Final Exam

(June 16, 2023, **2 hours**)

Econometrics (Spring 2023)

Professor: Wonmun Shin

Part I

True or False

(Total 40 points) Read each statement below carefully. Write a **T** if you think the statement is True. Write an **F** if you think the statement is false. Do not forget to state a reason (or concise explanation) for your choice.

1. When there exists a correlation between a regressor and regression errors, the OLS estimator is biased and inefficient because the classical assumptions are not satisfied. Nevertheless, the OLS estimator is still consistent when the sample is sufficiently large.

2. We can compare the coefficients of determination (R^2) between two models, $Y_i = \beta_1 + \beta_2 X_i + e_i$ and $\ln Y_i = \beta_1 + \beta_2 \ln X_i + e_i$ even if variables take different forms each other.

3. As the number of independent variables increases in a regression model, $\bar{R}^2$ (adjusted R^2) increases less than R^2.

4. IV (instrumental variable) regression means that we substitute X_{2i} (one explanatory variable causing an endogeneity issue) with Z_i (an instrument). That is, we estimate $Y_i = \beta_1 + \beta_2 Z_i + \beta_3 X_{3i} + e_i$ instead of estimating $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + e_i$ to obtain IV estimators.

5. Sample regression function (or fitted line) $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ obtained by least square estimation must pass through $(\bar{X}, \bar{Y})$.

6. If there is near multicollinearity in a regression model, then the F-statistic of overall significance test tends to be small though individual coefficients are significant.

7. Durbin-Watson test (DW test) can only be applied to detect the first-order autoregressive error process.

8. If error terms e_i are not normally distributed, the OLS estimators are not BLUE.

Part II

Short Questions

1. **(10 points)** In the simple linear regression $Y_i = \beta_1 + \beta_2 X_i + e_i$, we can obtain the OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$. *Student A* decides to use the demeaned values for X_i, considering X_is have large numbers. That is, *Student A* runs the following regression:

$$Y_i = \alpha_1 + \alpha_2 \left(X_i - \bar{X}\right) + u_i$$

where $\bar{X}$ is a mean of X_is. Obtain OLS estimators $\hat{\alpha}_1$ and $\hat{\alpha}_2$, and discuss whether they are different from $\hat{\beta}_1$ and $\hat{\beta}_2$ or not.

2. **(10 points)** Suppose that *Student B*, using wage data on 160 randomly selected male workers and 140 female workers, estimates the OLS regression:

$$\widehat{Wage}_i = \underset{(0.5)}{22} - \underset{(0.6)}{3} \cdot Female_i, \qquad R^2 = 0.02$$

where $Wage_i$ is measured in dollars per hour and $Female_i$ is a dummy variable that is equal to 1 if the worker is a female and 0 if the worker is a male. Define the ***gender wage gap*** as the difference in mean earnings between men and women. (i) Is there significant gender wage gap based on the above estimation result? (*If you need, please refer to the probability table on the last page.*) At the same time, *Student C* uses the same data but regresses $Wage_i$ on $Male_i$, a dummy variable that is equal to 1 if the worker is a male and 0 if the worker is a female. What are the regression results of *Student C*? (ii) Please fill the blanks below:

$$\widehat{Wage}_i = \underset{(0.49)}{[\quad]} + \underset{(0.61)}{[\quad]} \cdot Male_i, \qquad R^2 = [\quad]$$

3. **(10 points)** We relax the classical assumptions by allowing for different variances for different observations:

$$Var\left(e_i\right) = \sigma_i^2 \qquad \text{for } i = 1, 2, \cdots, n$$

When we faces the above heteroskedasticity, the usual formula for the variance of OLS estimator in $Y_i = \beta_1 + \beta_2 X_i + e_i$ is incorrect. As a result, the usual least squares standard error ($\sqrt{\widehat{Var}\left(\hat{\beta}_2\right)} = \sqrt{\frac{\hat{\sigma^2}}{\sum x_i^2}}$) is inconsistent, which might lead to wrong test results. Considering the problems caused by heteroskedasticity, *Student D* obtains **White estimator** (of standard error). Explain what is **White estimator** and discuss pros and cons of **White correction**. *(Tip: You don't have to write down the formula of White estimator.)*

4. **(10 points)** Consider a regression model passing through the origin, $Y_i = \beta X_i + e_i$. Let us assume that the regression errors are independent and identically distributed. Specifically, they follow normal distribution, *i.e.* $e_i \sim N\left(0, \sigma^2\right)$. Obtain the OLS estimator of the slope coefficient ($\hat{\beta}$), and discuss the efficiency of $\hat{\beta}$.

Part III

Long Questions

1. **(Total 15 points)** Suppose you want to analyze the effect of the size of houses on house prices, and you run a simple regression $Y_i = \beta_1 + \beta_2 X_i + e_i$ where Y_i is house price (measured in \$100) and X_i is house size (measured in m^2). The estimated model is followed:

$$\hat{Y}_i = 450 + 25X_i, \quad R^2 = 0.64$$

(a) **(5 points)** Compute the sample correlation coefficient between house price and house size.

(b) **(5 points)** Imagine that you want to change the measurement unit of house price from \$100 to \$10, holding the unit for house size m^2. How do your estimates ($\hat{\beta}_1$ and $\hat{\beta}_2$) change? Is R^2 in the new regression greater than before, or lower than before?

(c) **(5 points)** Imagine that you want to change the measurement unit of house size from m^2 to ft^2 (square feet), holding the unit for house price \$100. Note that 1 square feet is equal to 0.1 square meter. How do your estimates ($\hat{\beta}_1$ and $\hat{\beta}_2$) change? If the slope coefficient in the original regression (using m^2) were significant at 5%, then would the slope coefficient in the new regression (using ft^2) be still significant at 5%?

2. **(Total 20 points)** *Student E* has **monthly data** on demand for labor (L), aggregate output in current prices (Y), average wages at current prices (W), and consumer price index (P), for the manufacturing sector of a certain country for the period between **January 1993** and **April 2023**. *Student E* fits the following regression (standard errors in parentheses):

$$\widehat{\ln L} = \underset{(0.13)}{-3.15} + \underset{(0.09)}{0.42} \ln Y - \underset{(0.10)}{0.34} \ln W - \underset{(0.06)}{0.11} \ln P \qquad RSS = 10.0 \tag{1}$$

Next, *Student E* runs a regression $\ln L$ on real output (Y/P) and real wages (W/P):

$$\widehat{\ln L} = \underset{(0.15)}{-2.54} + \underset{(0.08)}{0.46} \ln\left(\frac{Y}{P}\right) - \underset{(0.07)}{0.30} \ln\left(\frac{W}{P}\right) \qquad RSS = 10.1 \tag{2}$$

Finally, *Student E* fits the following regression:

$$\widehat{\ln L} = \underset{(0.14)}{-2.55} + \underset{(0.09)}{0.45} \ln\left(\frac{Y}{P}\right) - \underset{(0.08)}{0.32} \ln\left(\frac{W}{P}\right) + \underset{(0.05)}{0.09} \ln\left(\frac{1}{P}\right) \qquad RSS = 9.9 \tag{3}$$

(a) **(5 points)** Give an economic interpretation of the slope coefficients in equation (2).

(b) **(5 points)** The regression model of equation (2) (*i.e.* $\ln L = \beta_1 + \beta_2 \ln(Y/P) + \beta_3 \ln(W/P) + e$) is a restricted version of the regression model of equation (1) (*i.e.* $\ln L = \beta_1 + \beta_2 \ln Y + \beta_3 \ln W + \beta_4 \ln P + e$). Specify the restriction.

(c) **(5 points)** Test the restriction that you obtained in (b), using <u>***F*-test**</u> (at 5% significance level). Please state the null and alternative hypotheses, the test statistic, the critical value, and your decision.

(d) **(5 points)** Test the restriction that you obtained in (b), using <u>**t-test**</u> (at 5% significance level). In order to do that, you need to use the estimation result in equation (3). Note that the regression model of equation (3) is $\ln L = \beta_1 + \beta_2 \ln (Y/P) + \beta_3 \ln (W/P) + \beta_4 \ln (1/P) + e$. Do not forget to state the null and alternative hypotheses, the test statistic and the critical value. Do you have the same conclusion in (c)?

3. **(Total 30 points)** A simple regression model is given by

$$Y_t = \beta_1 + \beta_2 X_t + e_t \qquad \text{for } t = 1, 2, \cdots, n \tag{4}$$

Note that the subscript t represents time unit (eg. year or month) and X_t are deterministic. The regression errors e_t follow AR(1) model:

$$e_t = 0.5 e_{t-1} + \nu_t \qquad \text{for } t = 1, 2, \cdots, n \tag{5}$$

where ν_t are uncorrelated random variables with constant variance, that is:

$$E(\nu_t) = 0, \qquad E(\nu_t^2) = 1, \qquad \text{and } E(\nu_t \nu_s) = 0 \ \text{ for } t \neq s$$

(a) **(2 points)** Calculate the value of $E(e_t)$ and $E(e_{t-1})$. Are they constant over time?

(b) **(2 points)** Calculate the value of $Var(e_t)$ and $Var(e_{t-1})$. Are they constant over time (*i.e.* homoskedastic)?

(c) **(3 points)** Calculate the covariance between e_{t-1} and ν_t. Do we have a concern of endogeneity issue in the given AR(1) model?

(d) **(3 points)** *Student F* argues that "*e_t is correlated with e_{t-1} as shown in equation (5), and therefore e_t is uncorrelated with e_{t-2} (the error at time $t-2$)*". Is this statement true? Explain.

(e) **(5 points)** Are the OLS estimators $\hat{\beta}_1$, $\hat{\beta}_2$ of the regression (4) the BLUE (best linear unbiased estimator)? Explain.

(f) **(5 points)** If you answered "No" in (e), state the GLS transformation of equation (4) to obtain the BLUE (here, you should delete some initial observations, but you don't have to take care of it). If you answered "Yes" in (e), then write down "Not Necessary".

(g) **(5 points)** *Student G* obtained the BLUE estimates, $\hat{\beta}_1^{BLUE}$ and $\hat{\beta}_2^{BLUE}$ based on 100 observations. The following numerical information is given to *Student G*:

$$Y_{100} = 70, \ \ X_{100} = 30, \ \ X_{101} = 40$$

$$\hat{\beta}_1^{BLUE} = -2, \ \ \hat{\beta}_2^{BLUE} = 2$$

Student G would like to predict Y_{101} (the value at time 101). Note that, when error terms follow AR(1) model, the prediction of Y_{101} would be:

$$\hat{Y}_{101} = \hat{\beta}_1^{BLUE} + \hat{\beta}_2^{BLUE} X_{101} + \rho \hat{e}_{100}$$

where ρ is the AR coefficient in equation (5). Calculate the predicted value $\hat{Y}_{101}$.

(h) **(5 points)** Now assume that the value of AR coefficient in equation (5) is 0 (zero), not 0.5. Based on the same 100 observations in (g), *Student H* estimated β_1 and β_2 and the estimates ($\hat{\beta}_1$ and $\hat{\beta}_2$) happened to be the same as in (g). What is the *Student H*'s predicted value $\hat{Y}_{101}$? Is it different from the *Student G*'s predicted value in (g)? Why do you think *Student H* get the same (or different) predicted value?

4. **(Total 15 points)** Consider the following model:

$$GDP_t = \beta_1 + \beta_2 M_t + \beta_3 M_{t-1} + \beta_4 \triangle M_t + e_t$$

where GDP_t is GDP (gross domestic product) at time t, M_t is money supply at time t, M_{t-1} is money supply at time $t-1$, and $\triangle M_t$ is a change in money supply between time t and $t-1$ (*i.e.* $\triangle M_t = M_t - M_{t-1}$). Money supply is exogenously given, and the regression error satisfies the conditions followed: $E(e_t)$ is zero, $Var(e_t)$ is homoskedastic, and there exists no autocorrelation in errors.

(a) **(5 points)** Give a proper interpretation of the slope coefficients β_2, β_3, and β_4. Do they economically make sense (*i.e.* is the above model reasonable from a economic perspective)?

(b) **(5 points)** Assuming you have the data to estimate the above model, would you succeed in estimating all the slope coefficients of the model? Why or why not?

(b) **(5 points)** Suppose that $\beta_3 M_{t-1}$ term is absent from the above model. Would your answer to (b) be the same?

- **Probability table for Final Exam**:
  - $P(F(1,+200) \geq 3.84) = 0.05$, $P(F(1,+200) \geq 6.63) = 0.01$ for F distribution with degrees of freedom $(1,+200)$. (+200 indicates for more than 200 degrees of freedom.)
  - $P(F(2,+200) \geq 3.00) = 0.05$, $P(F(2,+200) \geq 4.61) = 0.01$ for F distribution with degrees of freedom $(2,+200)$. (+200 indicates for more than 200 degrees of freedom.)
  - $P(F(3,+200) \geq 2.60) = 0.05$, $P(F(3,+200) \geq 3.78) = 0.01$ for F distribution with degrees of freedom $(3,+200)$. (+200 indicates for more than 200 degrees of freedom.)
  - $P(Z \geq 1.28) = 0.1$, $P(Z \geq 1.645) = 0.05$, $P(Z \geq 1.96) = 0.025$, $P(Z \geq 2.58) = 0.005$ where Z is a standard normal random variable.

(End of Exam, Total 5 Pages)