	94.4%	5804	−0.057	0.059	85.2%	10000
4	0.001	0.058	94.0%	7263	−0.060	0.059	83.2%	10000
5	−0.002	0.052	95.1%	8753	−0.059	0.058	82.6%	10000
6	0.0002	0.047	94.8%	10323	−0.062	0.058	81.3%	10000
7	0.002	0.043	94.9%	12016	−0.058	0.057	82.1%	10000
8	−0.002	0.040	93.7%	13883	−0.060	0.057	79.9%	10000
9	−0.0001	0.037	96.0%	16057	−0.054	0.056	83.0%	10000
10	−0.001	0.034	94.2%	18849	−0.045	0.055	87.5%	10000

Table 5.46: Bias, standard error and coverage for inverse probability weighted stratum-specific MR estimates with correct strata formed using the full exposure and instrument data and incorrect strata formed using the selected sample, in scenario A6 where selection correlation = 0.45 and selection probability = 40%.

Scenario A6

	Covariate adjusted MR				Simple MR			
Stratum	Bias	SE	Coverage	Stratum size	Bias	SE	Coverage	Stratum size
1	−0.049	0.056	85.7%	10000	−0.065	0.057	79.6%	10000
2	−0.056	0.056	80.5%	10000	−0.072	0.058	74.0%	10000
3	−0.055	0.057	84.5%	10000	−0.073	0.058	76.7%	10000
4	−0.058	0.057	83.3%	10000	−0.074	0.058	75.0%	10000
5	−0.056	0.057	82.5%	10000	−0.071	0.058	76.2%	10000
6	−0.057	0.057	83.0%	10000	−0.072	0.058	76.2%	10000
7	−0.053	0.057	84.8%	10000	−0.069	0.058	78.6%	10000
8	−0.056	0.056	82.7%	10000	−0.072	0.058	75.3%	10000
9	−0.051	0.056	84.9%	10000	−0.066	0.057	78.5%	10000
10	−0.044	0.055	88.1%	10000	−0.058	0.057	82.8%	10000

Table 5.47: Bias, standard error and coverage for stratum-specific MR estimates with and without covariate adjustment with incorrect strata formed using the selected sample, in scenario A6 where selection correlation = 0.45 and selection probability = 40%.

Scenario A7								
	Correctly stratified IPW-MR				IPW-MR			
Stratum	Bias	SE	Coverage	Stratum size	Bias	SE	Coverage	Stratum size
1	−0.015	0.199	93.9%	1257	−0.041	0.084	91.9%	10000
2	0.0001	0.122	94.1%	2325	−0.059	0.069	87.2%	10000
3	−0.003	0.095	95.6%	3515	−0.066	0.066	83.8%	10000
4	0.003	0.078	95.0%	4905	−0.072	0.065	80.9%	10000
5	0.001	0.067	93.9%	6572	−0.076	0.064	77.9%	10000
6	0.001	0.057	94.8%	8636	−0.078	0.063	76.8%	10000
7	0.001	0.049	95.0%	11253	−0.082	0.063	75.7%	10000
8	−0.0001	0.043	95.3%	14707	−0.077	0.062	76.9%	10000
9	−0.0002	0.036	95.3%	19558	−0.076	0.061	74.4%	10000
10	−0.001	0.030	94.8%	27272	−0.068	0.059	81.0%	10000

Table 5.48: Bias, standard error and coverage for inverse probability weighted stratum-specific MR estimates with correct strata formed using the full exposure and instrument data and incorrect strata formed using the selected sample, in scenario A7 where selection correlation = 0.45 and selection probability = 20%.

Scenario A7								
	Covariate adjusted MR				Simple MR			
Stratum	Bias	SE	Coverage	Stratum size	Bias	SE	Coverage	Stratum size
1	−0.053	0.056	86.5%	10000	−0.070	0.058	78.9%	10000
2	−0.062	0.058	82.2%	10000	−0.078	0.060	76.1%	10000
3	−0.068	0.059	80.7%	10000	−0.087	0.061	71.6%	10000
4	−0.071	0.060	79.2%	10000	−0.091	0.062	69.7%	10000
5	−0.075	0.061	76.9%	10000	−0.095	0.062	68.0%	10000
6	−0.075	0.061	76.6%	10000	−0.097	0.063	66.6%	10000
7	−0.079	0.061	74.0%	10000	−0.102	0.063	63.6%	10000
8	−0.075	0.061	78.3%	10000	−0.096	0.063	66.9%	10000
9	−0.075	0.061	76.6%	10000	−0.097	0.063	65.3%	10000
10	−0.070	0.061	81.2%	10000	−0.092	0.063	71.1%	10000

Table 5.49: Bias, standard error and coverage for stratum-specific MR estimates with and without covariate adjustment with incorrect strata formed using the selected sample, in scenario A7 where selection correlation = 0.45 and selection probability = 20%.

Scenario A8

	Correctly stratified IPW-MR				IPW-MR			
Stratum	Bias	SE	Coverage	Stratum size	Bias	SE	Coverage	Stratum size
1	−3.433	2.05	5.6%	2	0.213	75.7	83.5%	10000
2	−4.035	1108	38.3%	4	−0.575	1.002	73.9%	10000
3	−0.485	144	73.0%	12	−0.677	0.291	32.0%	10000
4	−3.319	2732	81.7%	45	−0.743	0.238	11.6%	10000
5	0.573	384	84.2%	151	−0.757	0.212	5.1%	10000
6	−0.290	31.3	82.2%	507	−0.728	0.194	4.0%	10000
7	−0.114	3.49	84.1%	1726	−0.689	0.178	2.0%	10000
8	−0.095	2.10	85.1%	6017	−0.620	0.162	2.9%	10000
9	−0.024	0.247	87.5%	21047	−0.532	0.145	3.7%	10000
10	−0.003	0.108	92.4%	70492	−0.411	0.125	7.4%	10000

Table 5.50: Bias, standard error and coverage for inverse probability weighted stratum-specific MR estimates with correct strata formed using the full exposure and instrument data and incorrect strata formed using the selected sample, in scenario A8 where selection correlation = 0.45 and selection probability = 5%.

Scenario A8

	Covariate adjusted MR				Simple MR			
Stratum	Bias	SE	Coverage	Stratum size	Bias	SE	Coverage	Stratum size
1	−0.604	0.185	8.3%	10000	−0.788	0.197	1.6%	10000
2	−0.702	0.207	7.9%	10000	−0.905	0.216	1.0%	10000
3	−0.753	0.214	5.4%	10000	−0.963	0.223	0.9%	10000
4	−0.743	0.212	5.6%	10000	−0.949	0.221	0.5%	10000
5	−0.712	0.206	6.3%	10000	−0.911	0.213	1.6%	10000
6	−0.658	0.195	8.3%	10000	−0.843	0.202	1.6%	10000
7	−0.603	0.181	7.7%	10000	−0.774	0.188	0.6%	10000
8	−0.527	0.164	9.6%	10000	−0.677	0.170	1.9%	10000
9	−0.442	0.145	12.1%	10000	−0.569	0.151	3.0%	10000
10	−0.329	0.121	21.3%	10000	−0.430	0.128	6.8%	10000

Table 5.51: Bias, standard error and coverage for stratum-specific MR estimates with and without covariate adjustment with incorrect strata formed using the selected sample, in scenario A8 where selection correlation = 0.45 and selection probability = 5%.

Scenario A9								
	Correctly stratified IPW-MR				IPW-MR			
Stratum	Bias	SE	Coverage	Stratum size	Bias	SE	Coverage	Stratum size
1	−0.016	0.310	92.2%	1577	−0.139	0.314	87.4%	10000
2	−0.016	0.135	95.0%	3527	−0.165	0.093	59.1%	10000
3	−0.001	0.090	95.0%	5684	−0.174	0.083	47.1%	10000
4	−0.001	0.068	94.4%	7862	−0.166	0.078	43.4%	10000
5	0.0002	0.055	95.5%	9928	−0.150	0.074	48.5%	10000
6	0.001	0.047	94.7%	11778	−0.143	0.072	47.4%	10000
7	0.002	0.042	95.1%	13339	−0.128	0.069	53.2%	10000
8	−0.001	0.039	94.9%	14589	−0.110	0.066	60.7%	10000
9	0.001	0.037	95.1%	15530	−0.095	0.063	68.4%	10000
10	−0.002	0.036	93.9%	16186	−0.078	0.060	74.0%	10000

Table 5.52: Bias, standard error and coverage for inverse probability weighted stratum-specific MR estimates with correct strata formed using the full exposure and instrument data and incorrect strata formed using the selected sample, in scenario A9 where selection correlation = 0.65 and selection probability = 60%.

Scenario A9								
	Covariate adjusted MR				Simple MR			
Stratum	Bias	SE	Coverage	Stratum size	Bias	SE	Coverage	Stratum size
1	−0.186	0.087	42.5%	10000	−0.244	0.090	21.8%	10000
2	−0.179	0.085	43.9%	10000	−0.231	0.088	24.9%	10000
3	−0.167	0.082	48.1%	10000	−0.215	0.084	27.8%	10000
4	−0.152	0.078	52.3%	10000	−0.195	0.081	33.4%	10000
5	−0.133	0.075	57.1%	10000	−0.171	0.077	39.1%	10000
6	−0.123	0.072	59.6%	10000	−0.159	0.074	41.1%	10000
7	−0.108	0.069	66.2%	10000	−0.139	0.071	48.4%	10000
8	−0.092	0.065	69.7%	10000	−0.118	0.067	57.8%	10000
9	−0.078	0.062	78.4%	10000	−0.100	0.064	65.5%	10000
10	−0.063	0.059	81.7%	10000	−0.082	0.061	72.1%	10000

Table 5.53: Bias, standard error and coverage for stratum-specific MR estimates with and without covariate adjustment with incorrect strata formed using the selected sample, in scenario A9 where selection correlation = 0.65 and selection probability = 60%.

Scenario A10

	Correctly stratified IPW-MR				IPW-MR			
Stratum	Bias	SE	Coverage	Stratum size	Bias	SE	Coverage	Stratum size
1	−0.421	157.016	91.2%	623	−0.136	0.357	90.2%	10000
2	−0.003	0.268	93.5%	1622	−0.193	0.111	57.1%	10000
3	−0.006	0.148	94.5%	3050	−0.201	0.096	45.5%	10000
4	0.004	0.106	94.5%	4950	−0.209	0.090	36.7%	10000
5	0.001	0.080	94.7%	7314	−0.206	0.086	34.0%	10000
6	−0.002	0.062	93.9%	10108	−0.195	0.083	35.2%	10000
7	−0.001	0.050	95.1%	13240	−0.175	0.079	41.6%	10000
8	−0.002	0.041	93.4%	16539	−0.162	0.076	41.5%	10000
9	−0.0005	0.035	94.3%	19796	−0.144	0.072	48.3%	10000
10	−0.001	0.031	96.7%	22757	−0.117	0.067	59.2%	10000

Table 5.54: Bias, standard error and coverage for inverse probability weighted stratum-specific MR estimates with correct strata formed using the full exposure and instrument data and incorrect strata formed using the selected sample, in scenario A10 where selection correlation = 0.65 and selection probability = 40%.

Scenario A10

	Covariate adjusted MR				Simple MR			
Stratum	Bias	SE	Coverage	Stratum size	Bias	SE	Coverage	Stratum size
1	−0.192	0.088	40.3%	10000	−0.249	0.092	22.0%	10000
2	−0.209	0.091	34.8%	10000	−0.268	0.094	18.5%	10000
3	−0.203	0.090	38.6%	10000	−0.261	0.093	18.1%	10000
4	−0.198	0.089	38.7%	10000	−0.253	0.091	21.4%	10000
5	−0.189	0.086	41.5%	10000	−0.241	0.089	21.5%	10000
6	−0.175	0.083	46.6%	10000	−0.223	0.086	26.6%	10000
7	−0.153	0.080	51.7%	10000	−0.198	0.082	32.7%	10000
8	−0.140	0.076	53.5%	10000	−0.180	0.079	36.2%	10000
9	−0.122	0.072	60.3%	10000	−0.157	0.074	43.1%	10000
10	−0.097	0.067	69.6%	10000	−0.127	0.069	55.7%	10000

Table 5.55: Bias, standard error and coverage for stratum-specific MR estimates with and without covariate adjustment with incorrect strata formed using the selected sample, in scenario A10 where selection correlation = 0.65 and selection probability = 40%.

Scenario A11

	Correctly stratified IPW-MR				IPW-MR			
Stratum	Bias	SE	Coverage	Stratum size	Bias	SE	Coverage	Stratum size
1	0.077	407.296	81.4%	27	−0.042	37.271	89.2%	10000
2	0.164	68.919	84.6%	124	−0.413	0.347	58.4%	10000
3	18.064	102191.100	87.2%	389	−0.484	0.181	21.7%	10000
4	−0.084	1.067	87.1%	1007	−0.490	0.156	10.6%	10000
5	−0.028	0.344	88.4%	2306	−0.470	0.142	8.6%	10000
6	−0.002	0.222	91.0%	4828	−0.451	0.131	7.4%	10000
7	0.003	0.148	91.7%	9300	−0.406	0.121	7.9%	10000
8	−0.004	0.097	92.9%	16512	−0.353	0.112	11.1%	10000
9	0.004	0.060	94.3%	26736	−0.304	0.102	14.0%	10000
10	−0.002	0.034	95.4%	38772	−0.237	0.090	23.1%	10000

Table 5.56: Bias, standard error and coverage for inverse probability weighted stratum-specific MR estimates with correct strata formed using the full exposure and instrument data and incorrect strata formed using the selected sample, in scenario A11 where selection correlation = 0.65 and selection probability = 20%.

Scenario A11

	Covariate adjusted MR				Simple MR			
Stratum	Bias	SE	Coverage	Stratum size	Bias	SE	Coverage	Stratum size
1	−0.443	0.149	14.0%	10000	−0.580	0.157	3.6%	10000
2	−0.496	0.157	10.3%	10000	−0.636	0.164	1.9%	10000
3	−0.493	0.156	9.9%	10000	−0.633	0.161	1.6%	10000
4	−0.462	0.150	12.0%	10000	−0.592	0.155	3.3%	10000
5	−0.426	0.142	14.8%	10000	−0.547	0.147	4.0%	10000
6	−0.396	0.133	15.0%	10000	−0.506	0.137	3.6%	10000
7	−0.348	0.123	19.0%	10000	−0.446	0.127	6.0%	10000
8	−0.295	0.112	25.4%	10000	−0.380	0.116	9.3%	10000
9	−0.249	0.101	31.6%	10000	−0.321	0.105	12.4%	10000
10	−0.188	0.088	41.5%	10000	−0.245	0.092	21.7%	10000

Table 5.57: Bias, standard error and coverage for stratum-specific MR estimates with and without covariate adjustment with incorrect strata formed using the selected sample, in scenario A11 where selection correlation = 0.65 and selection probability = 20%.

5.8 Discussion and Concluding Remarks

In this chapter, we have firstly considered selection bias in standard instrumental variables and Mendelian randomisation analyses, and examined how covariate adjustment and inverse probability weighting can help reduce or eliminate selection bias. We discuss in section 5.3 using simple path analysis that adjusting for measured covariates that are not colliders can reduce but cannot eliminate selection bias because of the existence of unmeasured confounders, and we demonstrate this result through simulation studies in section 5.5.

We then discuss in section 5.4.1 how we can apply inverse probability weighting (IPW) in the context of MR. Since it is well-established in the literature that IPW will induce bias when the propensity score model (i.e. the selection model in the context of selection bias) depends on missing or unmeasured data, we focus on the ideal situation where the data is missing at random and the selection model can be correctly specified and precisely estimated. We provide asymptotic guarantees of the inverse probability weighted MR estimator under regularity conditions, and note that its actual performance depends critically on the degree of selection. We argue that under severe selection, IPW is not guaranteed to reduce bias in finite sample, even if the selection model is fully specified and the sample size is reasonably large. More importantly, for any given sample size, it is always possible to specify a selection model such that IPW cannot eliminate selection bias. We demonstrate through simulation studies in section 5.5 and real data examples from UK Biobank in section 5.6 that while inverse probability weighted MR can eliminate selection bias under moderate selection, considerable finite sample bias exists under severe selection where the selection correlation is high and the selection probability is low (although the bias still vanishes as $n \longrightarrow \infty$, the sample size needs to be extremely large for the convergence to happen). Since the participation rate of UK Biobank is as small as 5.5%, it is likely that the selection effect for many traits is severe and the validity of results obtained from inverse probability weighted MR using participation weights will be subject to concerns.

Since small selection probabilities under severe selection will render extremely large weights, we discuss two potential remedial approaches including using stabilised weights and weight trimming in section 5.4.2. Both methods are not satisfactory, because stabilised weights are not different from normal inverse probability weights for MR, and weight trimming reduces the variance at the expense of increasing bias. Therefore, large weights seem to be unavoidable for inverse probability weighted MR. This is confirmed in our simulation studies and UK Biobank examples that standard errors can be exceedingly large under severe selection, rendering the estimates effectively useless. This is another disadvantage of using IPW to correct for selection bias in the context of MR.

Moreover, we note in section 5.4.3 that the commonly used first-order approximated standard error fails to provide valid confidence intervals for inverse probability weighted MR, and we suggest always using robust standard errors given by the asymptotic distributions in Theorem 8. Comparisons of the naive standard error and the robust standard error are also illustrated through simulations in section 5.5.

We then further consider selection bias in stratified nonlinear MR analyses and discuss whether covariate adjustment and inverse probability weighting can be applied to reduce or eliminate selection bias for stratum-specific MR estimates in section 5.7. We argue that covariate adjustment reduces selection bias for stratum-specific estimates as it similarly does for standard MR estimates, but IPW can no longer provide representative pseudo-population within strata even if the selection model is fully known, unless we can stratify on the overall population instead of the selected sample. This is because stratifying on the selected sample gives stratum memberships that are different from the true stratum memberships which would have been obtained from stratifying on the overall population had there been no selection. Since correct stratification requires full observations of the exposure and the instruments, the usefulness of IPW to correct for selection bias for stratum-specific estimates is limited. Even if we know the correct stratum memberships and perform IPW within correct strata, severe selection will still induce biases for inverse probability weighted stratum

specific estimates just as it would for inverse probability weighted overall MR estimates. We corroborate our argument through simulation studies in section 5.7.2.

Of course, there are a number of limitations in our analyses that can be considered in future research. The first limitation is that we did not provide a method for performing IPW with logistic regressions, but instead impose a linear model for the binary outcomes in our real data examples. Although various previous research (e.g. Hughes et al., 2019; Schoeler et al., 2023) also adopt this approach when analysing selection bias in the context of MR, this is not ideal as linear modelling is subject to model misspecification with binary outcomes. In addition, causal odds ratio is a common target of estimation, especially when testing for causal null hypothesis (Burgess and Thompson, 2021). We will discuss using IPW for binary outcomes in the context of MR in more details in Chapter 6 (section 6.2.4).

Another limitation is that we did not discuss exactly how to estimate the propensity score model for the inverse probability weights, as we assume for simplicity that the selection model is fully known. In practice, the selection model can depend on a large number of covariates so some type of model selection will be required. Alternatively, the selection model can depend on many exposures if the researchers are interested in the effect of a high-dimensional exposure. Although Theorem 8 allows multiple exposures and multiple covariates in general, IPW suffers from the curse of dimensionality and quickly becomes inefficient when the number of variables gets large. We will expand a bit more on this in Chapter 6 (section 6.2.5).

In summary, we have shown that when the selection model is correctly specified, inverse probability weighted MR can eliminate selection bias under linear structural equation models when the selection effect is moderate, but it often fails to do so when the selection is severe. As for stratified nonlinear MR, if we can only stratify on the selected sample instead of the full population, inverse probability weighting cannot mitigate selection bias for stratum-specific estimates even when the selection is only moderate. Consequently, researchers need to be careful of selection bias when running MR analyses, and cautions are required when using

IPW to correct for selection bias. Sensitivity analyses, such as negative controls analysis, on the possibility of selection bias are therefore recommended.

Chapter 6

Summary and Future Prospects

6.1 Summary of the Thesis

Recall that the title of the thesis is *Towards more accurate causal inference with instrumental variables and Mendelian randomisation analyses*. How much have we achieved towards this goal so far?

In Chapter 1, we start from the fundamental languages and concepts for causality, and introduce the three core IV assumptions that enable us to use instrumental variables to validly estimate causal effects even in the presence of unmeasured confounding as long as we are willing to accept additional modelling assumptions. We then introduce the classic 2SLS and GMM estimators that are widely applicable for general IV problems, and the IVW estimator that is more specific to MR applications (especially MR with summarised data). These classic IV methods identify either the average causal effect or the compliers average causal effect depending on whether we assume the linearity assumption or the monotonicity assumption.

The traditional modelling assumptions only enable us to identify some average version of the causal effects, which is often insufficient. How can we confidently assign a drug, for example, for patients with various levels of the exposure when the causal effect we estimate is only an average across the exposure distribution? This motivates us to relax the linearity assumption and impose a nonlinear structural model instead, hoping to estimate more accurate nonlinear causal effects so that, for example, we can identify threshold for relevant interventions. In Chapter 2, we firstly propose the nonlinear causal model that we are interested in, and then critically review some of the recent and important developments for estimating nonlinear causal effects using IVs. We start from the more traditional approach of using pre-defined basis functions to model or to approximate the nonlinear structural function (with either nonlinear 2SLS or control functions), and then review more recent approaches of using machine-learning algorithms to extend 2SLS or GMM nonparametrically. We also introduce methods developed in the context of MR, including polynomial MR and stratified nonlinear MR. We compare the methods through a simulation study, and observe that while the flexible machine-learning methods perform better when estimating a rather complicated nonlinear function, the MR approaches are more suitable for estimating simpler functions.

Another situation in which the machine-learning based IV methods can struggle is when the instruments are weak. This is a common problem for two-stage type methods that only the effect over instrument-predicted exposures can be reliably estimated, and the range of instrument-predicted exposure values can be very limited with weak instruments. Typically genetic variants only explain a very small proportion of the variance in the exposure distribution, so the nonparametric 2SLS or GMM methods using machine-learning algorithms are not suitable for MR applications. However, can we combine the flexibility of machine-learning methods with the stratification philosophy to estimate complex structural functions in the context of MR? This motivates us to develop nonparametric MR, a nonparametric extension to stratified nonlinear MR that requires only the rank-preserving assumption and can flexibly characterise complex exposure-outcome relationships by incorporating nonpara-

metric regressions that can be implemented using machine-learning methods such as kernel ridge regression. In Chapter 3, we develop the nonparametric MR framework with discrete IVs, which can be applied when using a single genetic variant as the instrument. In Chapter 4, we extend the framework to develop nonparametric MR with continuous IVs, which can be used in common situations where multiple genetic variants are combined to create a continuous genetic risk score. We note that although our nonparametric MR framework was developed when we keep MR applications in mind, our method equally applies to a wide range of IV problems outside of MR for which the instrument could be much stronger.

So far we have been assuming that the instrumental variables are all valid, but there are many situations in which the IV assumptions could be violated. One such example is the presence of selection bias. When individuals are selected into the study based on their exposure values, outcome values or values of other correlated variables, the no unmeasured confounding for the instrument and/or exclusion restriction assumptions will be violated, invalidating our IV analyses. In Chapter 5, we start from the linear IV model and consider covariate adjustment and inverse probability weighting (IPW) as potential solutions to reduce or eliminate selection bias. We firstly confirm that covariate adjustment can reduce but cannot eliminate selection bias. We then observe that while IPW could eliminate selection bias with moderate selection in IV analyses, it is not able to do so with severe selection even when the selection model is fully known and can be correctly estimated. This result has not been fully appreciated in existing literatures, as most existing literatures only mention the failure of IPW when the selection model is misspecified. We also observe that the usual first order delta approximation for standard errors in MR is appropriate for inverse probability weighted MR, and we provide a formula for calculating the robust standard error based on the asymptotic distribution of the inverse probability weighted MR estimator.

We then further consider in Chapter 5 how selection bias affects stratum-specific estimates from doubly-ranked stratification for stratified nonlinear MR. This is motivated by the ongoing debate around the stratified nonlinear MR approach that some implausible results

are generated by both residual and doubly-ranked stratifications in several negative control outcome analyses (e.g. Wade et al., 2023; Hamilton et al., 2024). We contribute to this debate by showing that when selection affects the overall MR estimate, the stratum-specific estimates will also be biased. In addition, we demonstrate that IPW cannot reduce selection bias for stratum-specific estimates even with only moderate selection unless we know the correct stratum memberships without selection. This deepens our understanding of stratified nonlinear MR, and suggests cautions for interpreting results from either residual or doubly-ranked nonlinear MR analyses.

There is still much more to be done in fields relevant to nonlinear IV and MR methods. In the following, we discuss several directions for future research. Some of the ideas have already been touched upon in previous chapters, and we summarise them below. We also expand on a couple of issues that are relevant but do not directly fit into previous chapters.

6.2 Prospects for Future Research

6.2.1 Bias from weak instruments for nonparametric 2SLS and GMM methods

We have mentioned that the machine-learning methods based on generalising 2SLS or GMM nonparametrically are prone to bias from weak instrument. However, we have not empirically explored the role of instrument strength for these methods in Chapter 2. It will be an important direction for future research to empirically understand weak instrument bias for nonparametric 2SLS and GMM methods, as this will guide instrument choices and study designs in practice.

6.2.2 Further development of nonparametric MR

As mentioned in Chapter 3 and 4, there are a number of directions we can take to develop our nonparametric MR framework further. The first direction is to consider other nonparametric regression methods, as local linear kernel estimation is shown to be suboptimal and kernel ridge regression is computationally demanding. The second direction is to explore alternative hyperparameter tuning procedures than standard cross-validation to speed up computation and improve performance. In addition, although we have defended our use of cross-validation by arguing that the nonparametric regressions do not directly estimate causal quantities, there could be other methods that are more suitable for IV problems. For example, Peysakhovich and Eckles (2018) propose an adaptive causal cross validation procedure for a regularised version of 2SLS for estimating causal effects using large collections of experimental data, and it might be beneficial to consider similar principles in our method. For nonparametric MR with continuous IVs, another important direction is to explore alternative models for estimating quantiles of the counterfactual exposure distributions that are more suitable for discrete or truncated exposures.

6.2.3 Causal effect heterogeneity with stratified nonlinear MR and nonparametric MR

A general direction for future research with nonlinear IV and MR methods is to explore heterogenous nonlinear causal relationships between the exposure and the outcome by considering nonlinear interactions between the exposure and other exogenous covariates. On the one hand, stratified nonlinear MR includes covariates by adjusting for them within stratum and ignoring their estimated coefficients when estimating the structural function via fractional polynomial regression. Although this approach can be interpreted as estimating structural function conditional on the covariates, it does not help in understanding effect

heterogeneity. To observe the effect of the covariates, researchers will need to stratify on the distribution of the covariates that they are interested in. For example, Burgess et al. (2024) stratified on sex and estimated nonlinear causal effects for men and women separately. However, stratification will not be feasible for high-dimensional covariates. On the other hand, we did not consider covariate adjustment for nonparametric MR in Chapter 3 and 4. Although we have defended this by arguing that covariate adjustment is not essential for valid causal estimates, again this means that we will have to stratify on the covariates to understand effect heterogeneity. Therefore, it will be an interesting future direction to develop a general framework that incorporates covariates nonlinearly and explicitly.

6.2.4 Using inverse-probability weighting to account for selection bias with a binary outcome in the context of MR

As mentioned in Chapter 5, an important direction for future research to account for selection bias is to incorporate IPW for binary outcome models. This has already been explored in contexts outside of IV or MR analyses. For example, in the context of survival analysis, Yu et al. (2023) propose an inverse probability weighted logistic regression estimator that adjusts for measured confounders and accounts for right-censoring. Specifically, their estimating equation is given by (with slight notational modification):

$$\sum_{r_i=1} \frac{V_i\big(Y_i - m_{ij}(\beta_j)\big)\mathbb{I}(X_i = j)}{\pi_{ij}(\alpha)\exp\big[-\Lambda_{ij}\big(\min(T_i, C_i)\big)\big]} = 0$$

where Y denotes the event, X denotes a discrete treatment taking j values, V denotes measured covariates, R denotes an indicator for censoring before treatment, T denotes event time, C denotes censoring time, $m_{ij}(\beta_j) = \text{expit}(V_i^T\beta_j)$ is the expit function (i.e. inverse logistic function), $\pi_{ij}(\alpha) = \mathbb{P}(X_i = j|V_i)$ is the propensity score function and $\Lambda_{ij}()$ is the cumulative hazard function of C_i for $X_i = j$.

However, similar approaches have not been explored for two-stage logistic regressions with

instrumental variables, so the use of weighted logistic regression to account for selection bias using IVs remains an open question. An alternative approach could be directly weighting the data to create representative pseudo-population, from which we can proceed to estimate the causal odds ratio. Either way, a potential challenge for using IPW with IVs for binary outcomes is the non-collapsible nature of the logistic model, as inverse probability weights change the distribution of the exposure.

6.2.5 High dimensional propensity score estimation

In Chapter 5, we assume the propensity score model (i.e. the selection model) is perfectly known. However, in practice, the propensity score can depend on multiple exposures and/or a very large number of covariates. Simple logistic regression becomes inefficient with large number of variables, and future work is needed to consider more sophisticated approaches for estimating the inverse probability weights in high-dimensional settings. There are recent efforts in the literature in this area, such as van der Laan (2014) who propose a "targeted IPW estimator" that incorporates data-adaptive approaches in estimating the weight function, and Ertefaie et al. (2023) who propose a nonparametric IPW estimator based on highly adaptive LASSO that achieves asymptotic efficiency. These methods are originally proposed to estimate treatment effect in the marginal structural models, and it will be an interesting future research topic to consider how to apply them in the context of IV and MR analyses.

6.2.6 Further investigation of selection bias in the context of stratified nonlinear MR

One driving motivation for the investigation of selection bias in Chapter 5 is the ongoing debate about stratified nonlinear MR as we discussed before. Nevertheless, our endeavours do not directly answer the question at the core of the debate of why we get biased results

with stratum-specific MR estimates while the overall MR estimates remain largely unbiased. In Chapter 5, we only considered scenarios where the selection is a deterministic logistic function for the whole population, and future work is needed to investigate other selection mechanisms where the selection effect might be stochastic and varies across the distribution of the exposure or the outcome. For example, a possible selection mechanism that we did not consider is when the selection depends on the exposure only at the tails but not at the middle of the exposure distribution. Subsequent investigations are also needed to consider how to eliminate such varying selection effects, especially when we do not know the variability of selection a priori.

6.2.7 Nonlinear IV and MR methods robust to invalid instruments

The no unmeasured confounding and exclusion restriction assumption for instrumental variables can often be violated in practice. Invalid IVs are common in MR analyses because of issues like pleiotropy and population stratification. Selection bias is another common example that invalidates IVs. While there are many robust IV and MR methods for linear IV models under various relaxed assumptions, such as the median method (Bowden et al., 2016), the mode-based method (Hartwig et al., 2017), the Lasso-based method (Kang et al., 2016; Windmeijer et al., 2018), and many others, how to handle invalid instruments in nonlinear causal models remains an open question.

When the exclusion restriction assumption is violated, for example, one way we can incorporate this in our nonlinear causal model is to modify our structural function into

$$Y = h(X, Z) + \epsilon_Y$$

and future work is required to develop methods that estimate h under this more general causal model.

6.3 Conclusion

In conclusion, this thesis contributes towards more accurate causal inference with instrumental variables and Mendelian randomisation analyses by firstly, reviewing and comparing recent developments of nonlinear IV methods; secondly, developing the nonparametric MR method with both discrete and continuous IVs; and thirdly, examining the usefulness of IPW in reducing selection bias for both linear and stratified nonlinear MR analyses. It is hoped that the methods, results and discussions presented in this thesis can facilitate applications of IV and MR analyses in practice and pave methodological foundations for future research.

Bibliography

Abadie, A., Angrist, J., and Imbens, G. W. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, 70(1):91–117.

Angrist, J., Imbens, G., and Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

Angrist, J. D. and Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442.

Angrist, J. D. and Pischke, J. (2009). *Mostly Harmless Econometrics: an Empiricist's Companion*. Princeton University Press.

Bacallado, S. (2019). Modern statistical methods course notes. Available at http://www.statslab.cam.ac.uk/~sb2116/modern_stat_methods.html.

Baiocchi, M., Small, D., Lorch, S., and Rosenbaum, P. (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *Journal of the American Statistical Association*, 105(492):1285–1296.

Bakhitov, E. and Singh, A. (2022). Causal gradient boosting: Boosted instrumental variables regression. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 604–605.

Banks, E., Herbert, N., Mather, T., Rogers, K., and Jorm, L. (2012). Characteristics of australian cohort study participants who do and do not take up an additional invitation to join a long-term biobank: the 45 and up study. *BMC Research Notes*, 5(1):655–656.

Bareinboim, E., Correa, J, D., Ibeling, D., and Iscard, T. (2022). On pearl's hierarchy and the foundations of causal inference. In Geffner, H., Dechter, R., and Halpern, J., editors, *Probabilistic and Causal Inference: The Works of Judea Pearl*, chapter 27. Association for Computing Machinery.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(11):463–482.

Bennett, A., Kallus, N., and Schnabel, T. (2019). Deep generalized method of moments for instrumental variable analysis. In *Proceedings of Advances in Neural Information Processing Systems 32*, pages 3564–3574.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blundell, R., Chen, X., and Kristensen, D. (2007). Semi-nonparametric iv estimation of shape-invariant Engel curves. *Econometrica*, 75(6):1613–1669.

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, 40(4):304–314.

Burgess, S. (2017). Estimating and contextualizing the attenuation of odds ratios due to non collapsibility. *Communications in Statistics - Theory and Methods*, 46(2):786–804.

Burgess, S. (2023). Violation of the constant genetic effect assumption can result in biased estimates for non-linear Mendelian randomization. *Human Heredity*, 88(1):79–90.

Burgess, S., Davies, N., Thompson, S., and EPIC-InterAct Consortium (2014). Instrumental variable analysis with a nonlinear exposure-outcome relationship. *Epidemiology*, 25(6):877–885.

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

Burgess, S., Sun, Y.-Q., Zhou, A., Buck, C., Mason, A. M., and Mai, X.-M. (2024). Body mass index and all- cause mortality in HUNT and UK Biobank studies: revised nonlinear Mendelian randomisation analyses. *BMJ Open*, 364:l1042–l1042.

Burgess, S. and Thompson, S. G. (2021). *Mendelian Randomization: methods for causal inference using genetic variants*. Chapman & Hall/CRC Interdisciplinary Statistics, 2nd edition.

Burgess, S., Thompson, S. G., and CRP CHD Genetics Collaboration. (2011). Avoiding bias from weak instruments in Mendelian randomization studies. *International Journal of Epidemiology*, 40(3):755–764.

Cai, S., Hartley, A., Mahmoud, O., Tilling, K., and Dudbridge, F. (2022). Adjusting for collider bias in genetic association studies using instrumental variable methods. *Genetic epidemiology*, 46(5-6):303–316.

Canan, C., Lesko, C., and Lau, B. (2017). Instrumental variable analyses and selection bias. *Epidemiology*, 28(3):396–398.

Canay, I. A., Santos, A., and Shaikh, A. M. (2013). On the testability of identification in some nonparametric models with endogeneity. *Econometrica*, 81(6):2535–2559.

Caponnetto, A. and De Vito, E. (2007). Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368.

Card, D. (1994). Earnings, schooling, and ability revisited. Working Paper 4832, National Bureau of Economic Research. Available at https://www.nber.org/papers/w4832.

Chen, L., Davey Smith, G., Harbord, R. M., and Lewis, S. J. (2008). Alcohol intake and blood pressure: a systematic review implementing a Mendelian randomization approach. *PLOS Medicine*, 5(3):e52–e52.

Chen, X. and Christensen, T. M. (2018). Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, 9(1):39–84.

Chernozhukov, V. and Hansen, C. (2005). An iv model of quantile treatment effects. *Econometrica*, 73(1):245–261.

Chernozhukov, V. and Hansen, C. (2006). Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics*, 132(2):491–525.

Chernozhukov, V. and Hansen, C. (2008). Instrumental variable quantile regression: A robust inference approach. *Journal of Econometrics*, 142(1):379–398.

Chernozhukov, V. and Hansen, C. (2013). Quantile models with endogeneity. *Annual Review of Economics*, 5:57–81.

Chernozhukov, V. and Hong, H. (2003). An mcmc approach to classical estimation. *Journal of Econometrics*, 115(2):393–346.

Cho, Y., Shin, S. Y., Won, S., Relton, C. L., Davey Smith, G., and Shin, M. J. (2015). Alcohol intake and cardiovascular risk factors: a Mendelian randomisation study. *Scientific Reports*, 5(1):1–13.

Cole, S. R. and Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168(6):656–664.

Darolles, S., Fan, Y., Florens, J. P., and Renault, E. (2011). Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. (2017). Training GANs with optimism. *arXiv preprint*.

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

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Royal Statistical Society*, 39(1):1–38.

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

Dudbridge, F., Allen, R. J., Sheehan, N. A., Schmidt, A. F., Lee, J. C., Jenkins, R. G., Wain, L. V., Hingorani, A. D., and Patel, R. S. (2019). Adjustment for index event bias in genome-wide association studies of subsequent events. *Nature Communications*, 10(1):1561–1561.

Ebrahim, S. and Davey Smith, G. (2008). Mendelian randomization: Can genetic epidemiology help redress the failures of observational epidemiology? *Human Genetics*, 123(1):15–33.

Ertefaie, A., Hejazi, N. S., and van der Lann, M. J. (2023). Nonparametric inverse probability weighted estimators based on the highly adaptive Lasso. *Biometrics*, 79(2):1029–1041.

Evans, R. and Didelez, V. (2015). Recovering from selection bias using marginal structure in discrete models. Paper presented at *Workshop on Advances in Causal Inference* at the 31st Conference on Uncertainty in Artificial Intelligence, Amsterdam, Netherlands. Available at http://ceur-ws.org/Vol-1504/uai2015aci_paper4.pdf.

Fry, A., Littlejohns, T. J., Sudlow, C., Doherty, N., Adamska, L., Sprosen, T., Collins, R., and Allen, N. E. (2017). Comparison of sociodemographic and health-related characteristics

of UK Biobank participants with those of the general population. *American Journal of Epidemiology*, 186(9):1026–1034.

Gill, P. E., Murray, W., and Wright, M. H. (2020). *Practical Optimization*. SIAM, Society for Industrial and Applied Mathematics.

Gkatzionis, A. and Burgess, S. (2019). Contextualizing selection bias in Mendelian randomization: how bad is it likely to be? *International Journal of Epidemiology*, 48(3):691–701.

Gkatzionis, A., Tchetgen Tchetgen, E. J., Heron, J., Northstone, K., and Tilling, K. (2024). Using instruments for selection to adjust for selection bias in Mendelian randomization. *Statistics in Medicine*, 43(22):4250–4271.

Greenland, S. (2003). Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*, 14(3):300–306.

Greenland, S. and Robins, J. M. (2009). Identifiability, exchangeability and confounding revisited. *Epidemiologic perspectives & innovations*, 6(1):4–4.

Gretton, A. (2019). Introduction to rkhs, and some simple kernel algorithms. Available at https://www.gatsby.ucl.ac.uk/~gretton/coursefiles/lecture4_introToRKHS.pdf.

Guo, Z. and Small, S. D. (2016). Control function instrumental variable estimation of nonlinear causal effect models. *Journal of Machine Learning Research*, 17:1–35.

Hamilton, F. W., Hughes, D. A., Spiller, W., Tilling, K., and Davey Smith, G. (2024). Non-linear Mendelian randomization: detection of biases using negative controls with a focus on BMI, vitamin D and LDL cholesterol. *European Journal of Epidemiology*, 39(5):451–465.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054.

Hansen, L. P., Heaton, J., and Yaron, A. (1996). Finite-sample properties of some alternative GMM estimators. *Journal of Business & Economic Statistics*, 14(3):262–280.

Hansen, L. P. and Singleton, K. J. (1982). Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica*, 50(5):1269–1286.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep IV: A flexible approach for counterfactual prediction. In *Proceedings of International Conference on Machine Learning*, pages 1414–1423.

Hartwig, F. P., Davey Smith, G., and Bowden, J. (2017). Robust inference in summary data Mendelian randomisation via the zero modal pleiotropy assumption. *International Journal of Epidemiology*, 46(6):1985–1988.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* Spriner, 2nd edition.

Hausman, J. A. (1977). Errors in variables in simultaneous equation models. *Journal of Econometrics*, 5(3):389–401.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46(6):1251–1271.

Heath, A. K., Kim, I. Y., Hodge, A. M., English, D., and Muller, D. C. (2019). Vitamin D status and mortality: a systematic review of observational studies. *International Journal of Environmental Research and Public Health*, 16:383–383.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Ecnometrica*, 47(1):153–161.

Heckman, J. J. and Vytlacil, E. J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. In *Proceedings of the National Academy of Sciences*, pages 4730–4734.

Hernán, M. A., Hernández-Diaz, S., and Robins, J. M. (2004). A structural approach to selection bias. *Epidemiology*, 15:615–625.

Hernán, M. A. and Robins, J. M. (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health*, 60(7):578–586.

Hernán, M. A. and Robins, J. M. (2024). *Causal Inference: What If*. CRC Press.

Higgins, H., Mason, A. M., Larsson, S. C., Gill, D., Langenberg, C., and Burgess, S. (2021). Estimating the population benefits of blood pressure lowering: A wide- angled Mendelian randomization study in UK Biobank. *Journal of the American Heart Association*, 10(17):e021098–e021098.

Hogan, J. W. and Lancaster, T. (2004). Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Statistical Methods in Medical Research*, 13(1):17–48.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Hughes, R. A., Davies, N. M., Davey Smith, G., and Tilling, K. (2019). Selection bias when estimating average treatment effects using one-sample instrumental variable analysis. *Epidemiology*, 30(3):350–357.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

Imbens, G. W. and Newey, W. K. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica*, 77(5):1481–1512.

Johnson, T. (2012). Efficient calculation for multi-snp genetic risk scores. presented at the American Society of Human Genetics Annual Meeting, San Francisco. Available at https://github.com/cran/gtx/blob/master/inst/doc/ashg2012.pdf.

Kang, H., Zhang, A., Cai, T. T., and Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144.

Keyes, K. M. and Westreich, D. (2019). UK Biobank, big data, and the consequences of non-representativeness. *The Lancet*, 393(10178):1297–1297.

Kook, L. and Pfister, N. (2024). Instrumental variable estimation of distributional causal effects. *arXiv preprint*.

Kress, R. (1989). *Linear Integral Equations*. Springer-Verlag.

Larsson, S. C., Burgess, S., Mason, A. M., and Michaëlsson, K. (2020). Alcohol consumption and cardiovascular disease: a Mendelian randomization study. *Circulation: Genomic and Precision Medicine*, 13(3):e002814–e002814.

Lee, B. K., L. J. and Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PLOS One*, 6(3):e18174–e18174.

Lee, S. (2007). Endogeneity in quantile regression models: A control function approach. *Journal of Econometrics*, 141(2):1131–1158.

Li, Q. and Racine, J. S. (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Liu, M., Jiang, Y., Wedow, R., Li, Y., Brazel, D. M., and Chen, F., . (2019). Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nature genetics*, 51(2):237–244.

Mahmoud, O., Dudbridge, F., Davey Smith, G., Munafo, M., and Tilling, K. (2022). A robust method for collider bias correction in conditional genome-wide association studies. *Nature Communications*, 13(1):619–619.

Mason, A. M. and Burgess, S. (2022). Software Application Profile: SUMnlmr, an R package that facilitates flexible and reproducible non-linear Mendelian randomization analyses. *International Journal of Epidemiology*, 51(6):2014–2019.

Metten, M.-A., Costet, N., Multigner, L., Viel, J.-F., and Chauvet, G. (2022). Inverse probability weighting to handle attrition in cohort studies: some guidance and a call for caution. *BMC Medical Research Methodology*, 12(1):45–55.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning*. MIT Press.

Muandet, K., Mehrjou, A., Lee, S. K., and Raj, A. (2020). Dual IV: A single stage instrumental variable regression. *arXiv preprint*.

Munafo, M. R., Tilling, K., Taylor, A. E., Evans, D. M., and Davey Smith, G. (2018). Collider scope: when selection bias can substantially influence observed associations. *International Journal of Epidemiology*, 47(1):226–235.

Nagelkerke, N., Fidler, V., Bernsen, R., and Borgdorff, M. (2000). Estimating treatment effects in randomized clinical trials in the presence of non-compliance. *Statistics in Medicine*, 19(14):1849–1864.

Nash, J. and Varadhan, R. (2011). Unifying optimization algorithms to aid software system users: optimx for R. *Journal of Statistical Software*, 43(9):1–14.

Nash, J. C. (2014). On best practice optimization methods in R. *Journal of Statistical Software*, 60(2):1–14.

Nashed, M. Z. and Wahba, G. (1974). Generalized inverses in reproducing kernel spaces: An approach to regularization of linear operator equations. *SIAM Journal on Mathematical Analysis*, 5(6):974–987.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometric*, 79(1):147–168.

Newey, W. K. (2013). Nonparametric instrumental variables estimation. *The American Economic Review*, 103(3):550–556.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.

Nocedal, J. and Wright, S. J. (1999). *Numerical Optimization*. Spriner.

Pearl, J. (2010). An introduction to causal inference. *The International Journal of Biostatistics*, 6(2):1–60.

Peters, J., Janzing, D., and Scholkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.

Peysakhovich, A. and Eckles, D. (2018). Learning causal effects from many randomized experiments using regularized instrumental variables. In *Proceedings of the 2018 World Wide Web Conference*, pages 699–707.

Press, W. H., T. S. A. V. W. T. and Flannery, B. P. (2007). *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, 3rd edition.

Richardson, T. S. and Robins, J. M. (2013). Single world intervention graphs (swigs): A unification of counterfactual and graphical approaches to causality. Working Paper Number 128. Available at http://www.csss.washington.edu/Papers/.

Robins, J. M. (1999). Marginal structural models versus structural nested models as tools for causal inference. In Halloran, E. M. and Berry, D., editors, *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 95–133. Springer-Verlag.

Robins, J. M., Hernán, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

Royston, P. and Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Journal of Applied Statistics*, 43(3):429–467.

Rubin, D, B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D, B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D, B. (1980). Comment on "Randomization analysis of experimental data: The Fisher randomization test". *Journal of the American Statistical Association*, 75(371):591–593.

Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2):147–177.

Schoeler, T., Speed, D., Porcu, E., Pirastu, N., Pingault, J. B., and Kutalik, Z. (2023). Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nature Human Behaviour*, 7(7):1216–1227.

Scholkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In Helmbold, D. and Williamson, B., editors, *Computational Learning Theory*, pages 416–426. Springer Berlin.

Seaman, S. R. and White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22(3):278–295.

Shalit, U., Johansson, F., and Sontag, D. (2016). Estimating individual treatment effect: generalization bounds and algorithms. *arXiv preprint*.

Singh, R., Sahani, M., and Gretton, A. (2019). Kernel instrumental variable regression. In *Proceedings of Advances in Neural Information Processing Systems 32*, pages 4593–4605.

Small, D. S. (2014). Commentary: Interpretation and sensitivity analysis for the localized average causal effect curve. *Epidemiology*, 25(6):886–888.

Sofianopoulou, E., Kaptoge, S. K., Afzal, S., Jiang, T., Dipender, G., and Gundersen, T. E., . (2024). Estimating dose–response relationships for vitamin D with coronary heart disease, stroke, and all-cause mortality: observational and Mendelian randomisation analyses. *The Lancet Diabetes & Endocrinology*, 12(1):e2–e11.

Splawa-Newyman, J., Dabrowska, D. M., and Speed, T. P. (1990). On the application of probability theory to agricultural experiments. Essay on principles. section 9. *Statistical Science*, 5(4):465–472.

Sriperumbudur, R., Fukumizu, K., and Lanckriet, G. (2010). On the relation between universality, characteristic kernels and rkhs embedding of measures. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, pages 773–780.

Staley, J. and Burgess, S. (2017). Semiparametric methods for estimation of a nonlinear exposure-outcome relationship using instrumental variables with application to Mendelian randomization. *Genetic Epidemiology*, 41(4):341–352.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines. Information Science and Statistics*. Springer.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., and Danesh, J. . (2015). UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):e1001779–e1001779.

Sulc, J., Sjaarda, J., and Kutalik, Z. (2022). Polynomial mendelian randomisation reveals non-linear causal effects for obesity-related traits. *HGG Advances*, 3(3):10124–10124.

Tanofsky-Kraff, M., Han, J. C., Anandalingam, K., Shomaker, L. B., Columbo, K. M., Wolkoff, L. E., Kozlosky, M., Elliott, C., Ranzenhofer, L. M., Roza, C. A., Yanovski, S. Z., and Yanovski, J. A. (2009). The FTO gene rs9939609 obesity-risk allele and loss of control over eating. *The American Journal of Clinical Nutrition*, 90(6):1483–1488.

Taylor, M., Tansey, K. E., Lawlor, D. A., Bowden, J., Evans, D. M., Davey Smith, G., and Timpson, N, J. (2017). Testing the principles of Mendelian randomization: Opportunities and complications on a genomewide scale. *bioRxiv preprint*.

Tian, H., Mason, A. M., Liu, C., and Burgess, S. (2023). Relaxing parametric assumptions for non-linear Mendelian randomization using a doubly-ranked stratification method. *PLOS Genetics*, 19(6):e1010823–e1010823.

Tudball, M. J., Bowden, J., Hughes, R. A., Ly, A., Munafo, M. R., Tilling, K., Zhao, Q., and Davey Smith, G. (2021). Mendelian randomisation with coarsened exposures. *Genetic Epidemiology*, 45(3):338–350.

van Alten, S., Domingue, B. W., Faul, J., Galama, T., and Marees, A. T. (2024). Reweighting UK Biobank corrects for pervasive selection bias due to volunteering. *International Journal of Epidemiology*, 53(3).

van der Laan, M. J. (2014). Targeted estimation of nuisance parameters to obtain valid statistical inference. *International Journal of Biostatistics*, 10(1):29–57.

VanderWeele, T. J. (2013). A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology*, 24(2):224–232.

Wade, K. H., Hamilton, F. W., Carslake, D., Sattar, N., Davey Smith, G., and Timpson, N. J. (2023). Challenges in undertaking nonlinear Mendelian randomization. *Obesity*, 31(12):2887–2890.

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

Wang, T. J., Pencina, M. J., Booth, S. L., Jacques, P. F., Ingelsson, E., Lainer, K., and et al. (2008). Vitamin D deficiency and risk of cardiovascular disease. *Circulation*, 117(6):e1010823–e1010823.

Wiatrak, M., A. S. and Nystrom, A. (2019). Stabilizing generative adversarial networks: A survey. *arXiv preprint*.

Windmeijer, F., Farbmacher, H., Davies, N., and Davey Smith, G. (2018). On the use of the Lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*, 114(527):1139–1350.

Xu, L., Chen, Y., Srinivasan, S., de Freitas, N., Doucet, A., and Gretton, A. (2021). Learning deep features in instrumental variable regression. In *Proceedings of International Conference on Learning Representation*.

Yu, Y., Zhang, M., and Mukherjee, B. (2023). An inverse probability weighted regression method that accounts for right-censoring for causal inference with multiple treatments and a binary outcome. *Statistics in Medicine*, 42(20):3699–3715.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. In *Proceedings of International Conference on Learning Representation*.

Zhao, Q. (2021). Lecture notes on causal inference. Available at https://www.statslab.cam.ac.uk/~qz280/teaching/causal-2023/notes-2021.pdf.

Appendix A

Appendix to the Thesis

A.1 Review on Kernel Methods

A.1.1 Construction of kernels

Recall the definition of an inner product.

Definition 2 (Inner Product)**.** Let $\mathcal{H}$ be a real vector space. An inner product on $\mathcal{H}$ is a map $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ that satisfies:

1. linearity: $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}} \ \ \forall \alpha_1, \alpha_2 \in \mathbb{R}, f_1, f_2, g \in \mathcal{H}$

2. symmetry: $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}} \ \ \forall f, g \in \mathcal{H}$

3. positive definiteness: $\langle f, f \rangle_{\mathcal{H}} \geq 0$ and $\langle f, f \rangle_{\mathcal{H}} = 0$ if and only if $f = 0$, $\forall f \in \mathcal{H}$

A kernel is simply an inner product between feature maps.

Definition 3 (Kernel (e.g. Steinwart and Christmann, 2008))**.** Let $\mathcal{X}$ be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *kernel* if there exists an inner product space $\mathcal{H}$ and a

feature map $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, x' \in \mathcal{X}$,

$$k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$$

Also note that a kernel must be symmetric, due to the symmetry property of the inner product, but a kernel needs not to be bilinear in general. Moreover, a kernel is positive definite in the following sense.

Definition 4 (Positive Definite Functions (e.g. **?**, **?**))**.** A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be positive definite if $\forall n \in \mathbb{N}, \alpha_1, ..., \alpha_n \in \mathbb{R}, x_1, ..., x_n \in \mathcal{X}$, we have

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0$$

We note that every inner product is a positive definite function, and so is a kernel.

Definition 5 (Kernel Matrix)**.** Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel. Let $n \in \mathbb{N}, x_1, ..., x_n \in \mathcal{X}$. A matrix K with entries $K_{ij} = k(x_i, x_j)$ is called the kernel matrix associated with k.

Because of the positive definiteness of kernels, kernel matrices are positive semi-definite. Due to this fact, we actually have an alternative characterisation/definition of kernels.

Definition 6 (Kernel (e.g., Bacallado, 2019))**.** Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a symmetric map. k is called a *kernel* if $\forall n \in \mathbb{N}, x_1, ..., x_n \in \mathcal{X}$, the matrix K with entries $K_{ij} = k(x_i, x_j)$ is positive semi-definite.

Under this alternative characterisation, our original definition of a kernel (Definition 3) can be proven as a theorem. We'll elaborate on this in appendix A.1.3.

We also have the following useful results about constructing/combining kernels (mainly taken from Gretton (2019) and Bacallado (2019)).

Theorem 9 (Sums of kernels are kernels)**.** *Let k_1, k_2 be kernels on $\mathcal{X}$ and let $\alpha \in \mathbb{R}^+$. Then $\alpha k_1, \alpha k_2, k_1 + k_2$ are all kernels on $\mathcal{X}$.*

Theorem 10 (Products of kernels are kernels)**.** *Let k_1, k_2 be kernels on $\mathcal{X}_1, \mathcal{X}_2$. Then $k_1 \times k_2$ is a kernel on $\mathcal{X}_1 \times \mathcal{X}_2$. If $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}$, then $k = k_1 k_2$ is a kernel on $\mathcal{X}$.*

Theorem 11 (Limits of kernels are kernels)**.** *Let $k_1, k_2, ..., k_n, ...$ be kernels on $\mathcal{X}$. If $\lim_{n\to\infty} k_n(x, x') = k(x, x')$ exists for all $x, x' \in \mathcal{X}$, then k is a kernel on $\mathcal{X}$.*

A.1.2 Examples of kernels

In practice, a kernel is something we choose. If we have chosen a kernel, we know there exist some feature maps (onto some feature space) for which the kernel characterises the inner product, and it is often not necessary to define the feature maps explicitly. In the following we list some of the most common kernels that may be chosen in practice.

1. *Linear kernel.* $k(x, x') = x^T x'$.

2. *Polynomial kernel.* $k(x, x') = (c + \langle x, x' \rangle)^d$, where $c \geq 0$ is some positive real number and $\langle x, x' \rangle$ is some inner product defined on $\mathcal{X}$. In particular, we can choose $k(x, x') = (1 + x^T x')^d$.

3. *Exponential kernel.* $k(x, x') = \exp(\langle x, x' \rangle)$.

4. *Gaussian kernel.*

$$k(x, x') = \exp\Big(- \frac{\|x - x'\|_2^2}{2\sigma^2} \Big)$$

Gaussian kernel is also known as *radial basis function kernel*, or simply *RBF kernel*, and is perhaps the most popular choice of kernel in practice.

A.1.3 Reproducing kernel Hilbert space (RKHS)

Recall the two alternative definitions of kernels (Definition 3 and 6). We have seen how we can start from Definition 3 and arrive at Definition 6. To see their equivalence, we'll elaborate

on how we can start from Definition 6 and arrive at Definition 3. The following is mainly taken from section 1.4.2 in Bacallado (2019).

Suppose we have a kernel k as defined in Definition 6. Define feature map $\phi : \mathcal{X} \to \mathcal{H}$ by $\phi(x) = k(\cdot, x)$. This way of writing the feature map is sometimes called the *canonical feature map*.

Let $\mathcal{H}$ be the vector space of functions that are linear combinations of the feature map. I.e.

$$\mathcal{H} = \left\{ f : f(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i),\ n \in \mathbb{N}, \alpha_1, ..., \alpha_n \in \mathbb{R} \right\}$$

This is certainly valid as $\phi(x) \in \mathcal{H}$. And we define the inner product on $\mathcal{H}$ by

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, x'_j) \tag{A.1}$$

where we write

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i),\ n \in \mathbb{N}, \alpha_1, ..., \alpha_n \in \mathbb{R} \tag{A.2}$$

$$g(\cdot) = \sum_{j=1}^{m} \beta_j k(\cdot, x'_j),\ m \in \mathbb{N}, \beta_1, ..., \beta_n \in \mathbb{R} \tag{A.3}$$

Then clearly if we take $f(\cdot) = k(\cdot, x)$ and $g(\cdot) = k(\cdot, x')$ we have

$$k(x, x') = \langle k(\cdot, x), k(\cdot, x') \rangle_{\mathcal{H}} = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}} \tag{A.4}$$

which is exactly what we have in Definition 3. However we need to make sure that (A.1) is a valid inner product and this can indeed be shown.

Note that if we take $f(\cdot)$ as (A.2) and $g(\cdot) = k(\cdot, x)$ we have

$$\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = \sum_{i=1}^{n} \alpha_i k(x_i, x) = \sum_{i=1}^{n} \alpha_i k(x, x_i) = f(x) \tag{A.5}$$

which turns out to be a very important property of the construction, known as the *reproducing property*.

Recall that a vector space is called complete when every Cauchy sequence converges and has a limit in the space, and a complete inner product space is called a *Hilbert Space*. Therefore, we can add the limits of all Cauchy sequences in $\mathcal{H}$ to make $\mathcal{H}$ a Hilbert space (with a slight abuse of notation). As a matter of fact, this "completed" $\mathcal{H}$ is a special type of Hilbert space known as the *Reproducing Kernel Hilbert Space*.

Definition 7 (Reproducing Kernel Hilbert Space (**?**))**.** Let $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$, and $\mathcal{H}$ is called a reproducing kernel Hilbert space if k satisfies:

1. $\forall x \in \mathcal{X},\ k(\cdot, x) \in \mathcal{H}$;

2. $\forall x \in \mathcal{X},\ \forall f \in \mathcal{H},\ \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$. This is known as the reproducing property.

In particular we note that, as $k(\cdot, x') \in \mathcal{H}$, we can take $f = k(\cdot, x')$ and obtain $\langle k(\cdot, x'), k(\cdot, x)\rangle_{\mathcal{H}} = k(x, x')$. Hence k, the reproducing kernel, is indeed a (positive definite) kernel. We'll state without proof the following important theorem about RKHS.

Theorem 12 (Moore-Aronszajn theorem (**?**))**.** *Every positive definite kernel k is associated with a unique RKHS $\mathcal{H}$.*

We should note that only the kernel is unique while the feature map needs not to be.

More importantly, as we saw earlier, given any kernel k, we can construct a RKHS $\mathcal{H}$, where $\mathcal{H}$ is the space of functions (and limits of Cauchy sequences) that are linear combinations of the canonical feature maps, such that k is the reproducing kernel of $\mathcal{H}$. But due to Theorem 12, the RKHS is unique, so our constructed $\mathcal{H}$ is *the* RKHS of k. This is to say that any functions in a RKHS can be written as a linear combination of feature maps:

$$f(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) \tag{A.6}$$

or the limits of Cauchy sequences, where we can let $n \to \infty$.

Let's consider the linear kernel as a simple illustrative example. Here $\mathcal{H} = \{f : f(x) = \beta^T x\}$ with $k(x, x') = x^T x'$. Given a function $f \in \mathcal{H} : f(x) = \beta^T x$, we can write

$$f(x) = k(\beta, x) = k(x, \beta) \;\; \forall x \in \mathbb{R}^p$$

Hence $f(\cdot) = k(\cdot, \beta)$ and therefore

$$\|f\|^2_{\mathcal{H}} = \langle f, f \rangle_{\mathcal{H}} = \langle k(\cdot, \beta), k(\cdot, \beta) \rangle_{\mathcal{H}} = k(\beta, \beta) = \beta^T \beta = \|\beta\|^2_2$$

A.1.4 Kernel ridge regression

We want to kernelise the usual ridge regression so that we can allow more general class of functions to fit the data. Specifically, we are interested in the following type of optimisation problem:

$$\hat{f} = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 + \lambda \|f\|^2_{\mathcal{H}} \tag{A.7}$$

where the penalisation controls the smoothness of the function. To solve this optimisation, we introduce the following very important theorem.

Theorem 13 (Representer Theorem (Scholkopf et al., 2001))**.** *let $C : \mathbb{R}^n \times \mathcal{X}^n \times \mathbb{R}^n \to \mathbb{R}$ be an arbitrary loss function, and $\Omega : [0, \infty) \to \mathbb{R}$ be strictly increasing. Let $x_1, ..., x_n \in \mathcal{X}, Y \in \mathbb{R}^n$ and $f : \mathcal{X} \to \mathbb{R}, f \in \mathcal{H}$ where $\mathcal{H}$ is a RKHS with reproducing kernel k. Then $\hat{f}$ minimises*

$$Q_1(f) = C\big(y, x_1, ..., x_n, f(x_1), ..., f(x_n)\big) + \Omega\big(\|f\|^2_{\mathcal{H}}\big) \tag{A.8}$$

over $f \in \mathcal{H}$ if and only if $\hat{f}(\cdot) = \sum_{i=1}^n \hat{\alpha}_i k(\cdot, x_i)$ where $\hat{\alpha} = (\hat{\alpha}_1, ..., \hat{\alpha}_n)^T \in \mathbb{R}^n$ minimises

$$Q_2(\alpha) = C\big(y, x_1, ..., x_n, K\alpha\big) + \Omega\big(\alpha^T K \alpha\big) \tag{A.9}$$

where K is the empirical kernel matrix with entries $K_{ij} = k(x_i, x_j), i, j = 1, ..., n$.

Sketch of proof (Bacallado, 2019). We decompose f into $f = u + v$ where $u \in U := \text{span}\{k(\cdot, x_1), ..., k(\cdot, x_n)\}$ and $v \in U^{\perp}$. We have

$$f(x_i) = \langle f, k(\cdot, x_i) \rangle_{\mathcal{H}} = \langle u + v, k(\cdot, x_i) \rangle_{\mathcal{H}} = \langle u, k(\cdot, x_i) \rangle_{\mathcal{H}} = u(x_i) \tag{A.10}$$

We also have

$$\Omega\big(\|f\|_{\mathcal{H}}^2\big) = \Omega\big(\|u\|_{\mathcal{H}}^2 + \|v\|_{\mathcal{H}}^2\big) \geq \Omega\big(\|u\|_{\mathcal{H}}^2\big) \tag{A.11}$$

(A.10) is not affected by v but $\Omega\big(\|f\|_{\mathcal{H}}^2\big)$ is not minimised if v is not zero. So the optimal f must have $v = 0$ and hence $\hat{f}(\cdot) = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$. Plug this into (A.8) we get (A.9). This completes the proof. □

Now we are ready to get a closed form solution to (A.7). We'll state the result as a theorem.

Theorem 14. *Let $x_1, ..., x_n \in \mathcal{X}, Y_1, ..., Y_n \in \mathbb{R}$ and $f : \mathcal{X} \to \mathbb{R}, f \in \mathcal{H}$ where $\mathcal{H}$ is a RKHS with reproducing kernel k. Then the solution to the optimisation problem (A.7) is given by*

$$\hat{f}(x) = \hat{\alpha}^T K_{Xx} \tag{A.12}$$

where $\hat{\alpha} = (K + n\lambda I_n)^{-1} Y$, K is the empirical kernel matrix with entries $K_{ij} = k(x_i, x_j), i, j = 1, ..., n$, and $K_{Xx} = \big(k(x, x_1), ..., k(x, x_n)\big)^T \in \mathbb{R}^n$ is the kernel vector.

Proof. By the representer theorem we know that the optimal f takes the form

$$f(\cdot) = \sum_{i=1}^n \alpha_i k(\cdot, x_i) \tag{A.13}$$

Plug (A.13) into (A.7) the objective function becomes:

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n \Big(Y_i - \sum_{j=1}^n \alpha_j k(x_i, x_j)\Big)^2 + \lambda \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(x_i, x_j) \\ &= \frac{1}{n}\|Y - K\alpha\|_2^2 + \lambda \alpha^T K \alpha \end{aligned} \tag{A.14}$$

where $\alpha = (\alpha_1, ..., \alpha_n)^T \in \mathbb{R}^n$ and K is the kernel matrix. Differentiate (A.14) w.r.t. α and set it to 0 we have:

$$\begin{aligned} &-\frac{2}{n}KY + \frac{2}{n}K^2\hat{\alpha} + 2\lambda K\hat{\alpha} \\ &= 2K\Big(-\frac{Y}{n} + (\frac{K}{n} + \lambda I_n)\hat{\alpha}\Big) = 0 \end{aligned}$$

Therefore

$$\hat{\alpha} = (K + n\lambda I_n)^{-1} Y$$

and

$$\hat{f}(x) = \sum_{i=1}^{n} \hat{\alpha}_i k(x, x_i) = \hat{\alpha}^T K_{Xx}$$

□

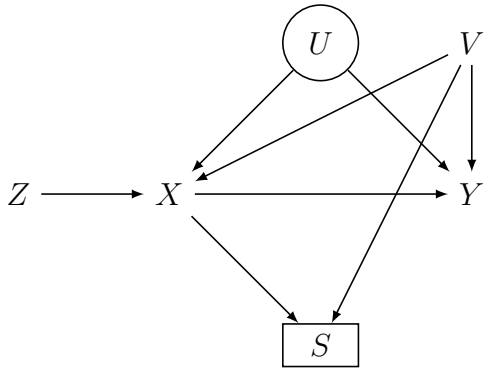
Note that, from the above theorem, we can write

$$\left(\hat{f}(x_1), ..., \hat{f}(x_n)\right)^T = K\hat{\alpha} = K(K + n\lambda I_n)^{-1} Y \tag{A.15}$$

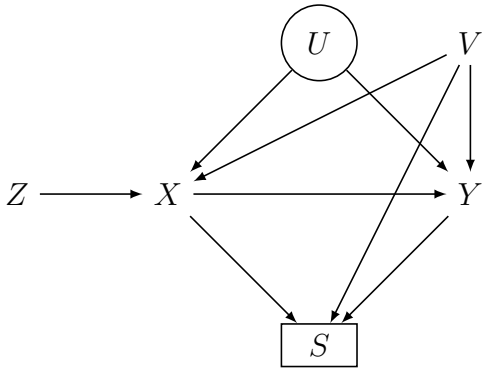
Note also that, by optimising (A.7), we are implicitly imposing constraint that $\|f\|_{\mathcal{H}}^2 \leq \infty$, which can be understood as a smoothness constraint.

In practice, there are two things we can choose: the kernel k, and the penalisation parameter λ. The kernel controls the smoothness of the class of functions we consider, while λ controls the tradeoff between smoothness and fitting error. To choose an optimal λ, one may use cross-validation on a held-out test.

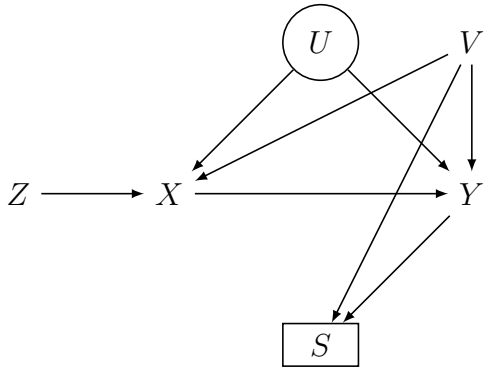
A.2 Additional Selection Mechanisms



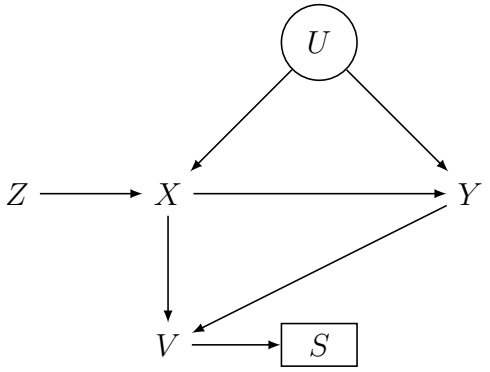
(E) selection according to both the exposure X and the covariate V



(F) selection according to the exposure X, the outcome Y and the covariate V



(G) selection according to both the outcome Y and the covariate V



(H) selection according to a downstream covariate V of both the exposure X and the outcome Y

Figure A.1: DAGs illustrating four more selection mechanisms. Z denotes the instruments, X denotes the exposures, Y denotes the outcome, V denotes measured covariates, U denotes unmeasured confounders and S denotes the selection indicator.