Mendelian randomization for binary disease outcomes

2SLS is for linear models, but disease outcomes in MR are mostly dichotomous!

- two ways to define causal effect
  - structural models by conditional on U
  - Potential outcomes: causal risk difference, causal risk ratio, and causal odds ratio
- 2SIV approximation methods for causal odds ratio when disease is rare.
- Consistent estimation by double-logistic models: pros and cons
- the impact of case-control sampling
- estimation when some IV are not valid

Structural models in econometrics

Structural models defining causal effect in the presence of unmeasured confounding

$$g\{E(Y|X,U)\} = \theta_0 + \theta_1 X + \theta_2 U$$

$$\theta_1 = g\{E(Y|X = x, U)\} - g\{E(Y|X = x-1, U)\}$$

- conditional on unknown U ???
- For linear or log-linear models, this is fine: population-average causal effect is the same as the conditional effect θ_1

Pearl's *do* operator

- The expectation of Y when X is manipulated (randomized) to x

$$E[Y|do(X = x)] \neq E[Y|X = x]$$

- Population causal effect in risk difference is

$$E[Y|do(X = x)] - E[Y|do(X = x - 1)]$$

- If U includes all confounding, then

$$\Pr(Y|X = x, U) = \Pr(Y|do(X = x), U)$$

Linear, log-linear, logistic models

$$E[Y|do(X=x)] = E_U\{E[Y|do(X=x),U]\} = \theta_0^* + \theta_1 x$$

$$E[Y|do(X=x)] = \int \exp(\theta_0 + \theta_1 x + \theta_2 u)\Pr(u)du = \exp(\theta_0^* + \theta_1 x)$$

$$E[Y|do(X=x)] = \int \frac{\exp(\theta_0 + \theta_1 x + \theta_2 u)}{1+\exp(\theta_0 + \theta_1 x + \theta_2 u)}\Pr(u)du \neq \frac{\exp(\theta_0^* + \theta_1 x)}{1+\exp(\theta_0^* + \theta_1 x}$$

- The population causal odds ratio is not the same as the conditional causal odds ratio due to non-collapsibility of logistic regression!
- we have no idea about what has been conditional upon, the stratum is not well defined!
- The causal effect for odds ratio in the presence of U is not well-defined!

2SIV approximation for logistic regression

Data generating models

$$X_i = \alpha_0 + \alpha G_i + \alpha_2 U_i + \epsilon_i,$$

$$Y_i \sim Bernoulli(p_i), \quad \log \frac{p_i}{1-p_i} = \theta_0 + \theta X_i + \theta_2 U_i$$

where U is the confounder and ϵ is random error

- Y is rare (most cancer endpoints), so that logistic model can be approximated by log-linear model
- population average risk ratio is approximated by the conditional risk ratio, and the conditional odds ratio

2SIV approximation

- 2SLS still works approximately

$$\begin{aligned} \text{Stage 1:} \quad & E(X_i|G_i) = \alpha_0^* + \alpha_1 G_i \\ \text{Stage 2:} \quad & \text{logit}E(Y_i|\hat{X}_i) = \theta_0^* + \theta_1 \hat{X}_i \end{aligned}$$

- $\text{COR} \approx OR(Y|G)^{1/\delta}$ where $\delta = E(X|G=1) - E(X|G=0)$.

As long as the error ϵ is independent of G given U

$$\begin{aligned} E(Y|G) &= E_{U|G} E_{X|G,U} E(Y|X,Z,U) \\ &\approx E_U E_{X|G,U} \exp(\theta_0 + \theta_1 X + \theta_2 U) \\ &= E_U \exp(\theta_0^* + \theta_2 U + \theta_1 \alpha_0 + \theta_1 \alpha_1 G + \theta_1 \alpha_2 U) \\ &= \exp\left\{\theta_0^{**} + \theta_1 \alpha_1 G\right\}, \end{aligned}$$

ϵ does not have to be normal or parametric!

Palmer et al (2008); Didelez et al (2010) Statistical Science.

2SIV provides a valid and consistent test for causal effect

$$X_i = \alpha_0 + \alpha_1 G_i + \alpha_2 U_i + \epsilon_i, \tag{1}$$

$$\text{logit}\left\{\text{pr}(Y_i = 1 | X_i, U_i)\right\} = \theta_0 + \theta_1 X_i + \theta_2 U_i, \tag{2}$$

Corollary in Dai et al (2014) AJE

Suppose data generating models are (1-2) and G is qualified as an instrumental variable, then the 2SIV estimand $\theta_1^* = 0$ if and only if the true causal effect $\theta_1 = 0$, and θ_1^* has the same sign as θ_1.

- Although the naïve 2SIV estimator for a causal odds ratio is generally not consistent, the corresponding testing procedure is valid and consistent for testing whether the causal effect $\theta_1 = 0$.
- This result holds regardless the disease is rare or not, thus establishes the general utility of 2SIV in Mendelian randomization analyses.

The control function (CF) estimator

The first stage regression error term contains some information about U, so

$$\begin{aligned} &\text{Stage 1:} && \hat{\epsilon}_i = X_i - \hat{X}_i \\ &\text{Stage 2:} && \text{Logit}(p_i) = \theta_0^* + \theta_1 \hat{X}_i + \theta_2 \hat{\epsilon}_i \end{aligned}$$

- For linear models, the control function estimator is equivalent to 2SLS
- For non-linear models, it captures some variability contained in U, thus reduce bias over the standard 2SIV estimator in last slide
- Sometimes called as "adjusted IV method" in the econ literature

Palmer et al (2008); Palmer et al (2011)

Potential outcomes

- Potential outcomes framework: let $Y(x)$ denote the potential outcome of Y when X is *experimentally* altered to an arbitrary value x within the set of all attainable values.
- define $Y_i(x = 0)$ and $Y_i(x = 1)$ for every subject
- average causal effect

$$E[Y_i(x = 1) - Y_i(x = 0)] = E[Y_i(x = 1)] - E[Y_i(x = 0)]$$

- Consistency assumption that $Y(x) = Y$ if $X = x$, so $\Pr(Y(x)|x) = \Pr(Y|x)$
- If X randomized, $x \perp Y(x)$, we can derive the causal effect is the ITT effect

$$E[Y_i|X_i = 1] - E[Y_i|X_i = 0]$$

Causal estimands in potential outcomes

- causal risk difference

$$E[Y(x)] - E[Y(x-1)]$$

- causal risk ratio

$$\frac{E[Y(x)]}{E[Y(x-1)]}$$

- causal odds ratio

$$\frac{E[Y(x)]/E[1-Y(x)]}{E[Y(x-1)]/E[1-Y(x-1)]}$$

The concept of $Y(X = 0)$ for a subject with value (Y=y,X=x)

- individual's "baseline" value (risk) for Y if X is altered to zero value.
- $Y(0)$ can be determined by other causal factors of Y than X
- confounding is introduced when observed X value is correlated with $Y(0)$: people with a higher "baseline" risk of colorectal cancer because of other causal factors happened to be more likely taking red meat.
- so $Y(0)$ can be viewed as the totality of other unmeasured causal effects on Y
- in RCT, X is orthogonal to individual "baseline" value (risk)
- represent a mathematical language to describe the confounding

Structural mean models (SMM)

Different from structural equation models, SMM does not assume data generating models.

- Postulate the causal effect model only

$$\frac{\Pr(Y=1|X,G)}{\Pr(Y(0)=1|X,G)} = \exp(\theta X)$$

$$\frac{\text{Odds}(Y=1|X,G)}{\text{Odds}(Y(0)=1|X,G)} = \exp(\theta X)$$

- This is a conditional causal effect, that does not depend on G (no interaction between G and X).
- Does not condition on the unknown confounding U. The effect in the well-defined, observed subset of the population
- Interpretation: if we are able to change X to zero, what would be the change (in ratio) of the disease odds for the subject

Vansteelandt & Goetghebeur (2003) JRSSB

Estimating equations for SMM

Use the independence between G and $Y(0)$ to construct estimating equation (GMM)

- Causal effect model

$$E(Y|X,G) - E[Y(0) = 1|X,G] = \theta X$$

- estimating equation

$$0 = \sum_{i}^{n} G_i(Y_i - \theta X_i)$$

- Semiparametric, there is no data-generating model on Y, the error distribution is not required.

Causal relative risk

$$\frac{\Pr(Y=1|X,G)}{\Pr(Y(0)=1|X,G)} = \exp(\theta X)$$

- Use $Y\exp(-\theta X)$ as predicted value of $Y(0)$. So

  Estimating function: $$0 = \sum_{i}^{n} \{G_i - E(G)\} Y_i exp(-\theta X_i)$$

- derivation

$$\begin{aligned} E\{G_i - E(G)\} Y_i exp(-\theta X_i) &= E\{G_i - E(G)\} E_{X_i|G_i} \Pr(Y_i(0)|X_i, G_i) \\ &= E\{G_i - E(G)\} \Pr(Y_i(0)|G_i) = 0 \end{aligned}$$

Vansteelandt & Goetghebeur (2003) JRSSB

Double-logistic regression for causal odds ratio estimates

- Causal odds ratio

$$\frac{\text{Odds}(Y = 1|X, G)}{\text{Odds}(Y(0) = 1|X, G)} = \exp(\theta X)$$

- We cannot subtract or divide from Y to get $Y(0)$ anymore
- logistic model for observed data

$$\text{Odds}(Y = 1|X, G) = \exp(\beta_0 + \beta_1 X + \beta_2 G)$$

$$0 = \sum_i^n \{G_i - E(G)\}\text{expit}\{\beta_0 + (\beta_1 - \theta)X_i + \beta_2 G\}$$

- the observed logistic function is preferably non-parametric, otherwise it might be possible that no θ satisfy the GMM equation

Vansteelandt & Goetghebeur (2003) JRSSB; Vansteelandt 2011 SS

The issues with SMM

- Is the association model compatible to the causal effect model?
- It can be fitted by R `gmm` function, but identifiability and convergence are often problematic
- The problem is that the gradient of estimating equations is not monotone.
- there is also issue of weak instrument

$$0 = \sum_{i}^{n} G_i(Y_i - \theta X_i)$$

$$0 = \sum_{i}^{n} \{G_i - E(G)\} Y_i exp(-\theta X_i)$$

$$0 = \sum_{i}^{n} \{G_i - E(G)\} \text{expit}\{\beta_0 + (\beta_1 - \theta)X_i + \beta_2 G\}$$

The issues with SMM

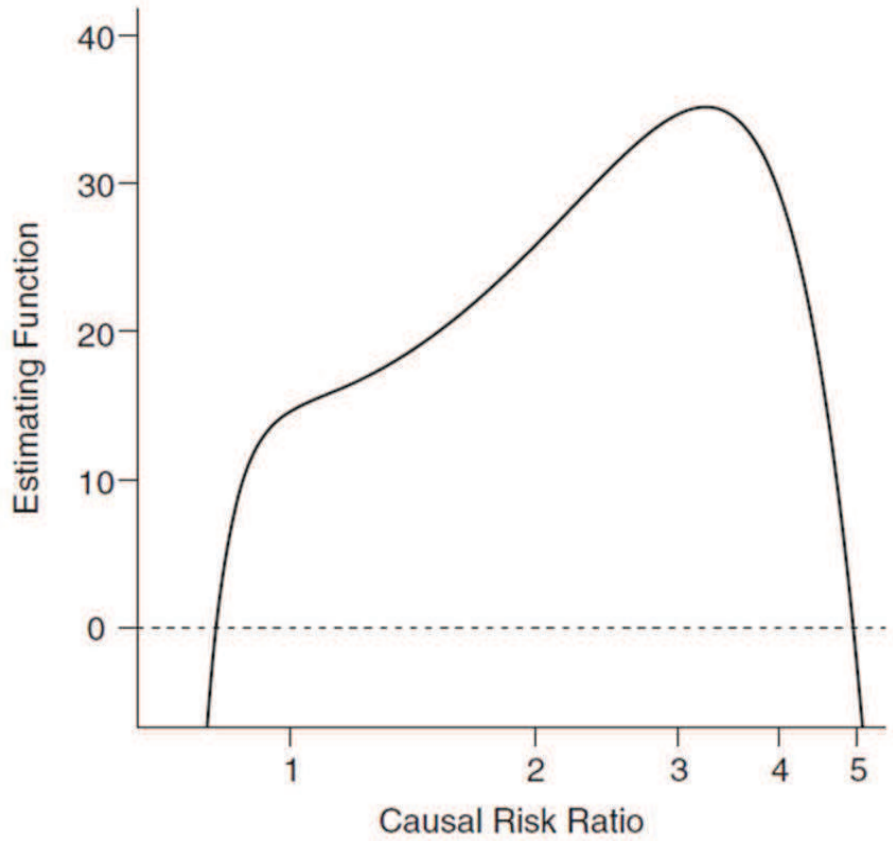


Figure 1. Estimating function for the example from Palmer et al. (20) demonstrating lack of identification. Two distinct parameter values for the causal risk ratio (0.81 and 4.95) satisfy the estimating equation $\sum_i y_i \exp(-\beta_1 x_i)(g_i - \bar{g}) = 0$, where $\bar{g}$ is the average value of G in the population.

Burgess S et al (2014) AJE

The impact of case-control sampling on two-stage IV estimators

The samples are composed of cases and controls (Y), intermediate phenotypes (X) and genotypes (G) were collected after case-control sampling.

$$E(X_i|G_i) = \alpha_0^* + \alpha_1 G_i$$

$$\begin{aligned} \text{2SLS:} \quad \text{logit}E(Y_i|\hat{X}_i) &= \theta_0 + \theta\hat{X}_i \\ \text{CF:} \quad \text{logit}E(Y_i|X_i, \hat{\epsilon}_i) &= \theta_0^* + \theta\hat{X}_i + \theta_2(X_i - \hat{X}_i) \end{aligned}$$

- The issue is to estimate $\hat{X}_i$ use case-control samples in the first stage.
- If X is associated with Y, then $X \sim G$ is assessed in a biased sample
- The regression of $X \sim G$ becomes a secondary trait association problem: see for example, Lin and Zeng (2009).
- If we correct for $\hat{\alpha}$ (and so $\hat{X}$), then no adjustment is needed in the second stage (prospective estimation under case-control sampling).

Methods accounting for the case-control sampling in MR

- Use controls only (rare disease)
  - inefficient
- Weighting estimating equation by inverse of sampling probability (IPW): Jiang et al (2006); Monsees et al (2009)
  - inefficient but robust
- Maximum likelihood estimator (MLE): Jiang et al (2006); Lin and Zeng (2009); Dai et al (2014).

$$\frac{Pr_{\boldsymbol{\eta}}(Y|X,G)Pr_{\alpha}(X|G)Pr_{f}(G)}{\int_{X,G} Pr_{\boldsymbol{\eta}}(Y|X,G)Pr_{\alpha}(X|G)Pr_{f}(G)dgdx}$$

  - efficient but not robust

Dai J et al (2014) AJE

Correcting for case-control sampling in structural mean models (SMM)

Bowden & Vansteelandt (2011) used inverse probability weighting

- Causal relative risk

$$0 = \sum_{i}^{n} \{G_i - \hat{E}_w(G)\} Y_i exp(-\theta)$$

  where $\hat{E}_w(G)$ is weighted estimate of the mean

- Causal odds ratio

$$\text{Odds}(Y = 1|X, G) = \exp(\beta_0 + \beta_1 X + \beta_2 G)$$

$$0 = \sum_{i}^{n} W_i \{G_i - \hat{E}_W(G)\} \text{expit}\{\beta_{0W} + (\beta_1 - \theta)X_i + \beta_2 G\}$$

Estimation when some genetic variants have pleiotropic effects

Suppose there are k genetic variants

$$\begin{aligned} X &= \alpha_0 + \sum_{j=1}^{k} \alpha_{1j} G_j + \varepsilon_1, \\ Y &= \theta_0 + \theta_1 X + \sum_{j=1}^{k} \gamma_j G_j + \varepsilon_2, \end{aligned}$$

then the reduce form $Y \sim G$ is

$$E(Y|G_j) = \beta_0 + \sum_{j=1}^{k} \beta_{1j} G_j = \beta_0 + \sum_{j=1}^{k} (\theta_1 \alpha_{1j} + \gamma_j) G_j$$

This implies

$$\beta_{1j} = \theta_1 \alpha_{1j} + \gamma_j$$

Bowden et al (2015) IJE

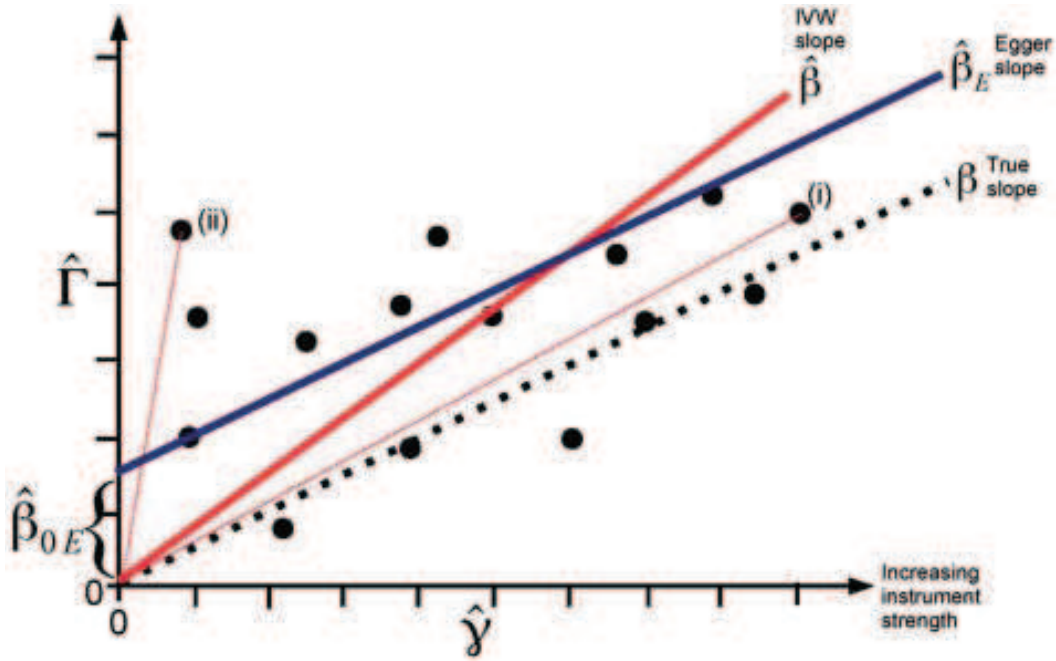
Weakened assumption: direct effects are not related to genetic effects on the exposure

- If γ_j all cancels out, IV estimator would be consistent
- If γ_j is not correlated with α_{1j} (Instrument strength Independent of Direct Effect)
- Estimate the slope of $\hat{\gamma}_j \sim \hat{\alpha}_{1j}$, it should be still consistent
- Eggers test for small study bias in meta-analysis: assesses whether the intercept term is different from zero. This will occur if the estimates from small studies are more skewed towards either high or low values compared with estimates from large studies.

Egger regression

Regress $\hat{\beta}_{1j}$ on $\hat{\alpha}_{1j}$

- The intercept indicates whether there are averaged direct effects (pleiotropy) .
- The slope is the causal effect estimate under potential pleiotropy.
- similar idea was used in Dai et al (2014) SIM



Bowden et al (2015) IJE

These developments are under linear models; Binary outcomes are not applicable!

Other ways to handle invalid instrument: Overidentifying conditions

- we are often dealing with one causal effect
- we have multiple instrumental potentially
- as long as 50% instruments are valid, then consistent estimation is still possible

Kang et al (2014) JASA

Summary

- Mendelian randomization studies can be useful to assess causality from exposure to disease outcome, with careful examination of assumptions.

- Important methodological works can be developed to further address diagnosis of violation assumption, estimation under direct effect, particularly for binary outcomes.

References

- Rubin DB. Estimating causal effects of treatment in randomized and non-randomized studies. Journal of Educational Psychology. 1974;66:688-701.
- Palmer TM, Thompson JR, Tobin JR, et al. Adjusting for bias and unmeasured confounding in Mendelian randomization studies with binary responses. Int. J. Epidemiol. 2008;37:1161-1168.
- Vansteelandt, S. and Goetghebeur, E. Causal inference with generalized structural mean models. Joural of Royal Statistical Society, Ser B. 2003;65:817-35.
- Didelez V, Meng S, Sheehan, NA. Assumptions of iv methods for observational epidemiology. Statistical Science. 2010;25:22-40.
- Vansteelandt S, Bowden J, Babanezhad M, et al. On instrumental variables estimation of causal odds ratio. Statistical Science. 2011;26:403-422.
- Burgess S, Granell R, Palmer TM, et al. Lack of identification in semiparametric instrumental variable models with binary outcomes. American Journal of Epidemiology. 2014; 180(1):111-119.
- Robins JM, Rotnitzky A. Estimation of treatment effects in randomized trials with non-compliance and a dichotomous outcome using structural mean models. Biometrika. 2004;91:
- Lin DY, Zeng D. Proper analysis of secondary phenotype data in case-control association studies. Genetic Epidemiology. 2009;33:256-265.
- Dai JY, Chan KC, Hsu L. Testing concordance of instrumental variable effects in generalized linear models with application to Mendelian randomization. Statistics in Medicine. 2014; 33(23):3986-4007.
- Dai JY, Zhang X. Mendelian randomization studies for a continuous exposure under case-control sampling. American Journal of Epidemiology. 2015;181(6):440-449.
- Bowden J, Vansteelandt SV. Mendelian randomization analysis of case-control data using structural mean models. Statistics in Medicine. 2011;30:678-694.