Mock Exam Questions (Midterm)

Instructions

This midterm contains 20 MC questions and 5 open questions, and will count for 40% of the final grade. All MC questions have equal weight (1pt). All open questions will have a weight of 3 pts each. The duration of this exam is 2 hours.

The exam covers the five core lectures: introduction to econometrics and data, the linear model I and II, and panel data I and II.

Multiple Choice Questions

Question 1

A retailer regresses weekly sales on the number of staff scheduled. Management wants to use the same regression for two purposes: first, to decide how many staff to schedule next week in order to raise sales; second, to forecast next week's sales given a roster that has already been fixed. Which statement is correct?

a) Both purposes require the coefficient on staffing to be unbiased.
b) Only the forecast requires an unbiased coefficient; the scheduling decision needs only a high R^2.
c) Only the scheduling decision requires an unbiased coefficient; the forecast needs the fitted values to be accurate, which they can be even if the coefficient is confounded.
d) Neither purpose requires any assumption about the error term, because OLS is always unbiased.

Question 2

A statistical agency surveys 3,000 households every year for ten years. Each year the households are drawn afresh from the population register, so almost no household appears in two different years. What kind of data is this, and what does its structure rule out?

a) Time series data, because there are ten time periods.
b) Panel data, but unbalanced, because the number of households differs slightly from year to year.
c) Panel data, which would allow the researcher to remove household-specific unobserved heterogeneity.
d) Pooled cross-sections, so household-specific unobserved heterogeneity cannot be removed by comparing a household with itself over time.

Question 3

After estimating $Wage_i = \hat{\beta}_0 + \hat{\beta}_1 Educ_i + \hat{u}_i$ by OLS, a student computes the sample correlation between $Educ_i$ and the residuals $\hat{u}_i$, finds that it is exactly zero, and concludes that education is exogenous. What is wrong with this reasoning?

a) Nothing is wrong; a zero sample correlation between a regressor and the residuals is the sample counterpart of the zero conditional mean assumption.
b) The correlation should have been computed between the residuals and the fitted values instead.
c) The OLS first-order conditions force this correlation to be zero in every sample, so the result carries no information about whether education is correlated with the unobserved error.
d) The reasoning is valid only asymptotically, so the student needs a larger sample before drawing the conclusion.

Question 4

Which statement correctly distinguishes the error term u_i from the residual $\hat{u}_i$?

a) They are the same quantity, computed on the population and on the sample respectively, so for a given observation they take the same value.
b) The error is the deviation of y_i from the population regression function and cannot be observed; the residual is the deviation from the estimated line and is computed from the data.
c) The residual cannot be observed because it depends on the unknown parameters, whereas the error is computed directly from the data.

d) The error is always larger in absolute value than the residual, because OLS minimises the sum of squared residuals.

Question 5

A marketing analyst estimates the following model, where `Sales` is measured in units sold and `AdSpend` in euros:

$$\widehat{\text{Sales}} = 240 + 38 \times \log(\text{AdSpend})$$

How should the coefficient 38 be interpreted?

a) A one-euro increase in advertising spending is associated with 38 more units sold.
b) A one-euro increase in advertising spending is associated with a 38% increase in sales.
c) A 1% increase in advertising spending is associated with about 0.38 more units sold.
d) A 1% increase in advertising spending is associated with 38 more units sold.

Question 6

Researcher A regresses wages on forty detailed job-history variables and reports $R^2 = 0.68$. Researcher B regresses wages on a randomly assigned training voucher and reports $R^2 = 0.06$. Which conclusion is justified?

a) A's model is better, because it accounts for more of the variation in wages.
b) B's model is misspecified, because an R^2 below 0.10 indicates important omitted variables.
c) The two cannot be compared without also knowing the F-statistics.
d) B's coefficient is the more credible estimate of a causal effect; R^2 measures explained variation and says nothing about whether a coefficient identifies a causal effect.

Question 7

A researcher has a fixed budget and must choose between surveying a larger number of firms drawn at random, and deliberately over-sampling firms with unusually high and unusually low marketing budgets while keeping the sample size unchanged. Which choice reduces the standard error of the estimated slope on marketing spend, and why?

a) Over-sampling the extremes, because the variance of the slope estimator falls as the sample variation in the regressor, SST_x, rises.
b) The larger random sample, because the variance of the slope estimator depends on n alone.

c) Neither, because the variance of the slope estimator depends only on the error variance σ^2.
d) Over-sampling the extremes, because concentrating on extreme firms reduces the error variance σ^2.

Question 8

A researcher estimates the effect of class size on test scores and omits parental involvement. Schools with more involved parents tend to have smaller classes, and parental involvement also raises test scores directly. The true effect of class size on scores is negative. How does the omission mislead the researcher?

a) The estimated coefficient is biased downward, so the harm done by large classes appears greater than it truly is.
b) The estimated coefficient is biased toward zero, so the harm done by large classes appears smaller than it truly is.
c) There is no bias, because both channels operate through parental involvement.
d) The estimated coefficient is biased upward, so large classes appear beneficial.

Question 9

A researcher estimates a simple regression of wages on years of education and obtains a slope of 4.5. Later work establishes that the true direct effect of ability on wages is 4.0, and that a regression of ability on education has a slope of 0.5. What does this imply about the true return to education?

a) 4.5, because omitted variable bias inflates the standard error rather than the coefficient.
b) 2.5
c) 6.5
d) 2.0

Question 10

A researcher studies whether holding an MBA raises earnings, using data on people hired by one consulting firm. The firm hires candidates who either hold an MBA or have exceptional prior experience. Within the hired sample, MBA holders earn less than non-holders. What is the most likely explanation?

a) The MBA genuinely lowers earnings for this population.
b) Prior experience is a confounder, and the regression should control for it.

c) Being hired is a collider: conditioning on it induces a negative association between the MBA and the other determinants of earnings, even if the MBA has no effect at all.
d) Prior experience is a mediator, so the regression recovers a direct effect rather than the total effect.

Question 11

A researcher models hourly wages by industry, using Manufacturing as the omitted reference category:

$$\widehat{\text{Wage}} = 21.40 - 3.20 \times \text{Retail} + 6.10 \times \text{Finance}$$

What is the estimated average wage difference between a worker in Finance and a worker in Retail?

a) €6.10
b) €2.90
c) €3.20
d) €9.30

Question 12

A researcher estimates $Wage_i = \beta_0 + \beta_1 Female_i + \beta_2 Educ_i + \beta_3 (Female_i \cdot Educ_i) + u_i$ and obtains $\hat{\beta}_1 = -4.00$ and $\hat{\beta}_3 = 0.15$. What is the estimated wage gap between a woman and a man who both have 16 years of education, in euros per hour?

a) -4.00
b) -1.60
c) -2.40
d) 0.15

Question 13

Which statement about heteroskedasticity-robust standard errors is correct?

a) They remove the bias that heteroskedasticity induces in the OLS coefficients.
b) They are valid only when the errors are also normally distributed.
c) They are always smaller than the default standard errors, which is why they are preferred.
d) They leave the coefficients untouched and change only the estimated variance, because heteroskedasticity does not bias the OLS coefficients in the first place.

Question 14

The within estimator used in fixed-effects models identifies a coefficient using which variation in the data?

a) Differences between units, after averaging each unit over time.
b) The difference between each pair of consecutive periods for every unit.
c) Deviations of each observation from that unit's own time-average, so only variation within units over time contributes.
d) The cross-section of units in the final period only.

Question 15

A researcher estimates a fixed-effects wage regression on a panel of individuals observed over eight years, and finds that the coefficient on years of schooling cannot be estimated. What is the reason?

a) Years of schooling is measured with error, and fixed effects amplify measurement error.
b) Schooling does not change within a person over the sample period, so the within transformation removes it along with the individual effect.
c) Fixed-effects estimation requires all regressors to be binary.
d) The sample contains too few individuals to identify an additional parameter.

Question 16

Fixed-effects estimation requires strict exogeneity. Which of the following situations violates that assumption?

a) The regressor is correlated with the time-invariant individual effect a_i.
b) The idiosyncratic errors are heteroskedastic across units.
c) Some units are observed in fewer periods than others.
d) A shock to the outcome in one period changes the value of the regressor in the following period.

Question 17

In a pooled OLS regression on panel data, the composite error is $v_{it} = a_i + u_{it}$, where $\text{Var}(a_i) = \sigma_a^2$, $\text{Var}(u_{it}) = \sigma_u^2$, and the two components are independent of each other and over time. What is $\text{Corr}(v_{it}, v_{is})$ for two periods $t \neq s$ of the same unit i?

a) $\dfrac{\sigma_a^2}{\sigma_a^2 + \sigma_u^2}$

b) Zero, because the idiosyncratic errors are independent over time.
c) σ_a^2, because that is the covariance between the two composite errors.
d) $\dfrac{\sigma_u^2}{\sigma_a^2 + \sigma_u^2}$

Question 18

The Hausman test compares the fixed-effects and random-effects estimators. What is its null hypothesis, and what follows if you fail to reject it?

a) H_0: the variance of the individual effect is zero. Failing to reject means pooled OLS is appropriate.
b) H_0: the fixed-effects and random-effects coefficients differ. Failing to reject means fixed effects should be used for its consistency.
c) H_0: the individual effect a_i is uncorrelated with the regressors. Failing to reject supports the random-effects estimator, which is the more efficient of the two.
d) H_0: the idiosyncratic errors are serially uncorrelated. Failing to reject means the first-difference estimator is preferred.

Question 19

The random-effects estimator applies the quasi-demeaning transformation $y_{it} - \theta\bar{y}_i$ to the data. What does the estimator converge to as θ approaches 1?

a) The fixed-effects (within) estimator.
b) The pooled OLS estimator.
c) The between estimator.
d) Nothing, because θ affects only the standard errors and not the point estimates.

Question 20

What does the Breusch-Pagan LM test for panel data examine, and what does a rejection imply?

a) H_0: the errors are homoskedastic. Rejection means robust standard errors should be reported.
b) H_0: the fixed-effects and random-effects estimates agree. Rejection means fixed effects should be used.
c) H_0: there is no individual-specific variance component ($\sigma_a^2 = 0$). Rejection means pooled OLS is inadequate and a panel estimator such as random effects is warranted.
d) H_0: the regressors are exogenous. Rejection means an instrumental variable is required.

Open Questions

Answer each question in full sentences. Where a question asks you to derive something, show the intermediate steps rather than only the final expression.

Question 1

This question uses a dataset containing wages and other characteristics for 526 working individuals. The key variables are `wage`, average hourly earnings in dollars, `educ`, years of education, and `exper`, years of potential work experience.

An economist runs a simple linear regression of hourly wage on years of education. The output is provided below.

Table 1: Regression of Wage on Years of Education

	(1)
(Intercept)	−0.905
	(0.685)
educ	0.541***
	(0.053)
Num.Obs.	526
R2	0.165

* p <0.1, ** p <0.05, *** p <0.01

Write down the estimated sample regression function, interpret the slope coefficient in the units of the data, and interpret the intercept. State whether the intercept has a meaningful real-world interpretation here, and justify your answer by reference to the sample the regression was estimated on.

Question 2

The regression in Question 1 omits a worker's experience. Suppose the population model that determines wages is

$$Wage = \beta_0 + \beta_1 Educ + \beta_2 Exper + u.$$

Reason through the consequences of leaving `exper` out. State the likely sign of β_2 and justify it economically; state the likely sign of the relationship between education and experience among working individuals of a given age, and justify that too; then combine the two using the omitted

variable bias formula to sign the bias in the estimated return to education. Conclude by saying whether the simple regression over- or understates the true return, and explain in one sentence why the sign of the bias, and not merely its presence, is what matters for the interpretation.

Question 3

Consider the model

$$Wage_i = \beta_0 + \beta_1 Female_i + \beta_2 Educ_i + \beta_3 (Female_i \cdot Educ_i) + u_i,$$

where $Female_i$ equals 1 for women and 0 for men, and $Educ_i$ is years of education.

Derive the separate regression equations implied by this model for men and for women, and use them to obtain both the marginal effect of an additional year of education for each group and the wage gap between a woman and a man with the same level of education. Explain what makes the gap depend on education, and state what the coefficients β_1 and β_3 each contribute to it. Finish by explaining what a test of $\beta_3 = 0$ would tell you, and why that is a different question from a test of $\beta_1 = 0$.

This answer model is very extensive, you do not have to be as extensive in your answer.

Question 4

An economist has a panel of employees observed over several years and wants the effect of training hours on productivity. Unobserved characteristics such as innate ability or motivation are constant over time, are not in the data, and plausibly affect both how much training an employee receives and how productive they are.

Explain precisely why this creates a statistical problem for pooled OLS, naming the assumption that fails and writing down the composite error that causes it to fail. Then show, algebraically, how the within transformation removes the offending term, and state what the fixed-effects estimator can no longer identify as a consequence. Finally, explain what a Hausman test would add here: say what it compares, what its null hypothesis asserts about this particular problem, and what you would conclude from a rejection.

This answer model is very extensive, you do not have to be as extensive in your answer.

Question 5

A researcher has annual data on 400 firms observed over ten years. They estimate a pooled OLS regression of firm profitability on R&D spending, report the default standard errors produced by the software, and conclude that the effect is significant because the t-statistic is 3.1.

Explain why the reported standard error is likely to be wrong. Write down the composite error for this setting, state what it implies about the correlation between two residuals belonging to the same firm in different years, and explain the direction in which the default standard error is therefore misleading and what that does to the t-statistic of 3.1. State what the researcher should report instead, at which level, and explain briefly why the number of firms rather than the number of firm-year observations governs how much independent information the data contain.

End of Exam