Hamed Nikookalam

PS 4

Table of Contents

Part 1: Reading

1. What is the difference between a confounder and a collider? How should you address each in your models?

A confounder is a variable that causes both your treatment X and your outcome Y. In a DAG, it usually looks like Z → Xand Z → Y. If you ignore such a variable, you mix up the effect of Xwith the effect of Z, which creates omitted variable bias. The solution is to adjust for true confounders (include them as controls in a regression) so that backdoor paths from X to Y running through Z are blocked.

A collider, by contrast, is a variable that is an effect of two (or more) variables, like X → Z ← Y. Z is hit from both sides, so it is a common effect, not a common cause. The articles and class slides emphasize that you should not adjust for colliders because conditioning on a collider opens a spurious path between its causes.

2. How can conditioning on a collider create bias?

In a collider structure $X \rightarrow Z \leftarrow Y$, $X$and Ycan be independent in the full population. When you condition on Z—by controlling for it, restricting the sample, or selecting cases with certain values of Z—You open a path between $X$and Y, making them statistically associated even if there is no causal relationship.

In the SAT–GPA example shows exactly this: admission status is a collider determined by SAT and GPA. Among all applicants, SAT and GPA may not be negatively related, but if you only look at students near or above the admissions cutoff, they appear negatively correlated—this is Berkson's paradox, a classic example of collider bias. This logic also underlies selection bias: when inclusion in the study is a collider influenced by both X and Y, analyzing only those selected cases produces biased estimates. Sampling bias, survivorship bias, and nonresponse bias are all special cases where our sample is systematically restricted (who gets sampled, who survives, who answers), effectively conditioning on a collider and making the observed relationships unrepresentative of the target population.

3. Why can't statistical summaries or correlations alone tell us whether to control for a variable?

Both articles stress that statistics alone can't reveal causal roles. The "causal quartet" in Causal Inference Is Not Just a Statistics Problem constructs four datasets that have identical summary

statistics and scatterplots, but in each one, the same variable plays a different causal role: collider, confounder, mediator, or part of an M-bias structure. Even though the tables and plots look the same, the true causal effect and the correct adjustment set differ across scenarios.

This means that simply seeing that a variable is correlated with X or Y (or has a small p-value) does not tell us whether we should control for it, because that variable might be a confounder (good to adjust), a mediator (adjusting changes the estimand), or a collider (adjusting induces bias). As Kuhle et al. argue, deciding what to adjust for requires substantive knowledge and a causal model (a DAG), not just automated variable selection or looking at correlations.

4. What is meant by a "kitchen sink" regression, and what is wrong with this approach to modeling?

A "kitchen sink" regression is a model where you throw in every available variable—or use stepwise procedures that add/drop variables based only on p-values or information criteria—without a clear causal rationale. The idea is "control for everything that might matter" and let the software decide.

Kuhle et al. show several problems with this approach. First, it ignores the direction of causal relationships: it treats all covariates as interchangeable, so you may adjust for mediators or colliders and thereby increase bias instead of reducing it. Second, estimates from such models often lack a clear causal interpretation, because you haven't specified what effect you're trying to estimate or what assumptions justify it. Third, including many variables based on significance inflates the type-I error rate, encourages data-mining, and risks overfitting and unstable coefficients, especially with limited sample size. Finally, kitchen-sink modeling sidelines domain expertise; the papers argue we should instead build models guided by DAGs and substantive theory about which variables are confounders that need to be controlled.

5. What is a "backdoor path" and how does multiple regression help block these paths?

A backdoor path is any path in a DAG from the treatment Xto the outcome Ythat starts with an arrow pointing into Xand carries association from common causes (confounders) rather than from the causal effect we care about. For example, $X \leftarrow Z \rightarrow Y$ is a simple backdoor path: Z is a confounder generating a spurious relation between X and Y. If backdoor paths are open, our estimate of the effect of X on Y will be biased because it mixes causal and non-causal association

Multiple regression can block these backdoor paths when we condition on the right set of variables—specifically, on confounders like Z but not on colliders. In the DAG framework used in both articles, we first use substantive knowledge to draw a graph, then identify a minimal sufficient adjustment set that closes all backdoor paths from X to Y while leaving the causal path

X → Y open. Including exactly that set as controls in the regression gives us an unbiased estimate of the causal effect (under the DAG's assumptions), whereas kitchen-sink or purely data-driven adjustment can leave some backdoor paths open or open new paths through colliders.

Part 2: Simulation

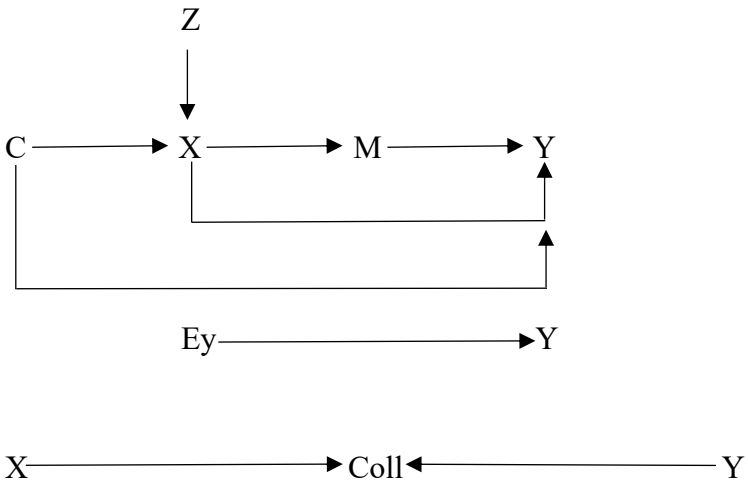
We study **how a pro-democracy social-media campaign affects people's willingness to participate in protest**.

Variables:

- X (treatment): exposure to the social-media campaign
- Y (outcome): willingness to protest (e.g., 0–10 scale)
- C (confounder): baseline political interest
    - more interested → more likely to see the campaign and more willing to protest
- M (mediator): activist network size after exposure
    - exposure grows your activist network, which increases protest willingness
- Coll (collider): invitation to a secret encrypted protest group
    - People with high X and high Y are more likely to be invited
- Ey (Y-only exogenous): risk-taking personality
    - affects protest willingness but not exposure
- Z (instrument / X-only exogenous): high-speed internet upgrade
    - makes exposure more likely, but has no direct effect on protest willingness

DAG:

political interest affects both exposure and protest willingness; the internet upgrade only affects exposure; exposure has both a direct effect on protest willingness and an indirect effect through activist networks; risk-taking only affects protest willingness; and the encrypted-group invitation is a collider caused by both exposure and willingness.



Arrows:

- **C → X, C → Y** (confounder)
- **Z → X** (instrument)
- **X → M → Y** (mediator path)
- **X → Y** (direct effect)
- **Ey → Y** (Y-only exogenous)
- **X → Coll ← Y** (collider)

Variables with no parents are purely random; all others are linear functions of their parents + error.

1. **Exogenous variables (no parents → generate first)**

   These are not caused by anything in the DAG:

   $$C \sim N(0, 1)$$

   $$Ey \sim N(0, 1)$$

   $$Z \sim N(0, 1)$$

2. **Treatment *X***

   Let political interest and the internet upgrade both increase exposure:

   $$X = 0.8*C + 0.6*Z + \varepsilon_X$$

3. **Mediator *M***

Network size mostly depends on exposure:

$$M = 0.7X + 0.3C + \varepsilon_M$$

4. **Outcome *Y***

Let Y depend on:

- Direct Effect of X
- mediator M
- confounder C
- Y-only exogenous Ey

$$Y = 0.5X + 0.6M + 0.8C + 0.7Ey + \varepsilon_Y$$

5. **Collider *Coll***

Collider is caused by both X and Y:

$$Coll = 0.7X + 0.7Y + \varepsilon_{coll}$$

Which, ε_X, ε_M, ε_Y, $\varepsilon_{coll} \sim N(0, 1)$ independently.

Code

```
set.seed(123)

n <- 2000   # sample size


## 1. Exogenous variables (no parents)

C  <- rnorm(n)   # confounder

Ey <- rnorm(n)   # Y-only exogenous variable

Z  <- rnorm(n)   # instrument (X-only exogenous)


## 2. Treatment

X  <- 0.8*C + 0.6*Z + rnorm(n)


## 3. Mediator
```

```
M  <- 0.7*X + 0.3*C + rnorm(n)


## 4. Outcome

Y  <- 0.5*X + 0.6*M + 0.8*C + 0.7*Ey + rnorm(n)


## 5. Collider

Coll <- 0.7*X + 0.7*Y + rnorm(n)


## Final dataset

dat <- data.frame(Y, X, C, M, Coll, Ey, Z)

head(dat)
```

output

Y	X	C	M	Coll	Ey	Z
0.82263	-1.0297	-0.5605	-0.3779	-0.2956	-0.5116	0.19655
1.96445	1.20238	-0.2302	2.5806	3.01772	0.23694	0.65011
1.93409	0.95682	1.55871	-0.5652	0.83692	-0.5416	0.671
-0.2253	-0.8176	0.07051	-0.2637	-0.2994	1.21923	-1.2842
-0.1555	-0.5084	0.12929	-0.5862	-0.2479	0.17414	-2.0261
4.07457	2.0872	1.71506	1.59604	5.27454	-0.6153	2.20533

1. Fit a model that recovers the direct effect of the treatment on the outcome variable. Which variables are necessary to recover the direct effect?

In the equations, the direct path $X \rightarrow Y$ has coefficient 0.5.
The total effect is larger because there is also a mediated path $X \rightarrow M \rightarrow Y$.

At the DAG:

There is a backdoor path $X \leftarrow C \rightarrow Y$.
→ We must adjust for C (the confounder) to block this.

We want the direct effect, not the total effect.
→ We must also adjust for M (the mediator) to block the path $X \rightarrow M \rightarrow Y$.

We must not adjust for the collider Coll (it is $X \rightarrow Coll \leftarrow Y$), because conditioning on a collider opens a spurious path.

The Y-only exogenous variable Ey and the instrument Z are not required for identification of the direct effect (they don't lie on any backdoor path). Ey can be added only for precision.

I estimate the following regression model:

$$Y_i = \beta_0 + \beta_X X_i + \beta_C C_i + \beta_M M_i + \varepsilon_i,$$

where β_X is the direct effect of the treatment X on the outcome Y.

With my simulated data, the estimated coefficient $\hat{\beta}_X$ will be close to the true direct effect (0.5).

Code

```
# Q1. Model for the direct effect of X on Y

mod_direct <- lm(Y ~ X + C + M, data = dat)

summary(mod_direct)
```

Output

Call:

lm(formula = Y ~ X + C + M, data = dat)

Residuals:

Min	1Q	Median	3Q	Max
-4.1869	-0.8459	0.0198	0.8247	3.5689

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.01627	0.02751	0.591	
X	0.43196	0.03118	13.856	
C	0.78593	0.03414	23.023	
M	0.65642	0.02715	24.182	

(Intercept) 0.554

X <2e-16 ***

C <2e-16 ***

M <2e-16 ***

Signif. codes:

0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.23 on 1996 degrees of freedom

Multiple R-squared: 0.7463, Adjusted R-squared: 0.7459

F-statistic: 1957 on 3 and 1996 DF, p-value: < 2.2e-16

To recover the direct effect of the treatment X on the outcome Y, I need to control for:

- C, the confounder, because it causally affects both X and Y and would otherwise open a backdoor path (X ← C → Y).

- M, the mediator, because I want the direct (non-mediated) effect of X on Y, so I must block the path X → M → Y.

I do not control the collider Coll (since conditioning on X → Coll ← Y would introduce spurious association), and the exogenous variables Ey (Y-only) and Z (instrument for X) are not necessary for identification of the direct effect, although Ey can be included just to improve precision.

2. Fit a model that recovers the total effect of the treatment on the outcome variable. How does your model change to estimate the total effect?

To estimate the total effect of treatment X on the outcome Y, I need to block confounding but keep both the direct and indirect (mediated) paths from X to Y.

In my DAG, the confounder C causes both X and Y, so I still need to control C to block the backdoor path X ← C → Y. However, I no longer control the mediator M, because I now want the total effect of X on Y, which includes both the direct path X → Y and the indirect path X → M → Y. Therefore, the regression model for the total effect is:

$Y_i = \beta_0 + \beta_X X_i + \beta_C C_i + \varepsilon_i,$

where β_X is now the total causal effect of X on Y.

Compared to the direct-effect model, the only change is that M is removed from the regression. I still do not control the collider Coll, and the exogenous variables Ey (Y-only) and Z (instrument for X) are not necessary for identification of the total effect.

Code

```
# Q2. Model for the TOTAL effect of X on Y

mod_total <- lm(Y ~ X + C, data = dat)

summary(mod_total)
```

Output

```
Call:

lm(formula = Y ~ X + C, data = dat)


Residuals:

    Min      1Q  Median      3Q     Max

-5.0171 -0.9289  0.0236  0.9361  4.1914


Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.007355   0.031272   0.235    0.814

X           0.919313   0.027040  33.998   <2e-16 ***

C           0.964611   0.037886  25.461   <2e-16 ***

---

Signif. codes:

0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.398 on 1997 degrees of freedom

Multiple R-squared:  0.672,    Adjusted R-squared:  0.6717

F-statistic:  2046 on 2 and 1997 DF,  p-value: < 2.2e-16
```

3. How do your results change when you control for the collider, the exogenous independent variable, or the instrument (individually, not all simultaneously)?

Starting from the total-effect model (Y ~ X + C), the coefficient on X is close to the true total effect of X on Y.

When I add the collider Coll to the model (Y ~ X + C + Coll), the estimated coefficient on X changes substantially and moves away from the true total effect. This happens because Coll is a collider on the path X → Coll ← Y, and conditioning on a collider opens a spurious association between X and Y (collider bias).

When I add the Y-only exogenous variable Ey instead (Y ~ X + C + Ey), the coefficient on X stays essentially the same as in the baseline total-effect model, but its standard error becomes smaller. Ey affects Y but not X and is not on any backdoor or collider path, so including it does not bias the estimate; it only improves precision.

Finally, when I add the instrument Z (Y ~ X + C + Z), the coefficient on X again remains very close to the baseline total effect. Z affects X but has no direct effect on Y, so adjusting for it is not necessary for identification in this OLS setup and does not systematically change the estimated causal effect, though it may slightly change the standard error.

Code

```
# Q3. How do results change when adding Coll, Ey, or Z?


# 1) Add the collider

mod_coll <- lm(Y ~ X + C + Coll, data = dat)

summary(mod_coll)


# 2) Add the exogenous Y-only variable

mod_Ey <- lm(Y ~ X + C + Ey, data = dat)

summary(mod_Ey)


# 3) Add the instrument (X-only exogenous)

mod_Z <- lm(Y ~ X + C + Z, data = dat)

summary(mod_Z)
```

Output

Call:

lm(formula = Y ~ X + C + Z, data = dat)

Residuals:

Min	1Q	Median	3Q	Max
-5.0166	-0.9321	0.0206	0.9345	4.2027

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.00725	0.03128	0.232	0.817	
X	0.91284	0.03137	29.103	<2e-16	***
C	0.96991	0.04006	24.211	<2e-16	***
Z	0.01484	0.03641	0.408	0.684	

Signif. codes:

0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.398 on 1996 degrees of freedom

Multiple R-squared: 0.672, Adjusted R-squared: 0.6715

F-statistic: 1363 on 3 and 1996 DF, p-value: < 2.2e-16

4. Given the reading and simulation results, how should you choose which variables to include in a model?

First, I need to decide whether I want the total effect of X on Y or the direct effect. If I want the total effect, I should adjust for pre-treatment confounders that cause both X and Y (like C in my DAG), but I should not control for mediators (like M) because that would block part of the effect I am trying to estimate. If I want the direct effect, I should still adjust for confounders like C, and now I should also adjust for the mediator M to block the indirect path X → M → Y.

Second, I should never control colliders, such as Coll in my simulation. The collider is caused by both X and Y, and the reading shows that conditioning on a collider opens a spurious path between its causes and creates "collider bias." My results confirm this: including Coll made the estimated effect of X move away from the true value. By contrast, adding purely exogenous variables that only affect Y (Ey) or only affect X (Z, the instrument) did not change the estimated causal effect of X on Y, it only affected precision. Overall, the right strategy is to use substantive knowledge to draw a DAG, identify a minimal sufficient adjustment set that blocks all backdoor paths without conditioning on colliders, and include exactly that set in the model rather than relying on kitchen-sink regressions or purely statistical variable selection.