Econometrics – 2020-2021
Midterm Exam, April 1, 2021.
(One session 2 hours without break)
Solutions

SECTION A
Question 1 OR 2 and obligatory Question 3.

Question 1. (11 marks) The student is interested in studying how the relative prices of wine influence its consumption in France. She's considering two alternative regression models - simple linear regression (1) and model in differences (2) - using data on wine consumption W_t (in billions of euro) subject to the relative price index of wine, P_t on the base of annual data (1995-2018), variable $t=1$ in 1995, $t=2$ in 1996 etc.

$$W_t = \alpha_1 + \alpha_2 P_t + u_t \qquad (1)$$

$$\Delta W_t = \gamma_1 + \gamma_2 \Delta P_t + v_t \qquad (2)$$

where $\Delta W_t = W_t - W_{t-1}$, $\Delta P_t = P_t - P_{t-1}$, and u_t, and v_t are disturbance terms.
Estimating these equations she obtained

$$\hat{W}_t = 15.6 - 0.03 P_t \qquad R^2 = 0.03$$
$$(3.86)\ (0.04) \quad DW = 0.11 \quad RSS = 22.4 \qquad (1e)$$

$$\Delta\hat{W}_t = 0.13 - 0.06 \Delta P_t \qquad R^2 = 0.29$$
$$(0.06)\ (0.02) \quad DW = 3.02 \quad RSS = 1.95 \qquad (2e)$$

(a) (11 marks) The supervisor advised the student to check the time series for stationarity and said that probably her estimated regression (1e) is spurious regression.

□ What is stationarity? Why it is important? Which model in level does difference model (2) correspond?
Solution: Definitions of stationarity, nonstationarity. Consequences.
Let $W_t = \alpha_0 + \alpha_1 t + \alpha_2 P_t + u_t$ then $W_{t-1} = \alpha_0 + \alpha_1 (t-1) + \alpha_2 P_{t-1} + u_{t-1}$ and so $\Delta W_t = \alpha_1 + \alpha_2 \Delta P_t + u_t - u_{t-1}$ which corresponds to $\Delta W_t = \gamma_1 + \gamma_2 \Delta P_t + v_t$ with $v_t = u_t - u_{t-1}$ of MA(1) type.
[5 mark]

□ The student argued that the equation (1e) is insignificant, so the regression cannot be spurious. Do you agree with her? What is spurious regression? How to tell whether some estimated regression based on time series is spurious?
Solution: The arguments of the student are not correct. The advise is of value.
[2 mark]

□ Explain how spurious regression can occur if the condition of stationary of time series is violated?
Solution: Let $Y_t = \pi_0 + \pi_1 X_t + v_t$, and $H_0 : \pi_1 = 0$ is correct, then $Y_t = \pi_0 + v_t$ and so $v_t \in I(1)$ since Y_t is $Y_t \in I(1)$ with corresponding consequences.
[4 mark]

(b) (11 marks) The researcher decided to test the time series for stationarity by constructing the following Dickey-Fuller equations

$$\Delta\hat{W}_t = 5.48 - 0.50 W_{t-1} + 0.07 t \qquad R^2 = 0.25$$
$$(2.17)\ (0.20) \quad (0.03) \qquad (3e)$$

$$\Delta\hat{P}_t = 5.48 - 0.05 P_{t-1} + 0.46 \Delta P_{t-1} - 0.08 t \qquad R^2 = 0.43$$
$$(8.09)\ (0.09) \quad (0.25) \quad (0.08) \qquad (5e)$$

$$\Delta^2\hat{W}_t = -1.15 \Delta W_{t-1} \qquad R^2 = 0.58$$
$$(0.23) \qquad (4e)$$

$$\Delta^2\hat{P}_t = -2.16 \Delta P_{t-1} \qquad R^2 = 0.19$$
$$(0.18) \qquad (6e)$$

□ Explore Dickey-Fuller equation models mathematically. How the model corresponding (3e) can be derived from the original AR specification? State null and alternative hypotheses and decision rule; explain the logic of Dickey-Fuller test based on ADF t-statistic (*not doing the test itself*).

Solution: Let $W_t = \beta_1 + \beta_2 W_{t-1} + \gamma t + u_t$ then $\Delta W_t = \beta_1 + (\beta_2 - 1)W_{t-1} + \gamma t + u_t$ with $H_0 : \beta_2 - 1 = 0$, $H_A : \beta_2 - 1 < 0$. ADF t table should be used. Conventional t-test for γ.

[5 marks]

□ What is the difference between the specification of equation (3e) and the specification of equation (5e)? How null and alternative hypotheses change? Does each of these specifications have any advantages (and possible disadvantages)?

Solution: $P_t = \beta_1 + \beta_2 P_{t-1} + \beta_3 P_{t-2} + \gamma t + u_t$ to control for AR(1), $\Delta P_t = \beta_1 + (\beta_2 + \beta_3 - 1)P_{t-1} - \beta_3 \Delta P_{t-2} + \gamma t + u_t$ with $H_0 : \beta_2 + \beta_3 - 1 = 0$, $H_a : \beta_2 + \beta_3 - 1 < 0$. Once again use ADF tables. This reduces df.

[6 marks]

(c) (11 marks) □ Interpret the results of the estimation of Dickey-Fuller equations using ADF t-statistic. What are your conclusions: which of the time series are stationary and which are non-stationary? Explain.

Solution: For (3e) $t_{W_{t-1}} = -2.5 > -3.62 = t^{5\%}_{ADF,crit}$(23, trend in model) - W_t is nonstationary, also significant trend ($t_{@trend} = 2.333 > 2.09 = t^{5\%}_{crit}(20)$). For (4e) $t_{\Delta W_{t-1}} = -5 < -3.77 = t^{1\%}_{ADF,crit}$(22, no trend) - stationary.
For (5e) $t_{\Delta P_t} = -0.56 > -3.63 = t^{5\%}_{ADF,crit}$(22, trend in model)– nonstationary, no trend $1.0 < 2.101 = t^{5\%}_{crit}(18)$.
For (6e) $t_{\Delta W_{t-1}} = -12 < -3.77 = t^{1\%}_{ADF,crit}$(22, no trend) – stationary

[11 marks]

Question 2. The student is interested in studying how the relative prices of wine influence its consumption in France. She's considering several alternative regression models using data on wine consumption W_t (in billions of euro) subject to the relative price index of wine P_t using annual data (1995-2018).
She starts from an ADL(1,1) model (1), and considers as its alternatives two simplified models: simple linear regression (2) and model in differences (3).

$$W_t = \beta_1 + \beta_2 P_t + \beta_3 P_{t-1} + \beta_4 W_{t-1} + u_t \quad (1)$$

$$W_t = \alpha_1 + \alpha_2 P_t + v_t \quad (2)$$

$$\Delta W_t = \gamma_1 + \gamma_2 \Delta P_t + w_t \quad (3)$$

where $\Delta W_t = W_t - W_{t-1}$, $\Delta P_t = P_t - P_{t-1}$, and u_t, v_t, and w_t are disturbance terms.
Estimating these equations she obtained

$$\hat{W}_t = 0.59 - 0.03P_t + 0.06P_{t-1} + 0.91W_{t-1} \qquad R^2 = 0.92 \quad (1e)$$
$$(1.82)\ (0.05)\quad (0.05)\quad (0.09)\quad DW = 3.08\quad RSS = 1.75$$

$$\hat{W}_t = 15.6 - 0.03P_t \qquad R^2 = 0.03 \quad (2e)$$
$$(3.86)\ (0.04)\quad DW = 0.11\quad RSS = 22.4$$

$$\hat{\Delta W}_t = 0.13 - 0.06\Delta P_t \qquad R^2 = 0.29 \quad (3e)$$
$$(0.06)\ (0.02)\quad DW = 3.02\quad RSS = 1.95$$

(a) (11 marks) □ Show that (2) and (3) are restricted versions of (1), what are the restrictions?

Solution: Eq. (2), $\beta_3 = \beta_4 = 0$.

Eq. (3) $\Delta W_t = \gamma_1 + \gamma_2 \Delta P_t + w_t$ or $W_t - W_{t-1} = \gamma_1 + \gamma_2 (P_t - P_{t-1}) + w_t$ and $W_t = \gamma_1 + \gamma_2 P_t - \gamma_2 P_{t-1} + W_{t-1} + w_t$ so $\beta_4 = 1$ and $\beta_3 = -\beta_2$.

[5 marks]

□ Test these restrictions. Which of these tests could be done using either R^2 or RSS and which can be done using only one of these indicators?

Solution: Eq. (2) both test give correct results $F = \frac{(0.92 - 0.03)/2}{(1-0.92)/(23-4)} = 106$ and $F = \frac{(22.4 - 1.75)/2}{1.75/(23-4)} = 112$ – significant as $F_{crit}^{1\%}(2,19) = 5.93$. Eq. (3) $F = \frac{(1.95 - 1.75)/2}{1.76/(23-4)} = 1.1$ – insignificant as $F_{crit}^{5\%}(2,19) = 3.52$. Eq. (2) and (3) have different dependent variables and so different *TSS* what makes their R^2 incomparable.

[6 marks]

(b) (11 marks) □ Test the equations (1e-3e) for autocorrelation in the following order: first test (2e), then (3e) and finally (1e).

Solution: Eq. (2): $DW = 0.11 < DW_L^{1\%}(24) = 1.04$ – positive autocorrelation.

Eq.(3) $4 - 3.02 = 0.98 < DW_L^{1\%}(23) = 1.02$ – negative autocorrelation.

Eq.(1) $h = (1 - 0.5 \cdot DW)\sqrt{\frac{n}{1 - n \cdot (s.e.(b_{Y_{t-1}}))^2}} = (1 - 0.5 \cdot 3.08)\sqrt{\frac{23}{1 - 23 \cdot (0.09)^2}} = -2.87 < 2.58 = N_{crit}^{1\%}$

– autocorrelation.

[6 marks]

□ Assuming in turn correctness of equations specification and taking into account your conclusions above explain what problems, if any, would be encountered if ordinary least squares were used to fit equations (2), (3), (1)? (*consider each of them in turn starting from* (2)*, then passing to* (3) *and finally* (1)). What are their comparative advantages and disadvantages?

Solution: Eq (2): Under positive autocorrelation detected in **b)** estimators of coefficients of (2) will be unbiased, consistent but inefficient. $R^2 = 0.03$ indicates on the misspecification of the equation.

Eq (3): The specification of equation (3) compared to (2) is richer, but now negative autocorelation detected in (b) makes estimators of its coefficients also unbiased, consistent but inefficient.

On the other hand possible non-stationarity of the initial series should also be taken into account. So using data based on differences in equation (3) could help to cope with the problem of nonstationarity.

Eq (1): As it was shown in (b) inclusion of lagged dependent variable into the right side of equation was not enough to get rid of autocorrelation. Under this autocorrelation together with autocorrelated disturbance term OLS estimates will be inconsistent.

[5 marks]

(c) (11 marks) The student decided to apply Common Factor Test to the model (1) estimating the restricted version of the equation (1)

$$\hat{W}_t = 0.54 - 0.05 P_t - 0.046 P_{t-1} + 0.92 W_{t-1} \qquad R^2 = 0.92 \qquad \text{(1Re)}$$
$$(1.03)\ (0.02) \quad (-) \quad (0.06) \quad DW = 3.00 \quad RSS = 1.81$$

□ What is the restriction? Why there is no standard error for coefficient of P_{t-1} in equation (1Re)?

Solution: Eq.(2) $W_t = \beta_1 + \beta_2 P_t + u_t$ and $u_t = \rho u_{t-1} + \varepsilon_t$ give $W_t = \beta_1 + \beta_2 P_t + \rho u_{t-1} + \varepsilon_t$. Lagging (2), multiplying by ρ and subtracting it from (2) we get $W_t = \beta_1 (1 - \rho) + \beta_2 P_t - \rho \beta_3 P_{t-1} + \rho W_{t-1} + \varepsilon_t$. This is restricted version of $W_t = \beta_1 + \beta_2 P_t + \beta_3 P_{t-1} + \beta_4 W_{t-1} + u_t$ with restriction $\beta_3 = -\beta_2 \cdot \beta_4$.

The standard error for coefficient of P_{t-1} is not calculated by the program as this coefficient is obtained by the multiplication of two estimated coefficients.

[4 mark]

□ Help her to run Common Factor test, giving test statistic and decision rule.

Solution: Test statistic $\chi^2 = n \cdot \ln\left(\frac{RSS_R}{RSS_U}\right)$ has χ^2-distribution with 1 degree of freedom (one restriction in our case). $\chi^2 = 23 \cdot \ln\left(\frac{1.81}{1.75}\right) = 0.77 < 3.84 = \chi^2_{crit}(1, 5\%)$ the null hypothesis $H_0: \beta_3 = -\beta_2 \cdot \beta_4$ is not rejected, so we choose restricted version of equation (1).

[3 mark]

□ The student told you that coefficient $\hat{\beta}_4 = 0.92$ of the variable W_{t-1} in (1Re) can be always interpreted as autocorrelation coefficient, so the estimate of autocorrelation coefficient is equal to 0.92. Comment the student's idea of running Common Factor Test in the situation under consideration and her conclusion.

Solution: The idea of doing a common factor test here was meaningless. To perform it, you must first make sure that there is no autocorrelation in the equation in question without restriction. But in b) using the Durbin test it was shown that equation (1e) is characterized by autocorrelation.
If the test were correct, the interpretation of the coefficient in front of the lag variable is also not unambiguous. As the restricted equation is obtained under assumption of the first order autocorrelation $u_t = \rho u_{t-1} + \varepsilon_t$ the interpretation depends on the test results. If null hypothesis $H_0: \beta_3 = -\beta_2 \cdot \beta_4$ (and so assumption of the first order autocorrelation $u_t = \rho u_{t-1} + \varepsilon_t$) is not rejected this coefficient can be interpreted as autocorrelation coefficient ρ. But if the null is rejected this coefficient express just the marginal effect of the lagged dependent variable on the dependent variable.

[4 mark]

Question 3. The student of the course 'Elements of Econometrics' 2019-20 decided to estimate influence of the attendance of seminars in Econometrics AT (in absolute numbers of seminars attended) in the first module – before October exam and in the second module – before December exam) and the quality of the home assignments HA (measured by the average grade per semester – from 0 to 100) on the grade EX (from 0 to 100) obtained at the corresponding exam. She asked the lecturer to provide her with data on all students for the period under consideration and generated the panel data as follows (all calculations were made on the basis of real data of autumn 2019 study for 209 students for whom complete data were available – balanced panel).
Let the model under consideration be

$$EX_{ij} = \beta_1 + \beta_2 HA_j + \beta_3 AT_{ij} + u_{ij};$$

where EX_{ij} - the result of the exam $i = 1, 2, 3, 4$ (EX_1 - Multiple choice test result in October 2019, EX_2 - Free Responce result in October 2019, EX_3 – MCh result in December 2019, EX_4– FR result in December 2019), $j = 1, 2, ...209$, HA_j is the average score for Home Assignments for the first semester for student j, AT_{1j} is the number of weeks when student j attended seminars in module 1, AT_{2j} is the number of seminars attended by student j in module 1 (including repeated attendance), AT_{3j} and AT_{4j} are the same for the module 2.
The student uses three alternative approaches to the evaluation of this model: 1) pooled OLS regression, 2) fixed effects panel regression model, and 3) random effects panel regression model. The results of the estimation are shown below

$$EX_{ij} = 20.42 + 0.37HA_j + 0.31AT_{ij} + e_{ij} \qquad R^2 = 0.2085$$
$$(1.37)\ (0.028) \qquad (0.26) \quad DW = 1.65 \quad RSS = 278905.4 \qquad \textbf{(1 - Pooled)}$$

$$EX_{ij} = 19.24 + 0.32HA_j + 0.65AT_{ij} + e_{ij} \qquad R^2 = 0.3314$$
$$(1.26)\ (0.026) \qquad (0.24) \quad DW = 1.96 \quad RSS = 235602.9 \qquad \textbf{(2 - Fixed)}$$

$$EX_{ij} = 19.35 + 0.32HA_j + 0.62AT_{ij} + e_{ij} \qquad R^2 = 0.2380$$
$$(2.21)\ (0.026) \qquad (0.25) \quad DW = 1.93 \quad RSS = 239680.7 \qquad \textbf{(3 - Random)}$$

(for simplicity, you can assume that the LSDV approach is used when evaluating the fixed effects model)

(a) (11) marks) □ The student remembers that the panel data are characterized by unobserved heterogeneity, but none of estimated equations (1)-(3) contained information about it. Explain to the student the differences between the three methods, stating the reason for the absence of unobserved heterogeneity term in the results.

Solution: Unobserved heterogeneity is a set of factors that determine the differences between panel data objects (in this case, different exams). This heterogeneity cannot be observed directly, by definition. At the same time, ignoring its presence may lead to a bias in the estimates and a decreasing quality of the model.
When the differences between exams are negligible, we can use pooled regression, which is based on the assumption that there is no unobserved heterogeneity. In panel data methods, this heterogeneity is accounted for in the form of differences in the regression intercepts for each of the objects. In the fixed effects method, these differences are attributed to certain fixed factors, whose influence on the position of the intercept is modeled by coefficients at dummy variables corresponding to the objects in question (four exams, to avoid being trapped in dummy variables, the model is set without a constant). In the random effects method, their presence is explained by a random phenomenon accounted for on a par with the random term of the equation. In both fixed and random effects methods, the heterogeneity can be seen in the table of fixed or random effects, and even calculate on this basis the comparative difficulty of the exams (see below).
[6 marks]

□ Consider LSDV fixed effect model, comment on the estimates of coefficients including constant term. Fixed effects estimates are also provided by computer program: $FE_1 = -0.12$, $FE_2 = -6.30$, $FE_3 = 11.86$, $FE_4 = -5.44$ (see coding of four exams 1-4 above). What conclusions about the comparative complexity of the examination tasks can be drawn from examining fixed effects? Explain your reasoning.
Solution: An increase in the average grade for homework by 1 adds 0.32 points to the exam grade (under the same attendance), attending an additional seminar increases the exam grade by 0.65 points (number of submitted HA being fixed). On average, a student who has not attended a single seminar and has not turned in a single homework assignment can expect to receive 19.24 points.
Fixed effects allow us to estimate how the regression plane shifts for each exam type (to get the regression equation for a particular exam, just add the value of the fixed effect to the LSDV equation constant, for example, for October FR (EX_2) the constant is equal to $19.24 - 6.30 = 12.94$, whereas for December MCh (EX_3) the regression equation constant is equal to $19.24 + 11.86 = 31.1$. We can see that all Free Response questions were more complicated than MCh, since their equation constants turn out to be smaller. It is interesting to note that the only positive fixed effect is for MCh test in December so it is probably less difficult.
[5 marks]

(b) (11 marks) □ Help the student to choose between equation (2) (fixed effects model), and equation (3) (random effects model) using Hausman test. The value of test statistic for the Hausman test was evaluated as 12.36. Comment on the idea of the test, assumptions used and results.
Solution: Darbin-Wu-Hausman (DWH) test is used to choose between fixed and random effects. It compares estimates of coefficients obtained by two alternative models. Under H_0 (absence of endogeneity expressed in correlation of a random term with explanatory variables and so there is no difference between coefficients obtained by two alternative models) both fixed effect and random effect models provide us with consistent

estimates. As $12.36 > 9.21 = \chi^2_{crit.,1\%}(2)$ null hypothesis is rejected. Fixed effect panel model should be chosen. This means that there are significant differences between examinations of different types and at different times, as well as the procedures of preparation and passing.

[4 marks]

□ Help the student to choose between models (1) and (2). Do appropriate F test and make a conclusion. Comment on the idea of the test used. What is your final choice between equations (1-3)?

Solution: This test evaluates the joint significance of the variables used in the LSDV method

To answer his question it is possible to compare RSS_{POOLED} with RSS_{LSDV}, or R^2_{POOLED} with R^2_{LSDV} using F-test:

$$F = \frac{(RSS_{POOLED} - RSS_{LSDV})/(4-1)}{RSS_{LSDV}/(4\cdot 209 - 4 - 2)} = \frac{(278905.4 - 235602.9)/3}{RSS_{LSDV}/830} = 50.85$$

or $F = \frac{(R^2_{LSDV} - R^2_{POOLED})/(4-1)}{(1-R^2_{LSDV})/(6\cdot 209 - 4 - 2)} = \frac{(0.3314 - 0.2085)/3}{(1-0.3314)/830} = 50.60$

(4 - number of crosssection objects (exams), 209 - number of students, 2 - number of variables) while $F^{1\%}_{crit}(3, 1000) = 3.80 \Rightarrow$ significant, choose fixed panel regression.

[7 marks]

(c) (11 marks) □ Estimate the significance of the regression equation (2) using F-statistic (indicate test statistics, degrees of freedom and critical values).

Solution: In fixed effect LSDV regression $R^2 = 0.3314$, so

$$F = \frac{R^2/5}{(1-R^2)/(4\cdot 209 - 4 - 2)} = \frac{0.3314/5}{(1-0.3314)/830} = 82.28 \text{ while } F^{1\%}_{crit}(5, 1000) = 3.04 \Rightarrow \text{significant}$$

[5 marks]

□ Answer to the main question of the research – does the attendance matter? Comment on the fact that only panel data models allow to state the significance of the coefficient of the attendance AT_{ij}.

Solution: Evaluate t-statistic for coefficient of AT_{ij}: Pooled: $0.31/0.26 = 1.19$, Fixed: $0.65/0.24 = 2.71$, Random $0.62/0.25 = 2.48$, while $t^{5\%}_{crit}(df = 830) \approx 1{,}96$, $t^{1\%}_{crit}(df = 830) \approx 2{,}58$, so only for Fixed effect model coefficient of AT_{ij} is significant at 1%, it is significant at 5% for Random effect model, and insignificant for Pooled regression. So only panel data models that take into account difficulty and other differences between exams allow for a significant positive contribution of attendance to a student's grade. In pooled models, the different difficulty and format of the exams masks it.

[6 marks]

SECTION B

(Question 4 **OR** Question 5).

Question 4. A student for her diploma paper evaluates the production function for the economy of a small European country whose main source of income is tourism. She has data on the total investments T in the tourism industry, G - gross domestic product (GDP), as well as the population P of the country for 47 years (in your answer the issues of stationarity of time series should be neglected). T and G are both measured in billions of euro, P is measured in millions. Hypothesizing that investments in tourism per capita depends on GDP per capita, she fits the regression (RSS = residual sum of squares):

$$\log\frac{\hat{T}}{P} = -3.74 + 1.19\log\frac{G}{P} \qquad R^2 = 0.9094,\ RSS = 14.26 \qquad (1)$$

She also runs the following regressions:

$$\log\hat{T} = -3.60 + 1.27\log G - 0.33\log P \quad R^2 = 0.9531 \quad RSS = 13.90 \qquad (2)$$

$$\log\frac{\hat{T}}{P} = -3.60 + 1.27\log\frac{G}{P} - 0.06\log P \quad R^2 = 0.9123, \quad RSS = 13.90 \qquad (3)$$

(a) (11 marks) □ What is the difference in economic meaning of the slope coefficients in equations (1) and (2)? Assume all coefficients are significant.

Solution: Simple vs. multiple regression: (1) is based on the values per capita, while (2) – on the use of absolute values. **(1):** the elasticity of investments in tourism per capita with respect to GDP per capita is 1.19. **(2):** The elasticity of investments in tourism with respect to GDP, controlling for population, is 1.27.

[4 marks]

□ Explain why coefficient of variable $\log\frac{G}{P}$ in equation (3) is equal to the coefficient of variable $\log G$ in equation (2).

Solution: $\beta_2(\text{eq.3}) = \dfrac{d\left(\frac{T}{P}\right)}{d\left(\frac{G}{P}\right)} \cdot \dfrac{\left(\frac{G}{P}\right)}{\left(\frac{T}{P}\right)} = \dfrac{\frac{1}{P}d(T)}{\frac{1}{P}d(G)} \cdot \dfrac{G}{T} = \dfrac{d(T)}{d(G)} \cdot \dfrac{G}{T} = \beta_2(\text{eq.2})$.

[4 marks]

□ Suggest possible explanation to the fact that the coefficient of variable $\log P$ in equation (2) is negative.

Solution: Population growth necessitates investments in housing, health care etc.

[3 marks]

(b) (11 marks) □ Demonstrate mathematically that equation (1) is a restricted version of equation (3), stating the restriction.

Solution: $\log\frac{T}{P} = \beta_1 + \beta_2\log\frac{G}{P} + u$ vs. $\log\frac{T}{P} = \beta_1 + \beta_2\log\frac{G}{P} + \beta_3\log P + u$. Restriction is $\beta_3 = 0$.

[3 marks]

□ Test the restriction, using an F test.

Solution: $F = \dfrac{(0.9122 - 0.9094)/1}{(1 - 0.9094)/44} = 1.4$ with critical value of $F^{5\%}(1, 44) = 4.08$ – not reject $\beta_3 = 0$.

[3 marks]

□ The supervisor told the student that her calculation of equation (3) was unnecessary; all of its coefficients can be clearly seen from equation (2). Comment on this.

Equation $\log T = \beta_1 + \beta_2\log G + \beta_3\log P + u$ is equivalent to

$\log T - \log P = \beta_1 + (\beta_2\log G - \beta_2\log P) + \beta_2\log P + \beta_3\log P - \log P + u$

and so to $\log\frac{T}{P} = \beta_1 + \beta_2\log\frac{G}{P} + (\beta_2 + \beta_3 - 1)\log P + u$ so exactly $-0.33 + 1.27 - 1 = -0.06$

[5 marks]

(c) (11 marks) □ Demonstrate mathematically that equation (1) is a restricted version of equation (2), stating the restriction.

Solution: The first specification $\log\frac{T}{P} = \beta_1 + \beta_2\log\frac{G}{P} + u$ may be rewritten as

$\log T - \log P = \beta_1 + \beta_2\log G - \beta_2\log P + u$ and this in turn as $\log T = \beta_1 + \beta_2\log G + (1 - \beta_2)\log P + u$. This is a restricted version of the more general specification $\log T = \beta_1 + \beta_2\log G + \beta_3\log P + v$ where $\beta_3 = 1 - \beta_2$. .

[6 marks]

□ Test the restriction, using an F test. Which of two F tests in (b) and (c) can be done only using RSS, and which using either RSS or R^2?

$F = \frac{(14.26 - 13.90)/1}{13.90/44} = 1.14 < 4.08 = F_{crit.}^{5\%}(1, 44)$ so the restriction $H_0 : \beta_3 = 1 - \beta_2$ is not rejected.
Two tests are numerically identical but the first one can be done using both RSS and R-squared, while the second one can be done only using RSS. Equations (1) and (2) have different dependent variables and so different TSS, so their R-squares are not comparable.
[5 marks]

Question 5. (By choice). The researcher studies the factors that determine people's satisfaction with their lives using the sample data from the Russian Longitudinal Monitoring of Economic Situation and Health (RLMS-HSE) - 13647 people for a certain year.

Description of variables:
Happy (binary dependent variable) - life satisfaction at the moment (1 - satisfied, 0 - dissatisfied);
AGE - respondent's age (number of full years);
MALE - Gender (1 - male, 0 - female);
EMP – employment (1 - yes, 0 - no);
REL - Religion Index according to the respondent's assessment (from 0 (atheist) to 5 - strictly observes all rites of his/her religion);
HEALTH - Health Index according to the respondent's assessment: (from 0 - very bad, to 5 - very good) ;
SMOKE - (1- smokes, 0 - not);
ALCO - (1 - consumes alcohol, 0 - not);
MAR - (1 - married, 0 - not married);
KIDS - number of children;
WAGE - salary per month in thousand rubles.

Regression	(1)		(2)		(3)		(4)	
	OLS		OLS		LOGIT		LOGIT	
	Coefficient	s.e.	Coefficient	s.e.	Coefficient	s.e.	Coefficient	s.e.
C	0.18	0.026	0.16	0.025	-1.38	0.12	-1.48	0.11
AGE	0.00056	0.00030			0.0018	0.0013		
ALCO	-0.028	0.0090	-0.025	0.0090	-0.12	0.039	-0.11	0.039
EMP	0.041	0.012	0.039	0.012	0.27	0.057	0.27	0.056
HEALTH	0.16	0.0067	0.16	0.0057	0.70	0.031	0.70	0.026
KIDS	0.023	0.0046			0.10	0.020		
MALE	-0.029	0.0092			-0.12	0.041		
MAR	-0.090	0.010			-0.40	0.045		
REL	-0.027	0.0044	-0.025	0.0043	-0.12	0.020	-0.11	0.019
SMOKE	0.095	0.010	0.080	0.0098	0.38	0.046	0.36	0.043
WAGE	0.0033	0.00029	0.0035	0.00028	0.018	0.0016	0.019	0.0016
R-squared	0.089		0.083					
McFadden R-squared					0.068		0.063	
F-statistic	133.35		204.76					
LR statistic					1293.10		1199.75	

(a) (11 marks) Consider regression (1) (all its coefficients except AGE are significant at 5%).
□ Give interpretation to the numerical values of some coefficients of regression (1) (*no more than 2, one of them should be of dummy variable*). In a couple of sentences, summarize in simple language the main results of the study that you find most interesting.
Solution: Interpretation of several coefficients for both qualitative and quantitative factors should be provided. For example, the probability of being happy for married people is 9 percentage points less than for nonmarried ones, but every child increases the probability of life satisfaction by 2.3 p.p. A student can notice and comment on some interesting findings like: according to the regressions marriage doesn't bring happiness but children compensate it; alcohol doesn't not make happy, but a cigarette does; men are usually less satisfied with life than women; religion doesn't add happiness, but the health does.
[4 marks]

□ Demonstrate mathematically that the heteroscedasticity is present in the data

Solution: Model is: $Happy_i = \beta_0 + \beta_1 AGE_i + \beta_2 ALCO_i + \beta_3 EMP_i + ... + u_i,\ i = 1, 2, ..., n$

$Happy_i = 1$ if a person is satisfied with life, $Happy_i = 0$ otherwise. As $Happy_i$ takes only two values 1 or 0, therefore u_i can take only two values: $1 - \beta_0 - \beta_1 AGE_i - ...$ when $Happy_i = 1$ and $-\beta_0 - \beta_1 AGE_i - ...$ when $Happy_i = 0$. Based on this we can write the probability distribution of u_i as

$Happy_i$	u_i	$f(u_i)$
1	$1 - \beta_0 - \beta_1 AGE_i - ...$	$+\beta_0 + \beta_1 AGE_i + ...+..$
0	$-\beta_0 - \beta_1 AGE_i - ...$	$1 - \beta_0 - \beta_1 AGE_i - ...$

As $E(u_i) = 0$ we can write $V(u_i)$ as

$$V(u_i) = E(u_i^2) = (1 - \beta_0 - \beta_1 AGE_i - ...)^2 \cdot (\beta_0 + \beta_1 AGE_i + ...) + (-\beta_0 - \beta_1 AGE_i - ...)^2 \cdot (1 - \beta_0 - \beta_1 AGE_i - ...) =$$
$$= (1 - \beta_0 - \beta_1 AGE_i - ...) \cdot (\beta_0 + \beta_1 AGE_i + ...) \times [(1 - \beta_0 - \beta_1 AGE_i - ...) + (\beta_0 + \beta_1 AGE_i + ...)] =$$
$$= (\beta_0 + \beta_1 AGE_i + ...) \cdot (1 - \beta_0 - \beta_1 AGE_i - ...) = E[Happy_i](1 - E[Happy_i]) = P_i(1 - P_i)$$

Hence the disturbance term is heteroscedastic (as P_i and so $P_i(1-P_i)$ is different for different points of the sample). This will make *OLS* estimators inefficient.

[7 marks]

(b) (11 marks) □ R-squared for regressions (1) and (2), and McFadden R-squared for regressions (3) and (4) all are very low. Can we conclude from this that the regressions are insignificant?

Solution: Low R-squared doesn't necessarily mean that the regression is bad, it simply means that there should be a lot of other factors influencing happiness that are not taken into account here.

For equations (1)-(4) it's possible just to use values of F-statistic/LR-statistic given in the table: $F(eq1) = 133.35)$, $F(eq2) = 204.76)$, $LR(eq3) = 1293.10)$, $LR(eq4) = 1199.75)$.

For (1) $F^{1\%}_{crit}(10, 1000) = 2.34$, for (2) $F^{1\%}_{crit}(6, 1000) = 2.82$, for (3) $\chi^{1\%}_{crit}(10) = 23.209$, for (4) $\chi^{1\%}_{crit}(6) = 16.812$ – all four equations are significant.

[4 marks]

□ Consider a group of demographic variables AGE, KIDS, MALE, MAR. Are these variables significant individually and considered together as a group of variables in regressions (1) and (3)? Explain your work.

Solution: Significance of the coefficients can be checked by t-tests in equation (1) and z-tests in equation (3). AGE is not significant, other variables in the group are significant in both regressions (it is given).

To test the significance of the group of variables AGE, KIDS, MALE, MAR in OLS equation:

$$F = \frac{(0.089 - 0.083)/4}{(1 - 0.089)/13636} = 22.45$$

– significant at 1% as $F > 4.65 = F^{1\%}_{crit}(4, 1000) > F^{1\%}_{crit}(4, 13636)$.

The same for equation (3) can be tested by LR statistic LR=2*(logL4-logL3). There is no information on log-likelihood in the table so one of the possible ways to calculate the necessary LR-statistic is:

LR4=2*(logL4-logL0)

LR3=2*(logL3-logL0)

LR=2*(logL4-logL3)= 2*(logL4-logL0)- 2*(logL3-logL0)=LR4-LR3=(1293.097-1199.752)=93.345

$\chi^{1\%}_{crit}(4) = 13.277$ => the group of variables is also significant at 1% s.l.

[7 marks]

(c) (11 marks) □ The researcher wants to use regression (3) to calculate the probability of being happy for a non-smoking, but alcohol consuming man of 40 years with a good health index (HEALTH = 3) and moderately religious (REL = 1), married with one child, and having job with a salary of 25 thousand rubles per month. Help him to do this.

Solution: For a man with job and indicated salary from the third equation

Z=-1.38+0.0018*40-0.12*1+0.27*1+0.7*3+0.1*1-0.12*1-0.4*1-0.12*1+0.38*0+0.018*25=0.852 .

So $P = 1/(1+\exp(-0.852)) = 0.7$

[3 marks]

□ How this probability changes in case of losing a job (keeping ALL other factors constant)?

Solution: Z=0.852, so $ME = [\exp(-0.852)/(1+\exp(-0.852))^2] \cdot (-0.27) = -0.057$ so losing a job means decreese in happiness by approximately 6 p.p. Alternatively: after losing a job
Z=-1.38+0.0018*40-0.12*1+0.27*0+0.7*3+0.1*1-0.12*1-0.4*1-0.12*1+0.38*0+0.018*25=0.582
So probability is $P = 1/(1+\exp(-0.582)) = 0.64$, $ME = 0.64 - 0.7 = -0.06$

[4 marks]

□ The reviewer, checking your calculation in the article submitted for publication, reasonably noticed that the man under consideration, having lost his job, at the same time lost his monthly salary of 25 thousand rubles, whereas the calculation performed earlier does not take this into account. Correct the previous result using any appropriate method.

Solution: If a man loses his job and salary:
Z=-1.38+0.0018*40-0.12*1+0.27*0+0.7*3+0.1*1-0.12*1-0.4*1-0.12*1+0.38*0+0.018*0=0.132
=> $P = 1/(1+\exp(-0.032)) = 0.533$. *Combined ME* $= -(0.533 - 0.7) = -0.17$, the real negative effect (17 p.p.) is almost three times greater.

[4 marks]