

Singapore Management University
Econ 107: Introduction to Econometrics
Practice Midterm Exam, Sep. 2020

1. (22 points, 2 points each) Multiple Choice Questions: choose the one alternative that best completes the statement or answers the question. **Please write your answer in the following box.**

1)	2)	3)	4)	5)		6)	7)	8)	9)	10)	11)

1) Daily stock prices for 200 stocks from 1 March 1990 to 1 March 2020 is an example of using
 A) time series data.
 B) panel data.
 C) cross-sectional data.
 D) experimental data.

2) In a linear regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$, which one of the following statements is **NOT true**?

- A) $\sum_{i=1}^n X_{3i} u_i = 0$.
 B) $\sum_{i=1}^n \hat{u}_i = 0$.
 C) $0 \leq R^2 = \frac{ESS}{TSS} \leq 1$.
 D) $\sum_{i=1}^n X_{3i} \hat{u}_i = 0$.

3) In a linear regression model $Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$, which one of the following statements is **true**?

- A) $\sum_{i=1}^n X_{4i} u_i = 0$.
 B) $\sum_{i=1}^n \hat{u}_i = 0$.
 C) $0 \leq R^2 = \frac{ESS}{TSS} \leq 1$.
 D) $\sum_{i=1}^n \hat{Y}_i \hat{u}_i = 0$.

4) In a linear regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$, which of the following statements about the assumption that $E(u_i | X_{1i}, X_{2i}, X_{3i}) = 0$ is **not correct**?

- A) It says that the conditional distribution of the error given the explanatory variables has a zero mean.
 B) It implies that u_i and X_{3i}^2 are uncorrelated.
 C) It implies that u_i^2 and X_{2i} are uncorrelated.
 D) It cannot hold if u_i and X_{1i} are correlated.

5) An example of a randomized controlled experiment is when
 A) random variables are controlled for by holding other factors constant.
 B) one U.S. state increases the cigarette tax and an adjacent state does not, and cigarette consumption differences are observed.
 C) households receive government bonus in 2017 but not in 2016.
 D) some 2nd graders in a specific elementary school are randomly assigned to attend piano lessons at school while others are not, and their end-of-year performance is compared.

6) In a linear regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$, which of the following factors makes $Var(\hat{\beta}_1)$ **smaller**?

- A) a smaller value of sample size
- B) a larger variation in Y_i
- C) a larger value of the error variance
- D) a larger variation in X_i

7) Which of the following statement is **true**?

- A) As you add irrelevant independent variables to a linear regression model, TSS will decrease.
- B) As you remove independent variables from a linear regression model, \bar{R}^2 will decrease.
- C) As n increases, the difference between \bar{R}^2 and R^2 increases.
- D) \bar{R}^2 is usually preferred to R^2 because it takes into account the degrees of freedom in the regression.

8) Consider the following multiple regression models A) to D) below. *Male* is a binary variable which takes on the value one if the individual is male, and is zero otherwise; *Female* = 1 if the individual is a female, and is zero otherwise; *Married* is a binary variable which is unity for married individuals and is zero otherwise, and *Single* = $1 - \text{Married}$. Regressing weekly earnings (*Earn*) on a set of explanatory variables, you will experience perfect multicollinearity in the following cases unless:

- A) $Earn = \beta_0 + \beta_1 \text{Married} + \beta_2 \text{Single} + \beta_3 \text{Schooling} + u$.
- B) $Earn = \beta_0 + \beta_1 \text{Male} + \beta_2 \text{Female} + \beta_3 \text{Schooling} + u$.
- C) $Earn = \beta_1 \text{Male} + \beta_2 \text{Female} + \beta_3 \text{Married} + \beta_4 \text{Schooling} + u$.
- D) $Earn = \beta_1 \text{Male} + \beta_2 \text{Female} + \beta_3 \text{Married} + \beta_4 \text{Single} + \beta_5 \text{Schooling} + u$.

9) Which of the following statements is **true** when two or more regressors in a multiple regression model are highly correlated?

- A) The OLS estimators of some slope coefficients are inconsistent.
- B) R^2 is close to zero.
- C) The sign of a slope coefficient estimate might be wrong.
- D) The OLS assumptions are violated.

10) You have collected data for the 50 U.S. states and estimated the following relationship between the change in the unemployment rate from the previous year ($\widehat{\Delta ur}$) and the growth rate of the respective state real GDP (g_{GDP}). The results are as follows

$$\widehat{\Delta ur} = \frac{2.81}{(0.12)} - \frac{0.23}{(0.04)} g_{GDP}, \quad n = 50, \quad R^2 = 0.36, \quad SER = 0.78.$$

Assuming that the estimator has a normal distribution, the 95% confidence interval for the slope is approximately the interval

- A) [2.57, 3.05]
- B) [-0.31, -0.15]
- C) [-0.31, 0.15]
- D) [-0.33, -0.13]

11) Given the information in the above question, we can also calculate the explained sum of squares (ESS) which is approximately given by

- A) 16.4268
- B) 29.2032
- C) 45.6300
- D) 17.1113

2. (8 points, 2 points each) Please first state whether you **agree or disagree** with each of the following claims. Then explain briefly why you agree or disagree with them.

(a) $0 \leq R^2 \leq 1$, when there is an intercept term in the linear regression model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$.

b) The OLS estimation is commonly used by applied researchers as the OLS estimators are always unbiased and consistent.

c) In the case of omitted relevant independent variables that may be correlated with other regressors, the confidence interval of a slope coefficient is generally valid.

d) For the simple linear regression (SLR) model: $Y_i = \beta_0 + \beta_1 X_i + u_i$, $i = 1, \dots, n$, the OLS estimator is consistent if (i) $Cov(X_i, u_i) = 0$ for $i = 1, 2, \dots, n$, (ii) (Y_i, X_i) , $i = 1, 2, \dots, n$, are IID, (iii) Large outliers are unlikely.

3. (10 points) Following Alfred Nobel's will, there are five Nobel Prizes awarded each year. These are for outstanding achievements in Chemistry, Physics, Physiology or Medicine, Literature, and Peace. In 1968, the Bank of Sweden added a prize in Economic Sciences in memory of Alfred Nobel. You think of the data as describing a population, rather than a sample from which you want to infer behavior of a larger population. The accompanying table lists the joint probability distribution between recipients in economics and the other five prizes, and the citizenship of the recipients, based on the 1969-2001 period.

Joint Distribution of Nobel Prize Winners in Economics and Non-Economics
Disciplines, and Citizenship, 1969-2001

	U.S. Citizen ($X = 0$)	Non-U.S. Citizen ($X = 1$)
Economics Nobel Prize ($Y = 0$)	0.118	0.049
Physics, Chemistry, Medicine, Literature, and Peace Nobel Prize ($Y = 1$)	0.345	0.488

Please answer the following questions.

- (a) (2 points) Compute and interpret $E(X)$.
- (b) (2 points) Calculate and interpret $E(X|Y = 1)$.
- (c) (2 points) Verify whether $E(X) = E[E(X|Y)]$ holds here.
- (d) (2 points) Calculate $\text{Var}(X|Y = 1)$.
- (e) (2 points) Are X and Y independent? Explain briefly.

4. **(22 points)** A researcher is using data for a sample of 274 male employees to investigate the relationship between hourly wage rates Y_i (measured in dollars per hour) and firm tenure X_i (measured in years). Preliminary analysis of the sample data produces the following sample information:

$$\begin{array}{llll} n & = & 274 & \sum_{i=1}^n Y_i = 1945.26 & \sum_{i=1}^n X_i = 1774.00 & \sum_{i=1}^n Y_i^2 = 18536.73 \\ \sum_{i=1}^n X_i^2 & = & 30608.00 & \sum_{i=1}^n X_i Y_i = 16040.72 & \sum_{i=1}^n \hat{u}_i^2 = 4105.297 \end{array}$$

Use the above sample information to answer all the following questions. Show explicitly all formulas and calculations.

- (a) (4 points) Compute OLS estimates of the intercept coefficient β_0 and the slope coefficient β_1 .
- (b) (4 points) Compute the value of R^2 , the coefficient of determination for the estimated OLS sample regression equation. Briefly explain what the calculated value of R^2 means.
- (c) (14 points) Test the significance for the slope coefficient at the 5% level using **three** approaches (2 points for each). State clearly the null hypothesis (2 points), the test statistics (2 points), and the conclusion (2 points). What are the assumptions you make (2 points)?

5. **(38 points)** A student considered the determination of house price in a city.

PART A: **(16 points)**

He has collected **88** observations on the following variables:

Y : price of the house measured in \$1000; X : size of the house measured in square feet. He runs the following linear regression model: $Y_i = \beta_0 + \beta_1 X_i + u_i$. The regression output is summarized in the left table as follows. Now the student would like to change the measurement units of Y from \$1000 to \$ (denoted as Y^*) and the measurement units of X from square feet to square meter (hint: 1 square foot = 0.0929 square meter) (denoted as X^*). We have $Y_i^* = \alpha_0 + \alpha_1 X_i^* + u_i^*$. The regression output is reported in the right table below.

Dependent variable: *price*.

Y(\$1000)				Y*(\$)			
Variable	Coeff	Std. error	t-stat	Variable	Coeff	Std. error	t-stat
c	39.96205	17.02263	2.347584	c	(A)		
$X(sqrft)$	0.137625	0.010945	12.57366	$X^*(sqrm)$	(B)	(C)	(D)
R^2	0.794070			R^2	(E)		
\bar{R}^2	0.789048			\bar{R}^2	(F)		
SER	36.39527			SER	(G)		
SSR	54309.25			SSR	(H)		

(2 points each) Assuming the errors are conditionally homoskedastic, fill in the blanks (in the places indicated by (A) - (H) in the above table). Demonstrate how you obtain each of your answers. **[Hint.** You don't need to draw the table. Just provide solutions and some intermediate steps.]

PART B: (22 points) The student collected data on more variables:

bdrms: number of bedrooms; *colonial*: =1 if the house is in colonial style, =0 otherwise

lotsize: size of lot in square feet. He has proposed two candidate models : Model 1 and Model 2. In Model 1, he thought that *sqrft* and *lotsize* may enter the model nonlinearly so that he included *sqrft*² and *lotsize*² as regressors, whereas in Model 2, he did not consider the nonlinear effect of *sqrft* and *lotsize* on the house price. The regression outputs are summarized in the following table.

Dependent variable: *price*.

	Model 1				Model 2			
Variable	Coeff	Std. error	t-stat	Prob.	Coeff	Std. error	t-stat	Prob.
<i>c</i>	63.953	74.425	0.859	0.393	-24.126	29.603	-0.815	0.417
<i>sqrft</i>	-0.0096	0.0620	(A)	0.877	0.124	0.013	9.314	0.000
<i>lotsize</i>	0.0117	0.0020	5.928	0.000	0.002	0.0006	3.230	0.002
<i>bdrms</i>	(B)	8.248	1.792	0.077	11.004	9.5153	1.156	0.250
<i>colonial</i>	26.379	13.491	1.955	0.054	13.715	14.637	0.937	0.351
<i>sqrft</i> ²	2.11E-05	1.29E-05	1.633	0.106				
<i>lotsize</i> ²	-1.11E-07	2.15E-8	-5.153	0.000				
<i>R</i> ²	0.770				(C)			
\bar{R}^2	0.753				0.660			
<i>SER</i>	(D)				59.877			
<i>SSR</i>	211135.7				297575.9			
<i>TSS</i>	(E)							

(a) (2 points) Interpret the coefficient estimate of *colonial* in Model 2.

(b) (10 points, 2 points each) Fill in the blanks (in the places indicated by (A), (B), (C), (D) and (E)) in the above table). Demonstrate how you obtain each of your answers. [**Hint.** You don't need to draw the table. Just provide solutions and some intermediate steps.]

(c) (10 points) Test whether the coefficient of *sqrft*² (denoted as β_{sqrft^2}) is positive at the 5% level using **two** approaches (2 points each). State clearly the null hypothesis (2 points), the test statistics (2 points), and the conclusion (2 points).