

Answers Mock Midterm

Answer Key

Multiple Choice Questions

1. **b)** The goal is to forecast a value (\hat{y}) as accurately as possible, which is the definition of a prediction task.
2. **d)** The data tracks the same cross-sectional units (individuals) over multiple time periods, which defines panel data.
3. **c)** One of the OLS First Order Conditions directly implies that the sum of the explanatory variable multiplied by the residual is zero ($\sum x_i \hat{u}_i = 0$), which means their sample covariance is zero.
4. **b)** In a log-log model, the slope coefficient is an elasticity, representing the percentage change in the dependent variable for a 1% change in the independent variable.
5. **b)** For males, the dummy variable `Female` is 0. Therefore, the predicted wage is $\widehat{\text{Wage}} = 22.50 - 3.75 \times (0) = 22.50$. The intercept represents the average value for the reference group.
6. **c)** R-squared measures the proportion of the sample variation in the dependent variable that is explained by the independent variable(s).
7. **c)** The omitted variable (“parental involvement”) has a positive effect on test scores ($\beta_2 > 0$). It is negatively correlated with the included variable, `ClassSize` ($\delta_1 < 0$). The bias is the product of these two effects: (positive) \times (negative) = negative bias. This negative bias will make the estimated coefficient for class size more negative (i.e., appear larger in magnitude) than the true effect.
8. **d)** The formula for the variance is $\widehat{\text{Var}}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{SST_x}$. To decrease the variance, one must either decrease the numerator (error variance) or increase the denominator. Increasing the variation in the explanatory variable (SST_x) increases the denominator, thus decreasing the overall variance.

9. **c)** The student must start with the formula relating the estimated coefficient to the true coefficient and the bias:

$$E[\hat{\alpha}_1] = \beta_1 + \text{Bias}$$

Then, use the specific formula for omitted variable bias: $\text{Bias} = \beta_2 \cdot \delta_1$

From the problem description:

- The true effect of the included variable is $\beta_1 = 2.5$.
- The true effect of the omitted variable is $\beta_2 = 4.0$.
- The coefficient from the auxiliary regression of the omitted variable on the included variable is $\delta_1 = 0.5$.

First, calculate the bias: $\text{Bias} = 4.0 \times 0.5 = 2.0$

Next, use this to find the expected value of the estimated coefficient: $E[\hat{\alpha}_1] = 2.5 + 2.0 = 4.5$

10. **c)** Use the formula of the estimated model to derive the expected wage for each category and then find the difference.

Expected wage for a Finance worker: For this person, **Finance** = 1 and **Retail** = 0.
 $E[\text{Wage}|\text{Finance}] = 18.50 - 2.10 \times (0) + 5.40 \times (1) = 18.50 + 5.40 = 23.90$

Expected wage for a Retail worker: For this person, **Retail** = 1 and **Finance** = 0.
 $E[\text{Wage}|\text{Retail}] = 18.50 - 2.10 \times (1) + 5.40 \times (0) = 18.50 - 2.10 = 16.40$

Calculate the difference: $\text{Difference} = E[\text{Wage}|\text{Finance}] - E[\text{Wage}|\text{Retail}]$ Difference = $23.90 - 16.40 = 7.50$

Alternative algebraic derivation: Let $\hat{\beta}_0 = 18.50$, $\hat{\beta}_1 = -2.10$, and $\hat{\beta}_2 = 5.40$. The difference is $(\hat{\beta}_0 + \hat{\beta}_2) - (\hat{\beta}_0 + \hat{\beta}_1) = \hat{\beta}_2 - \hat{\beta}_1 = 5.40 - (-2.10) = 5.40 + 2.10 = 7.50$.

11. **c)** The effect of a one-unit change in x_1 on y , holding x_2 constant.
12. **c)** A confounder, and failing to control for it will likely lead to omitted variable bias.
13. **b)** An F-test for the overall significance of the regression.
14. **c)** The standard errors for the coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ will be large.
15. **d)** The difference in the return to an additional year of education for females compared to males.
16. **c)** The lecture explicitly states that if $|\rho| = 1$, the series is non-stationary and is called a random walk, where shocks have a permanent effect.
17. **b)** The Long-Run Multiplier is defined as the total effect on Y after a permanent change in X has fully worked through the system, and it is calculated as the ratio of the sum of the distributed lag coefficients to one minus the sum of the autoregressive coefficients.

18. **c)** For an MA(q) model, any forecast for a horizon $h > q$ will only involve future error terms, whose conditional expectation is zero. Therefore, the forecast reverts to the process mean: $\hat{Y}_{t+h|t} = \mu$.
19. **c)** The “Within” estimator works by performing a “within transformation” or “de-meaning.” For each individual, it calculates the average of the dependent and independent variables over time and subtracts these individual-specific averages from the original data. This process removes the time-constant individual effect (a_i), as $\alpha_i - \bar{\alpha}_i = 0$.
20. **c)** The Fixed Effects model eliminates all variables that are constant over time for an individual. Since the firm’s industry sector does not change, it would be absorbed into the individual-specific fixed effect (a_i) and its coefficient cannot be estimated. The other variables (innovation, age, unemployment rate) all vary over time.

Open Questions

Question 1

a) Estimated Sample Regression Function (SRF). The estimated Sample Regression Function is:

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

b) Interpretation of the educ Coefficient. The estimated coefficient for **educ**, 0.541, indicates that for each additional year of education, an individual’s average hourly wage is predicted to increase by \$0.541, or approximately 54 cents, holding other factors constant.

c) Interpretation of the Intercept. The intercept, -0.905, represents the predicted average hourly wage for an individual with zero years of education. In this context, the intercept does not have a meaningful real-world interpretation because it is impossible to earn a negative wage, and having zero years of education is a significant extrapolation beyond the range of data for a sample of working individuals.

Question 2

a) Sign of the Coefficient on Experience (β_2). The likely sign of β_2 is **positive**. Holding education constant, additional years of work experience typically increase a worker’s skills and productivity, leading to higher wages.

b) Correlation Between Education and Experience. The likely correlation between **Educ** and **Exper** is **negative**. For a given age, time spent acquiring more education is time not spent in the labor force gaining experience. Therefore, individuals with more years of schooling tend to have fewer years of work experience.

c) Direction of the Omitted Variable Bias The simple model will likely **underestimate** the true effect of education on wages.

The bias in the estimated coefficient of **Educ** is determined by the product of two factors:

1. The effect of the omitted variable (**Exper**) on the dependent variable (**Wage**). (From part a, this is **positive**).
2. The correlation between the omitted variable (**Exper**) and the included variable (**Educ**). (From part b, this is **negative**).

Bias = (Effect of **Exper** on **Wage**) × (Correlation of **Exper** and **Educ**) Bias = (+) × (−) = **Negative Bias**

A negative bias means the estimated coefficient for education in the simple model will be smaller than its true value, thus underestimating the return to education.

Question 3

This answer model is very extensive, you don't have to be as extensive in your answer.

a) Specific Regression Equations for Males and Females. To find the specific regression equation for each group, we substitute the value of the **Female_i** dummy variable into the general model.

For males ($Female_i = 0$): We set $Female_i = 0$ in the original equation:

$$Wage_i = \beta_0 + \beta_1(0) + \beta_2 Educ_i + \beta_3(0 \cdot Educ_i) + u_i$$

The terms involving β_1 and β_3 drop out, simplifying the equation to:

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

For males, the intercept (the wage at zero years of education) is β_0 , and the slope (the return to an additional year of education) is β_2 .

For females ($Female_i = 1$): We set $Female_i = 1$ in the original equation:

$$Wage_i = \beta_0 + \beta_1(1) + \beta_2 Educ_i + \beta_3(1 \cdot Educ_i) + u_i$$

This simplifies to:

$$Wage_i = \beta_0 + \beta_1 + \beta_2 Educ_i + \beta_3 Educ_i + u_i$$

To make the structure clear, we can group the intercept and slope terms:

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

For females, the intercept is $(\beta_0 + \beta_1)$, and the slope is $(\beta_2 + \beta_3)$.

b) Marginal Effect of Education on Wages. The marginal effect of education is the partial derivative of the expected wage with respect to education. The expected wage equation is:

$$E[Wage_i] = \beta_0 + \beta_1 Female_i + \beta_2 Educ_i + \beta_3 (Female_i \cdot Educ_i)$$

Taking the partial derivative with respect to $Educ_i$:

$$\frac{\partial E[Wage]}{\partial Educ_i} = \beta_2 + \beta_3 Female_i$$

This shows that the marginal effect of education depends on the value of the $Female_i$ dummy.

For females ($Female_i = 1$): The marginal effect of an additional year of education on wages is:

$$\frac{\partial E[Wage]}{\partial Educ_i} = \beta_2 + \beta_3(1) = \beta_2 + \beta_3$$

Difference from the marginal effect for males: For males ($Female_i = 0$), the marginal effect of education is:

$$\frac{\partial E[Wage]}{\partial Educ_i} = \beta_2 + \beta_3(0) = \beta_2$$

The marginal effect for females differs from that of males by the value of the interaction coefficient, β_3 . If β_3 is statistically different from zero, it means that the economic return to an additional year of schooling is different for females and males.

- If $\beta_3 > 0$, females have a higher marginal return to education than males.
- If $\beta_3 < 0$, females have a lower marginal return to education than males.

c) Derivation and Interpretation of the Wage Gap. The wage gap is the difference in the expected wage between a female and a male who have the same level of education, $Educ$. We can derive this by subtracting the expected wage equation for males from the expected wage equation for females.

$$\text{Wage Gap} = E[Wage|Female = 1, Educ] - E[Wage|Female = 0, Educ]$$

Using the equations from part (a):

$$\text{Wage Gap} = [(\beta_0 + \beta_1) + (\beta_2 + \beta_3)\text{Educ}] - [\beta_0 + \beta_2\text{Educ}]$$

Distributing the negative sign and simplifying:

$$\text{Wage Gap} = \beta_0 + \beta_1 + \beta_2\text{Educ} + \beta_3\text{Educ} - \beta_0 - \beta_2\text{Educ}$$

The β_0 and $\beta_2\text{Educ}$ terms cancel out, leaving:

$$\text{Wage Gap} = \beta_1 + \beta_3\text{Educ}$$

Explanation of how the wage gap depends on education: The presence of the interaction term makes the wage gap a function of the level of education, rather than a constant value. The two coefficients in the gap equation have distinct interpretations:

- β_1 : This is the “baseline” wage gap at zero years of education ($\text{Educ} = 0$). It represents the difference in the starting wage between females and males, holding education constant at zero. A negative β_1 would indicate that females start at a lower wage.
- β_3 : This coefficient determines how the wage gap **changes** for each additional year of education.
 - If $\beta_3 > 0$, the wage gap becomes less negative (or more positive) as education increases. This means that with more education, the female wage “catches up” to or surpasses the male wage.
 - If $\beta_3 < 0$, the wage gap widens (becomes more negative) as education increases. This implies that the wage penalty for being female is larger at higher levels of education.
 - If $\beta_3 = 0$, the interaction term is irrelevant, and the wage gap is constant across all education levels, equal to β_1 .

Question 4

Part a)

- **One-Step-Ahead Forecast (for 2025, Q1):**
 - First, recognize that $t-1$ corresponds to 2024, Q4 and $t-2$ corresponds to 2024, Q3.
 - Calculation: $\text{Forecast}(\text{Q1}) = 0.2 + 1.2 \times (\text{INFLATION}_{2024, \text{Q4}}) - 0.4 \times (\text{INFLATION}_{2024, \text{Q3}})$
 $\text{Forecast}(\text{Q1}) = 0.2 + 1.2 \times (3.0) - 0.4 \times (2.5)$
 $\text{Forecast}(\text{Q1}) = 0.2 + 3.6 - 1.0 = 2.8$
 - The forecast for 2025 Q1 is **2.8%**.

- **Two-Step-Ahead Forecast (for 2025, Q2):**

- Crucially, use the **forecasted value** for $t - 1$ and the known actual value for $t - 2$.
- Calculation: $\text{Forecast}(Q2) = 0.2 + 1.2 \times (\text{Forecast}(Q1)) - 0.4 \times (\text{INFLATION}_{2024, Q4})$
 $\text{Forecast}(Q2) = 0.2 + 1.2 \times (2.8) - 0.4 \times (3.0)$
 $\text{Forecast}(Q2) = 0.2 + 3.36 - 1.2 = 2.36$
- The forecast for 2025 Q2 is **2.36%**.

Part b)

- **Forecast Errors:** Calculate the difference $e_t = Y_t - \hat{Y}_t$.

- Error for Q1: $e_1 = 3.5\%(\text{Actual}) - 2.8\%(\text{Forecast}) = 0.7$
- Error for Q2: $e_2 = 3.8\%(\text{Actual}) - 2.36\%(\text{Forecast}) = 1.44$

- **Root Mean Squared Error (RMSE) Calculation:**

- First, calculate the Mean Squared Error (MSE): $\text{MSE} = \frac{e_1^2 + e_2^2}{2} = \frac{0.7^2 + 1.44^2}{2}$ $\text{MSE} = \frac{0.49 + 2.0736}{2} = \frac{2.5636}{2} = 1.2818$
- Then, take the square root to get the RMSE: $\text{RMSE} = \sqrt{1.2818} \approx 1.132$

- **Interpretation:** The RMSE is approximately **1.13**. The student should explain that this value means that, on average, the model's forecasts for inflation over this two-quarter period were off by about **1.13 percentage points**. Since the RMSE is in the same units as the original variable, it provides a direct and interpretable measure of the typical size of a forecast error.

Question 5

This answer model is very extensive, you don't have to be as extensive in your answer.

Part a) Let the true relationship between productivity and training be represented by the following linear model:

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

Here, i indexes the employee and t indexes the time period. The term a_i represents the **unobserved heterogeneity**, such as an employee's innate ability or motivation. It is specific to each individual but constant over time. The term u_{it} is the idiosyncratic error.

The statistical problem of omitted variable bias arises if this unobserved effect a_i is correlated with an included explanatory variable. In this context, two conditions are likely to hold:

1. **Correlation with the regressor:** Employees with higher innate ability (a_i) may be more likely to receive or seek out training. $\text{Cov}(\text{Training}_{it}, a_i) \neq 0$

2. Correlation with the outcome: Innate ability (a_i) directly affects an employee's output. $\text{Cov}(\text{Productivity}_{it}, a_i) \neq 0$

If a simple OLS regression is performed, the model being estimated is $\text{Productivity}_{it} = \beta_0 + \beta_1 \text{Training}_{it} + v_{it}$, where the error term $v_{it} = a_i + u_{it}$ now contains the unobserved heterogeneity. For the OLS estimator of β_1 to be unbiased and consistent, the explanatory variable must be uncorrelated with the error term, requiring $\text{Cov}(\text{Training}_{it}, v_{it}) = 0$.

However, due to the correlation between training and ability, this condition is violated:

$$\text{Cov}(\text{Training}_{it}, v_{it}) = \text{Cov}(\text{Training}_{it}, a_i + u_{it}) = \text{Cov}(\text{Training}_{it}, a_i)$$

(assuming $\text{Cov}(\text{Training}_{it}, u_{it}) = 0$).

Since $\text{Cov}(\text{Training}_{it}, a_i) \neq 0$, the OLS assumption is violated. The resulting estimator for β_1 will be biased, capturing not only the effect of training but also spuriously attributing the effect of innate ability to the training variable.

Part b) The Fixed Effects (FE) estimator mechanically eliminates the time-invariant unobserved heterogeneity a_i from the model by applying the “**within transformation.**” This process consists of two steps:

Step 1: Calculate the time-average for each individual. For each employee i , we average the original equation over all T time periods:

$$\frac{1}{T} \sum_{t=1}^T \text{Productivity}_{it} = \frac{1}{T} \sum_{t=1}^T (\beta_0 + \beta_1 \text{Training}_{it} + a_i + u_{it})$$

Using the notation $\bar{x}_i = \frac{1}{T} \sum_{t=1}^T x_{it}$ for the time-average, this becomes:

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

Because a_i is constant over time for each individual, its average is simply itself: $\bar{a}_i = a_i$. The averaged equation is therefore:

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

Step 2: Subtract the time-averaged equation from the original equation (de-meaning). Subtracting the averaged equation (for individual i) from the original equation (for individual i at time t) yields:

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

The unobserved effect a_i is perfectly eliminated from the equation. Using the “de-meaned” or “within” notation where $\ddot{x}_{it} = x_{it} - \bar{x}_i$, the transformed model is:

$$\text{Productivity}_{it} = \beta_1 \text{Training}_{it} + \ddot{u}_{it}$$

By applying OLS to this transformed equation, we obtain the Fixed Effects estimator for β_1 . This estimator is consistent because the source of the endogeneity, a_i , has been removed. The coefficient β_1 is now estimated using only the variation *within* each employee’s productivity and training over time.