The International College of Economics and Finance
Econometrics. Mid-year exam. 2018 October 25. Part 2. (120 minutes).
SECTION A. Answer **ALL** questions **1-3** from this section.

1. [15 marks] You are asked to study the factors that determine employee monthly earnings E_i (in thousands of rubles) of a big corporation (60 observations in total), such as education ED_i, professional experience EXP_i (both in years), age AGE_i (also in years), and AGE_i^2 (in years squared):

$$E_i = \beta_1 + \beta_2 ED_i + \beta_3 EXP_i + \beta_4 AGE_i + \beta_5 AGE_i^2 + u_i; \quad i = 1, 2, ..., N. \qquad \textbf{(1)}$$

Let estimated equation be

$$\hat{E}_i = \underset{(800.86)}{-632.01} + \underset{(0.75)}{1.94} ED_i + \underset{(0.86)}{2.17} EXP_i + \underset{(55.95)}{42.36} AGE_i - \underset{(0.97)}{0.74} AGE_i^2 \quad R^2 = 0.18 \qquad \textbf{(1*)}$$

Another estimated equations is

$$\hat{E}_i = \underset{(14.69)}{-27.54} + \underset{(0.69)}{2.10} ED_i + \underset{(0.76)}{2.39} EXP_i \quad R^2 = 0.17 \qquad \textbf{(2*)}$$

(a) [3 marks] Outline briefly how to test whether education matters in determination of the monthly earnings using equation **(1*)**. How would you test whether experience influence positively on the earnings? Is it possible to test whether the earnings of all employees are constant independently of their education level, professional experience and age?

(b) [6 marks] Explain the meaning of coefficients $\hat{\beta}_2$ and $\hat{\beta}_3$ in **(1*)**. The coefficient $\hat{\beta}_3$ is a little bigger than $\hat{\beta}_2$. Can you conclude from here that spending a year acquiring the professional experience is more useful for future earnings in the corporation than a year of study? Why or why not? How can you test the hypothesis $\beta_3 = \beta_2$ against $\beta_3 \neq \beta_2$ using F-test? How can you test the hypothesis $\beta_3 = \beta_2$ against $\beta_3 > \beta_2$ using appropriate test?

(c) [6 marks] Explain the reason for inclusion variable of AGE_i^2 into the model? (*You may use a graphic illustration in addition to the text*). What is marginal effect of age in the model for the employee of 45 years old? Test whether age influence earnings and comment obtained results. Test whether this influence is linear (relationship between earnings and age is a linear function)?

2. [15 marks] A researcher is trying to build the best model for the total expenditure on private education $PRIV_t$ in USA (in billions of dollars) for 1993-2017 as a function of disposable personal income DPI_t (also in billions of dollars), relative price index for private education $PRELPRIV_t$ and $TIME_t$ (equal to 1 in 1993, 2 – in 1994 and so on).

(a) [5 marks] First a linear and log-linear models were estimated:

$$\hat{PRIV}_t = \underset{(11.81)}{29.72} + \underset{(0.0086)}{0.018} \cdot DPI_t - \underset{(0.14)}{0.38} \cdot PRELPRIV_t + \underset{(0.25)}{0.32} \cdot TIME_t \quad R^2 = 0.97, \; RSS = 5.54 \qquad \textbf{(1)}$$

$$\widehat{\log(PRIV_t)} = \underset{(1.66)}{6.92} + \underset{(0.0012)}{0.0021} \cdot DPI_t - \underset{(0.020)}{0.071} \cdot PRELPRIV_t + \underset{(0.036)}{0.067} \cdot TIME_t \quad R^2 = 0.94, \; RSS = 0.11 \qquad \textbf{(2)}$$

Give interpretation to the models and their coefficients (no justification for your interpretation is required). Discuss the significance of coefficient of DPI_t in equation **(2)** using 2-tail and 1-tail tests when needed. What assumptions should be taken to use of 1-tail test? Is it possible to choose between models **(1)** and **(2)** on the base of available information? Why or why not?

(b) [5 marks] The researcher also evaluates variable $PRIVZ_t$ dividing values of $PRIV_t$ by their geometric mean for the whole period $PRIVZ_t = PRIV_t / \sqrt[n]{PRIV_1 \cdot ... \cdot PRIV_n}$ and runs new regression

$$\hat{PRIVZ}_t = \underset{(1.18)}{2.97} + \underset{(0.0008)}{0.0018} \cdot DPI_t - \underset{(0.014)}{0.038} \cdot PRELPRIV_t + \underset{(0.025)}{0.032} \cdot TIME_t \quad R^2 = 0.97, \; RSS = 0.055 \qquad \textbf{(1*)}$$

Use Box-Cox test to choose between linear and logarithmic specifications of the model. Specify the formula for test statistic, its distribution and corresponding critical values. What is your conclusion? Which model can be selected as the best for further economic and econometric analysis? Explain your choice.

(c) [5 marks] The colleague of the researcher mentioned that equation **(1*)** cannot be compared with log-linear regression **(2)** using Box-Cox test as Zarembka transformation was applied only to linear regression, while equation **(2)** uses untransformed dependent variable, and recommended to apply Zarembka transformation also to **(2)** and use regression **(2*)** instead of **(2)**:

$$\widehat{\log(PRIVZ_t)} = \underset{(1.66)}{4.62} + \underset{(0.0012)}{0.0021}\cdot DPI_t - \underset{(0.020)}{0.071}\cdot PRELPRIV_t + \underset{(0.036)}{0.067}\cdot TIME_t \qquad R^2 = 0.94,\ RSS = 0.11 \qquad \textbf{(2*)}$$

Comment on her advice.

She also recommended to run double logarithmic regression **(3)**

$$\widehat{\log(PRIV_t)} = \underset{(5.19)}{-2.38} + \underset{(0.22)}{2.30}\cdot\log(DPI_t) - \underset{(1.13)}{2.24}\cdot\log(PRELPRIV_t) - \underset{(0.02)}{0.02}\cdot TIME_t \qquad R^2 = 0.98,\ RSS = 0.028 \qquad \textbf{(3)}$$

instead of **(2)** and compare it with regression **(1*)** using Box-Cox test. Comment on this advice, do recommended test and make your conclusion. Give interpretation to the coefficients of regression **(3).**

3. [15 marks] September 2018, elections for the mayor of Moscow were held. The winner was Sergei Sobyanin, his main competitor was Vadim Kumin (KPRF party). A student studying the factors influencing the victory of candidates got exit poll data from one of the polling stations. Each polled voter indicated the candidate for whom he voted (in this polling station all votes were distributed between Sobyanin and Kumin), and also reported how many campaign programs in favor of each of these two candidates (from 0 to 3) he had seen on television. His working hypothesis was that the decision of voters was affected by the number of views of campaign TV programs in favor of each of the candidates. As a result, the student has collected 16 observations: $(S=0, K=0, VS_1)$, $(S=1, K=0, VS_2)$, …, $(S=3, K=3, VS_{16})$, (where S is the number of views for Sobyanin, K is the number of views for Kumin, VS_i is the percentage of votes cast in each group for Sobyanin. For example in the group of voters with 3 TV views in favor of Sobyanin and with 1 TV view in favor of Kumin 96% of voters voted for Sobyanin. He also calculated the total number of TV views for each group $T_i = S_i + K_i$ and calculated sample covariance between S and K: $\mathrm{Cov}(S, K) = 0$.

Then he runs several regressions

$$\widehat{VS}_i = \underset{(5.74)}{67.35} + \underset{(3.06)}{6.85}S_i \qquad R^2 = 0.26 \qquad \textbf{(1)}$$

$$\widehat{VS}_i = \underset{(5.60)}{79.05} + \underset{(2.34)}{6.85}S_i - \underset{(2.34)}{7.80}K_i \qquad R^2 = 0.60 \qquad \textbf{(2)}$$

$$\widehat{VS}_i = \underset{(5.6)}{79.05} + \underset{(3.30)}{14.65}S_i - \underset{(2.34)}{7.80}T_i \qquad R^2 = 0.60 \qquad \textbf{(3)}$$

(a) [5 marks] Give interpretation to the coefficients of the variable S_i in equations **(1-3)**. Test the significance of the coefficients of these regressions and the significance of the equations as a whole. (*Do not compare coefficients or other characteristics of equations 1-3, you will be asked to do this later in* ***b*** *and* ***c***)

(b) [6 marks] Why the intercepts in equations **(2)** and **(3)** are identical while in **(1)** it is different? Explain why the coefficient of the variable T_i in equations **(3)** is identical to the coefficient of the variable K_i in equation **(2)**? Explain why the coefficient of the variable S_i in equations **(3)** is bigger than in equation **(2)**? Explain why the coefficients of the variable S_i in equations **(1)** and **(2)** are identical? (*You may use the formulas for estimators of regression coefficients or any other method for your explanations*).

(c) [4 marks] Explain why determination coefficients R^2 in equations **(2)** and **(3)** are identical, while in equation **(1)** R^2 is about three times less. Explain why in equation **(2)** the standard errors of both slope coefficients are identical. Suggest probable reason why the standard error of the coefficient for the variable S_i in equation **(2)** is less than corresponding standard error in equation **(1)**.

Bonus question: what drawbacks can you indicate in the student's study?

SECTION B. Answer **ONE** question from this section (**4 OR 5**).

4. [30 marks] At a lecture on econometrics the professor said to the students that OLS estimator for the coefficient α in a simple regression model without constant $Y_t = \alpha X_t + u_t$, $t = 1, 2, \ldots, T$ is given by

$\hat{\alpha}_{OLS} = \frac{\sum_{t=1}^{T} X_t Y_t}{\sum_{t=1}^{T} X_t^2}$. But this estimator is not the only one, there is another estimator of α, $\hat{\alpha}_1 = \frac{\overline{Y}}{\overline{X}} = \frac{\frac{1}{T}\sum Y_t}{\frac{1}{T}\sum X_t}$.

(a) [8 marks] Show that $\hat{\alpha}_1 = \frac{\overline{Y}}{\overline{X}}$ is also an unbiased estimator of α in $Y_t = \alpha X_t + u_t$. What assumptions are needed for this result being valid? Is $\overline{Y}/\overline{X}$ also a good estimator for β_2 in $Y_t = \beta_1 + \beta_2 X_t + u_t$? Explain.

(b) [7 marks] Derive population variance of the estimator $\hat{\alpha}_1 = \frac{\overline{Y}}{\overline{X}}$. What assumptions are used to derive your result?

(c) [8 marks] Giving a lecture an absent-minded professor by mistake wrote on a board the formula $\hat{\alpha} = \overline{Y_t / X_t}$ instead of $\hat{\alpha} = \overline{Y} / \overline{X}$ ($\overline{Y_t / X_t} = \frac{1}{T}\sum_{t=1}^{T}(Y_t / X_t)$). Does it change the conclusion that the estimator is unbiased? Show that population variance of $\hat{\alpha}_2 = \overline{Y_t / X_t}$ is given by the expression $\text{var}(\hat{\alpha}_2) = \frac{\sigma^2}{T^2}\sum_{t=1}^{T}(1/X_t^2)$. State the assumptions needed for your work.

(d) [7 marks] Which of the two estimators $\hat{\alpha}_1 = \overline{Y} / \overline{X}$ and $\hat{\alpha}_2 = \overline{Y_t / X_t}$ is more efficient? Explain what you understand by efficiency.
Hint: use inequality for the four means (quadratic, arithmetic, geometric, harmonic:

$$\sqrt{\frac{1}{T}\sum_{t=1}^{T} x_t^2} \geq \frac{\sum_{t=1}^{T} x_t}{T} \geq \sqrt[T]{x_1 \cdot x_2 \cdot \ldots \cdot x_T} \geq \frac{T}{(\sum_{t=1}^{T} \frac{1}{x_t})}$$

Question 5. [30 marks] A researcher has data on income *Y*, capital *K* and labor *L* indices for 1991-2017 related to the economy of some developing country (time series specific problems are out of consideration in this question). She estimated relationship between these variables using different production functions.

(a) [8 marks] First she runs a linear model

$$\hat{Y}_t = 10.83 + 0.22 \cdot K_t + 0.72 \cdot L_t \qquad R^2 = 0.95$$
$$(14.0)\ (0.025) \quad (0.12) \quad RSS = 3096.4 \qquad \textbf{(1)}$$

Noticing that the sum of the slope coefficients is close to unity she decided to test the restriction $\beta_K + \beta_L = 1$, and decided to run the restricted regression

$$Y = \alpha + \theta \cdot K + (1 - \theta) \cdot L + u, \qquad \textbf{(2)}$$

How equation **(2)** could be estimated using OLS (you may use results of this estimation $R^2 = 0.85$ and $RSS = 3137.9$)? How to test the restriction? Is it possible to run F-test, based on comparison of determination coefficients of restricted and unrestricted regressions? Why or why not? What is economic meaning of this restriction (if any)? Would you recommend to accept the restriction?

(b) [7 marks] Then the researcher runs auxiliary regression **(1a)**,

$$\hat{Y}_t = -51.36 + 0.48 \cdot K_t + 1.30 \cdot L_t - 0.003 \cdot (\hat{Y}_t^*)^2 \qquad R^2 = 0.96$$
$$(26.6)\ (0.10) \quad (0.24) \quad (0.001) \quad RSS = 2375 \qquad \textbf{(1a)}$$

where $(\hat{Y}_t^*)^2$ are squared estimated values of Y_t from the equation **(1)**: $\hat{Y}_t^* = 10.83 + 0.22 \cdot K_t + 0.72 \cdot L_t$

Comment on the meaning of researcher's actions, run the appropriate test and draw the conclusion. What implicit economic assumptions are used in the specification of equation **(1)**? To what extent can they be considered relevant to economic theory and practice? How to improve the specification of the equation **(1)**?

(c) [8 marks] Next she runs two logarithmic regressions

$$\log(Y) = 0.48 + 0.37 \cdot \log(K) + 0.53 \cdot \log(L) \qquad R^2 = 0.97$$
$$(0.39)\ (0.05) \quad (0.12) \quad RSS = 0.0763 \qquad \textbf{(3)}$$

and

$$\hat{\log(Y/L)} = -0.0050 + 0.33 \cdot \log(K/L) \qquad R^2 = 0.82$$
$$(0.0183\ \ (0.03) \quad RSS = 0.0811 \qquad \textbf{(4)},$$

Give the interpretation of both models and their parameters. Prove that equation **(4)** is a restricted version of the equation **(3)**. What is the restriction? Is it significant? What equation would you choose and why?

(d) [7 marks] Now the researcher introduces time (*T* equal to 1 in the first year of the period, 2 – in the second and so on), and estimates Cobb-Douglas (CD) production function using two different methods:

OLS applied to linearized model:

$$\hat{\log(Y)} = 2.53 - 0.067 \cdot \log(K) + 0.51 \cdot \log(L) + 0.03 \cdot T \qquad R^2 = 0.97$$
$$(1.17)\ \ (0.24) \quad (0.11) \quad (0.016) \quad RSS = 0.0664 \qquad \textbf{(5-OLS)}$$

NLS (Non-Linear Least Squares) applied to original model (all estimates except capital elasticity are significant) :

$$\hat{Y} = e^{3.8} \cdot K^{-0.046} \cdot L^{0.086} \cdot e^{0.008T} \qquad R^2 = 0.97 \qquad \textbf{(5-NLS)}$$

Comment on parameters of both equations, paying special attention to the coefficients of variables K and *T*. Why estimates of equations **(5-OLS)** and **(5-NLS)** are different?

GENERAL INSTRUCTIONS: Start answering each question 1-5 on the blank with corresponding number (ask for extra paper if necessary). Structure your answers in accordance with the structure of the questions. Testing hypotheses always state clearly null and alternative hypotheses provide critical value used for test, mentioning degrees of freedom and the significance level chosen for the test.