The International College of Economics and Finance
Econometrics – 2019-2020.
Midterm exam. 2019 October 24.
Part 2. Free Response Questions
(1 hour 50 minutes)

<u>General instructions.</u> Candidates should answer FOUR of the following FIVE questions: all 3 questions of the Section A and any 1 of the questions from Section B (questions 4-5). The weight of the Section A is 45% of the exam; one question chosen from the Section B adds 30%. You are advised to divide your time accordingly. Structure your answers in accordance with the structure of the questions. When testing hypotheses state clearly null and alternative hypotheses, provide critical value(s) used for the test, mentioning degrees of freedom, the significance level chosen for the test and the assumptions for the test to be valid.

SECTION A

Answer **ALL** questions from this section (questions **1-3**).
Each question in this section bears **15 marks**

Question 1.

A simple linear regression is considered, $\hat{Y} = b_1 + b_2 X$ estimated by OLS with determination coefficient R^2, where the variable X is considered to be non-stochastic (Model A assumptions).

(a) (4 points) □ Show that determination coefficient of the regression is always equal to the square of the sample correlation coefficient between X and Y: $R^2 = r_{X,Y}^2$.

Now assume that two regressors X_2 and X_3 are nonstochastic and $\text{Cov}(X_2, X_3) = 0$. A student estimates multiple regression $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$ obtaining

$$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3 \quad \text{with determination coefficient } R_1^2 \quad \textbf{(1)}$$

and two simple regressions

$$\hat{Y} = c_1 + c_2 X_2 \quad \text{with determination coefficient } R_2^2 \quad \textbf{(2)}$$

and

$$\hat{Y} = d_1 + d_3 X_3 \quad \text{with determination coefficient } R_3^2 \quad \textbf{(3)}$$

She noticed that

$$R_1^2 = R_2^2 + R_3^2$$

Show that it is always true on the assumptions under consideration.
The structure of the following questions will help you to choose the right way to carry out the necessary proof.

(b) (4 points) □ Using (a) get an expression for $R_2^2 + R_3^2$ in terms of sample variances and covariances 'Cov' and 'Var'.

□ Consider multiple regression (1). Using general formulas for estimating coefficients of a multiple regression by OLS obtain their expressions in simplified form taking into account the assumptions above.

(c) (7 points) □ Starting from definition of determination coefficient derive an expression for R_1^2 in terms of sample variances and covariances using the results obtained in (a) and (b), and compare it with expression for $R_2^2 + R_3^2$ obtained in (b).

□ Why can a proven property be considered as a generalization of a property $R^2 = r_{X,Y}^2$ for the simple linear regression $\hat{Y} = a_1 + a_2 X_2$?

Hint: for multiple regression model $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$ OLS estimators of β_2 and β_3 are

$$b_2 = \frac{\text{Cov}(X_2, Y)\text{Var}(X_3) - \text{Cov}(X_3, Y)\text{Cov}(X_2, X_3)}{\text{Var}(X_2)\text{Var}(X_3) - \left[\text{Cov}(X_2, X_3)\right]^2}$$ and

$$b_3 = \frac{\text{Cov}(X_3, Y)\text{Var}(X_2) - \text{Cov}(X_2, Y)\text{Cov}(X_2, X_3)}{\text{Var}(X_2)\text{Var}(X_3) - \left[\text{Cov}(X_2, X_3)\right]^2}$$

Question 2.

An employee of a real estate agency in a Russian city with a developed subway network is interested in estimating of the influence of the distance from the city center $CENTER_i$ (in kilometers) on the price of an two-room apartment in millions of rubles. Based on the data of 21 apartments sold during a period under consideration she runs a regression.

$$\underset{}{\hat{PRICE}_i} = \underset{(0.88)}{12.39} - \underset{(0.10)}{0.20} \cdot CENTER_i \qquad R^2 = 0.17 \quad RSS = 103.4 \qquad \textbf{(1)}$$

(a) (5 points) □ Is the regression slope coefficient significant (take into account that the realtor did not know exactly the sign of its coefficient before the regression calculation)?

□ Are the results of the estimation compatible with the hypothesis that true regression coefficient is positive?

□ Are the results of the estimation compatible with the hypothesis that true regression coefficient is 0.1?

□ How the conclusion on significance of the slope would change if the manager could use the assumption that the influence of the $CENTER_i$ on the apartment price is not positive?

The realtor, not satisfied with the obtained result, decided to take into account the additional factor – the distance to the nearest subway station $METRO_i$ (also in kilometers).

$$\hat{PRICE}_i = \underset{(0.97)}{13.71} - \underset{(0.09)}{0.22} \cdot CENTER_i - \underset{(0.25)}{0.58} \cdot METRO_i \qquad R^2 = 0.37 \quad RSS = 79.29 \qquad \textbf{(2)}$$

During the discussion at the workshop, the realtor received advice from a colleague to use Ramsey's test for this equation. Since the realtor was not experienced enough in econometrics, a colleague helped her calculate appropriate equation (using in the right side of (3) estimated values $\cdot \hat{PRICE}_i^*$ from equation (2):

$$\hat{PRICE}_i = \underset{(6.04)}{0.023} + \underset{(0.18)}{0.13} \cdot CENTER_i + \underset{(0.47)}{0.35} \cdot METRO_i + \underset{(0.033)}{0.07} \cdot (\hat{PRICE}_i^*)^2 \qquad R^2 = 0.51 \quad RSS = 60.64 \qquad \textbf{(3)}$$

Then the colleague helped her to estimate a new equation

$$\log \hat{PRICE}_i = \underset{(0.10)}{2.62} - \underset{(0.0095)}{0.019} \cdot CENTER_i - \underset{(0.026)}{0.059} \cdot METRO_i \qquad R^2 = 0.32 \quad RSS = 0.8448 \qquad \textbf{(4)}$$

and did Ramsey's test again (using in the right side of (5) estimated values $\cdot \log \hat{PRICE}_i^{**}$ from equation (4):

$$\log \hat{PRICE}_i = \underset{(1.53)}{0.62} + \underset{(0.039)}{0.030} \cdot CENTER_i + \underset{(0.11)}{0.084} \cdot METRO_i + \underset{(0.0088)}{0.012} \cdot (\log \hat{PRICE}_i^{**})^2 \quad R^2 = 0.39 \quad RSS = 0.7672 \qquad \textbf{(5)}$$

(b) (5 points) □ Help the realtor to understand the logic of her colleague in estimating these equations.

□ Explain what the Ramsey test is, what is the null hypothesis and what statistics it uses; use them to perform the necessary calculations.

□ What conclusions can be drawn from the results in this part of the study?

At the end of the study, the colleague estimated one more equation, to choose between linear and logarithmic functions

$$\hat{PRICE_i / GMEAN(PRICE)} = \underset{(0.092)}{1.29} - \underset{(0.0087)}{0.021} \cdot CENTER_i - \underset{(0.023)}{0.055} \cdot METRO_i \qquad R^2 = 0.37 \quad RSS = 0.5390 \qquad \textbf{(6)}$$

where $GMEAN(PRICE)$ **-** geometric mean of the price values in the sample.

Note that in the logarithmic regression, the Zarembka transformation is not obligatory for the comparability of equations, since it does not change the dependent variable due to the properties of logarithms.

(c) (5 points) □ Run the necessary test on the basis of the last equation.

□ Taking into account all the estimated multiple regression equations above and the previous conclusions, determine which form of dependence (linear or semi-logarithmic nonlinear) is the best for expressing the influence of the factors under consideration on the price of the apartment.

Question 3.

The student runs two production function models for the same data for some developing country: $t = 1,2,\ldots 30$, where y_t is per capita income, x_t is capital, and z_t is labor (all variables are index numbers)

$$\ln y_t = \alpha + \beta \ln x_t + 0.5 \ln z_t + v_{1t} \qquad \textbf{(1)}$$

$$\ln y_t = \alpha + \beta \ln x_t + \beta \ln z_t + v_{2t} \qquad \textbf{(2)}$$

given that x_t and z_t are deterministic sequences and $v_{1t} \sim iid(0,\sigma^2)$, $v_{2t} \sim iid(0,\sigma^2)$.

(a) (5 points) □ Explain how to find the least squares estimates of β. What are estimators of β for both equations (write out the explicit formulas using sample covariance and sample variance notation 'Cov' and 'Var')?

□ What are properties of these estimators assuming equations (1) and (2) to be in turn valid models?

(b) (5 points) Both regressions are the restricted versions of the general model

$$\ln y_t = \alpha + \beta \ln x_t + \gamma \ln z_t + u_t \qquad \textbf{(3)}.$$

□ What are the restrictions?

□ How these restrictions could be tested **using F-test**?

Describe the full test procedure, the test itself is not intended to be performed.

(c) (5 points) □ Is any of equations (1) or (2) characterized by the constant returns to scale restriction? If neither indicate appropriate equation corresponding to the restriction of the constant returns to scale.

□ Give interpretation to the restricted and unrestricted production functions.

□ How constant returns to scale restriction could be tested **using t-tests**?

Describe the full test procedure, the test itself is not intended to be performed.

SECTION B.

Answer **ONE** question from this section (**4 OR 5**).
Each question in this section bears **30 marks**

Question 4.

For the simple regression model $Y = \beta_1 + \beta_2 X + u$ estimated using OLS $Y_i = b_1 + b_2 X_i + e_i$ it is well known that the following properties of residuals e_i are true:

1) $\bar{e} = 0$;
2) $\sum X_i e_i = 0$;
3) $\sum \hat{Y}_i e_i = 0$;
4) $\bar{\hat{Y}} = \bar{Y}$.

Now consider multiple linear regression

$$Y = \beta_1 + \beta_2 X_2 + ... + \beta_k X_k + u \qquad \textbf{(1)}$$

Determine which of properties 1) - 4) remain true for the multiple regression in the original formulation (without any change), which should be formulated slightly differently (make the appropriate corrections in formulas and text if needed) and which cease to be true. Prove all the properties of the residuals for the multiple regression model (1) you specify. *Answer the questions in (a)-(f) in turn.*

(a) (5 points) □ Discuss whether property $\bar{e} = 0$ should be modified or not for the case of multiple regression model (1). Give justification for your wording and formulas (prove them).

(b) (5 points) □ Discuss whether property $\sum X_i e_i = 0$ should be modified or not for the case of multiple regression model (1). Give justification for your wording and formulas (prove them).

(c) (5 points) □ Discuss whether property $\sum \hat{Y}_i e_i = 0$ should be modified or not for the case of multiple regression model (1). Give justification for your wording and formulas (prove them).

(d) (5 points) □ Discuss whether property $\bar{\hat{Y}} = \bar{Y}$ should be modified or not for the case of multiple regression model (1). Give justification for your wording and formulas (prove them).

(e) (5 points) □ Explain mathematically from (d) the geometric meaning of the equality $\bar{\hat{Y}} = \bar{Y}$: multiple regression hyperplane $\hat{Y}_i = b_1 + b_2 X_{2i} + b_3 X_{3i} + ... + b_k X_{ki}$ passes through the middle point of the sample $(\bar{X}_2, ... \bar{X}_k, \bar{Y})$.

(f) (5 points) □ Show that for OLS estimation $Y_i = b_1 + b_2 X_{2i} + ... + b_k X_{ki} + e_i$ of the multiple linear regression $Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i$ is always true $TSS = ESS + RSS$.

□ Derive from here another form of this property $\text{Var}(Y_i) = \text{Var}(\hat{Y}_i) + \text{Var}(e_i)$

□ Show that $R^2 = 1 - \dfrac{RSS}{TSS}$.

Question 5.

A wine specialist is studying the factors that influence the price of Angevin Rose wine, produced from Grolleau grapes in the French Anjou region of the Loire Valley. She suggests that the price may be influenced by the ageing of the wine AGE_i (in years).

The researcher collected data on the average price P_i of a bottle of premium Angevin Rose wine (in euros) from 27 appellations (wine-making areas) in the Anjou region depending on its age and calculated two regressions

$$\begin{array}{ll} P_i = \underset{(11.30)}{87.83} + \underset{(0.76)}{0.82}AGE_i \quad R^2 = 0.045 & \qquad \textbf{(1)} \end{array}$$

and

$$\begin{array}{ll} P_i = \underset{(14.06)}{65.27} + \underset{(2.60)}{6.76}AGE_i - \underset{(0.10)}{0.23}AGE_i^2 \quad R^2 = 0.227 & \qquad \textbf{(2)} \end{array}$$

(a) (5 points) □ Comment on equations (1) and (2), explain the economic meaning of their coefficients.

□ Evaluate the significance of the coefficients of (1) and (2) and equations as a whole.

(b) (5 points) □ What are the advantages of the quadratic function (equation (2)) over the linear function (equation (1)) in explaining the price of Angevin wine? Give reasons for your answer.

□ How the price of a bottle of premium Angevin rose wine changes with additional aging of wine for one year for wines of 4 and 16 years of age?

The researcher also assumes that the quality of wine, and consequently its price, may be influenced by the average annual soil temperature $TEMP_i$ (in degrees Celsius) and the distance of the vineyard from the riverbed of the Loire $LOIRE_i$ (in kilometers) in the following way,

$$\begin{array}{ll} \hat{P}_i = \underset{(15.02)}{67.44} + \underset{(2.55)}{9.46}AGE_i - \underset{(0.12)}{0.39}AGE_i^2 + \underset{(0.047)}{0.0084}TEMP_i - \underset{(15.05)}{30.94}\log(LOIRE_i) \quad R^2 = 0.45 & \qquad \textbf{(3)} \end{array}$$

(c) (5 points) □ Interpret equation coefficients (3) for variables $TEMP_i$ and $\log(LOIRE)_i$.

□ Give a brief justification that your interpretation of the logarithmic coefficient is correct.

□ Why does the researcher use logarithms when including a variable distance from the Loire bed?

(d) (5 points) □ Are the temperature and distance variables significant individually and taken together (use additional assumptions if necessary)?

(e) (5 points) □ Explain theoretically whether it is possible to estimate the significance of the age of wine in equation (3) using the F-test, if not, which equation(s) should be additionally calculated.

□ How to perform the necessary test?

(f) (5 points) □ Sum up the research. Compare equations (1)-(2)-(3) in terms of their statistical quality. Which equation do you consider to be the worst? Which equation do you consider to be the best for further analysis. Argue each of your statements relating the ranking of equations.