Questions Mock Endterm

The questions in this mock exam are mostly based on lectures 5, 6, 7, and 8. In the real endterm, all course material is potentially relevant, including lectures 1 to 4, but the questions will be more heavily focused on the latter part of the course.

MC Questions

1. The Hausman test is used to compare the Fixed Effects (FE) and Random Effects (RE) models. What is the null hypothesis (H_0) of the Hausman test, and what is the appropriate modeling choice if you fail to reject the null?

    a) H_0: The variance of the individual-specific effect is zero ($\sigma_a^2 = 0$). If not rejected, choose Pooled OLS.
    b) H_0: The coefficients from the FE and RE models are the same. If not rejected, choose the FE model for its consistency.
    c) H_0: The unobserved individual effect (a_i) is uncorrelated with the explanatory variables (X_{it}). If not rejected, choose the more efficient RE model.
    d) H_0: There is serial correlation in the idiosyncratic error term (u_{it}). If not rejected, choose the First Differences (FD) model.

2. In a Pooled OLS model applied to panel data, the composite error term is defined as $v_{it} = a_i + u_{it}$, where a_i is the unobserved individual-specific effect and u_{it} is the idiosyncratic error. Assuming $\text{Var}(a_i) = \sigma_a^2$ and $\text{Var}(u_{it}) = \sigma_u^2$, what is the correlation between two error terms for the same individual i at different time periods t and s?

    a) Zero, because the error terms are assumed to be independent across time.
    b) $\frac{\sigma_a^2}{\sigma_a^2+\sigma_u^2}$, because the shared a_i component induces positive autocorrelation.
    c) σ_a^2, because the covariance between the error terms is equal to the variance of the individual effect.
    d) -0.5, which indicates that a First Differences estimator would be more appropriate.

3. A researcher is using panel data to investigate the determinants of firms' profits. One variable of interest is the country where the firm is headquartered, which does not change over the sample period. Which estimator is fundamentally unable to estimate the effect of the headquarter country, and why?

   a) The Random Effects estimator, because it requires all explanatory variables to be strictly exogenous.
   b) The Pooled OLS estimator, because it pools all data together and ignores the panel structure entirely.
   c) The Fixed Effects estimator, because its transformation subtracts the individual-specific mean from each variable ($\ddot{X}_{it} = X_{it} - \bar{X}_i$), which eliminates any variable that is constant over time.
   d) All panel data estimators are unable to estimate the effect of time-invariant variables.

4. What is a primary drawback of using the Linear Probability Model (LPM) for binary outcome data?

   a) The model's coefficients are difficult to interpret.
   b) It requires Maximum Likelihood Estimation, which is computationally intensive.
   c) Its predicted probabilities are not constrained to the interval.
   d) It cannot incorporate fixed effects for panel data analysis.

5. The Probit and Logit models are derived from a latent variable framework where an unobserved variable y_i^* determines the observed binary outcome y_i. What specific assumption about the error term (ϵ_i) in the latent model ($y_i^* = X_i\beta + \epsilon_i$) distinguishes the Probit model?

   a) The error term follows a Standard Logistic distribution.
   b) The error term follows a Student's t-distribution.
   c) The error term has a variance that depends on the independent variables (X_i).
   d) The error term follows a Standard Normal distribution.

6. How are the β coefficients in Logit and Probit models estimated?

   a) By using Ordinary Least Squares (OLS) to minimize the sum of squared residuals.
   b) By finding the parameters that maximize the joint probability of observing the actual data, a method known as Maximum Likelihood Estimation (MLE).
   c) By calculating the percentage of correctly predicted outcomes and choosing the coefficients that maximize this percentage.
   d) By using the F-test to determine the joint significance of the variables.

7. In a Logit or Probit model where the probability of an event is given by $P(y = 1|X) = F(\beta x_i)$, what is the correct formula for the marginal effect of a continuous variable x_k on this probability?

a) β
b) $F(X\beta) \cdot \beta$
c) $f(X\beta) \cdot \beta$
d) $\frac{1}{N} \sum_{i=1}^{N} \beta$

8. When assessing the goodness-of-fit of a binary outcome model, which of the following statements about McFadden's Pseudo R-squared (R^2_{McF}) is correct?

   a) It should be directly compared to the R-squared from an OLS model to determine which model is better.
   b) It is calculated as $1 - \frac{\ln L_{full}}{\ln L_{null}}$ and typically has much lower values than a conventional OLS R-squared.
   c) A value below 0.5 indicates a very poor model fit that should always be rejected.
   d) It represents the percentage of the variance in the dependent variable that is explained by the model.

9. The Tobit model is most appropriate for addressing which specific econometric issue?

   a) Sample selection bias, where data on the dependent variable is missing for a non-random group.
   b) Heteroskedasticity in the error term of a Linear Probability Model.
   c) A censored dependent variable, where the variable is continuous but has a large number of observations at a limit value (e.g., zero).
   d) Endogeneity caused by omitted variables.

10. According to the latent variable framework, the probability of the observed outcome being 1 is derived from the probability that the latent variable y_i^* is greater than zero. Given the model $y_i^* = \beta_0 + \beta_1 x_i + \epsilon_i$, this probability is expressed as:

    a) $P(y_i^* \leq 0)$
    b) $P(\epsilon_i > -(\beta_0 + \beta_1 x_i))$
    c) $P(\epsilon_i = \beta_0 + \beta_1 x_i)$
    d) $1 - F(\beta_0 + \beta_1 x_i)$

11. What is the "fundamental problem of causal inference" according to the potential outcomes framework?

    a) It is difficult to find a large enough sample size for empirical analysis.
    b) For any individual, we can only observe one potential outcome ($Y_i(1)$ or $Y_i(0)$), making the individual causal effect (τ_i) directly unobservable.
    c) The Average Treatment Effect (ATE) is always equal to zero.
    d) Correlation and causation are interchangeable concepts.

12. In the decomposition of the simple difference-in-means, $E[Y|T = 1] - E[Y|T = 0]$, the "selection bias" term is mathematically defined as:

a) $E[Y(1)|T=1] - E[Y(0)|T=1]$
b) $E[Y(1) - Y(0)]$
c) $E[Y(0)|T=1] - E[Y(0)|T=0]$
d) $\hat{Y}_{T,Post} - \hat{Y}_{T,Pre}$

13. In the standard Difference-in-Differences (DiD) regression model, which coefficient represents the DiD estimate of the causal effect? $Y_{it} = \beta_0 + \beta_1 Treat_i + \beta_2 Post_t + \beta_3 (Treat_i \times Post_t) + \epsilon_{it}$

a) β_0
b) β_1
c) β_2
d) β_3

14. The validity of the Difference-in-Differences (DiD) estimator relies on the crucial "parallel trends" assumption. What does this assumption state?

a) The treatment and control groups must have identical average outcomes in the pre-treatment period.
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.
c) The outcome for the control group must remain constant over both the pre- and post-treatment periods.
d) The treatment must have a positive and statistically significant effect on the outcome.

15. When using an event study design with multiple pre-treatment periods, how can researchers provide evidence for the parallel trends assumption?

a) By showing that the treatment effect is large and significant immediately at the time of the event ($k = 0$).
b) By demonstrating that the average outcomes for the treatment and control groups were identical in the first pre-treatment period.
c) By testing whether the estimated coefficients for the pre-treatment periods (δ_k for $k < 0$) are statistically indistinguishable from zero.
d) By including as many control variables as possible in the regression model.

16. In the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$, endogeneity is a problem because the Ordinary Least Squares (OLS) assumption $Cov(X_i, u_i) = 0$ is violated. A variable Z is considered a valid instrument to solve this problem if it satisfies which two core conditions?

a) **Relevance** ($Cov(Z_i, Y_i) \neq 0$) and **Exclusion Restriction** ($Cov(Z_i, u_i) = 0$)
b) **Relevance** ($Cov(Z_i, X_i) \neq 0$) and **Exclusion Restriction** ($Cov(Z_i, u_i) = 0$)
c) **Irrelevance** ($Cov(Z_i, X_i) = 0$) and **Inclusion Restriction** ($Cov(Z_i, u_i) \neq 0$)

d) **Exogeneity of X** ($Cov(X_i, u_i) = 0$) and **Relevance** ($Cov(Z_i, X_i) \neq 0$)

17. The Instrumental Variables (IV) estimator, under the necessary assumptions, does not estimate the Average Treatment Effect (ATE) for the whole population. Instead, it identifies the Local Average Treatment Effect (LATE). The LATE is the average causal effect for which specific subpopulation?

    a) The "Always-Takers," who take the treatment regardless of the instrument.
    b) The "Never-Takers," who do not take the treatment regardless of the instrument.
    c) The "Compliers," whose treatment status is changed by the instrument.
    d) The entire study population, including all subgroups.

18. A critical issue in IV analysis is the problem of "weak instruments," where the instrument is only weakly correlated with the endogenous variable. What is the common diagnostic test and rule of thumb used to detect weak instruments?

    a) A test of the exclusion restriction ($Cov(Z, u) = 0$), which should have a p-value less than 0.05.
    b) A first-stage F-statistic for the joint significance of the instruments, which should be greater than 10.
    c) The R-squared of the second-stage regression, which should be as high as possible.
    d) The t-statistic of the IV estimate for β_1 in the second stage, which should be greater than 1.96.

19. The Wald estimator is the simplest form of the IV estimator, used when both the instrument (Z) and the endogenous regressor (X) are binary. It is defined as a ratio of two effects. What is the correct formula?

    a) $\hat{\beta}_{\text{Wald}} = \frac{E[Y|X=1]-E[Y|X=0]}{E[Z|X=1]-E[Z|X=0]}$
    b) $\hat{\beta}_{\text{Wald}} = \frac{Cov(X,Y)}{Var(X)}$
    c) $\hat{\beta}_{\text{Wald}} = \frac{E[X|Z=1]-E[X|Z=0]}{E[Y|Z=1]-E[Y|Z=0]}$
    d) $\hat{\beta}_{\text{Wald}} = \frac{E[Y|Z=1]-E[Y|Z=0]}{E[X|Z=1]-E[X|Z=0]}$

20. In the general case with multiple instruments or control variables, the IV estimate is calculated using Two-Stage Least Squares (2SLS). Which of the following statements correctly describes this procedure?

    a) **Stage 1**: Regress the outcome Y on the instruments Z. **Stage 2**: Regress Y on the predicted values of Y from Stage 1.
    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.

c) **Stage 1**: Regress the outcome Y on the endogenous variable X to get a biased estimate. **Stage 2**: Regress the residuals from Stage 1 on the instruments Z to correct the bias.

d) **Stage 1**: Regress the instruments Z on the endogenous variable X. **Stage 2**: Regress the outcome Y on the residuals from Stage 1.

Open Questions

1. A city introduces a new public bike-sharing program to reduce traffic congestion. To evaluate its impact, you collect data on the average commute time (in minutes) from this city (the "Treatment Group") and a neighboring, similar city that did not introduce the program (the "Control Group"). The data is collected one year before the program's launch and one year after.

The average commute times are as follows:

	Before Program (Pre)	**After Program (Post)**
Treatment City	35 minutes	32 minutes
Control City	33 minutes	34 minutes

a) Calculate the "first difference" for the Treatment City (the change in average commute time over the two periods).

b) Calculate the "first difference" for the Control City. This represents the underlying trend in commute times.

c) Calculate the Difference-in-Differences (DiD) estimate for the effect of the bike-sharing program.

2. A company offers a voluntary professional development course designed to increase employee productivity. After the course, the company measures productivity on a scale of 1 to 100.

- The employees who **took the course** ($T = 1$) have an average productivity score of 85.
- The employees who **did not take the course** ($T = 0$) have an average productivity score of 70.

  A separate, rigorous analysis reveals that the **Average Treatment Effect on the Treated (ATT)** is 10 points. That is, for the specific group of employees who chose to take the course, the course itself caused their productivity to increase by an average of 10 points.

a) Calculate the simple difference-in-means of the productivity scores between those who took the course and those who did not.

b) Using the decomposition formula provided in the lecture, calculate the **Selection Bias**.

$$E[Y|T=1] - E[Y|T=0] = ATT + \text{Selection Bias}$$

c) Interpret the sign (positive or negative) and magnitude of the selection bias you calculated. What does this value tell you about the employees who chose to enroll in the course compared to those who did not, in terms of their potential productivity *without* the course? (Hint: Think about the definition of selection bias: Selection Bias = $E[Y(0)|T=1] - E[Y(0)|T=0]$).

3. A city wants to evaluate the causal effect of a new public transit subsidy program on employment. The program offers a 50% discount on monthly transit passes to eligible low-income residents. However, simply comparing the employment rates of those who used the subsidy ($X = 1$) versus those who did not ($X = 0$) would lead to a biased estimate due to self-selection (e.g., more motivated individuals might be more likely to both sign up for the subsidy and find a job).

   To overcome this, researchers use an instrumental variable. They randomly send an informational flyer and a pre-filled application form to a randomly selected group of eligible residents. This "encouragement" flyer serves as the instrument ($Z = 1$ for those who received it, $Z = 0$ for those who did not). After the study period, the researchers collect the following data:

- For the group that **received the encouragement flyer** ($Z = 1$):
  - The employment rate is 65%. ($E[Y|Z=1] = 0.65$)
  - The subsidy take-up rate is 40%. ($E[X|Z=1] = 0.40$)
- For the group that **did not receive the flyer** ($Z = 0$):
  - The employment rate is 55%. ($E[Y|Z=0] = 0.55$)
  - The subsidy take-up rate is 15%. ($E[X|Z=0] = 0.15$)

a) Calculate the Wald estimate for the causal effect of the transit subsidy program on the employment rate. Show your work.
b) Interpret the value you calculated for the **numerator** of the Wald estimator. What is the common name for this effect?
c) Interpret the value you calculated for the **denominator** of the Wald estimator. In the context of the potential outcomes framework, what does this value represent?

4. An economist is studying the market for avocados and wants to estimate the causal effect of price on the quantity demanded. They know that running a simple OLS regression of quantity on price will be biased due to simultaneity (i.e., unobserved demand shocks affect both price and quantity). To solve this problem, they use the amount of rainfall in the primary avocado-growing region as an instrumental variable (Z). The logic is that rainfall affects the avocado supply, which in turn affects the price, but rainfall does

not directly affect consumer demand for avocados. The economist runs three regressions using data where:

- Q_i is the quantity of avocados sold (in thousands of tons).
- P_i is the average price of an avocado (in dollars).
- Z_i is the seasonal rainfall (in mm).

The results are:

1. **OLS Regression**: $\hat{Q}_i = 150 + 10.5P_i$
2. **First Stage Regression:** $\hat{P}_i = 2.50 - 0.05Z_i$
3. **Reduced Form Regression:** $\hat{Q}_i = 120 + 2.0Z_i$

a) Look at the coefficient on rainfall (Z_i) in the First Stage regression. What does this coefficient tell you about the relationship between rainfall and avocado prices? Does the sign of the coefficient make economic sense? Explain why this result supports the **relevance** condition for an instrument.

b) Look at the coefficient on rainfall (Z_i) in the Reduced Form regression. What does this coefficient tell you about the overall relationship between rainfall and the quantity of avocados sold?

c) Using the coefficients from the First Stage and Reduced Form regressions, calculate the Two-Stage Least Squares (2SLS) estimate for β_1 in the structural model $Q_i = \beta_0 + \beta_1 P_i + u_i$. Show your calculation.

5. A financial analyst is modeling the probability that a loan applicant will default on a personal loan. The key predictor variable is the applicant's debt-to-income ratio (`dti`). The outcome variable, `default`, is 1 if the person defaulted and 0 otherwise. The analyst estimates a Linear Probability Model (LPM). The estimated LPM equation is:

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

a) Using this equation, calculate the predicted probability of default for an applicant with a `dti` of 0.5.

b) Now, calculate the predicted probability for an applicant with a `dti` of 2.0.

c) Based on these two calculations, explain the primary weakness of the LPM and interpret the meaning of the 0.70 coefficient.

End of Exam