Answers Mock Midterm

Answer Key

Multiple Choice Questions

1. **c)** The two goals place different demands on the same regression. Choosing a staffing level is a causal question: it requires that $\hat{\beta}_1$ measure what would happen to sales if staffing were changed, so the coefficient must be unbiased. Forecasting is a question about $\hat{y}$: if busy weeks are staffed more heavily, the fitted values can predict sales well even though the coefficient confounds staffing with demand.

2. **d)** Ten annual surveys of freshly drawn households are pooled cross-sections, not a panel. Because no household is observed twice, there is no within-household variation, and the fixed-effects logic of comparing a unit with itself over time is unavailable.

3. **c)** $\sum_i x_i \hat{u}_i = 0$ is one of the OLS first-order conditions. It therefore holds mechanically in every sample, whether or not the zero conditional mean assumption is true, and cannot serve as evidence for exogeneity. Exogeneity is a statement about the unobservable error u_i, not about the residual $\hat{u}_i$, and no residual-based calculation can test it.

4. **b)** The error $u_i = y_i - \beta_0 - \beta_1 x_i$ is the deviation from the population regression function and involves the unknown true parameters, so it is never observed. The residual $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ is the deviation from the fitted line and is computed once the estimates exist. For a given observation the two are generally different numbers.

5. **c)** This is a level-log model. Differentiating, $\Delta\text{Sales} = 38 \times \Delta\log(\text{AdSpend})$, and a 1% rise in advertising means $\Delta\log(\text{AdSpend}) \approx 0.01$. The predicted change in sales is therefore $38 \times 0.01 = 0.38$ units.

6. **d)** R^2 measures the share of the sample variation in the outcome that the regressors account for. It is silent on identification. B's voucher was randomly assigned, so its coefficient is credible as a causal effect despite the low R^2; A's forty controls raise the explained variation without establishing that any coefficient is causal.

7. **a)** The variance of the slope estimator is $\mathrm{Var}(\hat{\beta}_1) = \sigma^2/SST_x$. Spreading the sample over firms with very high and very low marketing budgets raises SST_x and so lowers the variance, even at unchanged n. Option d) is wrong because the sampling scheme does not change the error variance, and option b) is wrong because n enters only through SST_x.

8. **a)** Write the bias as $\beta_2\delta_1$, where β_2 is the effect of parental involvement on scores and δ_1 is the slope from regressing involvement on class size. Here $\beta_2 > 0$ and $\delta_1 < 0$, so the bias is negative. A negative bias added to a negative true effect makes the estimate more negative than the truth, so the estimate exaggerates the harm done by large classes.

9. **b)** The estimated slope from the short regression satisfies $E[\hat{\alpha}_1] = \beta_1 + \beta_2\delta_1$. Substituting the known quantities, $4.5 = \beta_1 + (4.0)(0.5) = \beta_1 + 2.0$, so $\beta_1 = 2.5$. The simple regression overstates the return to education by exactly the bias term of 2.0.

10. **c)** Being hired is caused both by holding the MBA and by having exceptional experience. Restricting the sample to those hired conditions on that common effect, which is the definition of a collider. Among the hired, someone without the MBA must have had unusually strong experience to get in, so the two determinants become negatively associated within the sample even if the MBA changes nothing.

11. **d)** Both coefficients are measured against the omitted Manufacturing category, so the comparison between two non-baseline groups is the difference of the two coefficients: $6.10 - (-3.20) = 9.30$. Equivalently, the predicted wages are $21.40 + 6.10 = 27.50$ in Finance and $21.40 - 3.20 = 18.20$ in Retail, and $27.50 - 18.20 = 9.30$.

12. **b)** The wage gap at a given level of education is $\beta_1 + \beta_3 Educ$, not β_1 alone. At 16 years, the estimate is $-4.00 + 0.15 \times 16 = -4.00 + 2.40 = -1.60$. Option a) mistakes the intercept shift for the gap at every education level.

13. **d)** Heteroskedasticity does not bias the OLS coefficients; it invalidates the usual variance formula. Robust standard errors therefore change only the estimated variance, leaving every point estimate exactly as it was. They are not always larger or always smaller than the default, and they require no normality.

14. **c)** The within transformation subtracts each unit's own time-average, $\ddot{x}_{it} = x_{it} - \bar{x}_i$, so identification comes entirely from movement within units over time. Option b) describes the first-difference estimator, which is a different transformation, and option a) describes the between estimator.

15. **b)** For an individual whose schooling is complete, years of schooling takes the same value in all eight years. The within transformation subtracts the person's own mean, which is that same value, so the regressor becomes identically zero and is dropped along with a_i. Fixed effects cannot identify the effect of anything that does not move within a unit.

16. **d)** Strict exogeneity requires the idiosyncratic error to be uncorrelated with the regressors in every period, past, present and future. Feedback from this period's shock into next

period's regressor violates exactly that. Option a) describes correlation with a_i, which is the problem fixed effects is designed to solve rather than an assumption it needs.

17. **a)** The covariance is $\text{Cov}(v_{it}, v_{is}) = \text{Var}(a_i) = \sigma_a^2$, because the shared component is the only term the two periods have in common, while the total variance of each composite error is $\sigma_a^2 + \sigma_u^2$. The correlation is therefore $\sigma_a^2/(\sigma_a^2 + \sigma_u^2)$, which is strictly positive whenever $\sigma_a^2 > 0$. This is why pooled OLS on panel data needs clustered standard errors.

18. **c)** Both estimators are consistent when a_i is uncorrelated with the regressors, but only random effects is efficient; when the correlation is present, random effects becomes inconsistent while fixed effects remains consistent. The null is therefore that the correlation is absent, and failing to reject it leaves the more efficient random-effects estimator available.

19. **a)** Quasi-demeaning nests the familiar estimators as special cases of θ. At $\theta = 0$ nothing is subtracted and the estimator is pooled OLS; at $\theta = 1$ the full unit mean is subtracted, which is exactly the within transformation, so random effects coincides with fixed effects. In practice θ lies between the two and rises with σ_a^2 relative to σ_u^2.

20. **c)** The Breusch-Pagan LM test asks whether there is any individual-specific variance component at all. Under $\sigma_a^2 = 0$ the composite error collapses to u_{it} and pooled OLS is adequate. Rejecting means the units genuinely differ in a persistent way, so a panel estimator is called for. Note that this is a different question from the one the Hausman test answers, which is whether that component is correlated with the regressors.

Open Questions

Question 1

The estimated sample regression function is

$$\widehat{\text{wage}} = -0.905 + 0.541 \cdot \text{educ}.$$

The slope says that comparing two individuals who differ by one year of education, the one with more schooling earns on average \$0.541 more per hour, or roughly 54 cents. Note that this is a comparison between individuals, not an effect holding other factors fixed: the regression contains no other regressors, so nothing is being held constant. Any difference in experience, ability or occupation that happens to accompany education is included in this number.

The intercept, -0.905, is the predicted hourly wage of an individual with zero years of education. It has no meaningful real-world interpretation here for two reasons. A negative wage is not economically possible, and more importantly the sample consists of working individuals, essentially none of whom has zero years of schooling. The intercept is an extrapolation far outside the range of education the data actually contain, so the fitted line is not credible there even though the arithmetic is valid.

Question 2

The likely sign of β_2 is positive. Holding education fixed, additional years in the labour market build skills and job-specific knowledge, and wages typically rise with experience.

The likely relationship between education and experience is negative. Among individuals of a given age, years spent in full-time education are years not spent working, so the two compete for the same calendar time. A regression of `exper` on `educ` would therefore have a negative slope, $\delta_1 < 0$.

Combining the two through the omitted variable bias formula,

$$E[\hat{\beta}_1^{\text{short}}] = \beta_1 + \beta_2\delta_1 = \beta_1 + (+) \times (-) < \beta_1.$$

The bias is negative, so the simple regression understates the true return to education.

The sign matters, and not merely the fact that bias exists, because a signed bias converts a flawed estimate into a usable bound. Knowing the estimate is biased downward tells us the true return is at least as large as 0.541, which supports a conclusion about the direction of the effect; knowing only that "there is bias" would leave the estimate uninformative in either direction.

Question 3

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

Setting $Female_i = 0$ leaves

$$Wage_i = \beta_0 + \beta_2 Educ_i + u_i,$$

so men have intercept β_0 and slope β_2. Setting $Female_i = 1$ and collecting terms,

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

so women have intercept $\beta_0 + \beta_1$ and slope $\beta_2 + \beta_3$.

Differentiating the conditional expectation with respect to education,

$$\frac{\partial E[Wage]}{\partial Educ} = \beta_2 + \beta_3 Female,$$

which equals β_2 for men and $\beta_2 + \beta_3$ for women. The return to a year of schooling is thus allowed to differ by gender, and β_3 is the size of that difference.

Subtracting the male equation from the female equation at a common level of education gives the gap:

$$E[Wage \mid Female = 1, Educ] - E[Wage \mid Female = 0, Educ] = \beta_1 + \beta_3 Educ.$$

The gap depends on education precisely because the interaction permits different slopes. Without β_3, the two lines would be parallel and the gap would be the constant β_1 at every level of schooling. The two coefficients play distinct roles: β_1 is the gap evaluated at zero years of education, which is the vertical distance between the two intercepts, while β_3 is the rate at which the gap changes per additional year of education. If $\beta_3 > 0$ the gap narrows as education rises; if $\beta_3 < 0$ it widens.

Testing $\beta_3 = 0$ asks whether the return to education is the same for women and men, that is, whether the gap is constant across education levels. Testing $\beta_1 = 0$ asks only whether the gap vanishes at zero years of education, a point outside the range of the data. The two are different questions, and the second is much the weaker one: β_1 could be zero while a substantial gap exists at every education level actually observed, provided $\beta_3 \neq 0$.

Question 4

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

Write the model with the unobserved component made explicit:

$$\text{Productivity}_{it} = \beta_0 + \beta_1 \text{Training}_{it} + a_i + u_{it},$$

where a_i collects the time-invariant characteristics such as ability and motivation. Pooled OLS does not have a_i in the data, so it estimates a model whose error is the composite $v_{it} = a_i + u_{it}$.

The assumption that fails is exogeneity of the regressor with respect to that composite error. Because more able or more motivated employees are also more likely to receive training,

$$\text{Cov}(\text{Training}_{it}, v_{it}) = \text{Cov}(\text{Training}_{it}, a_i) \neq 0,$$

and the pooled OLS estimate of β_1 absorbs part of the productivity difference that ability alone would have produced. This is unobserved heterogeneity, and it biases the estimate in the direction of the correlation, here upward.

Averaging the model over the T periods of employee i, and using $\bar{a}_i = a_i$ because a_i does not vary over time,

$$\overline{\text{Productivity}}_i = \beta_0 + \beta_1 \overline{\text{Training}}_i + a_i + \bar{u}_i.$$

Subtracting this from the original equation eliminates the offending term:

$$\text{Productivity}_{it} - \overline{\text{Productivity}}_i = \beta_1(\text{Training}_{it} - \overline{\text{Training}}_i) + (u_{it} - \bar{u}_i).$$

The term $a_i - a_i$ is identically zero, so β_1 is now estimated from within-employee variation alone.

The cost is that the transformation removes every regressor that is constant within an employee, not only a_i. Fixed effects therefore cannot identify the effect of gender, of schooling completed before the sample begins, or of any other fixed characteristic, and it cannot recover a_i itself as a parameter of interest.

The Hausman test compares the fixed-effects estimator with the random-effects estimator. Random effects also uses the between variation and is the more efficient of the two, but it is consistent only if a_i is uncorrelated with the regressors, which is exactly the condition this scenario calls into question. The null hypothesis is that $\text{Cov}(\text{Training}_{it}, a_i) = 0$. A rejection says the correlation is real, so random effects is inconsistent and fixed effects should be preferred despite its inability to identify time-invariant regressors.

Question 5

The composite error for this setting is

$$v_{it} = a_i + u_{it},$$

where a_i is whatever persistently distinguishes one firm from another, such as management quality, market position or accounting practice, and u_{it} is the year-specific shock.

Two residuals belonging to the same firm in different years share a_i, so they are positively correlated:

$$\text{Corr}(v_{it}, v_{is}) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_u^2} > 0 \qquad (t \neq s).$$

The default standard error assumes the 4,000 firm-year observations are 4,000 independent pieces of information. Positive within-firm correlation means they are not: much of what makes a firm profitable in one year also makes it profitable in the next, so the data contain less independent variation than the sample size suggests. The default formula therefore understates the true sampling variability, standard errors come out too small, and every t-statistic is inflated.

The reported value of 3.1 is an overstatement of the evidence, and a correctly computed statistic could easily fall below conventional significance thresholds.

The researcher should report standard errors clustered at the firm level, which allows arbitrary correlation between observations of the same firm while maintaining independence across firms.

The number of firms rather than the number of firm-year observations governs the precision because firms, not firm-years, are the independent draws. Observing one firm for ten years tells us a great deal about that firm and rather little about the population of firms. Adding years to an existing firm mostly adds repeated information about the same a_i, whereas adding firms adds genuinely new draws, which is why cluster-robust inference depends on the number of clusters.