Final Exam

Instructions: This is a closed book exam. You may use a calculator. The use of other electronic devices, such as cellphones, laptops, or tablets, is prohibited. There are 40 points on this 75-minute exam. The possible points for each question are shown in parentheses following the question number. Answers should be written in English. Please write legibly, and do not forget to write your name and student id number on the answer sheet. Good luck!

Part I: Multiple choice questions. Select one or two choices that are correct unless stated otherwise. You do not have to explain your answer.

1. (2 points) Suppose that you estimate the following multiple linear regression model using an i.i.d. sample,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i,$$

where X_{1i} is the variable of interest. Omitted variable bias in $\hat{\beta}_1$

a. is always present as long as $R^2 < 1$.
b. is always present as long as $E(u_i \mid X_{1i}, X_{2i}, \ldots, X_{ki}) \neq 0$.
c. is always present but is negligible in almost all economic examples.
d. occurs if the omitted variable is correlated with X_{1i} but is not a determinant of Y_i.
e. None of the above

2. (2 points) The OLS residual $\hat{u}_i$ from the multiple regression model,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i,$$

a. cannot be calculated because there is more than one explanatory variable.
b. is an estimator of the standard deviation of the regression error.
c. is the difference between Y_i and its OLS predicted value.
d. has zero sample mean.
e. None of the above.

3. (2 points) Suppose that you estimate the following multiple linear regression model,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i.$$

A 95% confidence set for (β_1, β_2)

a. contains the true population values of the coefficients in 95% of randomly drawn samples.
b. is the union of two 95% confidence intervals for β_1 and β_2.
c. is an ellipse centered at the OLS estimates $(\hat{\beta}_1, \hat{\beta}_2)$.
d. cannot be computed because we do not observe the population values of (β_1, β_2).
e. None of the above.

4. (2 points) An economist determines which control variables to include in a multiple linear regression model
   a. by adding all variables that (s)he has collected data for.
   b. so that $E(u_i \mid X_i, W_i) = E(u_i \mid W_i)$ where X_i is (are) the variable(s) of interest and W_i is (are) the control variable(s).
   c. so that $\bar{R}^2$ gets maximized.
   d. by thinking through the application, referring to economic theory, and using judgment.
   e. None of the above.

5. (2 points) Suppose that you estimate the following multiple linear regression model,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,$$

where X_{1i} and X_{2i} are highly correlated. In this case,
   a. the linear regression model cannot be estimated by OLS.
   b. one of the four Least Squares assumptions[1] is violated.
   c. β_1 and/or β_2 will be imprecisely estimated.
   d. $\hat{\beta}_1$ and/or $\hat{\beta}_2$ will suffer from omitted variable bias.
   e. None of the above

6. (2 points) Consider the multiple linear regression model, $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$. Suppose that you want to test $H_0 : \beta_1 + \beta_2 = 2$ vs. $H_1 : \beta_1 + \beta_2 \neq 2$. You can transform the regression so that the restriction in the null hypothesis becomes a restriction on a single coefficient in an equivalent regression. Which of the following regression will do the job?
   a. Estimate $Y_i = \beta_0 + \beta_1 (X_{1i} - X_{2i}) + \gamma X_{2i} + u_i$ and test whether $\gamma = 2$.
   b. Estimate $Y_i = \beta_0 + \beta_1 (X_{1i} + X_{2i}) + \gamma X_{2i} + u_i$ and test whether $\gamma = 2$.
   c. Estimate $Y_i - 2X_{2i} = \beta_0 + \gamma X_{1i} + \beta_2 (X_{2i} - X_{1i}) + u_i$ and test whether $\gamma = 0$.
   d. Estimate $Y_i + 2X_{2i} = \beta_0 + \beta_1 (X_{1i} - X_{2i}) + \gamma X_{2i} + u_i$ and test whether $\gamma = 0$.
   e. Estimate $Y_i - 2X_{2i} = \beta_0 + \beta_1 (X_{1i} - X_{2i}) + \gamma X_{2i} + u_i$ and test whether $\gamma = 0$.

[1] The four least squares assumptions are 1) zero conditional mean for the error term, 2) all independent and dependent variables being i.i.d., 3) all independent and dependent variables having finite fourth moments, and 4) no perfect multicollinearity.

7. (2 points) Stationarity means that
   a. the error term in a linear regression model is serially correlated.
   b. the probability distribution of a time series variable does not change over time.
   c. a time series has a persistent upward trend.
   d. historical relationships might not be reliable guides to the future.
   e. None of the above

Part II: Longer questions. Show your work and explain your answers.

8. (7 points) Suppose Y_t follows the stationary AR(1) model $Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$, where $|\beta_1| < 1$ and u_t is i.i.d. with $E(u_t) = 0$ and $\text{var}(u_t) = \sigma_u^2$.

   (a) (1 points) Compute $E(u_t \mid Y_{t-1}, Y_{t-2}, \ldots)$.

   (b) (2 points) Compute the mean and variance of Y_t.

   (c) (3 points) Compute the first, second, and *p*-th autocorrelations of Y_t.

   (d) (1 points) Suppose $Y_T = 1$. Compute $Y_{T+1|T} = E(Y_{T+1} \mid Y_T, Y_{T-1}, \ldots)$.

9. (19 points) An economic study investigates whether college graduates earn more than those without a college degree. This study uses observational data from the August 2020 Economically Active Population Survey (EAPS) on South Korean households. The analysis is restricted to a random sample of 30–59 year old men who are full-time wage workers. Use the following R outputs to answer the questions.

   Variables in the dataset eaps:

earn	monthly earnings in 10,000 South Korean won (KRW)
colgrad	1 if a worker is a college graduate; 0 otherwise
age	age in years

   R outputs:

```
> library(sandwich)
> library(lmtest)
> library(car)
>
> eaps$age3039 <- as.numeric(eaps$age>=30 & eaps$age<40)
> eaps$age4049 <- as.numeric(eaps$age>=40 & eaps$age<50)
> eaps$age5059 <- as.numeric(eaps$age>=50 & eaps$age<=59)
>
> nrow(eaps)
[1] 8033
> nrow(subset(eaps, colgrad==1))
[1] 3800
```

```
> reg1 <- lm(earn ~ 1, data = eaps)
> summary(reg1)

Call:
lm(formula = earn ~ 1, data = eaps)

Residuals:
    Min      1Q  Median      3Q     Max 
-354.35 -119.35  -46.35   80.65 1930.65 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  369.349      1.952   189.2   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 174.9 on 8032 degrees of freedom

> 
> reg2 <- lm(earn ~ colgrad, data = eaps)
> coeftest(reg2, vcov. = sandwich)

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)    
(Intercept) 320.0428     1.9983 160.154 < 2.2e-16 ***
colgrad     104.2304     3.8148  27.323 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(reg2)$r.sq
[1] 0.08850989

> 
> reg3 <- lm(earn ~ colgrad + age4049 + age5059, data = eaps)
> coeftest(reg3, vcov. = sandwich)

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)    
(Intercept) 262.1387     2.9739  88.147 < 2.2e-16 ***
colgrad     110.9976     3.7588  29.530 < 2.2e-16 ***
age4049      66.0433     3.9421  16.753 < 2.2e-16 ***
age5059      92.1513     4.6557  19.793 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(reg3)$r.sq
[1] 0.1342542

> 
> lht(reg3, "age4049=age5059", vcov = sandwich)
Linear hypothesis test

Hypothesis:
age4049 - age5059 = 0

Model 1: restricted model
Model 2: earn ~ colgrad + age4049 + age5059
```

```
Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)    
1   8030                        
2   8029  1 30.257 3.901e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> lht(reg3, c("age4049=0","age5059=0"), vcov. = sandwich)
Linear hypothesis test

Hypothesis:
age4049 = 0
age5059 = 0

Model 1: restricted model
Model 2: earn ~ colgrad + age4049 + age5059

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)    
1   8031                        
2   8029  2 242.34 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> lht(reg3, c("colgrad=0","age4049=0","age5059=0"), vcov. = sandwich)
Linear hypothesis test

Hypothesis:
colgrad = 0
age4049 = 0
age5059 = 0

Model 1: restricted model
Model 2: earn ~ colgrad + age4049 + age5059

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)    
1   8032                        
2   8029  3 393.19 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(a) (6 points) Fill out the following table.

	Sample mean ($\overline{earn}$)	Sample standard deviation (s_{earn})	Number of observations (n)
All	(0.5 points)	(1 point)	(0.5 points)
Those without a college degree ($colgrad = 0$)	(0.5 points)	(1 point)	(0.5 points)
College graduates ($colgrad = 1$)	(0.5 points)	(1 point)	(0.5 points)

(b) (4 points) Based on the estimation results of the linear regression model,

$$earn_i = \beta_0 + \beta_1 colgrad_i + \beta_2 age4049_i + \beta_3 age5059_i + u_i ,$$

report and interpret $\hat{\beta}_1$. Is β_1 larger than 1,000,000 KRW at the 1% significance level? Explain.

(c) (2 points) Compute $\overline{R}^2$ of the regression model in (b). Report to four decimal places.

(d) (2 points) Explain why *age3039* is not included as a regressor in the regression model in (b).

(e) (2 points) Test $H_0 : ESS = 0$ vs. $H_1 : ESS \neq 0$ based on the regression model in (b). (*ESS* stands for the explained sum of squares.)

(f) (3 points) Do the regression results suggest that there is a college premium in the South Korean labor market? Explain why or why not. (Think about whether the Conditional Mean Independence assumption is likely to hold. Try to provide a specific example.)