Pset4

Gabriel Wang

December 2, 2025

1 Part 1: Reading

Question 1

A collider is a variable that is influenced by two or more other variables, even though those variables may not be causally related to each other. In a DAG, this structure is represented as X → Z ← Y. Correctly identifying a collider requires understanding, or at least reasonably assuming, the data-generating mechanism using domain knowledge and visualizing it with a directed acyclic graph (DAG). If Z occurs after both the exposure and the outcome, we can check whether Z indeed happens after X and Y in the time order. Colliders should not be adjusted for, because conditioning on a collider opens a spurious path between X and Y.

A confounder is a variable that causally affects both the exposure and the outcome, creating a non-causal association between them. In a DAG, this is expressed as $X \leftarrow Z \rightarrow Y$. Again, understanding the underlying data-generating mechanism and using a DAG is essential. A confounder typically occurs before the exposure and also influences the outcome, so we can check whether Z appears before X in the time sequence. Confounders should be adjusted for to block the backdoor path. Additional tools such as the change-in-estimate approach or regularization methods (e.g., penalized maximum likelihood, Lasso, ridge regression) may help decide whether to include Z, but these methods must still be guided by causal reasoning rather than purely statistical criteria.

Question 2

Two variables that are connected through a collider may be uncorrelated in the population, but once we condition on the collider, we can induce spurious and even contradictory associations between them. Even if the two variables are truly related, conditioning on a collider can still distort the estimation of the average treatment effect (ATE).

Question 3

Statistical summaries or correlations alone cannot tell us whether to control for a variable because they ignore the underlying causal structure.

First, they do not reveal the direction of relationships. Without assumptions about the data-generating mechanism, we cannot distinguish confounders, mediators, or colliders.

Second, purely statistical models often produce effect estimates with no meaningful causal interpretation, leading to misinterpretation of coefficients.

Third, using significance tests to select variables inflates the overall Type I error rate—multiple testing increases the chance of falsely identifying a variable as important.

Fourth, data-driven selection can lead to overfitting and unstable models, where many "significant" variables are simply noise specific to the dataset.

Finally, correlations ignore domain knowledge; without substantive understanding or DAGs, we cannot determine whether a variable should or should not be controlled.

Question 4

A "kitchen sink" regression refers to throwing every available variable into a model and selecting variables based on p-values or information criteria. This approach is flawed because it ignores the directionality of relationships, making it impossible to distinguish confounders, mediators, and colliders; it produces effect estimates with no meaningful causal interpretation; it inflates the Type I error rate due to extensive multiple testing; and it leads to unstable, overfit models driven by noise rather than causal structure. Also it wasted resources of researches.

Question 5

A “backdoor path” is a non-causal path that creates a spurious association between two variables, even when no direct causal relationship exists. It typically arises when a third variable is a common cause of both the exposure and the outcome (a confounder).

Multiple regression helps block these backdoor paths by conditioning on the confounding variable—represented as “boxing” the variable in a DAG. Once we adjust for the confounder in the regression model, the backdoor path is closed, and the estimated association between the exposure and outcome becomes unbiased, assuming no other confounders remain unaccounted for.

2 Part 2: Simulation

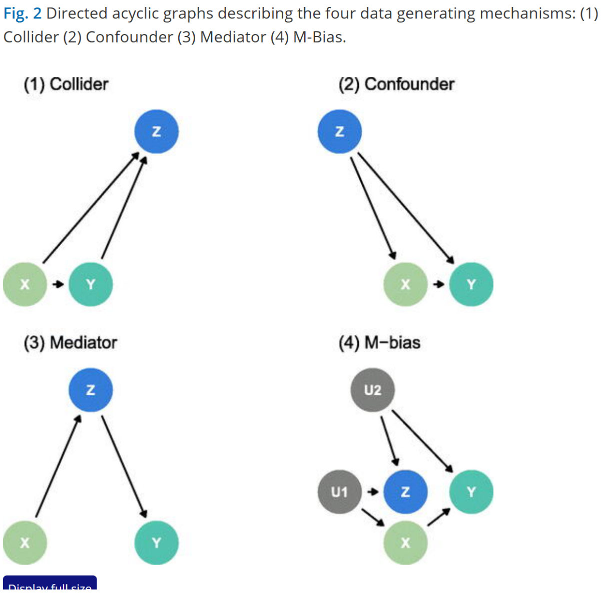

Fig. 2 Directed acyclic graphs describing the four data generating mechanisms: (1) Collider (2) Confounder (3) Mediator (4) M-Bias.

Figure 1: Enter Caption

Question 1

```
#question1
set.seed(1)
n <- 100

# 1. Exogenous variables
C <- rnorm(n)   # Confounder
W <- rnorm(n)   # Instrument
E <- rnorm(n)   # Exogenous independent variable

# 2. Treatment X (affected by confounder C and instrument W)
X <- 0.6*C + 0.8*W + rnorm(n)

# 3. Mediator M (affected by X)
M <- 0.7*X + rnorm(n)

# 4. Outcome Y (affected by X, M, and C; NOT W)
Y <- 1.0*X + 0.5*M + 0.4*C + rnorm(n)

# 5. Collider Z (affected by X and Y)
Z <- X + Y + rnorm(n)

Direct_Effect <- lm(Y~X+M+C)
stargazer::stargazer(Direct_Effect)
```

Table 1:

	Dependent variable:
	Y
X	0.967^{***} (0.105)
M	0.445^{***} (0.085)
C	0.331^{***} (0.117)
Constant	-0.033 (0.098)
Observations	100
R^2	0.809
Adjusted R^2	0.803
Residual Std. Error	0.969 (df = 96)
F Statistic	135.865^{***} (df = 3; 96)
Note:	*p<0.1; **p<0.05; ***p<0.01

Direct effect is 0.96749.

Question 2

```
Total_Effect <- lm(Y~X+C)
print(Total_Effect)
stargazer::stargazer(Total_Effect)
```

Table 2:

	Dependent variable:
	Y
X	1.327^{***} (0.090)
C	0.390^{***} (0.132)
Constant	-0.061 (0.110)
Observations	100
R^2	0.754
Adjusted R^2	0.749
Residual Std. Error	1.094 (df = 97)
F Statistic	148.999^{***} (df = 2; 97)
Note:	*p<0.1; **p<0.05; ***p<0.01

The total effect is 1.32671.

Question 3

```r
colliderlm <- lm(Y~X+Z)
Exogenouslm <- lm(Y~X+E)
Instrumentlm <- lm(Y~X+W)

stargazer::stargazer(colliderlm,Exogenouslm,Instrumentlm )
```

Table 3:

	Dependent variable:		
	Y		
	(1)	(2)	(3)
X	0.205	1.427***	1.423***
	(0.124)	(0.088)	(0.104)
Z	0.524***		
	(0.047)		
E		-0.022	
		(0.111)	
W			0.008
			(0.142)
Constant	0.087	-0.027	-0.027
	(0.076)	(0.114)	(0.115)
Observations	100	100	100
R^2	0.883	0.732	0.732
Adjusted R^2	0.880	0.727	0.727
Residual Std. Error (df = 97)	0.755	1.142	1.142
F Statistic (df = 2; 97)	365.499***	132.723***	132.655***
Note:	*p<0.1; **p<0.05; ***p<0.01		

When conditioning on the collider (Column 1), the estimate of the treatment effect is severely biased due to opening a backdoor path between X and Y through Z.

Conditioning on an exogenous independent variable (Column 2) does not change the estimate, because E is unrelated to both X and Y and therefore does not violate any causal identification assumptions.

Conditioning on the instrument (Column 3), while still producing an approximately correct coefficient due to the simulated DGP, is theoretically inappropriate because it violates the exclusion restriction and destroys the rationale for using W as an instrument.

Question 4

Based on the reading and my simulation results, I should include only **confounders**—variables that causally affect both the treatment X and the outcome Y. Adjusting for confounders blocks backdoor paths and reduces bias.

I should **not** include **mediators**, unless I aim to estimate the **direct effect**. Conditioning on mediators blocks part of the treatment's causal pathway and prevents identification of the **total effect**.

I should also avoid conditioning on **colliders**, which opens non-causal backdoor paths and introduces bias.

Similarly, I should **not** control for **instruments**. Conditioning on an IV violates the exclusion restriction and undermines IV identification.

Thus, I should **control only for confounders** and avoid mediators, colliders, and instruments unless required by a specific identification strategy.