Average Causal Effect in Observational Studies III

Xinzhou Guo

HKUST

(Credited to Zhichao Jiang)

March 11, 2024

RCT

Observational + Ignorability (all confounders)

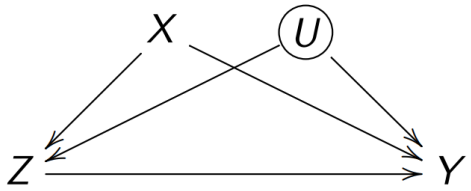
$Z \perp (Y_0, Y_1) \mid X$

Latent

Identification assumptions in observational studies

- Identification assumptions:
  - ignorability: $Z_i \perp \{Y_i(1), Y_i(0)\} \mid \mathbf{X}_i$
  - overlap: $0 < \text{pr}\,(Z_i = 1 \mid \mathbf{X}_i) < 1$ for all $\mathbf{X}_i$
- Possible existence of observed and unobserved confounders

$$\mathbb{E}\{Y_i(z) \mid Z_i = z, \mathbf{X}_i\} \neq \mathbb{E}\{Y_i(z) \mid Z_i = z, \mathbf{X}_i, \mathbf{U}_i\}$$



- Latent ignorability: $Z_i \perp \{Y_i(1), Y_i(0)\} \mid (\mathbf{X}_i, U_i)$ – when will it happen?
- Not directly testable from the observed data but sensitivity can be analyzed; i.e. quantify the evidence in the presence of U.

$E(Y(1) \mid Z=1, X) \neq E(Y \mid Z=1, X)$

not identifiable

$E(Y(1) \mid Z=1, X, U) = E(Y \mid Z=1, X, U)$

Analysis based on (Y, Z, X)

↓

evidence (p-value, ...)

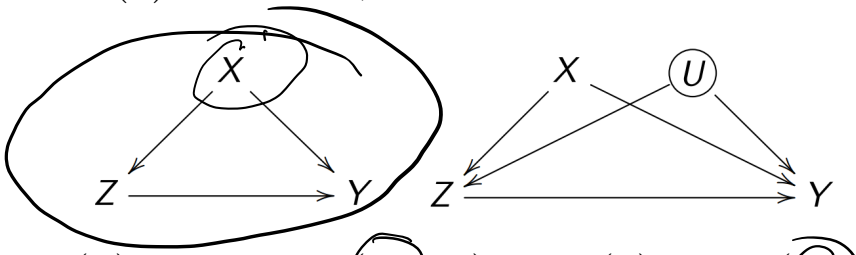
how trustworthy?

sensitivity analysis

p-value sensitivity analysis

Causal diagram

- Pearl (1995) introduce causal diagram. Textbook: Pearl (2000)
- Useful for visualizing causal relationship
- Examples: $\epsilon_Z \perp \epsilon_Y(z)$ for $z = 0, 1$



$$X \sim F_X(x), \quad Z = f_Z(X, \epsilon_Z), \quad Y(z) = f_Y(X, z, \epsilon_Y(z))$$

$$X \sim F_X(x), \quad U \sim F_U(u), \quad Z = f_Z(X, U, \epsilon_Z)$$

$$Y(z) = f_Y(X, U, z, \epsilon_Y(z))$$

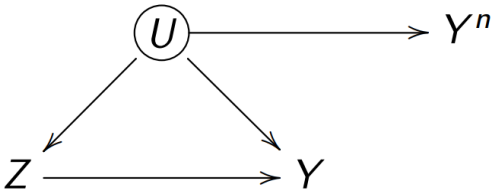
Is it $Z_i \perp \{Y_i(1), Y_i(0)\} \mid \mathbf{X}_i$?

Z decided on ✓

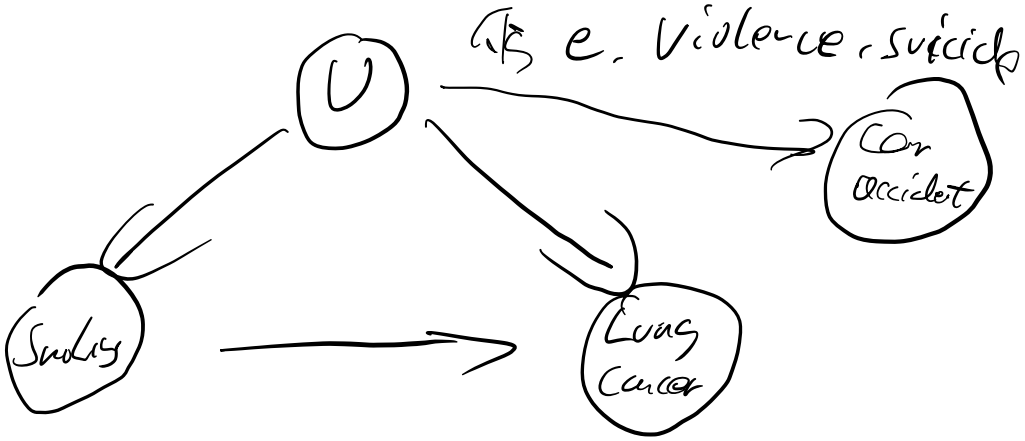
$Y(1)$ . . . ✓

$Y(0)$ d . . . ✓

Negative outcome

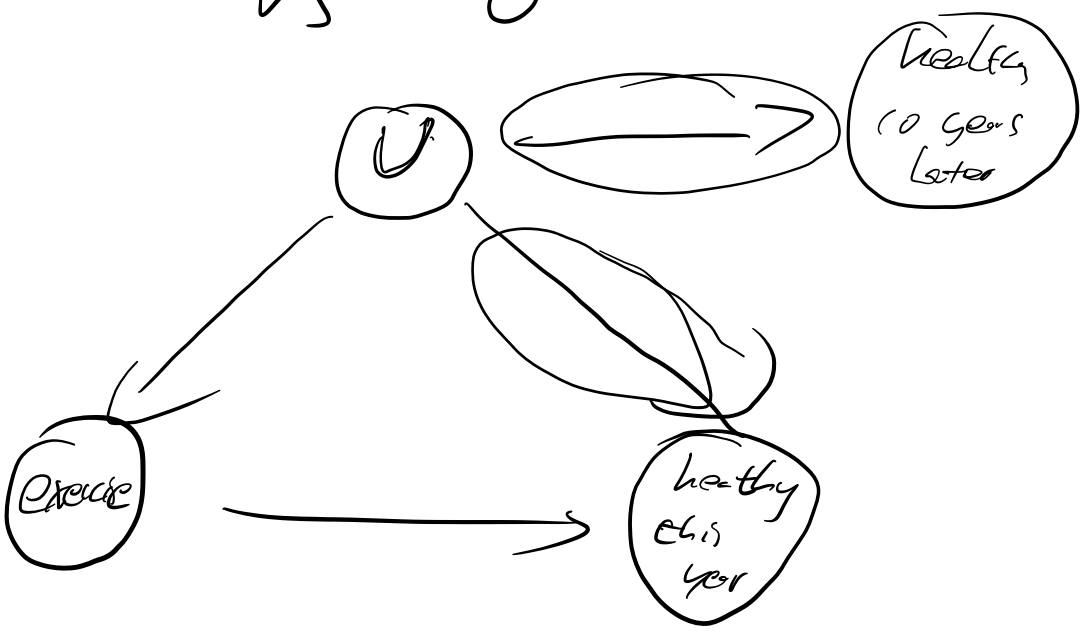


- Negative outcome Y_i^n : similar to Y_i in terms of confounding
  - if $Z_i \perp \{Y_i(1), Y_i(0)\} \mid \mathbf{X}_i$, then $Z_i \perp \{Y_i^n(1), Y_i^n(0)\} \mid \mathbf{X}_i$
  - $\mathbb{E}\{Y_i^n(1) - Y_i^n(0)\}$ = known value (often 0)
- Assessing ignorability $\rightsquigarrow$ estimate ACE on the negative outcome assuming U does not exist and compare it with the known value
  - the effect of smoking on car accident → violence, suicide (Cornfield et al., 1959)
  - lagged outcome (Imbens and Rubin, 2015)
  - Lipsitch et al., (2010), "Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies"
  - the effect of known effects (Rosenbaum, 1989)

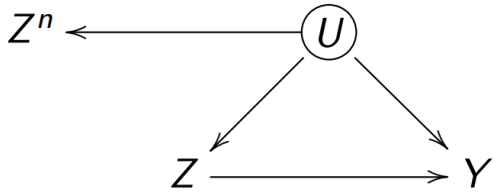


Car accident ~ Smoking

VS



Negative treatment



- Negative treatment Z_i^n : similar to Z_i in terms of confounding
  - if $Z_i \perp \{Y_i(1), Y_i(0)\} \mid \mathbf{X}_i$, then $Z_i^n \perp \{Y_i(1), Y_i(0)\} \mid \mathbf{X}_i$
  - $\mathbb{E}\{Y_i\,(Z_i^n = 1) - Y_i\,(Z_i^n = 0)\}$ = known value
- Assessing ignorability $\rightsquigarrow$ estimate ACE on the negative treatment assuming U does not exist and compare it with the known value
- Examples: Sanderson et al. (2018) "Negative control exposure studies in the presence of measurement error: implications for attempted effect estimate calibration"

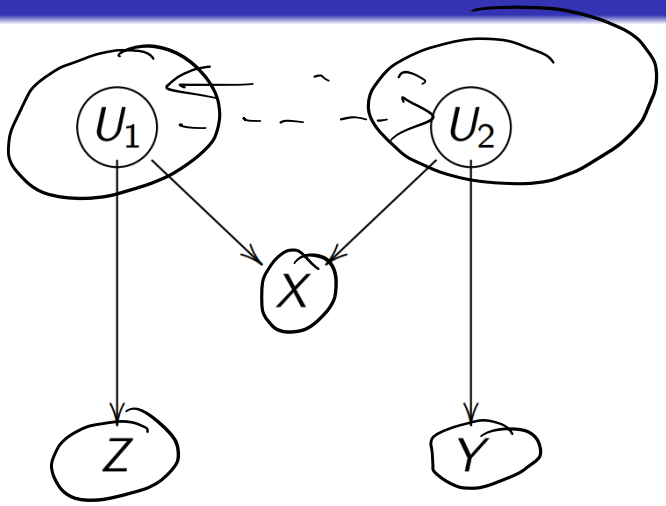
Negative treatment

Exposure	Negative control exposure	Outcome(s)
Maternal smoking	Paternal smoking	Offspring outcomes: Inattention/hyperactivity[15,20] Obesity/adiposity[16,22–24]
		Blood pressure[17]
		Gestational diabetes[21]
		ADHD symptoms[19]
		Cognitive development[18]
		Offspring psychotic symptoms[46]
Maternal psychosocial stress	Paternal psychosocial stress	Offspring vascular function[54]
Maternal smoking during pregnancy	Maternal smoking after pregnancy	Offspring respiratory outcomes[39]
		Offspring psychotic symptoms[46]
Maternal alcohol consumption during pregnancy	Maternal alcohol consumption before pregnancy	Offspring ADHD symptoms[40]
Maternal BMI/obesity	Paternal BMI	Offspring BMI/adiposity[26–33]
		Offspring cognitive and psychomotor development[55]
Length of pre-birth inter-pregnancy interval	Length of post-birth inter-pregnancy interval	Risk of schizophrenia in the offspring[56]
Folic acid supplements in pregnancy	Other supplements in pregnancy	Autism spectrum disorders[37]
		Language development delays[38]
Prescription for trimethoprim 1–3 months before pregnancy	Prescription for trimethoprim 13–15 months before pregnancy	Offspring congenital malformation[57]
Air pollutant exposure during pregnancy	Air pollutant exposure before and after pregnancy	Offspring autism spectrum disorder[41]
Exposure to childhood infections	Hospital attendance for broken bones	Multiple sclerosis later in life[58]
Adherence to prescribed statins and beta blockers	Adherence to other prescribed medication	Long-term mortality after acute myocardial infarction[59]
Vaccination during flu season	Vaccination outside flu season	Mortality and hospitalization from flu[60]
Swimmers' exposure to bacteria in water	Non-swimmers	Gastrointestinal illnesses after an increase in bacteria levels in water[61]

Problem of over-adjustment

- Rosenbaum (2002): "there is no reason to avoid adjustment for a variable describing subjects before treatment"
- Rubin (2007): typically, the more conditional an assumption, the more acceptable it is.
- VanderWeele and Shpitser (2011) called this the "pre-treatment criterion"
- Pearl disagrees

M-bias



- Linear model:

$$\begin{aligned} X &= aU_1 + bU_2 + \epsilon_X \\ Z &= cU_1 + \epsilon_Z \\ Y(z) &= dU_2 + \epsilon_Y \\ (U_1, U_2, \epsilon_X, \epsilon_Z, \epsilon_Y) &\sim \text{ i.i.d. } N(0,1) \end{aligned}$$

- Unadjusted estimator is proportional to $\text{cov}(Z, Y) = 0 \rightsquigarrow$ unbiased
- Adjusted estimator is proportional to $\rho_{ZY|X} = -abcd \rightsquigarrow$ biased
- More details in Ding and Miritrix (2015)

$$\frac{cov(cU_1, \frac{da}{b}U_1)}{}$$

$$Y \sim Z + X$$

$$\beta_Z = \frac{cov(Z, Y \mid X)}{Var(Z \mid X)}$$

$$Y \sim Z$$

$$\beta_Z = \frac{cov(Z, Y)}{Var(Z)}$$

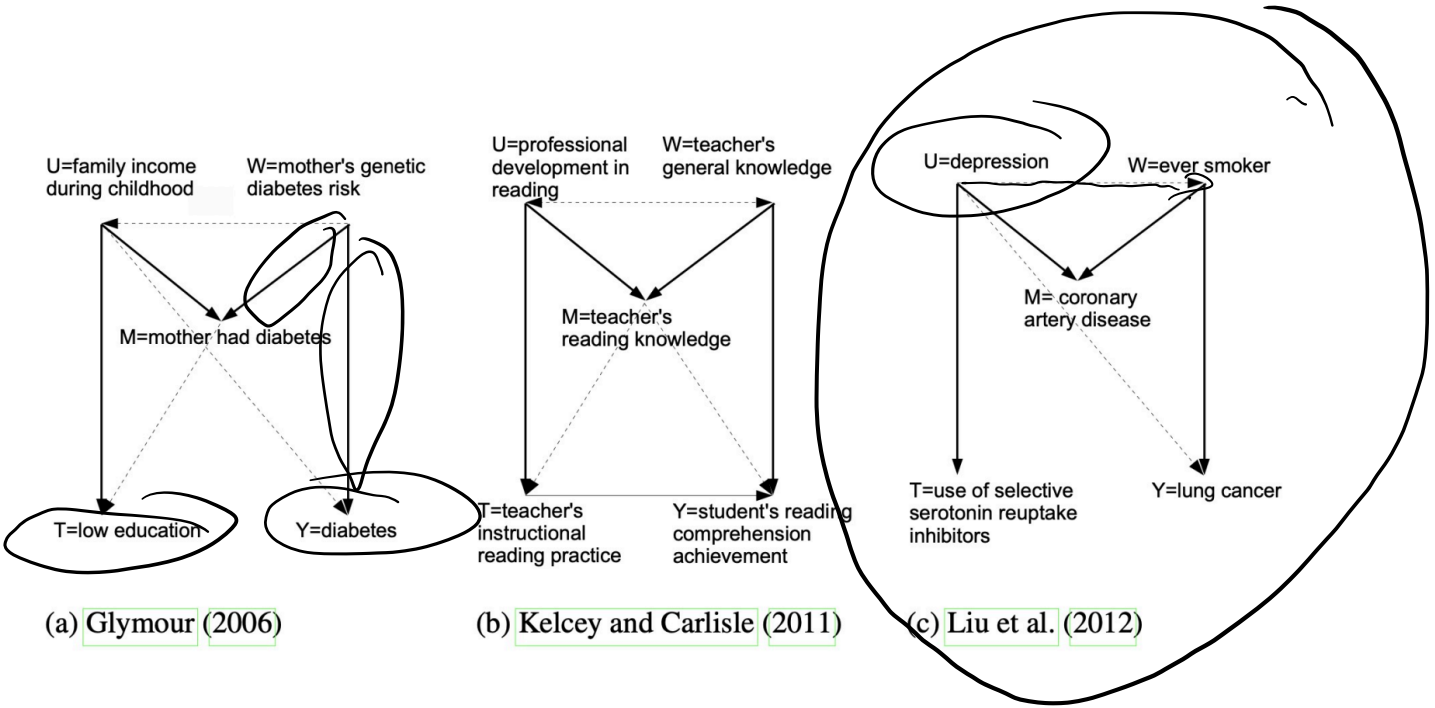
$$cov(Z, Y \mid X)$$

$$= cov(cU_1, dU_2 \mid aU_1 + bU_2 = x)$$

$$= cov\left(\frac{c(x - bU_2)}{a}, dU_2\right)$$

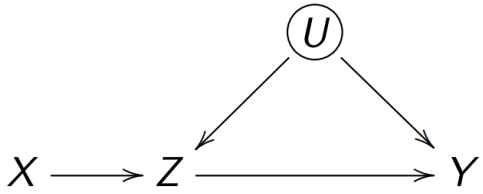
$$= -\frac{cbd}{a}$$

M-structures with deviations



(a) Glymour (2006) (b) Kelcey and Carlisle (2011) (c) Liu et al. (2012)

Z-bias



- Linear model:

$$Z = aX + bU + \epsilon_Z$$
$$Y(z) = \tau z + cU + \epsilon_Y$$
$$(U, X, \epsilon_Z, \epsilon_Y) \sim \text{ i.i.d. } N(0,1)$$

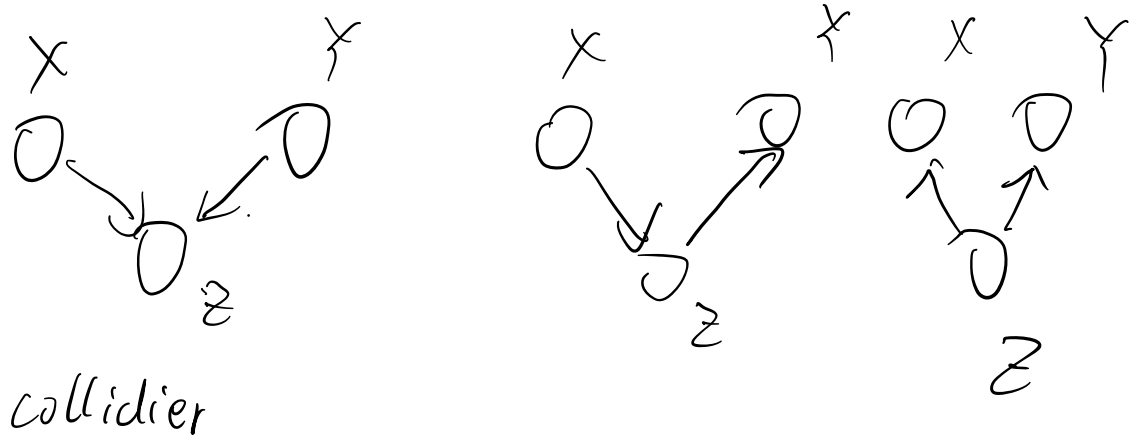
- Unadjusted estimator $\tau_{\text{unadj}} = \tau + \frac{bc}{a^2+b^2+1}$
- Adjusted estimator $\tau_{\text{adj}} = \tau + \frac{bc}{b^2+1}$ (how about adjusting both X and U)
- More details in Ding et al. (2017) but X needs to be used in another way

$$\operatorname{cov}(Y(z), z) = \operatorname{cov}(\tau z + cU, z)$$
$$= \tau \operatorname{Var} z + bc \operatorname{Var} U)$$

$$\frac{\operatorname{cov}(Y(z), z)}{\operatorname{Var}(z)} = \frac{\operatorname{cov}(\tau z + cU, z)}{\operatorname{Var}(aX + bU + \varepsilon_z)}$$

$$= \tau + \frac{\operatorname{cov}(cU, aX + bU)}{a^2 + b^2 + 1}$$

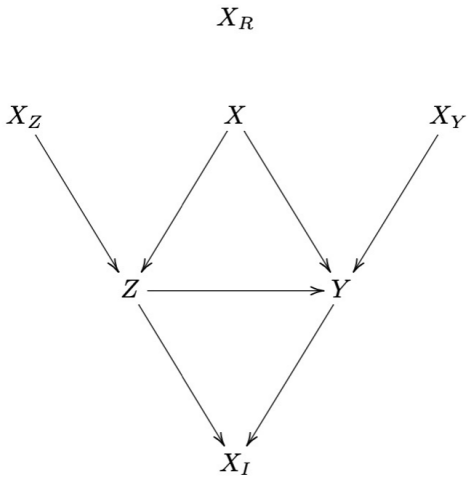
$$= \tau + \frac{bc}{a^2 + b^2 + 1}$$



collidier

What covariates should be adjusted?

Rule out M-bias and assume we already adjust for X



- Adjust for X to remove bias
- Adjust for X_Y to improve prediction
- Adjusting for X_Z or X_R will increase variability
- Adjusting for X_I will introduce bias

What if ignorability is not possible?

- Better design $\rightsquigarrow$ instrumental variable (next topic)
- Partial identification $\rightsquigarrow$ bounds on ACE
- Sensitivity analysis: how results would change under certain types of latent confounding
  - nonparametric sensitivity analysis, e.g., Cornfield condition
  - parametric sensitivity anlaysis

Partial identification

- Identification $\rightsquigarrow$ point estimation
- Partial identification
  - a parameter is θ partially identifiable if the observed data distribution is compatible with multiple values of θ
  - bounds on parameter: all the possible values that are compatible with the observed value
- Cochran (1953) uses the idea of worse-case bounds in surveys with missing data
- Manski (1990) applies the idea to causal inference

Bounds on ACE

- Outcome bounded in $[l, u] \rightsquigarrow [l - u, u - l]$ – too wide
- (Improved) lower and upper bound on ACE – why?

$$\begin{aligned}
&\mathbb{E}\left(Y_i \mid Z_i = 1\right) \Pr\left(Z_i = 1\right) + l \Pr\left(Z_i = 0\right) - u \Pr\left(Z_i = 1\right) \\
&\qquad - \mathbb{E}\left(Y_i \mid Z_i = 0\right) \Pr\left(Z_i = 0\right) \\
&\mathbb{E}\left(Y_i \mid Z_i = 1\right) \Pr\left(Z_i = 1\right) + u \Pr\left(Z_i = 0\right) - l \Pr\left(Z_i = 1\right) \\
&\qquad - \mathbb{E}\left(Y_i \mid Z_i = 0\right) \Pr\left(Z_i = 0\right)
\end{aligned}$$

- Bounds are still too wide to be informative and what if the outcome is unbounded?
- Stronger assumptions $\rightsquigarrow$ tighter bounds
  - other possible assumptions: $\text{cor}(Y(1), Y(0)) \geqslant 0$, $Z = \mathbf{1}(Y(1) - Y(0) \geqslant 0)$.
  - statistical inference: confidence interval on the bounds or the true parameter (Imbens and Manski, 2004) – intersection bound

Sensitivity analysis for latent confounding

- Sensitivity analysis assesses the robustness of study conclusions
- How much does the key identification assumption needs to be violated in order for an empirical finding to disappear?
  - What is the difference between sensitiivity analysis and covariate balance test?
  - we do not assume ignorability with X but assume latent ignorability

$$Z_i \perp \{Y_i(1), Y_i(0)\} \mid (X_i, U_i)$$

  - nonparametric sensitivity analysis for binary outcome $\rightsquigarrow$ Cornfield condition
  - parametric sensitivity analysis

Smoking and lung cancer

- Z: smoking
- Y : lung cancer
- Doll and Hill (1950) find that the relative risk of cigarette smoking on lung cancer was 9 after adjusting for many observed covariates X
- Fisher (1957) criticizes their result because it is possible that a hidden gene (U) simultaneously causes cigarette smoking and lung cancer although the true causal effect of cigarette smoking on lung cancer is absent

Sensitivity analysis (Cornfield et al. 1959. J. Natl. Cancer Inst.)

- Question: How important does a confounder have to be to explain away the observed association? $\rightsquigarrow$ robustness of findings

The magnitude of the excess lung-cancer risk among cigarette smokers is so great that the results can not be interpreted as arising from an indirect association of cigarette smoking with some other agent or characteristic, since this hypothetical agent would have to be at least as strongly associated with lung cancer as cigarette use, no such agent has been found or suggested.

Cornfield Condition

- Causal and observed relative risks:

$$RR_{ZY} = \frac{\Pr\{Y_i(1) = 1\}}{\Pr\{Y_i(0) = 1\}}, \quad RR_{ZY}^{\text{obs}} = \frac{\Pr(Y_i = 1 \mid Z_i = 1)}{\Pr(Y_i = 1 \mid Z_i = 0)}$$

- Assume $\{Y_i(1), Y_i(0)\} \perp Z_i \mid U_i$ – how about observing some confounders?
- Magnitude of binary unobserved confounder:

$$RR_{UY} = \frac{\Pr(Y_i = 1 \mid U_i = 1)}{\Pr(Y_i = 1 \mid U_i = 0)}, \quad RR_{ZU} = \frac{\Pr(U_i = 1 \mid Z_i = 1)}{\Pr(U_i = 1 \mid Z_i = 0)}$$

  which measure the association to outcome and treatment.
- How large do RR_{UY} and RR_{ZU} have to be in order for $RR_{ZY}^{obs} > 1$ and $RR_{ZY} = 1$ (no treatment) hold at the same time; i.e. explain away?

Cornfield Conditions

Theorem

Suppose that U is binary and $Z \perp Y \mid U$; i.e. $RR_{ZY} = 1$. Assume

$$RR_{ZY}^{\text{obs}} > 1, \quad RR_{ZU} > 1, \quad RR_{UY} > 1 \text{ (not substantial)}.$$

We have

$$RR_{ZY}^{obs} \leqslant \frac{RR_{ZU} \times RR_{UY}}{RR_{ZU} + RR_{UY} - 1}$$

- Define $h(w_1, w_2) = w_1 w_2 / (w_1 + w_2 - 1)$ for $w_1 > 1$ and w_2
  - $h(w_1, w_2) \leqslant \min(w_1, w_2)$
  - $h(w_1, w_2) \leqslant w^2/(2w - 1)$, where $w = \max(w_1, w_2)$
- Cornfield condition: $\min(RR_{ZU}, RR_{UY}) \geqslant RR_{ZY}^{\text{obs}}$

$$RR_{ZY}^{OLS} = \frac{P(Y=1|Z=1)}{P(Y=1|Z=0)}$$

$$= \frac{\{P(U=1|Z=1)\,P(Y=1|Z=1,U=1) + P(U=0|Z=1)\,P(Y=1|Z=1,U=0)\}}{\{P(U=1|Z=1)\,P(Y=1|U=1) + P(U=0|Z=1)\,P(Y=1|U=0)\}}$$

$$= \frac{\tau_1 \cdot RR_{UY} + 1 - \tau_1}{\tau_0 RR_{UY} + 1 - \tau_0}$$

$$= \frac{(RR_{UY}-1)\tau_1 + 1}{(RR_{UY}-1)RR_{ZU}\tau_1 + 1}$$

$$\tau_1 = P(U=1|Z=1)$$

$$\tau_0 = P(U=1|Z=0)$$

$$\frac{\tau_1}{\tau_0} \doteq RR_{ZU}$$

E-value

- $\max\left(RR_{ZU}, RR_{UY}\right) \geqslant RR_{ZY}^{\text{obs}} + \sqrt{RR_{ZY}^{\text{obs}}\left(RR_{ZY}^{\text{obs}} - 1\right)}$
- E-value (VanderWeele and Ding, 2017): $RR_{ZY}^{\text{obs}} + \sqrt{RR_{ZY}^{\text{obs}}\left(RR_{ZY}^{\text{obs}} - 1\right)}$
  - the maximum of the confounding measures RR_{UY} and RR_{ZU} need to be at least as large as the E-value to explain away the observed relative risk
  - p-values: sampling uncertainty – interpret p-value from balance test?
  - E-value: confounding bias also known as general Cornfield condition

Smoking and Lung Cancer

	Lung cancer	No lung cancer
Smoker	397	78577
Non-smoker	51	108778

- $RR_{ZY}^{obs} = 10.73$ with 95% confidence interval $[8.02, 14.36]$ (Hammond and Horn. 1958. JAMA)
- Cornfield condition:

$$\min\left(RR_{ZU}, RR_{UY}\right) \geqslant 10.73$$

- Generalized Cornfield condition:

$$\max\left(RR_{ZU}, RR_{UY}\right) \geqslant RR_{ZY}^{obs} + \sqrt{RR_{ZY}^{obs}\left(RR_{ZY}^{obs} - 1\right)} = 20.95$$

- For the lower confidence interval, this number equals 15.52

Extensions

- Non-zero true causal effect

$$RR_{ZY} \leqslant RR_{ZY}^{obs} \times \frac{RR_{ZU} + RR_{UY} - 1}{RR_{ZU} \times RR_{UY}}$$

  - modify the definition – why?

$$RR_{UY} = \max_{z} \Pr(Y = 1 \mid Z = z, U = 1) / \Pr(Y = 1 \mid Z = z, U = 0)$$

  - upper bound on RR_{ZY} with two sensitivity parameters
  - sensitivity parameters are used to quantify the difference between observed and true in the presence of unmeasured confounders and can be unobservable
- Discrete U : Ding and VanderWeele (2016)

Parametric sensitivity analysis

- Include U in regression models (Rosenbaum and Rubin, 1983)

$$\begin{aligned}\text{logit } \Pr\{Y_i(z) \mid \mathbf{X}_i, U_i\} &= \beta_0 + \beta_Z z + \beta_X \mathbf{X}_i + \beta_U U_i \\ \text{logit } \Pr(Z_i = 1 \mid \mathbf{X}_i, U_i) &= \alpha_0 + \alpha_X X_i + \alpha_U U_i\end{aligned}$$

  - binary U with sensitivity parameters $\Pr(U = 1 \mid \mathbf{X}_i)$ – why?
  - can use other models and sensitivity parameters
  - becomes very arbitrary sometimes

Sensitivity analysis for ACE

- Sensitivity parameters $\epsilon_1(X)$ and $\epsilon_0(X)$ – what is the interpretation?

$$\frac{\mathbb{E}\{Y(1) \mid Z=1, X\}}{\mathbb{E}\{Y(1) \mid Z=0, X\}} = \epsilon_1(X)$$
$$\frac{\mathbb{E}\{Y(0) \mid Z=1, X\}}{\mathbb{E}\{Y(0) \mid Z=0, X\}} = \epsilon_0(X)$$

- $\epsilon_1(X) = \epsilon_0(X) = 1$ implies ignorability
- observed data provide no information on the sensitivity parameters

Sensitivity analysis for ACE

- Outcome regression formula

$$\mathbb{E}\{Y(1) \mid Z=1\} = \mathbb{E}\{\mu_1(X)/\epsilon_1(X) \mid Z=1\}$$
$$\mathbb{E}\{Y(0) \mid Z=0\} = \mathbb{E}\{\mu_0(X)\epsilon_0(X) \mid Z=0\}$$

- Inverse probability weighting formula

$$\mathbb{E}\{Y(1)\} = \mathbb{E}\left\{w_1(X)\frac{ZY}{e(X)}\right\}$$
$$\mathbb{E}\{Y(0)\} = \mathbb{E}\left\{w_0(X)\frac{(1-Z)Y}{1-e(X)}\right\}$$

  - $w_1(X) = e(X) + \{1-e(X)\}/\epsilon_1(X)$ and $w_0(X) = e(X)\epsilon_0(X) + \{1-e(X)\}$

- For different $\epsilon_1(X)$ and $\epsilon_0(X)$, we can estimate ACE; what if $\epsilon_1(X) = \epsilon_0(X) = 0$?

Summary

- Assessing ignorability: negative outcome, negative treatment
- When ignorability fails
  - partial identification
  - sensitivity analysis
- Other strategies
  - instrumental variable method
  - twin study, DID, etc.

Suggested readings

- Sensitivity analysis
  - parametric: Rosenbaum and Rubin, "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome"
  - non-parametric: Ding's book, Ding and VanderWeele, "Sensitivity Analysis Without Assumptions"
- Bradford Hill criteria: nine criteria to provide epidemiologic evidence of a causal relationship
  - strength $\rightsquigarrow$ cornfield condition, E-value
  - consistency $\rightsquigarrow$ meta-analysis, invariant prediction (Peters, Buhlmann and Meinshausen, 2016)
  - Specificity $\rightsquigarrow$ specificity score