Sensitivity to Unobserved Confounding in Studies with Factor-structured Outcomes

Alexander Franks

8/8/23

Slides and Paper

- Slides: afranks.com/talks
- *Sensitivity to Unobserved Confounding in Studies with Factor-structured Outcomes*, (JASA, 2023) https://arxiv.org/abs/2208.06552
- Joint work with Jiajing Zheng (formerly UCSB), Jiaxi Wu (UCSB) and Alex D'Amour (Google)
- Please look at the paper for full set of assumptions and technical details

Causal Inference From Observational Data

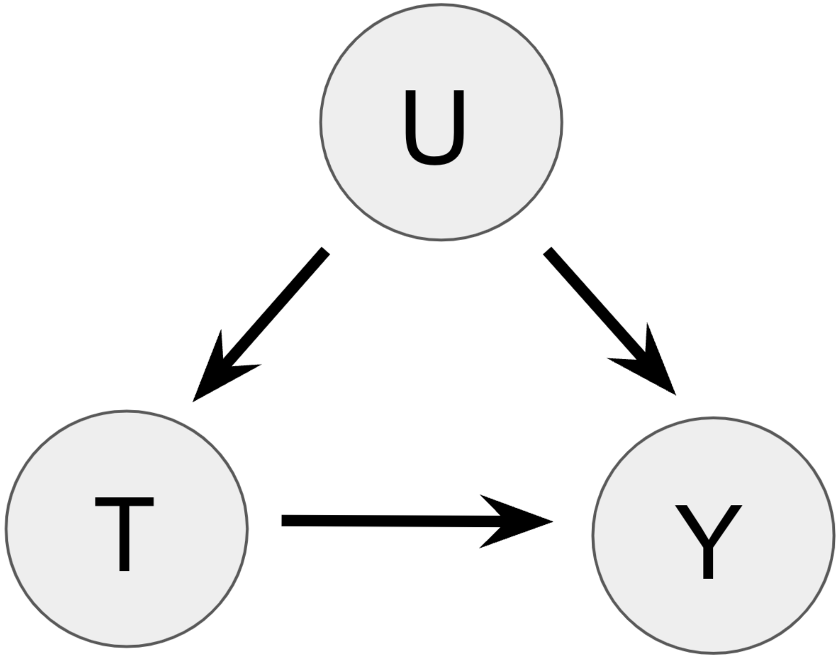
- Consider a treatment T and outcome Y
- Interested in the population average treatment effect (PATE) of T on Y:

$$E[Y|do(T=t)] - E[Y|do(T=t')]$$

- In general, the PATE is not the same as

$$E[Y|T=t] - E[Y|T=t']$$

Confounders

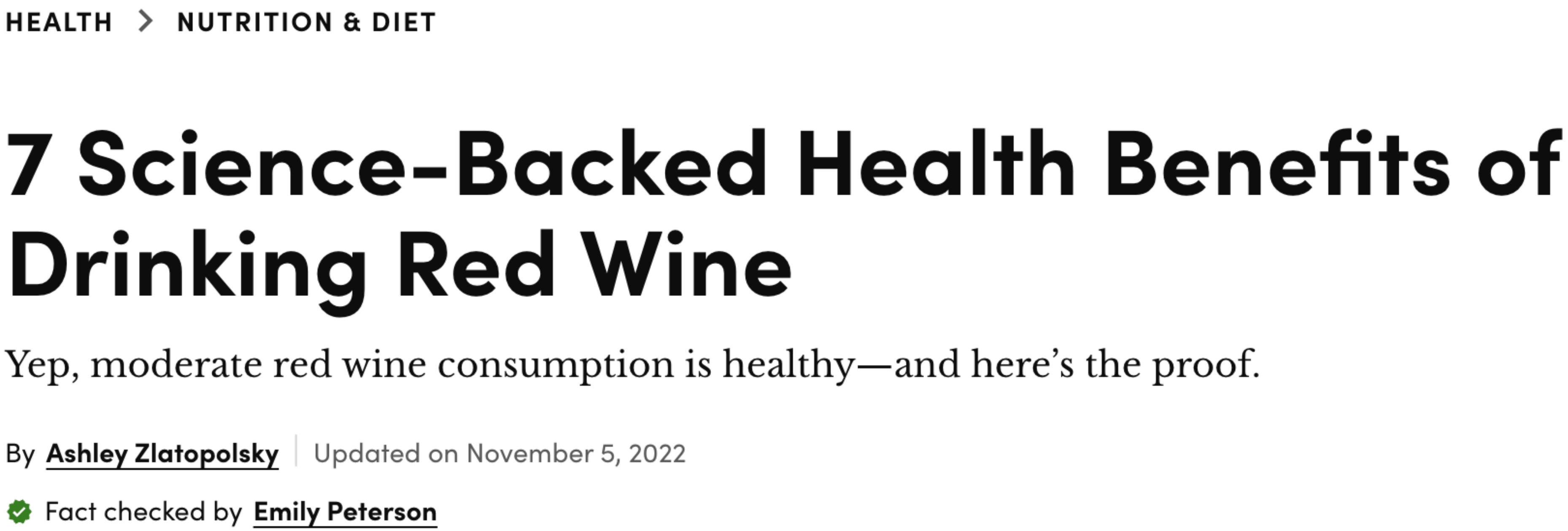


Need to control for U to consistently estimate the causal effect

Confounding bias

- Observed data regression of T on Y fails because the distribution of U varies in the two treatment arms
- We try to condition on as many *observed* confounders as possible to mitigate potential confounding bias
- Commonly assumed that there are “no unobserved confounders” (NUC) but this is unverifiable
- Sensitivity analysis is a tool for assessing the impacts of violations of this assumption

A Motivating Example



HEALTH > NUTRITION & DIET

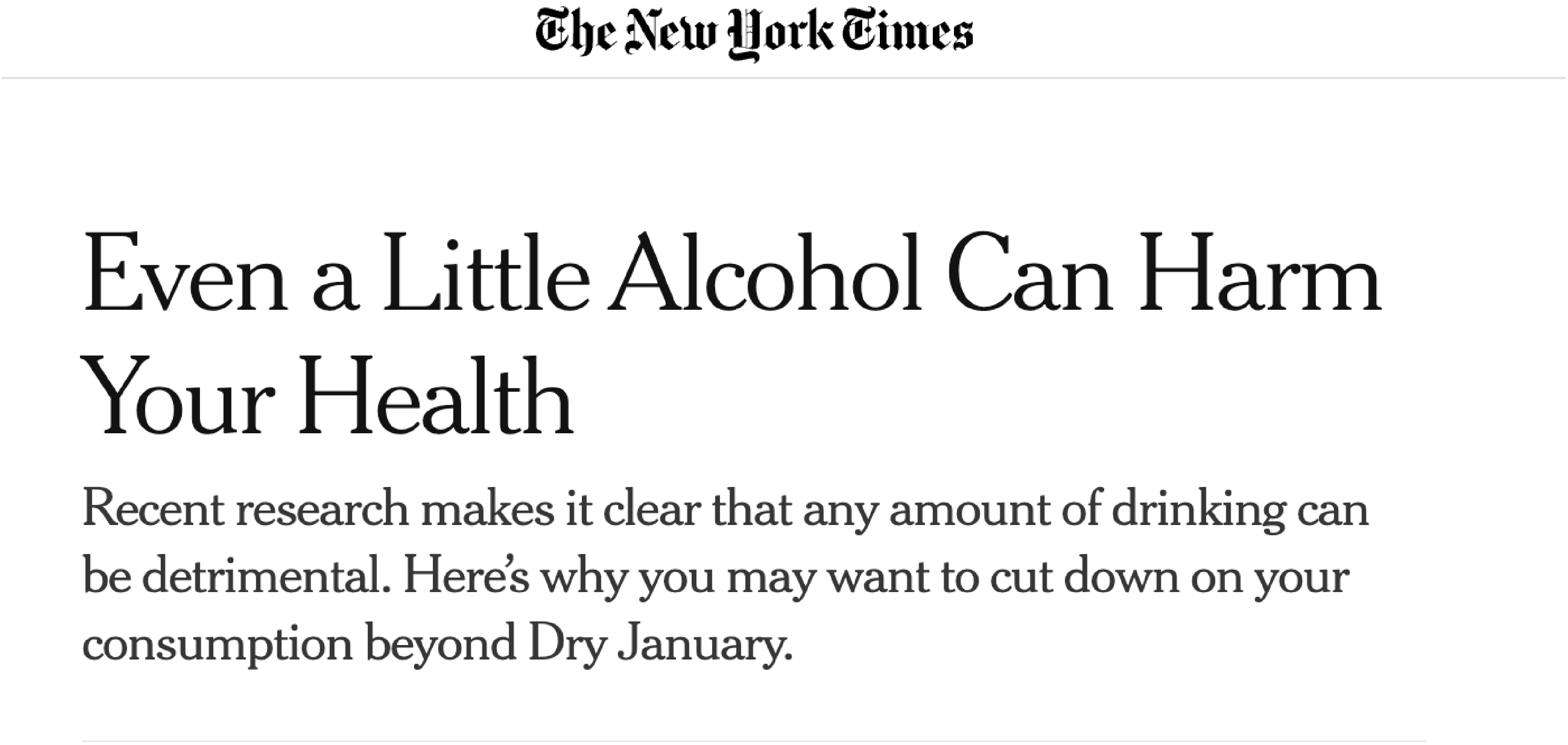
7 Science-Backed Health Benefits of Drinking Red Wine

Yep, moderate red wine consumption is healthy—and here's the proof.

By Ashley Zlatopolsky | Updated on November 5, 2022

✓ Fact checked by Emily Peterson

A Motivating Example



The New York Times

Even a Little Alcohol Can Harm Your Health

Recent research makes it clear that any amount of drinking can be detrimental. Here’s why you may want to cut down on your consumption beyond Dry January.

The Effects of Light Alcohol Consumption

- Observational data from the National Health and Nutrition Examination Study (NHANES) on alcohol consumption.
- Light alcohol consumption is positively correlated with blood levels of HDL (“good cholesterol”)
- Define “light alcohol consumption’’ as 1-2 alcoholic beverages per day
- Non-drinkers: self-reported drinking of one drink a week or less
- Control for age, gender and indicator for educational attainment

HDL and alcohol consumption

```r
summary(lm(Y[, "HDL"] ~ drinking + X))
```

```

Call:
lm(formula = Y[, "HDL"] ~ drinking + X)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.0855 -0.6127 -0.0512  0.6389  4.2383 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.225550   0.091105   2.476 0.013412 *  
drinking    0.597399   0.091917   6.499 1.11e-10 ***
Xage        0.006409   0.001452   4.415 1.09e-05 ***
Xgender     0.689557   0.049426  13.951  < 2e-16 ***
Xeduc       0.194338   0.051161   3.799 0.000152 ***
```

What must be true for this correlation to be non-causal?

Blood mercury and alcohol consumption

```r
summary(lm(Y[, "Methylmercury"] ~ drinking + X))
```

```
Call:
lm(formula = Y[, "Methylmercury"] ~ drinking + X)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.3570 -0.7363 -0.0728  0.6242  4.1127 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.442044   0.096385   4.586 4.91e-06 ***
drinking     0.364096   0.097244   3.744 0.000188 ***
Xage         0.008186   0.001536   5.330 1.14e-07 ***
Xgender     -0.062664   0.052290  -1.198 0.230966    
Xeduc        0.269815   0.054126   4.985 6.95e-07 ***
```

But… no plausible causal mechanism in this case

Residual Correlation

```r
hdl_fit <- lm(Y[, "HDL"] ~ drinking + X)
mercury_fit <- lm(Y[, "Methylmercury"] ~ drinking + X)

cor.test(hdl_fit$residuals, mercury_fit$residuals)
```

```

	Pearson's product-moment correlation

data:  hdl_fit$residuals and mercury_fit$residuals
t = 3.7569, df = 1437, p-value = 0.0001789
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.04718758 0.14953581
sample estimates:
      cor 
0.0986225 
```

Residual correlation might be indicative of confounding bias

Sensitivity Analysis

- NUC unlikely to hold exactly. What then?
- Calibrate assumptions about confounding to explore range of causal effects that are plausible
- Robustness: quantify how “strong” confounding has to be to nullify causal effect estimates

- Many methods for single outcome analyses.

See e.g. (Cinelli and Hazlett 2020)

Multi-outcome Sensitivity Analysis

- If we measure multiple outcomes, is there prior knowledge that we can leverage to strengthen causal conclusions?

- What might residual correlation in multi-outcome models mean for potential for confounding?

- How do results change when we assume a priori that certain outcomes cannot be affected by treatments?
  - Null control outcomes (e.g. alcohol consumption should not increase mercury levels)

A Structural Equation Model

- U (m-vector) and X are possible causes for T (scalar) and Y (q-vector)
- X are observed but U are not.

$$\begin{aligned} \mathrm{U} &= \epsilon_{\mathrm{U}} \\ \mathrm{T} &= \mathrm{f}_{\epsilon}(\mathrm{X}, \mathrm{U}) \\ \mathrm{Y} &= \mathrm{g}(\mathrm{T}, \mathrm{X}) + \Gamma\Sigma_{\mathrm{u}|\mathrm{t},\mathrm{x}}^{-1/2}\mathrm{U} + \epsilon_{\mathrm{y}}, \end{aligned}$$

- This SEM is compatible with the observed data having factor structured residuals, $\mathrm{Cov}(\mathrm{Y}|\mathrm{T}, \mathrm{X}) = \Gamma\Gamma' + \Lambda$

A Sensitivity Specification

- Propose a sensitivity parameterization governed by a single m-vector, ϱ, the partial correlation vector between T and U
- Define $0 \leq R^2_{T \sim U|X} := \frac{||\varrho||_2^2}{\sigma^2_{t|x}} < 1$ to be the squared norm of the partial correlation between T and U given X
- Confounding bias is a function of factor loadings, Γ, and sensitivity vector, ϱ

Multi-outcome Assumptions

- Multi-outcome assumptions:

Assumption (Factor confounding)

All potential confounding is reflected in correlation among outcomes.

Assumption (Factor identifiability)

Factor loadings are identifiable (up to rotation) (Anderson and Rubin 1956)

Bounding the Omitted Variable Bias

Theorem (Bounding the bias for outcome Y_j)

Given the structural equation model, sensitivity specification and given assumptions, the squared omitted variable bias for the PATE of outcome Y_j is bounded by

$$\text{Bias}_j^2 \leq \frac{(t_1 - t_2)^2}{\sigma_{t|x}^2} \left(\frac{R^2_{T \sim U|X}}{1 - R^2_{T \sim U|X}} \right) \| \Gamma_j \|_2^2$$

- The bound on the bias for outcome j is proportional to the norm of the factor loadings for that outcome
- A single sensitivity parameter, $R^2_{T \sim U|X}$, shared across all outcomes

Bounding the Omitted Variable Bias

Theorem (Bounding the bias for outcome Y_j)

Given the structural equation model, sensitivity specification and given assumptions, the squared omitted variable bias for the PATE of outcome Y_j is bounded by

$$\text{Bias}_j^2 \leq \frac{(t_1 - t_2)^2}{\sigma_{t|x}^2} \left(\frac{R^2_{T\sim U|X}}{1 - R^2_{T\sim U|X}} \right) \|\Gamma_j\|_2^2$$

- The bound on the bias for outcome j is proportional to the norm of the factor loadings for that outcome
- A single sensitivity parameter, $R^2_{T\sim U|X}$, shared across all outcomes

Bounding the Omitted Variable Bias

Theorem (Bounding the bias for outcome Y_j)

Given the structural equation model, sensitivity specification and given assumptions, the squared omitted variable bias for the PATE of outcome Y_j is bounded by

$$\mathrm{Bias}_j^2 \leq \frac{(t_1 - t_2)^2}{\sigma_{t|x}^2} \left(\frac{R^2_{T\sim U|X}}{1 - R^2_{T\sim U|X}} \right) \| \Gamma_j \|_2^2$$

- The bound on the bias for outcome j is proportional to the norm of the factor loadings for that outcome
- A single sensitivity parameter, $R^2_{T\sim U|X}$, shared across all outcomes

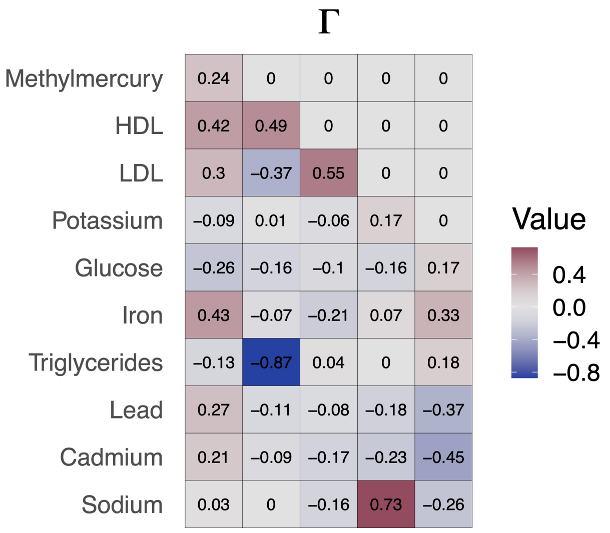
Null Control Outcomes

- Null control outcomes are outcomes which we assume are not caused by the treatment
  - e.g. methylmercury in drinking example
- Theory tells us how null control assumptions change the sensitivity/robustness to additional unmeasured confounding.

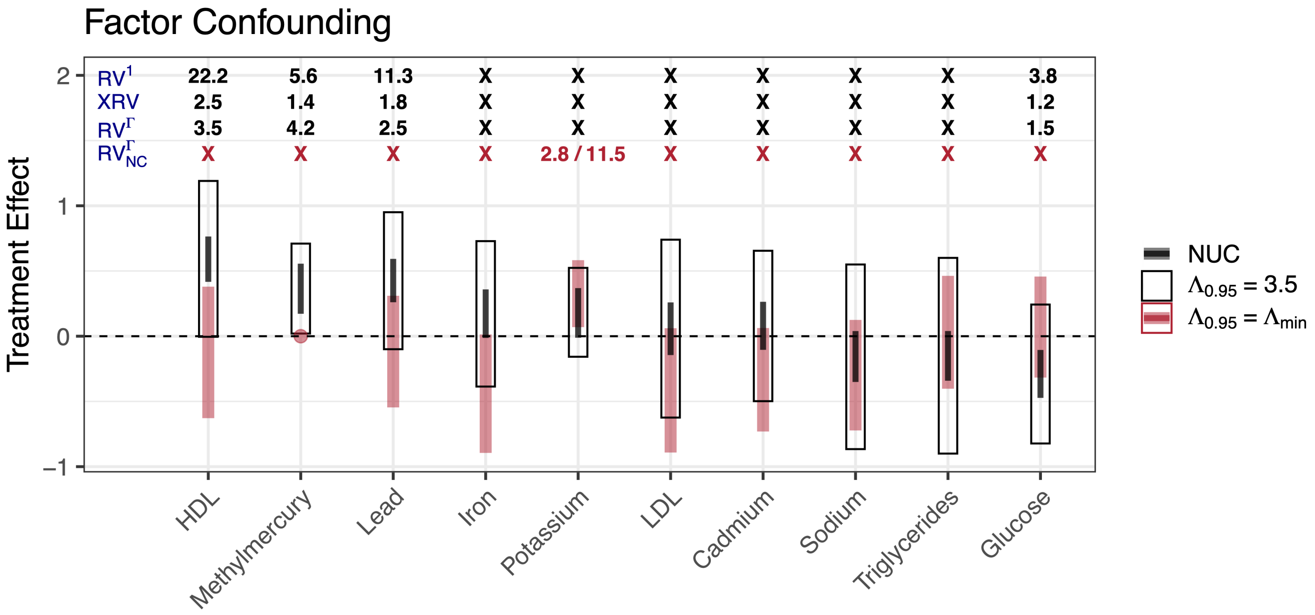
The effects of light drinking

- Measure ten different outcomes from blood samples:
  - natural: HDL, LDL, triglycerides, potassium, iron, sodium, glucose
  - environmental toxicants: mercury, lead, cadmium.
- Measured confounders: age, gender and indicator for highest educational attainment
- Residual correlation in the outcomes might be indicative of additional confounding bias

Results: NHANES alcohol study



Methylmercury	0.24	0	0	0	0
HDL	0.42	0.49	0	0	0
LDL	0.3	−0.37	0.55	0	0
Potassium	−0.09	0.01	−0.06	0.17	0
Glucose	−0.26	−0.16	−0.1	−0.16	0.17
Iron	0.43	−0.07	−0.21	0.07	0.33
Triglycerides	−0.13	−0.87	0.04	0	0.18
Lead	0.27	−0.11	−0.08	−0.18	−0.37
Cadmium	0.21	−0.09	−0.17	−0.23	−0.45
Sodium	0.03	0	−0.16	0.73	−0.26



	HDL	Methylmercury	Lead	Iron	Potassium	LDL	Cadmium	Sodium	Triglycerides	Glucose
RV^1	22.2	5.6	11.3	X	X	X	X	X	X	3.8
XRV	2.5	1.4	1.8	X	X	X	X	X	X	1.2
RV^Γ	3.5	4.2	2.5	X	X	X	X	X	X	1.5
RV^Γ_{NC}	X	X	X	X	2.8 / 11.5	X	X	X	X	X

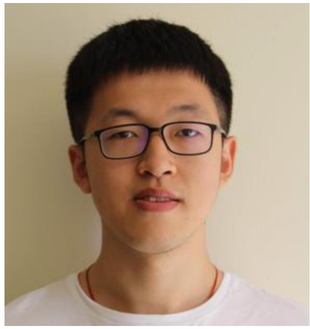
Takeaways

- Prior knowledge unique to the multi-outcome setting can help inform assumptions about confounding
- Sharper sensitivity analysis, when assumptions hold
- Negative control assumptions can potentially provide strong evidence for or against robustness

Comments

- Slides: afranks.com/talks
- *Sensitivity to Unobserved Confounding in Studies with Factor-structured Outcomes*, (JASA, 2023) https://arxiv.org/abs/2208.06552
- Identification with multiple treatments multiple outcomes
  - Collaboration on effects of pollutants on multiple heath outcomes
- Sensitivity analysis for more general models / forms of dependence.

Thanks!



- Jiaxi Wu (top, UCSB)
- Jiajing Zheng (middle, formerly UCSB)
- Alex D’Amour (bottom, Google Research)

References

Anderson, Theodore W., and Herman Rubin. 1956. "STATISTICAL INFERENCE IN FACTOR ANALYSIS." In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 5: Contributions to Econometrics, Industrial Research, and Psychometry*, 3.5:111–50. University of California Press.

Cinelli, Carlos, and Chad Hazlett. 2020. "Making Sense of Sensitivity: Extending Omitted Variable Bias." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82 (1): 39–67.

A Sensitivity Specification

- Interpretable specification for $\mu_{u|t,x}$ and $\Sigma_{u|t,x}$ parameterized by a single m-vector, ϱ

$$\mu_{u|t,x} = \frac{\varrho}{\sigma^2_{t|x}}\left(t - \mu_{t|x}\right) \ ,$$

$$\Sigma_{u|t,x} = I_m - \frac{\varrho\varrho'}{\sigma^2_{t|x}},$$