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

General instructions. 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: $R^2 = r_{X,Y}^2$.

$$R^2 = \frac{\text{Var}(\hat{Y})}{\text{Var}(Y)} = \frac{\text{Var}(b_1 + b_2 X)}{\text{Var}(Y)} = b_2^2 \frac{\text{Var}(X)}{\text{Var}(Y)} = \left(\frac{\text{Cov}(X,Y)}{\text{Var}(X)}\right)^2 \frac{\text{Var}(X)}{\text{Var}(Y)} = \frac{\text{Cov}^2(X,Y)}{\text{Var}(X)\text{Var}(Y)} = r_{X,Y}^2$$

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

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

and two simple regressions

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

and

$\hat{Y} = d_1 + d_3 X_3$ with determination coefficient R_3^2 **(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'.

$$R_2^2 + R_3^2 = r_{Y,X_2}^2 + r_{Y,X3}^2 = \frac{\text{Cov}^2(X_2,Y)}{\text{Var}(Y)\text{Var}(X_2)} + \frac{\text{Cov}^2(X_3,Y)}{\text{Var}(Y)\text{Var}(X_3)} \qquad \textbf{(4)}$$

□ 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.

As $\text{Cov}(X_2, X_3) = 0$ the formulas from 'hint' below collapse to $b_2 = \frac{\text{Cov}(X_2,Y)}{\text{Var}(X_2)}$, $b_3 = \frac{\text{Cov}(X_3,Y)}{\text{Var}(X_3)}$.

(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).

$$R_1^2 = \frac{\mathrm{Var}(\hat{Y})}{\mathrm{Var}(Y)} = \frac{\mathrm{Var}(b_1 + b_2X_2 + b_3X_3)}{\mathrm{Var}(Y)} = b_2^2\frac{\mathrm{Var}(X_2)}{\mathrm{Var}(Y)} + b_3^2\frac{\mathrm{Var}(X_3)}{\mathrm{Var}(Y)} + 2b_2b_3\frac{\mathrm{Cov}(X_2,X_3)}{\mathrm{Var}(Y)} =$$

$$= \left(\frac{\mathrm{Cov}(X_2,Y)}{\mathrm{Var}(X_2)}\right)^2\frac{\mathrm{Var}(X_2)}{\mathrm{Var}(Y)} + \left(\frac{\mathrm{Cov}(X_3,Y)}{\mathrm{Var}(X_3)}\right)^2\frac{\mathrm{Var}(X_3)}{\mathrm{Var}(Y)} = \frac{\mathrm{Cov}^2(X_2,Y)}{\mathrm{Var}(X_2)\mathrm{Var}(Y)} + \frac{\mathrm{Cov}^2(X_3,Y)}{\mathrm{Var}(X_3)\mathrm{Var}(Y)},$$ what coincides with expression 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 $\hat{Y} = a_1 + a_2X_2$?

From (a) this is equivalent to $R^2 = r_{Y,X_2}^2 + r_{Y,X_3}^2$.

Hint: for multiple regression model $Y = \beta_1 + \beta_2X_2 + \beta_3X_3 + u$ OLS estimators of β_2 and β_3 are

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

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

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.

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

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

Let $PRICE_i = \beta_1 + \beta_2 CENTER_i + u_i$. $t = \frac{-0.20}{0.10} = -2.00$, $|-2.00| < t_{crit}^{5\%}(19) = 2.11$, cannot reject $H_0: \beta_2 = 0$ under $H_a: \beta_2 \neq 0$.

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

CI: $(-0.20 - 2.11 \cdot 0.10; -0.20 + 2.11 \cdot 0.10)$ or $(-0.411; 0.011)$ is a set of all null-hypotheses compatible with the data, here some positive null hypotheses are not rejected.

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

$|t| = \left|\frac{-0.20 - 0.1}{0.10}\right| = |-3.00| = 3 > t_{crit}^{1\%}(19) = 2.861$, reject $H_0: \beta_2 = 0.1$ under $H_a: \beta_2 \neq 0.1$.

□ **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?**

$t = \frac{-0.20}{0.10} = -2.00$ **is the same, but** $t_{crit}^{5\%}(1\text{-side}, 19) = 1.729$**, so the slope coefficient is significant.**

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).

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

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 $\hat{PRICE}_i^*$ from equation (2):

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

Then the colleague helped her to estimate a new equation

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

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

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

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

A colleague tries to improve the quality of the multiple regression equation by including nonlinearity there. To com proposed variants the Ramsey test is applied in turn to the linear and semi-logarithmic equations.

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

Ramsey test: run the regression: $Y_j = \beta_1 + \sum_{j=2}^{k} \beta_j X_j + u$ ⇒ save the fitted variables of Y : $\hat{Y}_j = b_1 + \sum_{j=2}^{k} b_j X_j$ ⇒ add $\hat{Y}_j^2$ to the original specification: $Y_j = \beta_1 + \sum_{j=2}^{k} \beta_j X_j + \beta_{\hat{Y}_j^2} \hat{Y}_j^2 + u$ ⇒ t-test is applied for: $H_0 : \beta_{\hat{Y}_j^2} = 0$; $H_a : \beta_{\hat{Y}_j^2} \neq 0$. If reject some kind of nonlinearity may be present.

Logic of the study: applying the Ramsey test to equation (2) and calculating equation (3), $t = \frac{0.07}{0.033} = 2.121$, $t_{crit}^{5\%}(21-3) = t_{crit}^{5\%}(18) = 2.101$ ⇒ significant; applying the Ramsey test to equation (4) and calculating equation (5): $t = \frac{0.012}{0.088} = 1.36$, the same t_{crit} ⇒ insingnificant ⇒ no further improvement needed.

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

The use of non-linear regression is preferable.

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

$$\underset{(0.092)\ (0.0087)}{\hat{PRICE}_i / GMEAN(PRICE) = 1.29 - 0.021 \cdot CENTER_i} \underset{(0.023)}{- 0.055 \cdot METRO_i} \qquad \underset{RSS = 0.6990}{R^2 = 0.37} \qquad \textbf{(6)}$$

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.

From (4) and (6) using Box-Cox test $\chi^2 = \frac{n}{2}\ln\frac{RSS(\text{bigger})}{RSS(\text{smaller})} = \frac{21}{2}\ln\frac{0.8448}{0.6990} = 1.99 < \chi^2_{crit}(1, 5\%) = 3.84$

$\Rightarrow$ insingnificant.

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

Box-Cox test does not allow to choose between equations (2) or (4), but Ramsey RESET test rejects the linear form of dependence in favor of the nonlinear form. So finally one should recommend equation (4).

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')?

Both regressions are simple linear regression models, so $\hat{\beta} = \frac{\text{Cov}(Y, X)}{\text{Var}(X)}$.

Preliminary transformations:

For (1) $\ln y_t - 0.5 \ln z_t = \alpha + \beta \ln x_t + u_t$

For (2) $\ln y_t = \alpha + \beta(\ln x_t + \ln z_t) + v_t$.

So for (1): $\hat{\beta} = \frac{\text{Cov}(\ln y_t - 0.5 \ln z_t, \ln x_t)}{\text{Var}(\ln x_t)} = \frac{\text{Cov}(\ln y_t, \ln x_t) - 0.5\,\text{Cov}(\ln z_t, \ln x_t)}{\text{Var}(\ln x_t)}$

for (2): $\hat{\beta} = \frac{\text{Cov}(\ln y_t, \ln x_t + \ln z_t)}{\text{Var}(\ln x_t + \ln z_t)} = \frac{\text{Cov}(\ln y_t, \ln x_t) + \text{Cov}(\ln y_t, \ln z_t)}{\text{Var}(\ln x_t) + \text{Var}(\ln z_t) + 2\,\text{Cov}(\ln x_t, \ln z_t)}$.

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

Both estimators of β are BLUE under assumption that models (1) and (2) are correspondingly true 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?

Resrictions are $\gamma = 0.5$ for (1), and $\gamma = \beta$ for (2).

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

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

To test $\gamma=\beta$ we use $F=\frac{(RSS_2 - RSS_3)/1}{RSS_3/(30-3)}$ where RSS_2 and RSS_3 are the sums of squared residuals for the models (2) and (3) (or rather $F=\frac{(R_3^2 - R_2^2)/1}{(1-R_3^2)/(30-3)}$). The same for $\gamma=0.5$ (use RSS_1 and RSS_3)

(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.

Neither of restrictions in **(b)** $\gamma=0.5$ or $\gamma=\beta$ is constant returns to scale restriction, this one is $\beta+\gamma=1$ where the corresponding model is

$$\ln y_t=\alpha+\beta\ln x_t+(1-\beta)\ln z_t+w_t \quad (4)$$

□ Give interpretation to the restricted and unrestricted production functions.

Equivalent form for (4) is $y_t=e^{\alpha}x_t^{\beta}z_t^{1-\beta}e^{w_t}$ or $\frac{y_t}{z_t}=e^{\alpha}\left(\frac{x_t}{z_t}\right)^{\beta}e^{w_t}$ (relation between capital-income ratio $\frac{x_t}{z_t}$ and labor productivity $\frac{y_t}{z_t}$, β is elasticity of income with respect to capital. The parameters β and γ of unrestricted production function $\ln y_t=\alpha+\beta\ln x_t+\gamma\ln z_t+u_t$ are elasticities of income with respect to capital and labor correspondingly.

□ 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*.

To use *t*-test for testing restriction $\gamma+\beta=1$ one should transform the model (one of the possible approaches):

$$\ln y_t=\alpha+\beta\ln x_t+\gamma\ln z_t+u_t$$
$$\ln y_t=\alpha+\beta\ln x_t+(\gamma\ln x_t-\gamma\ln x_t)+(\ln x_t-\ln x_t)+\gamma\ln z_t+u_t$$
$$\ln y_t-\ln x_t=\alpha+(\beta\ln x_t+\gamma\ln x_t-\ln x_t)+(\gamma\ln z_t-\gamma\ln x_t)+u_t$$
$$\ln y_t-\ln x_t=\alpha+(\beta+\gamma-1)\ln x_t+\gamma(\ln z_t-\ln x_t)+u_t$$

Now we use t-test to test coefficient of the variable $\ln x_t$ for significance.

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

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

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).

(a) and (b) follow from OLS:

Theoretical regression $Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u$, OLS estimation: $\hat{Y}_i = b_1 + b_2 X_{2i} + b_3 X_{3i} + \ldots + b_k X_{ki}$

Residuals $e_i = Y_i - \hat{Y}_i = Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i} - \ldots - b_k X_{ki}$

OLS principle: $S = \sum e_i^2 \to \min$, $S = \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - b_1 - b_2 X_{2i} - \ldots - b_k X_{ki})^2 \to \min$

FOC:

$$\frac{\partial S}{\partial b_1} = -\sum 2(Y_i - b_1 - b_2 X_{2i} - \ldots - b_k X_{ki}) = 0 \Rightarrow \bar{e} = 0$$

$$\frac{\partial S}{\partial b_2} = -\sum 2(Y_i - b_1 - b_2 X_{2i} - \ldots - b_k X_{ki}) X_{2i} = 0 \Rightarrow \sum X_{2i} e_i = 0$$

.....

$$\frac{\partial S}{\partial b_k} = -\sum 2(Y_i - b_1 - b_2 X_{2i} - \ldots - b_k X_{ki}) X_{ki} = 0 \Rightarrow \sum X_{ki} e_i = 0.$$

(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).

Small changes are required in the proof compared to the case of simple regression model:

$$\sum \hat{Y}_i e_i = \sum (b_1 + b_2 X_{2i} + \ldots + b_k X_{ki}) e_i = b_1 \sum e_i + b_2 \sum X_{2i} e_i + \ldots + b_k \sum X_{ki} e_i = 0 + 0 + \ldots + 0 = 0$$

(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).

$$\bar{\hat{Y}} = \frac{\sum \hat{Y}_i}{n} = \frac{\sum (Y_i + e_i)}{n} = \frac{\sum Y_i}{n} + \frac{\sum e_i}{n} = \bar{Y} + 0 = \bar{Y}$$

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

Let $f(X) = b_1 + b_2 X_{2i} + b_3 X_{3i} + \ldots + b_k X_{ki}$ be regression hyperplane in k-dimensional linear space. Substitute middle values $(\overline{X}_2, \ldots \overline{X}_k)$ into this linear function

$$f(X) = b_1 + b_2 \overline{X}_{2i} + b_3 \overline{X}_{3i} + \ldots + b_k \overline{X}_{ki} = \frac{n b_1}{n} + b_2 \overline{X}_2 + \ldots + b_k \overline{X}_2 = \frac{\sum b_1}{n} + b_2 \frac{\sum X_{2i}}{n} + \ldots + b_k \frac{\sum X_{ki}}{n}$$

$$= \frac{\sum (b_1 + b_2 X_{2i} + \ldots + b_k X_{ki})}{n} = \frac{\sum \hat{Y}_i}{n} = \overline{\hat{Y}} = \overline{Y}$$

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

By definition $TSS = \sum (Y_i - \overline{Y})^2$; $ESS = \sum (\hat{Y}_i - \overline{Y})^2$; $RSS = \sum e_i^2$

$$\sum (Y_i - \overline{Y})^2 = \sum ([\hat{Y}_i + e_i] - \overline{Y})^2 = \sum ([\hat{Y}_i - \overline{Y}] + e_i)^2 = \sum (\hat{Y}_i - \overline{Y})^2 + \sum e_i^2 + 2\sum ([\hat{Y}_i - \overline{Y}] e_i) =$$
$$= \sum (\hat{Y}_i - \overline{Y})^2 + \sum e_i^2 + 2\sum \hat{Y}_i e_i - 2\overline{Y} \sum e_i = \sum (\hat{Y}_i - \overline{Y})^2 + \sum e_i^2$$

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

$$\sum (Y_i - \overline{Y})^2 = \sum (\hat{Y}_i - \overline{Y})^2 + \sum e_i^2 \Leftrightarrow \frac{1}{n}\sum (Y_i - \overline{Y})^2 = \frac{1}{n}\sum (\hat{Y}_i - \overline{Y})^2 + \frac{1}{n}\sum e_i^2 \Leftrightarrow$$
$$\mathrm{Var}(Y_i) = \mathrm{Var}(\hat{Y}_i) + \mathrm{Var}(e_i)$$

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

By definition $R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$

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

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

and

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

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

For equation (1) the intercept 87.83 (euros) shows expected price of a bottle of young, unripe wine. The coefficient at variable AGE_i of 0.82 indicates that each year the wine aged adds on average 82 eurocents to

the price of the wine (marginal effect of ageing). For equation (2): the marginal effect $\frac{dP_i}{dAGE_i} = 6.76 - 2 \cdot 0.23 AGE_i$ (decreasing marginal effect of ageing).

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

Let $P_i = \beta_1 + \beta_2 AGE_i + u_i$. $t = \frac{0.82}{0.76} = 1.08 < t_{crit}^{5\%}(25) = 2.06 \Rightarrow$ cannot reject $H_0 : \beta_2 = 0$.

All coefficients of equation (2) are significant at 5% ($t_{const.} = 4.62, t_{AGE} = 2.6, t_{AGE^2} = 2.3$ while $t_{crit}^{5\%}(24) = 2.064, t_{crit}^{1\%}(24) = 2.797$). The group of variables AGE and AGE^2 responsible for ageing of wine is significant at 5% level: $F = \frac{R^2/2}{(1-R^2)/24} = \frac{0.227}{1-0.227} * 12 = 3.52$ while $F_{crit}^{5\%}(2, 24) = 3.40$.

(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.

In equation (2) $P_i = \beta_1 + \beta_2 AGE_i + \beta_3 AGE_i^2 + u_i$ the presence of a square term allows to take into account the nonlinear nature of the dependence of the wine quality on its age, with the increase in the aging of wine its price increases due to the positive coefficient at the linear term but the square term has an increasing negative impact.

□ 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?

From equation (2) $\frac{dP_i}{dAGE_i} = 6.76 - 2 \cdot 0.23 AGE_i$. So $\left.\frac{dP_i}{dAGE_i}\right|_{AGE=4} = 6.76 - 2 \cdot 0.23 \cdot 4 = 4.92$ so the wine aged 4 years adds to its value 4.92 euro per year. $\left.\frac{dP_i}{dAGE_i}\right|_{AGE=16} = 6.76 - 2 \cdot 0.23 \cdot 16 = -0.6$, so the wine aged 16 years loses 60 cents in its value per year.

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,

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

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

With a rise in soil temperature by one degree Celsius, the wine adds an average of 0.84 euro cents to the price. The increase in the distance from the Loire by one percent of the existing distance leads to a decrease in the price of wine by $30.94/100 \approx 0.31$ euros.

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

Let the relationship under consideration be $Y = a + b \cdot \log X$. Then $dY = d(a + b \cdot \log X) = b \cdot \frac{dX}{X}$.

From here $b = \frac{dY}{\frac{dX}{X}}$ or $\frac{b}{100} = \frac{dY}{\frac{dX}{X} * 100(\%)}$. Setting $\frac{dX}{X} \cdot 100\% = 1\%$ we obtain $\frac{b}{100} = dY$. So if we divide *b* by 100, the resulting number will show the increase of *Y*, provided *X* increases by one percent

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

As the distance from the Loire bed increases, the marginal effect of this factor decreases: $\frac{\partial Y}{\partial X} = \frac{\partial(a + b \cdot \log X)}{\partial X} = \frac{b}{X}$.

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

To test $H_0 : \beta_{TEMP} = 0$ $t = \frac{0.0084}{0.047} = 0.18$, obviously H_0 is not rejected. To test $H_0 : \beta_{\log(LOIRE)} = 0$ $t = \frac{-30.94}{15.05} = -2.056$, while $t_{crit}^{5\%}(22) = 2.074$, so we cannot reject H_0. If we assume that the coefficient in front of the variable $\log(LOIRE_i)$ cannot be positive, then $t_{crit}^{5\%}(1-side, 22) = 1.717$, and now the coefficient would be significant.

To test $H_0 : \beta_{TEMP} = 0, \beta_{\log(LOIRE)} = 0$: $F = \frac{(0.45 - 0.23)/2}{(1 - 0.45)/(27 - 5)} = 4.4$, while $F_{crit}^{5\%}(2, 22) = 3.44$ and $F_{crit}^{1\%}(2, 22) = 5.72$, so H_0 is rejected only at 5% level.

(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.

It is impossible to estimate the significance of both age variables in equation (3) based only on the information in this equation: if AGE increases its square AGE^2 also increases. To perform the necessary test, it is necessary to estimate additionally the restricted equation of the type

$$\hat{P}_i = \gamma_1 + \gamma_2 TEMP_i + \gamma_3 \log(LOIRE_i) + u_i$$

which does not contain both variable connected with the ageing of wine and to get R_R^2 for it .

□ How to perform the necessary test?

Then one need to perform F-test comparing it with $R_U^2 = 0.45$ for equation (3):

$$F = \frac{(R_U^2 - R_R^2)/2}{(1 - R_U^2)/22} \sim F(2, 22)$$

(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 and selection of equations.

Equation (1) is obviously the worst of the equations: its coefficient is insignificant. Adding a square of age of wine makes all variables (age and its square as well as both factors together) significant. The third equation includes also two additional variables what are significant taken together. So it is considered the most suitable for further analysis.