# Answers Mock Endterm

## MC Answers

1. c) The Hausman test's null hypothesis is that the unobserved individual effects are not correlated with the regressors. Under this assumption, the Random Effects estimator is both consistent and more efficient (i.e., has smaller variance) than the Fixed Effects estimator, making it the preferred choice.
2. b) The presence of the time-invariant individual effect $a_i$ in the error term for all periods for a given individual means the errors are correlated over time. The derivation shows $\text{Corr}(v_{it}, v_{is}) = \frac{\text{Cov}(v_{it}, v_{is})}{\text{Var}(v_{it})} = \frac{\sigma_a^2}{\sigma_a^2+\sigma_u^2}$. This autocorrelation necessitates the use of clustered standard errors.
3. c) The Fixed Effects (or "within") estimator is designed to eliminate omitted variable bias from unobserved individual effects ($a_i$) by demeaning the data. A consequence of this transformation is that any variable that does not change over time for a given individual (like gender, race, or a firm's headquarter location) becomes zero and its effect cannot be estimated.
4. c) The LPM can produce nonsensical predictions, such as probabilities less than 0 or greater than 1.
5. d) The Probit model assumes the error term is distributed as $\epsilon_i \sim N(0, 1)$, and its Cumulative Distribution Function ($\Phi$) is used to model the probability.
6. b) MLE is the standard estimation technique for these models; it finds the parameter values that make the observed data "most likely".
7. c) The marginal effect is the derivative of the probability with respect to $x_k$. By the chain rule, this is the derivative of the CDF, which is the PDF ($f(\cdot)$), multiplied by the coefficient $\beta_k$. The effect is not constant and depends on the values of all X variables.
8. b) McFadden's $R^2$ compares the log-likelihood of the full model to a null model (intercept only). The lecture notes explicitly state that values are much lower than OLS R-squared and that a value of 0.2-0.4 can indicate a very good fit.
9. c) The Tobit model is designed for censored data, such as hours worked or expenditure on a specific good, where many observations can be clustered at zero.
10. b) The derivation is $P(y_i = 1) = P(y_i^* > 0) = P(\beta_0 + \beta_1 x_i + \epsilon_i > 0)$, which rearranges to $P(\epsilon_i > -(\beta_0 + \beta_1 x_i))$.

11. b) For any individual, we can only observe one potential outcome ($Y_i(1)$ or $Y_i(0)$), making the individual causal effect ($\tau_i$) directly unobservable.
12. c) $E[Y(0)|T=1] - E[Y(0)|T=0]$
13. d) $\beta_3$
14. b) In the absence of the treatment, the average outcome for the treated group would have evolved in the same way as the average outcome for the control group.
15. c) By testing whether the estimated coefficients for the pre-treatment periods ($\delta_k$ for $k < 0$) are statistically indistinguishable from zero.
16. b) Relevance ($Cov(Z_i, X_i) \neq 0$) and Exclusion Restriction ($Cov(Z_i, u_i) = 0$). A valid instrument must be correlated with the endogenous variable it is instrumenting for (Relevance) and be uncorrelated with the error term of the main equation to ensure it only affects the outcome through the endogenous variable (Exclusion Restriction).
17. c) The "Compliers," whose treatment status is changed by the instrument.
18. b) A first-stage F-statistic for the joint significance of the instruments, which should be greater than 10.
19. d) $\hat{\beta}_{\text{Wald}} = \frac{E[Y|Z=1]-E[Y|Z=0]}{E[X|Z=1]-E[X|Z=0]}$. The Wald estimator correctly identifies the causal effect by scaling the Intent-to-Treat effect (the effect of the instrument on the outcome, in the numerator) by the share of compliers (the effect of the instrument on treatment take-up, in the denominator).
20. b) Stage 1: Regress the endogenous variable $X$ on the instruments $Z$ and exogenous controls $W$. Stage 2: Regress the outcome $Y$ on the *predicted values* $\hat{X}$ from Stage 1 and the exogenous controls $W$.

## Open Questions Answers

1.

**a) First Difference (Treatment City):**

The first difference for the treatment group is the change in the average outcome for this group before and after the treatment.

$$\Delta Y_{\text{Treatment}} = Y_{\text{Treatment, Post}} - Y_{\text{Treatment, Pre}}$$
$$\Delta Y_{\text{Treatment}} = 32 - 35 = -3 \text{ minutes}$$

**b) First Difference (Control City):**

This calculation shows the change in the average outcome for the control group over the same period.

$$\Delta Y_{\text{Control}} = Y_{\text{Control, Post}} - Y_{\text{Control, Pre}}$$

$$\Delta Y_{\text{Control}} = 34 - 33 = 1 \text{ minute}$$

**c) Difference-in-Differences (DiD) Estimate:**

The DiD estimate is the difference between the change in the treatment group and the change in the control group. It isolates the effect of the treatment by removing the underlying trend observed in the control group.

$$\text{DiD} = \Delta Y_{\text{Treatment}} - \Delta Y_{\text{Control}}$$
$$\text{DiD} = (-3) - (1) = -4 \text{ minutes}$$

The DiD estimate suggests that the bike-sharing program caused a **4-minute reduction** in the average commute time.

2.

**a) Simple Difference-in-Means:**

This is the straightforward comparison of the average outcomes between the treated and untreated groups.

$$E[Y|T = 1] - E[Y|T = 0] = 85 - 70 = 15$$

**b) Selection Bias Calculation:**

The simple difference in means can be broken down into the Average Treatment Effect on the Treated (ATT) and selection bias. By rearranging the provided formula, we can solve for the selection bias.

$$\text{Selection Bias} = (E[Y|T = 1] - E[Y|T = 0]) - \text{ATT}$$
$$\text{Selection Bias} = 15 - 10 = 5$$

**c) Interpretation of Selection Bias:**

The selection bias is **positive** and has a magnitude of **5 points**. This indicates that the employees who chose to take the professional development course would have had a productivity score that was, on average, 5 points higher than those who did not take the course, even if they had not participated in the program. This suggests that the employees who enrolled in the course were already more motivated, skilled, or productive to begin with.

3.

**a) Wald Estimate Calculation:**

The Wald estimate is a way to calculate the causal effect of a treatment using an instrumental variable. It is the ratio of the change in the outcome associated with the instrument to the change in the treatment uptake associated with the instrument.

$$\text{Wald Estimate} = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[X|Z=1] - E[X|Z=0]}$$

$$\text{Wald Estimate} = \frac{0.65 - 0.55}{0.40 - 0.15} = \frac{0.10}{0.25} = 0.40$$

The estimated causal effect of the transit subsidy on the employment rate is **0.40**, or a 40 percentage point increase.

**b) Numerator Interpretation:**

The numerator, $E[Y|Z=1] - E[Y|Z=0] = 0.10$, represents the **Intention-to-Treat (ITT) effect**. This is the effect of being encouraged (receiving the flyer) on the outcome (employment rate), regardless of whether the individual actually took up the subsidy.

**c) Denominator Interpretation:**

The denominator, $E[X|Z=1] - E[X|Z=0] = 0.25$, represents the effect of the instrument (encouragement) on the treatment take-up rate. In the context of potential outcomes, this value represents the proportion of **"compliers"**—individuals who only take up the subsidy if they receive the encouragement flyer.

4.

**a) First Stage Regression Interpretation:**

The coefficient on rainfall ($Z_i$) in the first stage regression is **-0.05**. This indicates that a 1 mm increase in seasonal rainfall is associated with a **$0.05 decrease** in the average price of an avocado. This makes economic sense, as more rainfall generally leads to a better avocado harvest (increased supply), which in turn lowers the price. This result supports the **relevance** condition for an instrumental variable because it shows that the instrument (rainfall) is correlated with the endogenous variable (price).

**b) Reduced Form Regression Interpretation:**

The coefficient on rainfall ($Z_i$) in the reduced form regression is **2.0**. This signifies that a 1 mm increase in seasonal rainfall is associated with an increase of **2,000 tons** in the quantity of avocados sold. This captures the total effect of the instrument on the outcome.

**c) 2SLS Estimate Calculation:**

The 2SLS estimate for the causal effect of price on quantity can be calculated by dividing the reduced-form coefficient by the first-stage coefficient.

$$\hat{\beta}_{2SLS} = \frac{\text{Reduced Form Coefficient}}{\text{First Stage Coefficient}}$$

$$\hat{\beta}_{2SLS} = \frac{2.0}{-0.05} = -40$$

The 2SLS estimate indicates that a $1 increase in the price of an avocado leads to a decrease in the quantity demanded by **40,000 tons**.

5.

**a) Predicted Probability for dti = 0.5:**

To find the predicted probability, we insert the `dti` value into the estimated equation.

$$\hat{P}(\text{default} = 1) = -0.15 + 0.70 \cdot (0.5)$$

$$\hat{P}(\text{default} = 1) = -0.15 + 0.35 = 0.20$$

The predicted probability of default is **20%**.

**b) Predicted Probability for dti = 2.0:**

$$\hat{P}(\text{default} = 1) = -0.15 + 0.70 \cdot (2.0)$$

$$\hat{P}(\text{default} = 1) = -0.15 + 1.40 = 1.25$$

The predicted probability of default is **125%**.

**c) Weakness of LPM and Coefficient Interpretation:**

The calculation for a `dti` of 2.0, which results in a predicted probability of 125%, highlights the primary weakness of the Linear Probability Model: **it can produce predicted probabilities outside the logical 0 to 1 range**.

The coefficient of **0.70** on `dti` is interpreted as the change in the probability of default for a one-unit increase in the debt-to-income ratio. Specifically, for every 1-unit increase in an applicant's `dti`, the probability of them defaulting on the loan is predicted to increase by **70 percentage points**.