Econometrics III: Applied Methods — NovaSBE, 2026

Problem Set 1 Solutions: Exercise C (OVB)

C.1 The OVB Formula

1.1 *(1 pt)* Show that the probability limit of the OLS estimator $\hat{\tilde{\beta}}$ from the short regression is:

$$\text{plim } \hat{\tilde{\beta}} = \beta + \gamma \cdot \delta$$

where δ is defined as the slope coefficient from the population regression of the omitted variable Z on the included regressor X. This is the *omitted variable bias (OVB) formula.*

Solution:

The OLS estimator satisfies plim $\hat{\tilde{\beta}} = \text{Cov}(X, Y)/\text{Var}(X)$.
Substituting $Y = \alpha + \beta X + \gamma Z + \varepsilon$:

$$\begin{aligned} \text{Cov}(X, Y) &= \text{Cov}(X, \alpha + \beta X + \gamma Z + \varepsilon) \\ &= \beta \cdot \text{Var}(X) + \gamma \cdot \text{Cov}(X, Z) \end{aligned}$$

Therefore:

$$\text{plim } \hat{\tilde{\beta}} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \beta + \gamma \cdot \frac{\text{Cov}(X, Z)}{\text{Var}(X)} = \beta + \gamma \cdot \delta$$

1.2 *(0.5 pts)* Interpret each component of the OVB formula. Based on this, explain the conditions under which there would be no bias even when Z is omitted from the regression.

Solution:

Components:

- γ = effect of the omitted variable Z on the outcome Y
- δ = relationship between the omitted variable Z and the included regressor X

No bias if either:

1. $\gamma = 0$: the omitted variable doesn't affect Y (so omitting it is harmless)
2. $\delta = 0$: the omitted variable is uncorrelated with X (so its omission doesn't contaminate the estimate of β)

C.2 Application: Returns to Education

2.1 *(0.5 pts)* What sign do you expect for γ? What sign do you expect for δ? Based on these predictions, what is the expected sign of the omitted variable bias when ability is omitted?

Solution:

- $\gamma > 0$: Higher ability raises wages (more productive workers earn more)
- $\delta > 0$: Higher ability people get more education (able people find schooling easier/more rewarding)

Therefore: Bias $= \gamma \cdot \delta > 0$ is **upward**. The short regression overestimates the true return to education.

2.2 *(0.5 pts)* A researcher argues: "The bias from omitting ability is probably small because ability only explains a small fraction of the total variation in wages." Is this argument correct? Explain your answer.

Solution:

No, this argument is incorrect.

OVB depends on γ (effect of ability on wages) and δ (correlation between ability and education), *not* on the R^2 of ability in the wage equation.

A variable can explain little overall variance in Y (low partial R^2) but still cause large bias if it is strongly correlated with X. The magnitude of bias depends on the product $\gamma \cdot \delta$, not on how much variance ability explains.

2.3 *(0.5 pts)* Suppose $\gamma = 0.3$ (a one standard deviation increase in ability raises log wages by 0.3), $\delta = 0.15$ (each year of education is associated with 0.15 SD higher ability), and the true return to education is $\beta = 0.06$ (6% per year). Calculate what the short regression (omitting ability) would estimate. Comment.

Solution:

$$\text{plim}\ \hat{\tilde{\beta}} = \beta + \gamma \cdot \delta = 0.06 + 0.3 \times 0.15 = 0.06 + 0.045 = \boxed{0.105}$$

The short regression estimates a 10.5% return to education instead of the true 6%. The bias (4.5 percentage points) is substantial—it overstates the true effect by 75%.

C.3 OVB in Practice (R)

3.1 *(0.5 pts)* Estimate the "short" regression of `wage` on `education` only (omitting ability). Report the coefficient on education and comment.

Solution:

```
short_reg <- lm(wage ~ education, data = ovb_data)
```

Coefficient on education: $\hat{\beta} \approx 0.112$

This is biased upward relative to the true effect of 0.08, consistent with positive OVB from omitting ability.

3.2 *(1 pt)* Now estimate the "long" regression that includes `ability` as a control. Report the coefficient on education. Then verify that the OVB formula holds empirically: (i) estimate $\hat{\gamma}$, (ii) estimate $\hat{\delta}$, (iii) compute $\hat{\gamma} \times \hat{\delta}$, and confirm this equals the difference between your short and long regression coefficients.

Solution:

```
long_reg <- lm(wage ~ education + ability, data = ovb_data)
aux_reg <- lm(ability ~ education, data = ovb_data)
```

Results:

- Long regression coefficient on education: $\hat{\beta}_{\text{long}} \approx 0.078$ (close to true 0.08)
- $\hat{\gamma} \approx 0.255$ (coefficient on ability in long regression)
- $\hat{\delta} \approx 0.134$ (coefficient from regressing ability on education)

OVB verification:

- Theoretical bias: $\hat{\gamma} \times \hat{\delta} = 0.255 \times 0.134 = 0.034$
- Actual bias: $\hat{\beta}_{\text{short}} - \hat{\beta}_{\text{long}} = 0.112 - 0.078 = 0.034$ ✓

The OVB formula holds exactly in finite samples for OLS.

3.3 *(1 pt)* The variable `ability_proxy` is a noisy measure of true ability (it equals ability plus measurement error). Estimate the regression of `wage` on `education` and `ability_proxy`. Is the coefficient on education closer to the true value of 0.08 than in the short regression? Discuss what this implies for empirical research where we can only observe imperfect proxies for confounders.

Solution:

```
proxy_reg <- lm(wage ~ education + ability_proxy, data = ovb_data)
```

Coefficient on education with proxy: $\hat{\beta} \approx 0.089$
Comparison:

- Short (no control): 0.112 (biased)
- With noisy proxy: 0.089 (partially corrected)
- With true ability: 0.078 (close to true 0.08)

Implication: The proxy reduces but doesn't eliminate bias. Measurement error attenuates the proxy's coefficient, leaving residual confounding. In practice, proxies help but don't fully solve the OVB problem—we should be cautious about claiming causal estimates even when controlling for available proxies.

C.4 Multiple Omitted Variables

4.1 *(1 pt)* Derive the bias in $\hat{\hat{\beta}}$ when both Z_1 and Z_2 are omitted from the regression.

Solution:

Following the same derivation as before, with $Y = \alpha + \beta X + \gamma_1 Z_1 + \gamma_2 Z_2 + \varepsilon$:

$$\text{Cov}(X, Y) = \beta \cdot \text{Var}(X) + \gamma_1 \cdot \text{Cov}(X, Z_1) + \gamma_2 \cdot \text{Cov}(X, Z_2)$$

Therefore:

$$\text{plim } \hat{\hat{\beta}} = \beta + \gamma_1\delta_1 + \gamma_2\delta_2$$

where $\delta_j = \text{Cov}(X, Z_j)/\text{Var}(X)$ is the coefficient from regressing Z_j on X.
Bias $= \gamma_1\delta_1 + \gamma_2\delta_2$

4.2 *(1 pt)* Is it possible for the biases from two omitted variables to cancel each other out? Under what conditions would this happen? Is this possibility a good reason to be unconcerned about omitted variable bias in practice? Explain why or why not.

Solution:

Yes, biases can cancel if $\gamma_1\delta_1 = -\gamma_2\delta_2$.
This requires one bias term to be positive and the other negative (e.g., if $\gamma_1 > 0, \delta_1 > 0$ but $\gamma_2 > 0, \delta_2 < 0$).
This is NOT reassuring for several reasons:

1. It requires *exact* cancellation—a knife-edge condition that is unlikely to hold precisely
2. We cannot verify cancellation without knowing the true parameters (which we don't)
3. Additional omitted variables may not conveniently cancel
4. Even approximate cancellation would be coincidental, not something we can rely on

C.5 Bad Control: Mediators

5.1 *(1 pt)* Suppose the structural equations are:

$$\begin{aligned} M_i &= \alpha_M + \theta X_i + \nu_i \\ Y_i &= \alpha_Y + \phi M_i + \eta_i \end{aligned}$$

where $\mathbb{E}[\nu_i|X_i] = 0$ and $\mathbb{E}[\eta_i|X_i, M_i] = 0$. What is the total causal effect of X on Y? Express your answer in terms of the structural parameters.

Solution:

The causal chain is $X \to M \to Y$. Substituting the first equation into the second:

$$\begin{aligned} Y_i &= \alpha_Y + \phi(\alpha_M + \theta X_i + \nu_i) + \eta_i \\ &= (\alpha_Y + \phi\alpha_M) + \phi\theta X_i + (\phi\nu_i + \eta_i) \end{aligned}$$

Total causal effect of X on Y: $\boxed{\theta \cdot \phi}$
This is the product of the effect of X on M (which is θ) and the effect of M on Y (which is ϕ).

5.2 *(1 pt)* Now suppose a researcher estimates the regression:

$$Y_i = \tilde{\alpha} + \tilde{\beta} X_i + \tilde{\phi} M_i + u_i$$

What does $\tilde{\beta}$ estimate in this regression? Should a researcher who wants to estimate the total effect of the nutrition program control for M? Why or why not?

Solution:

$\tilde{\beta}$ estimates the **direct effect** of X on Y, holding M constant.

In this setup, there is no direct effect—the entire effect of X on Y operates through M. Therefore $\tilde{\beta} \approx 0$.

No, a researcher who wants the total effect should **not** control for M. Controlling for the mediator "blocks" the causal path and removes exactly the effect we want to measure.

This is a "bad control" because M is caused by X (it's post-treatment). The policy question "should we implement this program?" requires the total effect, not the direct effect.

5.3 *(1 pt) (R)* Use the dataset `health_intervention.csv`, which contains `treatment` (0/1), `health_status` (0-100 scale, measured after treatment), and `test_score`. Estimate (i) the regression of test score on treatment only, and (ii) the regression of test score on treatment controlling for health status. Report both coefficients on treatment and explain why they differ. Which coefficient answers the policy question "should we implement this program?"

Solution:

```
total_effect <- lm(test_score ~ treatment, data = health_data)
bad_control <- lm(test_score ~ treatment + health_status, data = health_data)
```

Results:

- Without health status control: treatment coefficient ≈ 4.75 (total effect)
- With health status control: treatment coefficient ≈ 0.17 (direct effect ≈ 0)

Why they differ: The program works *through* health—nutrition improves health, which improves cognition. When we control for health status, we block this pathway and find no remaining effect.

Policy question: The first coefficient (without control) answers "should we implement this program?" Policymakers care about the total effect of the intervention, not whether there's a direct effect bypassing health.

C.6 Bad Control: Colliders

6.1 *(0.5 pts)* If you estimate the regression $Y_i = a + bX_i + e_i$ in the full population, what coefficient do you expect for $\hat{b}$? Why?

Solution:

$\hat{b} \approx 0$

Since talent (X) and attractiveness (Y) are independent in the population, $\text{Cov}(X, Y) = 0$. The OLS coefficient estimates this population covariance (scaled by variance), so it should be approximately zero.

6.2 *(0.5 pts)* Now suppose you only observe famous actors (i.e., you condition on C being above some threshold). Among this selected sample of famous actors, would you expect talent and attractiveness to be positively correlated, negatively correlated, or uncorrelated? Provide an intuitive explanation for your answer.

Solution:

Negatively correlated.

Intuition: To become famous, you need talent OR attractiveness (or both). Among famous people:

- If someone lacks talent, they must have compensated with attractiveness to become famous
- If someone lacks attractiveness, they must have compensated with talent

This creates a negative correlation in the selected sample, even though no correlation exists in the population. This is sometimes called "Berkson's paradox" or "selection bias."

6.3 *(1.5 pts)* Show algebraically that conditioning on C (either by restricting the sample or by controlling for C in a regression) induces a negative correlation between X and Y, even though they are unconditionally independent. You may use the fact that for jointly normal variables, $\text{Cov}(X, Y|C) = \text{Cov}(X, Y) - \text{Cov}(X, C)\text{Cov}(Y, C)/\text{Var}(C)$.

Solution:

Given $C = \alpha + \lambda_1 X + \lambda_2 Y + \xi$ with X, Y, ξ independent:

- $\text{Cov}(X, Y) = 0$ (given)
- $\text{Cov}(X, C) = \lambda_1 \text{Var}(X) > 0$
- $\text{Cov}(Y, C) = \lambda_2 \text{Var}(Y) > 0$

Using the formula:

$$\begin{aligned}
\text{Cov}(X, Y|C) &= \text{Cov}(X, Y) - \frac{\text{Cov}(X, C)\text{Cov}(Y, C)}{\text{Var}(C)} \\
&= 0 - \frac{\lambda_1 \text{Var}(X) \cdot \lambda_2 \text{Var}(Y)}{\text{Var}(C)} \\
&= -\frac{\lambda_1 \lambda_2 \text{Var}(X)\text{Var}(Y)}{\text{Var}(C)} < 0
\end{aligned}$$

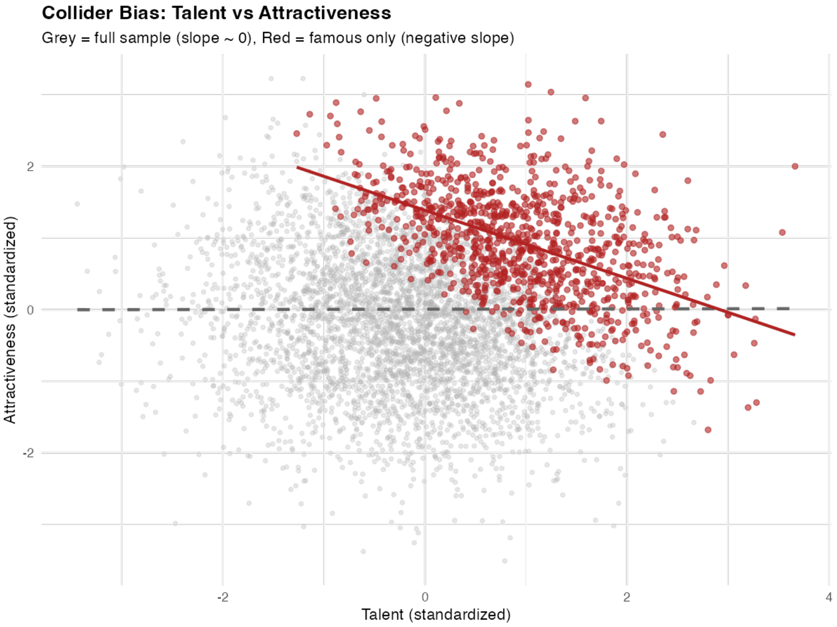
Since $\lambda_1, \lambda_2 > 0$, conditioning on the collider C induces a **spurious negative correlation**.

6.4 *(1 pt) (R)* Use the dataset `collider_simulation.csv`, which contains `talent` (standardized), `attractiveness` (standardized), `fame` (continuous score), and `is_famous` (indicator for top 20% of fame). Create a scatter plot of attractiveness (y-axis) against talent (x-axis). Plot the full sample in grey, then overlay the "famous" subsample (`is_famous == 1`) in a distinct color. Add regression lines for each group. Describe what you observe.

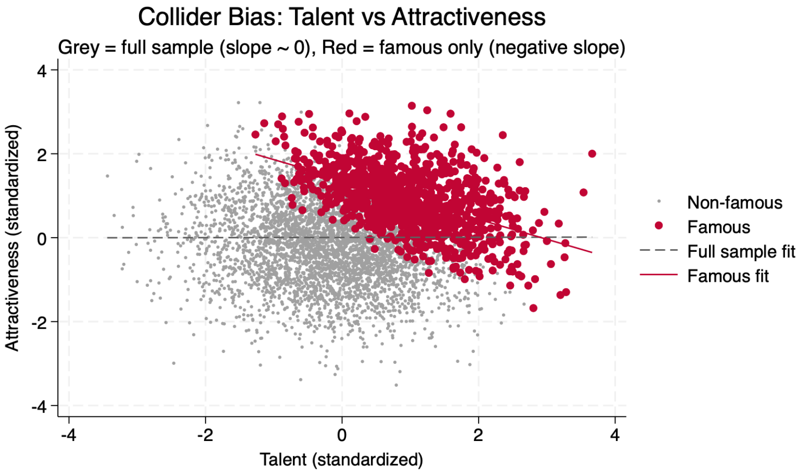
Solution:

```
library(ggplot2)
ggplot(collider_data, aes(x = talent, y = attractiveness)) +
  geom_point(data = subset(collider_data, is_famous == 0),
             color = "grey70", alpha = 0.3) +
  geom_point(data = subset(collider_data, is_famous == 1),
             color = "firebrick", alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "grey40", linetype = "dashed") +
  geom_smooth(data = subset(collider_data, is_famous == 1),
              method = "lm", se = FALSE, color = "firebrick")
```

R Output:



Stata Output:



Observation:

- Grey points (full sample): No pattern, regression line is essentially flat (slope ≈ 0)
- Red points (famous subsample): Clear negative pattern, regression line has steep negative slope

The collider bias is visually apparent: talent and attractiveness are uncorrelated overall, but strongly negatively correlated among the famous.

6.5 *(1.5 pts) (R)* Regress attractiveness on talent in three ways: (a) using the full sample, (b) restricting to the famous subsample only, (c) using the full sample but controlling for fame. Report the coefficient on talent from each regression. Explain why (b) and (c) give similar results, and why both differ from (a).

Solution:

```
# (a) Full sample
full_reg <- lm(attractiveness ~ talent, data = collider_data)

# (b) Famous only
famous_reg <- lm(attractiveness ~ talent,
                 data = subset(collider_data, is_famous == 1))

# (c) Controlling for fame
control_reg <- lm(attractiveness ~ talent + fame, data = collider_data)
```

Coefficients on talent:

- (a) Full sample: ≈ 0.002 (essentially zero)
- (b) Famous only: ≈ -0.475 (strong negative)
- (c) Controlling for fame: ≈ -0.809 (also strongly negative)

Explanation:

- (a) is correct: talent and attractiveness are independent in the population
- (b) and (c) both condition on the collider (fame), inducing spurious negative correlation
- (b) conditions by sample restriction; (c) conditions by regression control—both create similar bias
- (c) is even more negative because it conditions on the continuous fame variable rather than a binary threshold

Lesson: Controlling for a collider in a regression creates similar bias to restricting the sample based on the collider.

C.7 Deciding What to Control For

7.1 *(0.5 pts)* You are estimating the effect of a job training program (X) on wages (Y). You have data on workers' pre-program education levels (Z). Workers with more education are both more likely to enroll in training and have higher wages, regardless of training.

Solution:

Yes, control for education.

Education is a **confounder**: it affects both the treatment (training enrollment) and the outcome (wages), and it is determined *before* treatment. Controlling for it blocks the backdoor path and reduces omitted variable bias.

7.2 *(0.5 pts)* You are estimating the effect of a new teaching method (X) on student test scores (Y). You have data on hours spent studying (Z), which is measured after the teaching method is implemented. Students exposed to the new method tend to study more hours.

Solution:

No, don't control for study hours.

Study hours is a **mediator**: teaching method → study hours → test scores.

Controlling for study hours would block part of the causal effect we want to measure. If the new method works partly by motivating students to study more, we want to capture that effect.

7.3 *(0.5 pts)* You are estimating the effect of hospital quality (X) on patient mortality (Y). You have data on whether the patient was admitted to the ICU (Z). Both sicker patients (who have higher mortality risk) and patients at low-quality hospitals are more likely to be admitted to the ICU.

Solution:

No, don't control for ICU admission.

ICU admission is a **collider**: it is caused by both patient severity (which affects mortality) and hospital quality (what we're trying to estimate).

Controlling for ICU admission would induce a spurious correlation between hospital quality and unobserved severity, biasing our estimate. Among ICU patients, low-quality hospitals would appear to have healthier patients (because their patients didn't need to be that sick to get admitted), making low-quality hospitals look artificially better.