

Selection Bias in Mendelian Randomization

Apostolos Gkatzionis

MRC Integrative Epidemiology Unit, University of Bristol

19-2-2021

Email: apostolos.gkatzionis@bristol.ac.uk

Joint work with Paul Newcombe (BSU, Cambridge),
Steve Burgess (BSU, Cambridge), Kate Tilling (IEU, Bristol).

- 1 An Introduction to Mendelian Randomization
- 2 Selection Bias in Mendelian Randomization
 - Structure of Bias
 - Magnitude of Bias - Simulations
- 3 Adjustments for Selection Bias
 - Instruments for Selection
 - MR Inference with Instruments for Selection

- 1 An Introduction to Mendelian Randomization
- 2 Selection Bias in Mendelian Randomization
 - Structure of Bias
 - Magnitude of Bias - Simulations
- 3 Adjustments for Selection Bias
 - Instruments for Selection
 - MR Inference with Instruments for Selection

- *Does chocolate consumption increase coronary heart disease risk?*
- Intuitively, chocolate → obesity.
- But observational studies: No, it protects against CHD!

Cardiac risk factors and prevention
Original article

Habitual chocolate consumption and risk of cardiovascular disease among healthy men and women FREE

Chun Shing Kwok^{1, 2}, S Matthijs Boekholdt¹, Marleen A H Lentjes⁴, Yoon K Loke³, Robert N Luben⁴, Jessica K Yeung⁶, Nicholas J Wareham⁷, Piyo K Myint¹, Kay-Tee Khaw⁴

[Author affiliations +](#)

Abstract

Objective To examine the association between chocolate intake and the risk of future cardiovascular events.

Methods We conducted a prospective study using data from the European Prospective Investigation into Cancer (EPIC)-Norfolk cohort. Habitual chocolate intake was quantified using the baseline food frequency questionnaire (1993–1997) and cardiovascular end points were ascertained up to March 2008. A systematic review was performed to evaluate chocolate consumption and cardiovascular outcomes.

Results A total of 20 951 men and women were included in EPIC-Norfolk analysis (mean follow-up 11.3±2.8 years, median 11.9 years). The percentage of participants with coronary heart disease (CHD) in the highest and lowest quintile of chocolate consumption was 9.7% and 13.8%, and the respective rates for stroke were 3.1% and 5.4%. The multivariate adjusted HR for CHD was 0.88 (95% CI 0.77 to 1.01) for those in the top quintile of chocolate consumption (16–99 g/day) versus non-consumers of chocolate intake. The corresponding HR for stroke and cardiovascular disease (cardiovascular disease defined by the sum of CHD and stroke) were 0.77 (95% CI 0.62 to 0.97) and 0.86 (95% CI 0.76 to 0.97). The propensity score matched estimates showed a similar trend. A total of nine studies with 157 809 participants were included in the meta-analysis. Higher compared to lower chocolate consumption was associated with significantly lower CHD risk (five studies; pooled RR 0.71, 95% CI 0.56 to 0.92), stroke



Cardiac risk factors and prevention
Original research article

Chocolate consumption and risk of cardiovascular diseases: a meta-analysis of prospective studies

Yongcheng Ren¹, Yu Liu¹, Xi-Zhuo Sun¹, Bing-Yuan Wang^{1, 2}, Yang Zhao^{2, 3}, De-Chen Liu^{2, 3}, Dong-Dong Zhang², Xue-Jiao Liu², Rui-Yuan Zhang^{2, 3}, Hao-Hang Sun², Fei-Yan Liu², Xu Chen², Cheng Cheng², Lei-Lei Liu², Qiong-Gui Zhou², Ming Zhang², Dong-Sheng Hu^{1, 2}

[Author affiliations +](#)

Abstract

Objective Studies investigating the impact of chocolate consumption on cardiovascular disease (CVD) have reached inconsistent conclusions. As such, a quantitative assessment of the dose–response association between chocolate consumption and incident CVD has not been reported. We performed a systematic review and meta-analysis of studies assessing the risk of CVD with chocolate consumption.

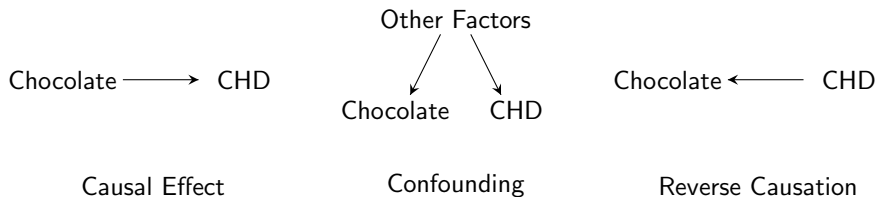
Methods PubMed and EMBASE databases were searched for articles published up to 6 June 2018. Restricted cubic splines were used to model the dose–response association.

Results Fourteen publications (23 studies including 405 304 participants and 35 093 cases of CVD) were included in the meta-analysis. The summary of relative risk (RR) per 20 g/week increase in chocolate consumption was 0.982 (95% CI 0.972 to 0.992, $P=30.4%$, $n=11$) for CVD (heart failure: 0.995 (0.981 to 1.010, $P=36.3%$, $n=5$); total stroke: 0.956 (0.932 to 0.980, $P=25.5%$, $n=7$); cerebral infarction: 0.952 (0.917 to 0.988, $P=0.0%$, $n=4$); haemorrhagic stroke: 0.931 (0.871 to 0.994, $P=0.0%$, $n=4$); myocardial infarction: 0.981 (0.964 to 0.997, $P=0.0%$, $n=3$); coronary heart disease: 0.986 (0.973 to 0.999, $n=11$). A non-linear dose–response ($P_{non-linearity}=0.001$) indicated that the most appropriate dose of chocolate consumption for reducing risk of CVD was 45 g/week (RR 0.890; 95% CI 0.849 to 0.932).



Correlation and Causation

- Observational studies can only detect correlation.
- And correlation does not imply causation!
- In particular, correlation can admit one of three explanations:



- To distinguish between these, use causal inference.

Three main approaches for causal inference:

1 Clinical Trials

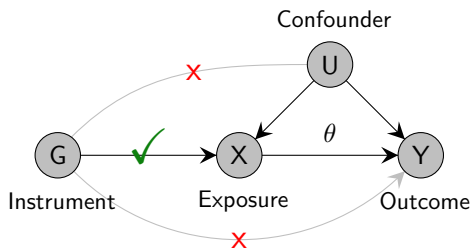
- "Gold standard" for causal inference when feasible.
- But often infeasible or unethical.
- E.g. randomization for chocolate consumption??

2 Adjustments in Observational Studies

- If all confounders are observed, add them to the model.
- But we cannot be sure that all confounders are observed.

3 Instrumental Variables Analysis.

Instrumental Variable Analysis



Idea: find a variable G that satisfies the assumptions:

- $G \rightarrow X$
- $G \perp\!\!\!\perp U \mid X$.
- $G \perp\!\!\!\perp Y \mid X, U$.

G is called an instrumental variable and can be used to assess causality.

IV Analysis - Estimation

Most common approach: Two-Stage Least Squares (2SLS).

- 1st stage: model $X \sim G$. E.g. for linear regression

$$X_i = \alpha_X + G_i^T \beta_X + \epsilon_{1i}$$

and compute fitted values \hat{X}_i .

- 2nd stage: model $Y \sim \hat{X}$

$$Y_i = \alpha_Y + \hat{X}_i \theta + \epsilon_{2i}$$

- Intuition: \hat{X} is the "component of X that is determined by G ", so $Y \sim \hat{X}$ is unconfounded.
- Generalization: Two-Stage Residual Inclusion (2SRI).

Alternative approaches exist, e.g. express as a structural equation model and use MLE.

Mendelian randomization is the use of genetic variants as instrumental variables to assess the existence of a causal relationship between exposure X and outcome Y .

- Popularized by Davey Smith & Ebrahim (2003).
- Genetic data are not affected by environmental confounders so are ideal as instruments!
- Random allocation of DNA at conception works in a similar way as randomization in clinical trials.

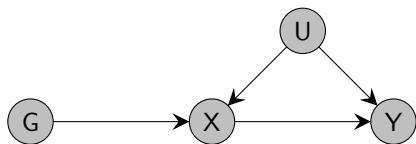
Genome-Wide Association Studies

GWAS: most common type of genetic studies.

- Collect DNA samples from 1000s of individuals.
- Identify points in their DNA chain where differences exist (SNPs).
- G_{ij} : how many copies of a base pair sequence individual i has at SNP j (0/1/2).
- For each SNP G_j , fit $X \sim G_j$ and assess which SNPs affect X .

MR typically uses data from existing GWAS. Complications:

- GWAS studies typically only report summary statistics $\hat{\beta}_j$ and standard errors $\hat{\sigma}_j$ per SNP.
- So MR has to rely only on these summary statistics.
- This is very restrictive!!
- Moreover, X and Y may not even be measured in the same GWAS.



- Summary data can be estimated reliably because $G - X$ and $G - Y$ are unconfounded.
- Want to conduct MR analysis with summary data: $\hat{\beta}_{Xj}, \hat{\sigma}_{Xj}, \hat{\beta}_{Yj}, \hat{\sigma}_{Yj}$.
- With a single SNP G , the 2SLS estimate is

$$\hat{\theta} = \frac{\hat{\beta}_Y}{\hat{\beta}_X}, \quad \text{Var}(\hat{\theta}) = \frac{\hat{\sigma}_Y^2}{\hat{\beta}_X^2} + \frac{\hat{\beta}_Y^2 \hat{\sigma}_X^2}{\hat{\beta}_X^4}$$

which can be computed with summary statistics.

MR with Summary Data - Multiple SNPs

- With P independent SNPs, $G = (G_1, \dots, G_P)$, use the Inverse Variance Weighted (IVW) estimator:

$$\hat{\theta}_{IVW} = \frac{\sum_j \hat{\beta}_{Yj} \hat{\beta}_{Xj} \hat{\sigma}_{Yj}^{-2}}{\sum_j \hat{\beta}_{Xj}^2 \hat{\sigma}_{Yj}^{-2}}, \quad \text{Var}(\hat{\theta}_{IVW}) = \frac{1}{\sum_j \hat{\beta}_{Xj}^2 \hat{\sigma}_{Yj}^{-2}}$$

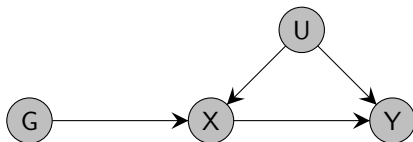
- With correlated SNPs:

$$\hat{\theta}_{IVW} = (\hat{\beta}_X^T \Omega^{-1} \hat{\beta}_X)^{-1} \hat{\beta}_X^T \Omega^{-1} \hat{\beta}_Y, \quad \text{Var}(\hat{\theta}_{IVW}) = (\hat{\beta}_X^T \Omega^{-1} \hat{\beta}_X)^{-1}$$

where $\Omega_{jk} = \hat{\sigma}_{Yj} \hat{\sigma}_{Yk} \rho_{jk}$.

- Intuition:
 - Meta-analysis of SNP-specific estimates.
 - As a Least Squares fit from the (weighted) regression $\hat{\beta}_{Yj} \sim \hat{\beta}_{Xj}$.

Violations of IV Assumptions



- $U \rightarrow G$ should not happen.
 - It can happen with population stratification but GWAS studies typically account for this.
- $G \rightarrow X$ can be controlled by selecting suitable SNPs from a GWAS.
 - But if $G \rightarrow X$ is weak, we have weak instrument bias.
- $G \rightarrow U$ or $G \rightarrow Y$ is a concern.
 - "Pleiotropy" or "exclusion restriction".
 - Formally untestable.

Active area of research in recent years. Approaches for selecting valid SNPs and obtaining unbiased causal effect estimates include:

- Median-based estimation (MR-median).
- Kernel density estimation (MR-MBE).
- Outlier detection and deletion (MR-Presso).
- L1-penalization (sisVIVE, MR-Lasso).
- Robust regression (MR-robust, MR-Raps).
- Bayesian variable selection (MR-Beside, JAM-MR, Berzuini et al).
- Mixture models (ConMix, MR-Mix).
- G-estimation (MR-Genius).
- Etc

Active areas of research:

- Multivariable MR: jointly model multiple (correlated) X_j .
- Clustering in MR (MR-clust): identify SNPs with similar biological functions.
- Cis-MR: use SNPs from a single gene region, assess the suitability of the gene as a drug target, inform clinical trials.

Genetic databases have started making individual-level data available: can use IV methods for individual-level data?

- Nonlinear MR.
- Network analysis.
- Machine learning?

Back to Our Example

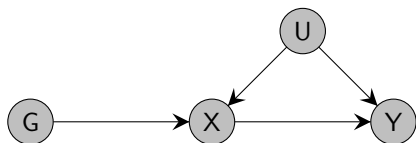
Does chocolate intake increase CHD risk?

Analysis using the MR-Base website:

method	↕↕	nsnp↕↕	b↕↕	se↕↕	pval↕↕
MR Egger		15	-0.8507	0.9024	0.363
Weighted median		15	0.3138	0.2269	0.1666
Inverse variance weighted		15	0.2195	0.1702	0.1971
Weighted mode		15	0.3306	0.4005	0.4229

Effect is in the risk-increasing direction, but not statistically significant!

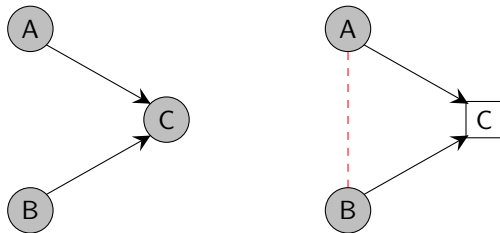
- 1 An Introduction to Mendelian Randomization
- 2 Selection Bias in Mendelian Randomization
 - Structure of Bias
 - Magnitude of Bias - Simulations
- 3 Adjustments for Selection Bias
 - Instruments for Selection
 - MR Inference with Instruments for Selection



- Like most epidemiological studies, MR is susceptible to selection bias.
- Examples:
 - 1 Sample not representative of the study population.
 - 2 Assessing the causal effect of exposures on disease progression.
 - 3 Survival bias in elderly populations.
- Aim: quantify selection bias in Mendelian randomization.

Collider Bias

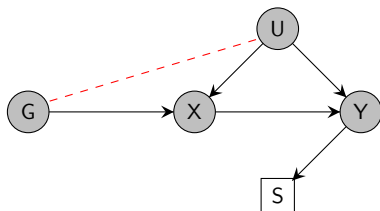
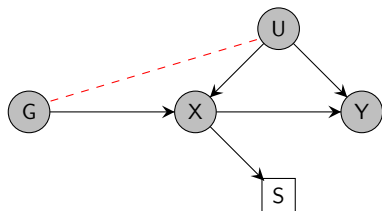
- Selection bias in MR arises as a result of collider bias.
- Two random variables that are independent of each other will become dependent when conditioning on a common effect (the collider).



- A, B marginally independent.
- But A, B not independent conditional on C.

Selection Bias In MR

- Let $S \in \{0, 1\}$ denote selection into the study.
- If $X \rightarrow S$ or $Y \rightarrow S$, then S is a collider (common effect) of G, U .



- Even if $G \perp\!\!\!\perp U$, we will have $G \not\perp\!\!\!\perp U \mid S$, which violates one of the IV assumptions.

- We conducted a simulation study to assess the impact of selection bias in MR.
- Our initial simulation setting was:

$$G_i, U_i \sim N(0, 1)$$

$$X_i = \alpha_G G_i + \alpha_U U_i + \sqrt{1 - \alpha_G^2 - \alpha_U^2} \epsilon_{Xi}$$

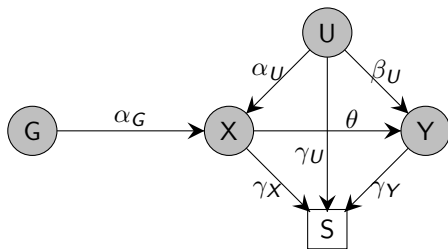
$$Y_i = \theta X_i + \beta_U U_i + \sqrt{1 - \theta^2 - \beta_U^2} \epsilon_{Yi}$$

$$S_i \sim \text{Bernoulli}(\pi_i) \quad , \quad \text{logit}(\pi_i) = \gamma_0 + \gamma_X X_i + \gamma_U U_i + \gamma_Y Y_i$$

$$\epsilon_{Xi}, \epsilon_{Yi} \sim N(0, 1)$$

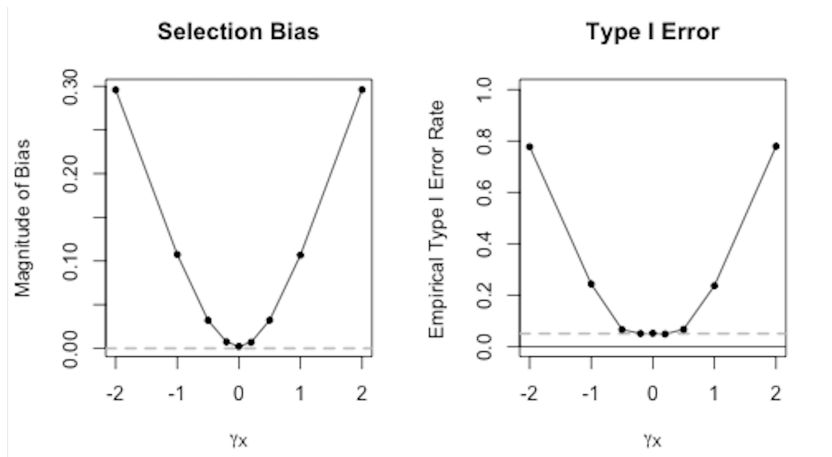
Simulation Model (Graph)

In the form of a causal diagram:



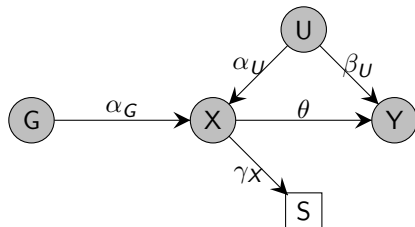
- $\alpha_G = \sqrt{0.02}$ (2% genetic variation in X).
- $\alpha_U = \beta_U = \sqrt{0.5}$.
- $\beta_X = 0$ (no X – Y causal effect).
- Initially, $\gamma_Y = \gamma_U = 0$.
- We varied the selection effect parameter γ_X .

Simulation Results - Baseline Scenario



Bias is symmetric in γ_X and fairly weak for small and moderate values of the selection effect.

Further Simulations

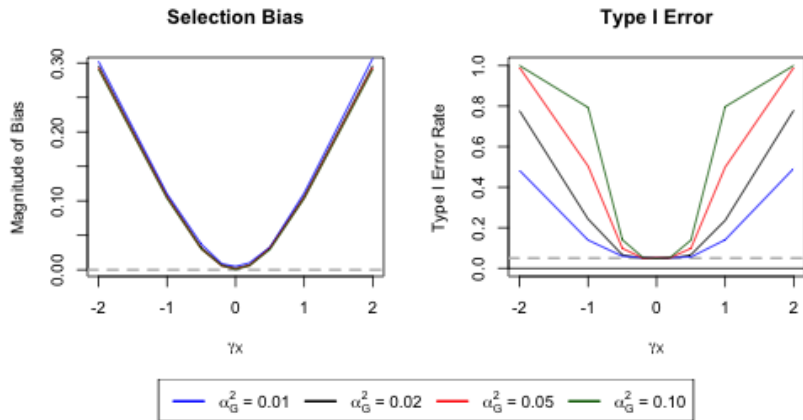


We then varied in turn:

- The proportion α_G of genetic variation in X .
- The confounder-exposure effect α_U .
- The confounder-outcome effect β_U .
- The causal effect θ .
- The structure of the causal diagram.

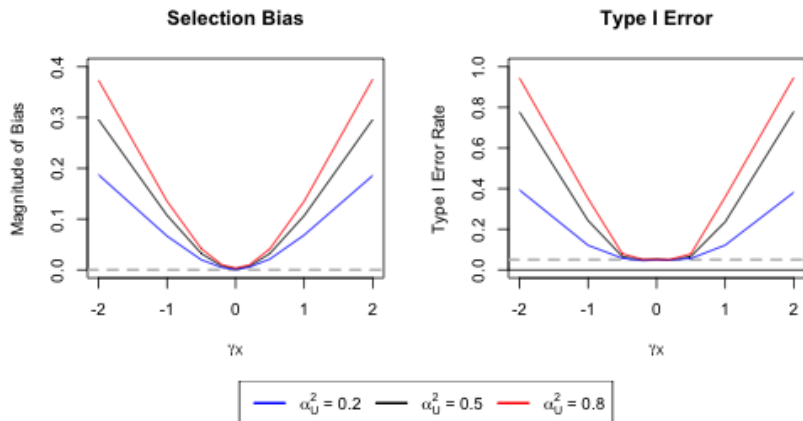
Simulation Results - Instrument Strength

Instrument strength α_G has no impact on causal effect estimates. It does, however, affect Type I error rates: a stronger instrument yields smaller standard errors.



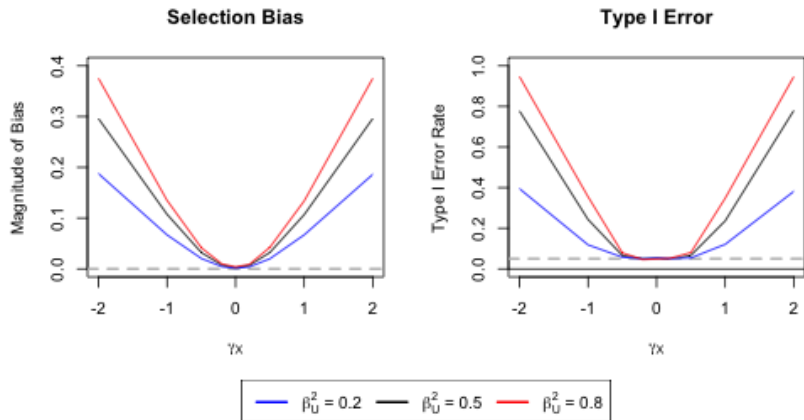
Simulation Results - Confounder-Exposure Association

The strength α_U of the $U - X$ association does impact the magnitude of selection bias, with more confounding associated with larger biases.



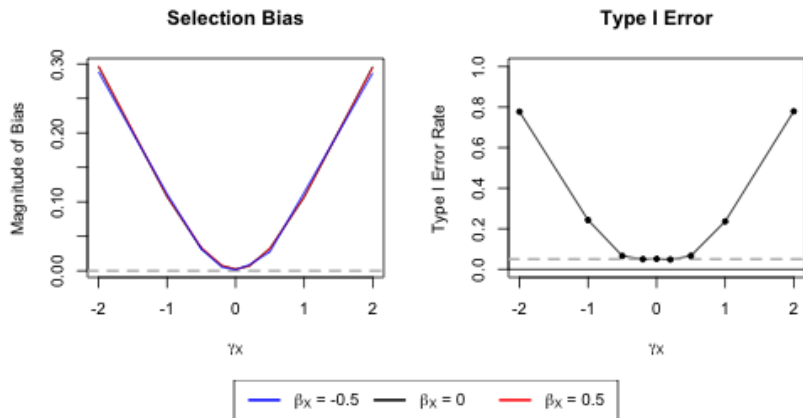
Simulation Results - Confounder-Outcome Association

The same applies to the $U - Y$ association parameter β_U . A strong confounder effect is associated with larger selection bias.

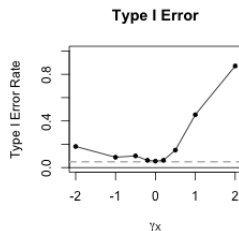
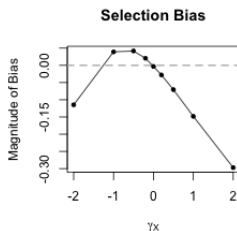
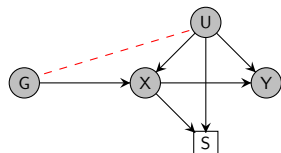


Simulation Results - Causal Effect

The magnitude of the true causal effect β_X does not affect selection bias (at least not when $X \rightarrow S$).

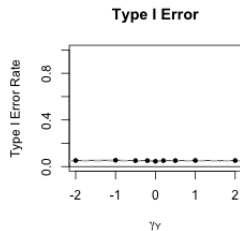
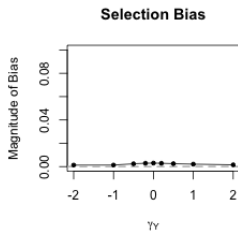
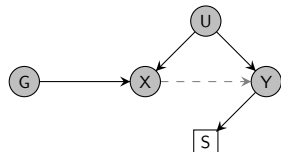


Different Causal Diagrams I



When the confounder also has a direct effect on selection, the bias is no longer symmetric in γ_X . Its direction depends on the relative strengths of the $U \rightarrow S$ and $U \rightarrow X \rightarrow S$ effects.

Different Causal Diagrams II



When selection depends on the outcome the magnitude of the causal effect does have an impact on selection bias. In particular, if the true $X - Y$ causal effect is null, there is no bias.

Also, the bias does not affect case-control studies when cases and controls are sampled at random from the respective populations.

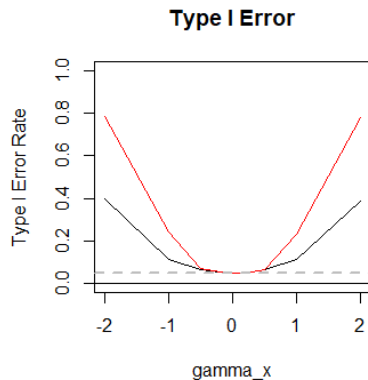
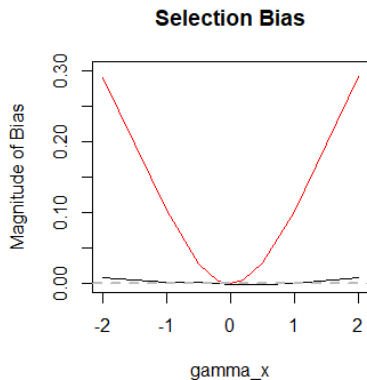
If individual-level data are available, Inverse Probability Weighting (IPW) can be used to remove selection bias.

- Model $\mathbb{P}(S = 1|G, X, Y)$, possibly using data from a separate sample.
- Compute $\pi_i = \mathbb{P}(S_i = 1|G_i, X_i, Y_i)$ for individuals in the study.
- Weight individual i by $\frac{1}{\pi_i}$ when computing causal effect estimates.

Can adjust for selection bias, provided that the selection model is correctly specified.

Simulation Results - IPW

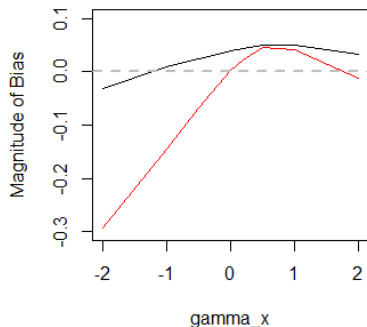
With a correctly specified model, IPW eliminates bias as expected. Type I error rates are improved, though not nominal.



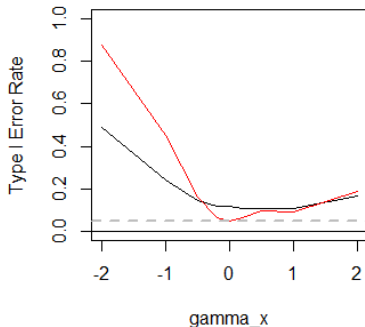
Simulation Results - IPW

When the IPW model is misspecified (here: have a $X \rightarrow U$ effect that is not accounted for) IPW can behave worse than unadjusted estimates for small selection effects.

Selection Bias



Type I Error



- 1 An Introduction to Mendelian Randomization
- 2 Selection Bias in Mendelian Randomization
 - Structure of Bias
 - Magnitude of Bias - Simulations
- 3 Adjustments for Selection Bias
 - Instruments for Selection
 - MR Inference with Instruments for Selection

Selection Bias and Missing Data

- Selection bias can be viewed as a missing data problem.
- E.g. consider an observational study of $Y \sim X$.
- We fully observe X_i but have missing data for Y_i .
- IPW or imputation requires that $\mathbb{P}(S = 1)$ depends only on observed data (data missing at random - MAR).
- But if e.g.

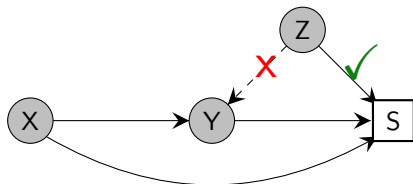
$$\mathbb{P}(S_i = 1) = f(X_i, Y_i)$$

we cannot use IPW, since we have missing data for Y (data missing not at random - MNAR).

- IV analysis can be used to adjust for selection bias with MNAR data (Tchetgen Tchetgen & Wirth, 2017).

Instrument for Selection

Idea: use an instrumental variable Z for the selection process S .



The instrument Z must be fully observed and must satisfy the following conditions:

- 1 IV relevance: $Z \rightarrow S \mid X$.
- 2 Exclusion restriction: $Z \perp\!\!\!\perp Y \mid X$.
- 3 Selection bias is homogeneous on the scale of the parameter of interest.

Plus additional modelling assumptions.

Homogeneity Assumption - Linear Regression

- Often, the estimand of interest is the mean effect $\mathbb{E}(Y|X) = \mu(X)$.
E.g. in linear regression

$$Y|X = X^T\beta + \epsilon \quad , \quad \epsilon \sim N(0, \sigma^2)$$

- In this context, the quantity

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

represents selection bias.

- Homogeneity assumption (on an additive scale) implies that

$$\mathbb{E}(Y|S = 1, X, Z) - \mathbb{E}(Y|S = 0, X, Z) = \delta(X)$$

(does not depend on Z).

- Instrument affects missing status but not the magnitude of bias.

Homogeneity Assumption - Linear Regression

- Some algebra then yields:

$$\mathbb{E}(Y|X, Z, S = 1) = \mu(X) + \delta(X) [1 - \pi(X, Z)]$$

where $\pi(X, Z) = \mathbb{P}(S = 1|X, Z)$ is the propensity score

- $\mu(X)$ cannot be estimated directly due to missing data, but $\mathbb{E}(Y|X, Z, S = 1)$ can.
- Under modelling assumptions for δ , π , can use MLE to estimate $\mu(X)$.
- E.g. if $\mu(X) = X^T \beta$, $\delta(X) = X^T \eta$, $\text{logit } \pi(X, Z) = (X \ Z)^T \alpha$, the likelihood to be maximized is

$$\begin{aligned} \ell(\theta) = \sum_i & (S_i \log \phi(Y_i - \mathbb{E}(Y_i|X_i, Z_i, S_i = 1); 0, \sigma^2) \\ & + S_i \log \pi(X_i, Z_i; \alpha) + (1 - S_i) \log(1 - \pi(X_i, Z_i; \alpha))) \end{aligned}$$

which only depends on observed data.

Homogeneity Assumption - Logistic Regression

- For logistic regression, the quantity of interest is the Odds Ratio

$$\mu(X) = \text{logit } \mathbb{P}(Y = 1|X)$$

- Homogeneity assumption in the Odds Ratio scale:

$$\log \left(\frac{\mathbb{P}(Y = 1|S = 1, X, Z)}{\mathbb{P}(Y = 0|S = 1, X, Z)} \bigg/ \frac{\mathbb{P}(Y = 1|S = 0, X, Z)}{\mathbb{P}(Y = 0|S = 0, X, Z)} \right) = \omega(X)$$

does not depend on Z .

- The relationship between the full-data and observed-data regression is

$$\text{logit } \mathbb{P}(Y = 1|X, Z, S = 1) = -\log \left(\lambda(X, Z)e^{\omega(X)} + 1 - \lambda(X, Z) \right) + \mu(X) + \omega(X)$$

where $\lambda(X, Z) = \mathbb{P}(S = 1|X, Z, Y = 0)$.

- Once again, this can be fitted by MLE.

Homogeneity Assumption - Poisson Regression

- For Poisson regression, the estimand is

$$\mu(X) = \log \mathbb{E}(Y|X)$$

- The homogeneity assumption states that

$$\frac{\mathbb{E}(Y|S = 1, X, Z)}{\mathbb{E}(Y|S = 0, X, Z)} = \nu(X)$$

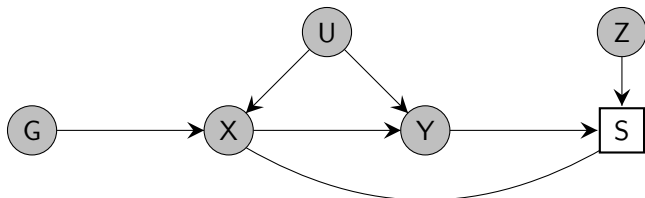
does not depend on Z .

- And the observed-data regression curve satisfies

$$\begin{aligned} \log \mathbb{E}(Y|X, Z, S = 1) &= -\log(\nu(X)\pi(X, Z) + 1 - \pi(X, Z)) \\ &\quad + \mu(X) + \log \nu(X) \end{aligned}$$

which can be fitted by MLE.

Extension to Mendelian Randomization



- Same idea can be used in MR (with individual-level data).
- Use one instrument (G) for inference and another (Z) for selection.
- Z can be either genetic or non-genetic.

MR with a single instrument for inference:

- The causal effect is estimated using the ratio estimate

$$\hat{\theta} = \frac{\hat{\beta}_Y}{\hat{\beta}_X}$$

where $\hat{\beta}_X$ is obtained from a $X \sim G$ regression and $\hat{\beta}_Y$ from a $Y \sim G$ regression.

- Can implement the "IV for selection" method for each regression, get selection-adjusted estimates $\hat{\beta}_X, \hat{\beta}_Y$.
- The method's assumptions extend directly.
- Since X, Y are modelled separately, we can have missing values for either X or Y (or both).

MR with multiple Instruments for inference:

- Can repeat the "single instrument" procedure for each SNP, get selection-adjusted summary statistics, then use summary-statistics methods such as IVW.
- This would also allow the use of summary-level pleiotropy-robust methods.
- But can be slow for many SNPs, and summary-level methods require a two-sample framework.
- Combine with Two-Stage Least Squares: can implement the "IV for selection" as part of either the 1st-stage or 2nd-stage regression.
- Causal effect estimation is fine.
- But not clear how to adjust standard errors for 1st-stage uncertainty.
- Bootstrap?

Observational studies:





- Assess the method's robustness to various assumptions.
- When will the homogeneity assumption hold in practice? Can it be replaced?

Mendelian randomization:

- Simulations ongoing. Results suggest that the method can adjust for selection bias but yields causal effect estimates with considerably wider CIs.
- Use of structural equation models to implement the method in the 2SLS framework.

Applications:

- Selection bias in Covid-19 research.

-  Davey Smith, G. and S. Ebrahim (2003).
'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?
International Journal of Epidemiology 32(1), 1 – 22.
-  Gkatzionis, A. and S. Burgess (2018).
Contextualizing selection bias in Mendelian randomization: how bad is it likely to be?
International Journal of Epidemiology 48(3), 691–701.
-  Tchetgen Tchetgen, E. and K. Wirth (2017).
A general instrumental variable framework for regression analysis with outcome Missing Not at Random
Biometrics 73, 1123–1131.
-  Slob, E. and S. Burgess (2020).
A comparison of robust Mendelian randomization methods using summary data
Genetic Epidemiology 44, 313–327.