In-Class Exercise 10

For this in-class exercise, work with your group of 2-3 people, to answer the following questions. These questions are not necessarily easy and sometimes they will not have a clear "correct" answer. The goal is to get you thinking about the material we've learned. Some of these questions may require you to discuss and debate with your group members to come up with an answer or can cover topics that we have not yet covered in class.

Be prepared to share your answers with the class and add to the discussion.

After class submit your a do-file with your answers in comments to Moodle for grading. You will be graded as a group on your submission. Only one group member needs to submit the assignment, but make sure add all group member names.

Your group constitutes a team of economists hired by the Department of Health. You have been given a dataset of 500 hospitals to determine if increased funding (expenditure) actually saves lives. You have the following variables:

- mortality_rate: Deaths per 1,000 admissions (Outcome Y).
- expenditure: Spending per patient in thousands (D).- doctors: Number of MDs on staff.
- nurses: Number of RNs on staff.
- beds: Total number of beds.
- tech_index: A score (1-100) of the hospital's technological equipment.

Problem 1

You decide to control for hospital size by including doctors, nurses, and beds in your regression. However, you find that the standard errors for these variables are massive, and their t-statistics are insignificant, even though you know staff size matters.

A colleague points out that large hospitals always have more of all these things.

What is the specific econometric term for this problem?

If doctors and nurses have a correlation of 0.85, why does this make it difficult for the model to isolate the effect of doctors specifically?

Propose a data transformation (e.g., creating a new variable or dropping one) to solve this without losing the information about hospital size.

! Solution

1. **Term: Multicollinearity** (or High Multicollinearity).
2. **Why difficult?** When two variables move in lock-step (correlation 0.85), the regression cannot tell *which* one is causing the change in Y. It can't "hold nurses constant" to change doctors, because in the data, whenever doctors increase, nurses increase too. This inflates the variance of the coefficient estimates ($Var(\hat{\beta})$).
3. **Transformation:**
   - Create a composite index (e.g., `total_staff = doctors + nurses`).
   - Remove the variable.

Problem 2

You plot the residuals of your model against hospital size. You notice that for massive hospitals, the residuals are tightly clustered around zero (predictable outcomes), but for tiny rural hospitals, the residuals vary wildly (huge errors).

1. Which specific assumption (A1-A6) is being violated here?
2. Your boss asks, "Does this bias our coefficient estimates?" How do you answer? If the coefficients aren't biased, what is the specific problem with your hypothesis tests?
3. You decide to use "Weighted Least Squares." Which hospitals should receive less weight in the regression: the small ones with high variance or the large ones with low variance? Why?

! Solution

1. **Assumption: Homoskedasticity** ($Var(u \mid x) = \sigma^2$). The error variance is *not* constant; it depends on hospital size (x). This is **Heteroskedasticity**.
2. **Bias vs. Inference:**
   - **Bias:** No. OLS coefficients remain unbiased.
   - **Problem:** The Standard Errors are **wrong/biased**. This means T-statistics and Confidence Intervals are invalid, leading to incorrect hypothesis tests.
3. **WLS:** The **small hospitals (high variance)** should receive **less weight**.
   - Logic: We trust the data points from large hospitals more because they are more precise (smaller variance). We down-weight the "noisy" observations from small hospitals.

Problem 3

You run a simple regression of `mortality_rate` on `tech_index`.

$$Mortality = \beta_0 + \beta_1 TechIndex + \varepsilon$$

Shockingly, you find that $\beta_1 > 0$. The model suggests that better technology kills more people.

What might be driving this result?

1. Let $Cov(Tech, Severity) > 0$ What is the interpretation of this?
2. Let $\beta_{severity} > 0$ What is the interpretation of this?
3. Using the OVB formula, show mathematically why your estimate $\tilde{b}_1$ is positive, even if the true effect of technology (β_1) is negative (saving lives).

! Solution

1. **Cov(Tech, Severity) > 0:** High-tech hospitals attract the sickest (most severe) patients because they are referral centers.
2. $\boldsymbol{\beta_{severity} > 0}$**:** Sicker patients are more likely to die (higher mortality rate).
3. **OVB Formula:**

$$E[\tilde{\beta}_{tech}] = \beta_{tech} + \beta_{severity} \cdot \frac{Cov(tech, severity)}{Var(tech)}$$

   - We expect True $\beta_{tech} < 0$ (Tech saves lives).
   - Bias term: Positive $\times$ Positive = **Positive Bias**.
   - If the Bias is large enough, it overwhelms the negative true effect, flipping the sign to positive ($\tilde{\beta}_{tech} > 0$).

Problem 4

You cannot measure "Patient Severity" directly (it is unobserved). However, you have data on "ICU Admissions Rate" (Z). You plan on using it as a proxy for Severity in your regression.

To use ICU Admissions as a valid proxy for Severity, two conditions must be met. What are they? Go through each of them and discuss whether they are likely to hold in this context. How "strong" are these assumptions likely to be?

! Solution

1. **Correlation Condition:** The proxy (Z) must be strongly correlated with the unobserved variable (X^*).
   - *Check:* Very likely. Sicker patients are admitted to the ICU. Strong assumption holds.
2. **Redundancy (Ignorability) Condition:** Once we control for Severity (X^*), the proxy (Z) should not affect Mortality (Y) directly ($Cov(u, Z) = 0$ in the structural equation). Alternatively, the measurement error ($Z - X^*$) should be uncorrelated with Treatment (D).
   - *Check:* This is harder. Does going to the ICU affect mortality *other than* because you are sick? Maybe (better care in ICU saves you, or hospital-acquired infections in ICU kill you). If Z affects Y directly, it's an imperfect proxy, but likely better than nothing.

Problem 5

To truly fix the bias, you need an Instrumental Variable (Z) for `tech_index` (D). You need something that affects a hospital's technology budget but has nothing to do with the patients' health.

Your team finds data on "Senator Tenure"—the number of years the local district's senator has been in office. The theory: Senior senators bring home more "pork barrel" grants for hospital equipment upgrades.

Explain why Senator Tenure must be correlated with `tech_index`. Explain why Senator Tenure must not be correlated with the error term (e.g., patient health) in the mortality regression.

Does "Senator Tenure" pass the weirdness test? Why?

Can you think of a reason the Exclusion Restriction might fail?

! Solution

1. **Relevance (First Stage):** Senior senators have more political clout to secure federal grants. If they direct money to hospitals, `tech_index` increases. (Check with F-stat > 10).
2. **Exclusion (Independence):** A senator's time in office shouldn't directly make people healthier or sicker, except through the funding they bring.
3. **Weirdness Test:** Yes, it passes. It seems "weird" (uncorrelated) to think that the number of years a politician has sat in Washington would biologically cause a patient in their district to have a heart attack.
4. **Failure of Exclusion:** If senior senators bring money for *everything* (better roads, better schools, better nutrition programs, more doctors), then Tenure affects Mortality through channels *other than* just the Tech Index. Additional funding for nurse salaries (not tech) would violate the restriction.

Problem 6

When you run the IV regression using "Senator Tenure" as an instrument for `tech_index`, you find that the estimated effect of technology on mortality is similar to the proxy estimate from Problem 4. Does this add confidence that your results are correct? Why or why not?

! Solution

Yes, it adds confidence. - We used two very different methods (Proxy Variable vs. Instrumental Variable) relying on different assumptions to solve the same OVB problem. - If both methods produce a similar estimate (e.g., Tech reduces mortality), it suggests the result is robust and not an artifact of one specific flawed assumption. Triangulation strengthens causal claims.

Problem 7

If you had relied on the "Naive" OLS regression from Part 3, what dangerous policy recommendation might you have made to the Ministry of Health? How does the IV or Proxy approach change that recommendation?

! Solution

- **Naive Policy:** "Technology correlates with death. We should **defund** equipment purchases to save lives." (Disastrous).
- **Correct Policy (IV/Proxy):** "Once we account for the fact that sick people go to high-tech hospitals, we see technology actually saves lives. We should **invest** in more equipment."