REVIEW QUESTIONS – HANDOUT 1

Introduction to Panel Data

Master in Data Science for Decision Making
Barcelona School of Economics

Instructions: These questions review the main concepts from our first class. Try to answer them without looking at your notes first. Questions marked with (*) require more detailed written responses.

PART A: CONCEPTUAL QUESTIONS

Question 1: Correlation vs. Causation

Consider the following statement: "Countries with higher coffee consumption have higher GDP per capita. Therefore, increasing coffee consumption will increase GDP."

a) Does this statement confuse correlation with causation? Explain.

b) Provide at least two alternative explanations for the observed correlation that don't involve a causal relationship from coffee to GDP.

c) What would you need to establish causality in this relationship?

Question 2: The Conditional Expectation

a) Why do econometricians typically focus on estimating $E(y|X)$ rather than other features of the distribution of y given X?

b) Under what condition is $E(y|X)$ guaranteed to be linear in X?

c) If the true relationship is $E(y|X) = X^2$ but you estimate $E(y|X) = \beta X$, what are you actually estimating? Does β have a meaningful interpretation?

Question 3: Assumptions and Interpretation (*)

Consider the model: $y = X\beta + \varepsilon$

a) State the two key assumptions (Assumptions 1 and 2 from class) that allow OLS to provide consistent, causally interpretable estimates.

b) For each assumption, explain what happens if it's violated:

- What happens to the properties of $\hat{\beta}$?
- What does $\hat{\beta}$ measure in each case?

c) Which violation is more problematic from a causal inference perspective, and why?

Question 4: Omitted Variable Bias – Formula

Recall the omitted variable bias formula:

$$\hat{\beta} - \beta = \gamma \times \frac{\text{Cov}(\eta, X)}{\text{Var}(X)}$$

where η is the omitted variable and γ is its true coefficient.

a) Explain in words what each component of this formula represents.

b) Under what condition does omitting a variable NOT cause bias?

c) If $\gamma > 0$ and $\text{Cov}(\eta, X) < 0$, will $\hat{\beta}$ overestimate or underestimate β?

PART B: APPLIED QUESTIONS

Question 5: Predicting the Direction of Bias (*)

You want to estimate the effect of class size on student test scores. You run the regression:

$$\text{TestScore}_i = \beta_0 + \beta_1 \text{ClassSize}_i + u_i$$

using cross-sectional data on schools.

a) What omitted variables might be in u_i? Name at least three.

b) For each omitted variable you identified:

- What is the likely sign of its correlation with ClassSize?
- What is the likely sign of its effect on TestScore (i.e., the sign of γ)?
- What is the implied direction of bias on $\hat{\beta}_1$?

c) Given your analysis, do you expect $\hat{\beta}_1$ to overstate or understate the true causal effect of class size on test scores?

Question 6: When Panel Data Helps

For each of the following research questions, explain:

(i) Whether panel data would be helpful, and why

(ii) What specific unobserved heterogeneity panel data would help control for

(iii) Whether cross-sectional data would be sufficient, and under what assumptions

a) Does training programs increase worker productivity?

b) Do higher temperatures reduce agricultural output?

c) Does smoking cause lung cancer?

d) Do minimum wage increases reduce employment?

Question 7: Panel Data Structure

Consider a dataset tracking 500 firms over 8 years (2016–2023).

a) What is N and what is T?

b) Is this a balanced or unbalanced panel? How would you know?

c) Is this a "short panel" or "long panel"? What are the implications for estimation?

d) If 50 firms exit the dataset in 2020 and never return, is the panel still balanced? How many total observations would you have?

e) Give an example of a variable that varies only in the i dimension, only in the t dimension, and in both dimensions.

Question 8: Fixed Effects Model – Conceptual (*)

You estimate the following model on a panel of countries (i) observed over years (t):

$$\text{GDPgrowth}_{it} = c_i + \beta_1 \text{Investment}_{it} + \beta_2 \text{Education}_{it} + \varepsilon_{it}$$

a) What does c_i represent? Give three concrete examples of what might be captured by c_i.

b) Why can't you include a time-invariant variable like "Landlocked" (dummy for whether the country is landlocked) in this regression?

c) Explain in your own words how the "within transformation" removes c_i from the equation.

d) After the within transformation, what variation in the data identifies β_1?

Question 9: Fixed vs. Random Effects

Consider a study on the effect of advertising expenditure on firm sales using panel data.
Model: $\text{Sales}_{it} = c_i + \beta_1 \text{Advertising}_{it} + \varepsilon_{it}$

a) Under what assumption on c_i would a Random Effects estimator be consistent?

b) Why might this assumption be problematic in this context? (Hint: think about what c_i might contain)

c) If you're unsure whether the Random Effects assumption holds, which estimator should you use and why?

d) What is the key advantage of Fixed Effects over Random Effects in terms of robustness?

Question 10: Empirical Application (*)

You have panel data on 1,000 students observed in grades 9, 10, and 11. You want to study whether having access to a computer at home affects math test scores.

$$\text{TestScore}_{it} = c_i + \beta_1 \text{Computer}_{it} + \beta_2 \text{ParentEducation}_i + \varepsilon_{it}$$

where:

- TestScore_{it} = student i's math score in grade t
- $\text{Computer}_{it} = 1$ if student i has computer access in grade t
- ParentEducation_i = years of parental education (time-invariant)
- c_i = student fixed effect

a) Can you include ParentEducation_i in a fixed effects regression? Why or why not?

b) What unobserved student characteristics might c_i capture that could bias a cross-sectional OLS estimate?

c) State the strict exogeneity assumption (FE1) for this model. What does it require about the relationship between ε_{it} and Computer_{it}?

d) Write out what the within-transformed regression would look like. What identifies β_1 in the within estimator?

e) Suppose some students who don't have computers in grade 9 get them in grade 10, while others never get them. Which students contribute to the identification of β_1?

BONUS QUESTION (Optional, for extra practice)

Question 11: Dynamics

Suppose you want to study whether people's consumption today depends on their consumption yesterday (habit formation):

$$\text{Consumption}_{it} = c_i + \beta_1 \text{Income}_{it} + \gamma \text{Consumption}_{it-1} + \varepsilon_{it}$$

a) Why can't you estimate this model with a standard fixed effects estimator?

b) What is the endogeneity problem created by including $\text{Consumption}_{it-1}$?

c) This course focuses on static panels. Can you speculate why dynamic panels require different estimation methods?

ANSWER KEY NOTES FOR STUDENTS

- Questions 1–4 test conceptual understanding from the "9 Questions" review
- Questions 5–7 test understanding of omitted variable bias and panel data basics
- Questions 8–10 test understanding of fixed effects models
- Questions marked (*) require written explanations – practice articulating concepts clearly!
- If you struggle with any question, review the corresponding section in the handout

Good luck!

Solutions will be posted after the next class. For now, try to work through these on your own or in study groups. Be prepared to discuss your answers in the next session.