The International College of Economics and Finance
Econometrics 2019-2020. First Semester Exam, December 26.
Part 2. (2 hours). Answer all questions (1,2,3) from section A and one (4 or 5) - from section B.
Solutions Outline

IMPORTANT: Start answering each question on the form with the desired question 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.

SECTION A. Answer **ALL** questions **1-3** from this section.

Question 1. [15 marks] Consider the simple linear regression model:

$$Y_i = \beta_1 + \beta_2 X_i + u_i,\ i = 1, ..., n.$$

We assume that the errors $\{u_i\}_{i=1}^n$ are independent random variables with zero mean. The regressor $\{X_i\}_{i=1}^n$ is non-stochastic (fixed under repeated sampling). Under these conditions, the OLS estimator for β_2, $\hat{\beta}_2^{OLS}$, is unbiased. (You are not asked to derive $\hat{\beta}_2^{OLS}$).

(a) [2 marks] □ What is unbiasedness of an estimator.
□ Explain the concept of unbiasedness in simple words without using mathematical notations.

Unbiasedness: $\mathrm{E}(\hat{\beta}_2) = \beta_2$.
We are correct on average in repeated samples, we will not make systematic errors when estimating β.

(b) [6 marks] □ Let Z_i is some non-stochastic variable different from X_i and Y_i.
Consider two other estimators for the slope β_2:

$$\hat{\beta}_2^{*} = \frac{\sum (Z_i - \overline{Z}) Y_i}{\sum (Z_i - \overline{Z}) X_i} \text{ and } \hat{\beta}_2^{**} = \frac{\sum (Z_i - \overline{Z}) Y_i}{\sum (Z_i - \overline{Z}) Z_i}, \text{ where } \overline{Z} = \frac{1}{n}\sum Z_i .$$

Indicate whether $\hat{\beta}_2^{*}$ and $\hat{\beta}_2^{**}$ are unbiased estimators for β_2.

First $\hat{\beta}_2^{*} = \beta_2 + \frac{\sum (Z_i - \overline{Z}) u_i}{\sum (Z_i - \overline{Z}) X_i}$; $\mathrm{E}(\hat{\beta}_2^{*}) = \beta_2 + \frac{\sum (Z_i - \overline{Z}) \mathrm{E}(u_i)}{\sum (Z_i - \overline{Z}) X_i} = \beta_2$ ..(from $Y_i = \beta_1 + \beta_2 X_i + u_i$ and $\sum (Z_i - \overline{Z}) = 0$). Second $\hat{\beta}_2^{**} = \beta_2 \frac{\sum (Z_i - \overline{Z}) X_i}{\sum (Z_i - \overline{Z}) Z_i} + \frac{\sum (Z_i - \overline{Z}) u_i}{\sum (Z_i - \overline{Z}) Z_i}$; $\mathrm{E}(\hat{\beta}_2^{**}) = \beta_2 \frac{\sum (Z_i - \overline{Z}) X_i}{\sum (Z_i - \overline{Z}) Z_i} \neq \beta_2$.

(c) [7 marks]) □ Transforming formulas for $\hat{\beta}_2^{*}$ and $\hat{\beta}_2^{**}$ show that $\hat{\beta}_2^{*}$ is in fact some instrumental variable estimator of β_2, and $\hat{\beta}_2^{**}$ is OLS estimator of slope coefficient in some regression different from regression (1).

$$\hat{\beta}_2^{IV} = \frac{\sum (Z_i - \overline{Z})(Y_i - \overline{Y})}{\sum (Z_i - \overline{Z})(X_i - \overline{X})} = \frac{\sum (Z_i - \overline{Z}) Y_i}{\sum (Z_i - \overline{Z}) X_i} = \hat{\beta}_2^{*}.$$

$$\hat{\beta}_2^{**} = \frac{\sum (Z_i - \overline{Z}) Y_i}{\sum (Z_i - \overline{Z}) Z_i} = \frac{\sum (Z_i - \overline{Z})(Y_i - \overline{Y})}{\sum (Z_i - \overline{Z})^2} = \hat{\gamma}_2^{OLS} \text{ for } Z_i = \gamma_1 + \gamma_2 Y_i + u_i .$$

□ Let $n \to +\infty$. Are $\hat{\beta}_2^{*}$ and $\hat{\beta}_2^{**}$ consistent estimators of β_2 (no derivation expected, just conclude from your work in b-c).

Being IV estimator $\hat{\beta}_2^{*}$ is consistent estimator of β_2.
$\hat{\beta}_2^{**} = \hat{\gamma}_2^{OLS}$ in the regression $Z_i = \gamma_1 + \gamma_2 Y_i + u_i$ but not in $Y_i = \beta_1 + \beta_2 X_i + u_i$ so $\hat{\beta}_2^{**}$ is biased and inconsistent estimator of β_2 unless the case when $\gamma_2 = \beta_2$.

□ Briefly indicate how you would choose between the three estimators, $\hat{\beta}_2^{OLS}$, $\hat{\beta}_2^{*}$ and $\hat{\beta}_2^{**}$.

$\hat{\beta}_2^{OLS}$ is BLUE and should, therefore, be chosen over $\hat{\beta}_2^{*}$ (and of course over $\hat{\beta}_2^{**}$).
(detailed explanations required in a,b,c).

Question 2. [15 marks] The following simultaneous equations model is considered:

$$Y = \beta_1 + \beta_2 X + u \qquad (1)$$
$$X = \alpha_1 + \alpha_2 Y + v \qquad (2)$$

where X and Y are endogenous variables, and u and v are identically and independently distributed disturbance terms with zero means. The sample consists of n observations (X_i, Y_i).

(a) [4 marks] □ Derive reduced form system of equations for the system above.

For (1) $Y = \frac{1}{1-\alpha_2\beta_2}(\beta_1 + \beta_2\alpha_1 + u + \beta_2 v)$. For (2) $X = \frac{1}{1-\alpha_2\beta_2}(\alpha_1 + \alpha_2\beta_1 + v + \alpha_2 u)$.

□ Using reduced form system show that in equations (1)-(2) Gauss-Markov conditions (GMC) are violated

v is a part of Y, so in $X = \alpha_1 + \alpha_2 Y + v$ Y correlates with v, and so GMC are violated. The same for (2).

(b) [6 marks] □ Show that OLS estimator $\hat{\alpha}_2^{OLS}$ of α_2 is inconsistent.

$$\hat{\alpha}_2^{OLS} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)} = \alpha_2 + \frac{\mathrm{Cov}(\frac{1}{1-\alpha_2\beta_2}(\beta_1 + \beta_2\alpha_1 + u + \beta_2 v), v)}{\mathrm{Var}(\frac{1}{1-\alpha_2\beta_2}(\beta_1 + \beta_2\alpha_1 + u + \beta_2 v))} = \alpha_2 + \frac{(1-\alpha_2\beta_2)(\mathrm{Cov}(u,v) + \beta_2\,\mathrm{Cov}(v,v))}{\mathrm{Var}(u) + \beta_2^2\,\mathrm{Var}(v) + 2\beta_2\,\mathrm{Cov}(u,v)}$$

$$\mathrm{plim}\,\hat{\alpha}_2^{OLS} = \alpha_2 + (1-\alpha_2\beta_2)\frac{\mathrm{cov}(u,v) + \beta_2\,\mathrm{var}(v)}{\mathrm{var}(u) + \beta_2^2\,\mathrm{var}(v) + 2\beta_2\,\mathrm{cov}(u,v)} = \alpha_2 + (1-\alpha_2\beta_2)\frac{\beta_2}{1+\beta_2^2}$$

(c) [5 marks] □ What can be said on the identification of the second equation?

Order condition: $G - 1 = 2 - 1 = 1$. The number of missed variables in second equation is 0 - underidentified.

To make the second equation exactly identified, an additional instrument is introduced into the equation system - an exogenous variable Z that correlates with a variable Y but does not correlate with a random term v of the second equation.

$$Y = \beta_1 + \beta_2 X + \beta_3 Z + u \qquad \textbf{(1*)}$$
$$X = \alpha_1 + \alpha_2 Y + v \qquad \textbf{(2)}$$

□ Show that the instrumental variable estimator $\hat{\alpha}_2^{IV}$ based on the instrument Z is consistent.

$$\hat{\alpha}_2^{IV} = \frac{\mathrm{Cov}(X,Z)}{\mathrm{Cov}(Y,Z)} = \alpha_2 + \frac{\mathrm{Cov}(v,Z)}{\mathrm{Cov}(Y,Z)};\ \mathrm{plim}\,\hat{\alpha}_2^{IV} = \alpha_2 + \frac{\mathrm{plimCov}(v,Z)}{\mathrm{plimCov}(Y,Z)} = \alpha_2 + \frac{\mathrm{cov}(v,Z)}{\mathrm{cov}(Y,Z)} = \alpha_2.$$

□ The researcher decides to use two-stage least squares (TSLS) hoping to obtain a more efficient estimator of α_2. First he fits OLS regression

$$\hat{Y} = h_1 + h_2 Z \qquad (3)$$

saves the fitted values, and uses them as an instrument for Y in equation (2). Demonstrate that obtained TSLS estimator $\hat{\alpha}_2^{TSLS}$ is the same as $\hat{\alpha}_2^{IV}$.

$$\hat{\alpha}_2^{TSLS} = \frac{\mathrm{Cov}(X,\hat{Y})}{\mathrm{Cov}(Y,\hat{Y})} = \frac{\mathrm{Cov}(X, h_1 + h_2 Z)}{\mathrm{Cov}(Y, h_1 + h_2 Z)} = \frac{\mathrm{Cov}(X,Z)}{\mathrm{Cov}(Y,Z)} = \hat{\alpha}_2^{IV}$$

(detailed explanations required in a,b,c).

Question 3. [15 marks] An ICEF student during non-study time works in a record company. It was not a banner year but there appeared potential clients – two musical bands. In case of successful release of their albums in December company turns a profit, which allows to make bonus payment for its employees and happily celebrate the New Year. There are data on 200 albums on the base of which the student wants to analyze and predict the success of future albums.

Success – dependent binary variable which takes value 1 if number of sales of an album in the week after release is greater than 200 thousands.

Budget – the amount (in thousands of US dollars) spent promoting the album before release.

Airplay – number of times songs from an album were played on radio in the week before release.

Rank – people's estimate of the attractiveness and stylishness of the performing musicians out of 10

The student estimates different model specifications (standard errors in parentheses):

	(i) OLS	(ii) Logit	(iii) Logit
Constant	-0.688 (0.1286)	-9.9075 (1.783)	-10.911 (2.2275)
Budget	0.00047 (0.00005)	0.00419 (0.0007)	0.0042 (0.0007)
Airplay	0.0185 (0.0021)	0.1483 (0.0245)	0.2222 (0.0958)
Airplay^2	-	-	-0.0012 (0.0015)
Rank	0.0589 (0.0181)	0.5047 (0.2081)	0.5082 (0.2074)
R^2	0.522312		
McFadden R^2		0.5163	0.5186
RSS	23.874	20.520	20.309
LogLikelihood	-71.238	-67.027	-66.703

One band **A** plans to spend \$10000 on advertising and order 55 plays on radio in the week before release, the majority of people in focus-group found them "gorgeous" giving rank 7. The other band **B** is the band of young students who can afford to spend only \$2000 on advertising and order 25 plays on radio but people liked the way they look even much more and ranked them 10.

(a) [3 marks] □ Give interpretations to the coefficients of model (i).

The coefficient of ***RANK*** 0,059: every additional point in ranking (other factors being equal) increases the chances of success by 5,9 p.p.; in the same manner for the other factors.

□ What's the difference between models (ii) and (i)?

LPM (i) – OLS. In (ii) the chance of success is $F(z)=1/(1+e^{-z})$, where $z_i=\beta_0+\beta_1 BUDGET_i+\beta_2 AIRPLAY_i+\beta_2 RANK_i+u_i$, - MLE.

□ What are advantages and disadvantages of each type of models?

LPM: predictions outside the range $[0,1]$, the error term is heteroskedastic, and is not normal, so usual tests could not be applied, the constant marginal effect is not realistic.

Logit: $F(z)=1/(1+e^{-z})\in[0,1]$. MLE properties of estimators: consistent, asymptotically efficient and asymptotically normally distributed. Marginal effects depend on characteristics of individuals.

(b) [6 marks] □ According to model (ii) what are the chances for each band to issue a successful album?

For Band A: $z_i=-9.908+0.004\cdot 10+0.148\cdot 55+0.505\cdot 7=1.822 \Rightarrow P=1/(1+e^{-1.822})=0.861 \Rightarrow$ **86,1%**

For Band B: $z_i=-9.908+0.004\cdot 2+0.148\cdot 25+0.505\cdot 10=-1.146 \Rightarrow P=1/(1+e^{1.146})=0.241 \Rightarrow$ **24,1%**

□ Investigate the marginal effects of budget, airplay and ranking. Compare the results with ones obtained in model (i).

$\frac{\partial p}{\partial X_i} = \frac{dp}{dZ} \cdot \frac{\partial Z}{\partial X_i} = \frac{e^{-Z}}{(1+e^{-Z})^2} \cdot \beta_i$, For Band A: $\frac{e^{-1.822}}{(1+e^{-1.822})^2} = 0.1198$ For Band B: $\frac{e^{1.146}}{(1+e^{1.146})^2} = 0.183$

For Band A:

(ME) of Budget $= 0.1198 \cdot 0.0042 = 0.0005 \Rightarrow$ **0.05** p.p. (**0.047** in LPM model)

ME(Airplay): $= 0.1198 \cdot 0.148 = 0.0178 \Rightarrow$ **1.78** p.p. (**1.85** in LPM model)

ME(Rank): $= 0.1198 \cdot 0.505 = 0.061 \Rightarrow$ **6.1** p.p. (**5.89** in LPM model). The differences are small.

For Band B:

ME(Budget) **0.077** p.p. (**0.047** in LPM model)

ME(Airplay): **2.7** p.p. (**1.85** in LPM model)

ME(Rank): **9.24** p.p. (**5.89** in LPM model). The differences are quite large.

□ What are the maximum possible marginal effects of budget, airplay and ranking according to model (ii)?

Max ME(Budget): $\frac{e^0}{(1+e^0)^2} \cdot \beta_1 = \frac{1}{4} \cdot 0.0042 = 0.001 \Rightarrow$ **0.1** p.p.. For Airplay: **3.7** p.p., For Ranking **12.6** p.p.

(c) [6 marks] Explore the model (iii).

□ What was the logic behind to include the variable $AIRPLAY_i^2$?

Negative sign of $AIRPLAY_i^2$ along with positive sign of $AIRPLAY_i$ - probably at some point people can get tired and bored of listening the same song - the marginal effect of airplays on radio can become negative

□ Investigate the marginal effect of radio airplay according to model (iii) for both groups.

$\frac{\partial z}{\partial (AIRPLAY_i)} = 0.22 - 2 \cdot 0.0012 AIRPLAY_i$

For Band A:, $f(1.2794) = 0.1703$ ME: $0.1703 \cdot (0.22 - 2 \cdot 0.0012 \cdot 55) = 0.0153 \Rightarrow$ **1.5** p.p. (**1.85** in LPM)

For Band B: $f(-1.0156) = 0.1952$ ME: $0.1952 \cdot (0.22 - 2 \cdot 0.0012 \cdot 25) = 0.0317 \Rightarrow$ **3.2** p.p. (**1.85** in LPM)

□ Check the significance of quadratic term $AIRPLAY_i^2$ with the help of LR test. What other way you could do that?

$LR = -2(\log L_R - \log L_U) = -2(-67.027 + 66.703) = 0.648 < 3.84 = \chi^2_{crit,5\%}(1) \Rightarrow$ insignificant.

The alternative way z-test on $AIRPLAY_i^2$: $|z| = \frac{|-0.0012|}{0.0015} = 0.8 < 1.96 = z_{crit,5\%} \Rightarrow$ insignificant.

□ Compare the results of model (i)-(iii) and make conclusions – what model the student should continue to work with, and what recommendations he should provide the bands with to make their albums successful?

LOGIT model without quadratic term is the best. Choose model (ii).

Recommendations. Band A – improve the image (marginal effect of ranking is the biggest one). Band B – use more radio airplays (if they order 20 airplays more (to 45) their chance for success increases to 86%).

(More detailed explanations in a, b, c are welcome).

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

Question 4. [30 marks] The researcher is studying the effect of professional trainings on the increase of sales in chain of kitchen stores in Moscow and Moscow region. The target is the number of contracts (variable S) sold by a particular seller during the year. The experience of seller, measured by the full number of years worked in that retail chain (variable E from 0 to 5), and the results of the IQ test , which each seller passes on hiring (variable IQ, the average value for all categories of sellers was equal to about 100), are considered as independent variables. Seller's participation in professional trainings is characterized by a dummy variable T (equal to one for trained sellers and zero for other sellers).

(a) [8 marks] Based on the data characterizing 2000 sellers, the researcher calculates the following equations

$$\hat{S}_i = -20.4 + 5.2E_i + 1.8IQ_i \quad RSS = 2735345 \qquad \textbf{(1)}$$

$$\hat{S}_i = -22.8 + 5.3E_i + 1.7IQ_i + 19.0T_i \quad RSS = 2579858 \qquad \textbf{(2)}$$

$$\hat{S}_i = -18.1 + 4.8E_i + 1.7IQ_i + 14.3T_i + 1.6E_i * T_i + 0.1IQ_i * T_i \quad RSS = 2577011 \qquad \textbf{(3)}$$

□ Explain the economic meaning of all equation (1) coefficients.

A manager with sales experience greater by one year than his colleague sells an average of 5.2 contracts more, provided their IQ are equal. Almost the same for IQ.

□ The researcher has noticed that the constant and coefficients of variables E and IQ in equations (2) and (3) insignificantly differ from corresponding coefficients of equation (1) and has concluded that the trainings are ineffective. Do you agree with this conclusion? Explain your reasoning.

From (1) and (2): $F = \dfrac{(2735345 - 2579858)}{2579858} \cdot (2000 - 4) = 120.3$, $F_{crit}^{1\%}(1, 1000) = 6.66 \Rightarrow$ variable T in (2) is significant. From (1) and (3): $F = \dfrac{(2735345 - 2577011)/3}{2577011} \cdot (2000 - 6) = 40.8$, $F_{crit}^{1\%}(3, 1000) = 3.80 \Rightarrow$ variables T $E*T$ and $IQ*T$ in (3) are significant as group, so trainings are effective.

□ What implicit assumptions are used in the specification of equation (2)? Why does the researcher calculate equation (3)? Is it possible to check the validity of these assumptions based on a comparison of equations (2) and (3)? What equation would you suggest to choose for further analysis?

Specification of eq. (2) is based on the implicit assumption that training is equally effective for managers with and without selling experience and for managers with high and low levels of IQ.

From (2) and (3): $F = \dfrac{(2579858 - 2577011)/2}{2577011} \cdot (2000 - 6) = 1.1$, $F_{crit}^{5\%}(2, 1000) = 3.00 \Rightarrow$ assumption is valid $\Rightarrow$ choose eq. (2).

(b) [7 marks] In addition, the researcher decided to estimate regressions in the specification of equation (1) separately for sellers who did not participate in the training (Equation 4, 1373 observations) and for sellers who received professional training (Equation 5, 627 observations):

$$\hat{S}_i = -18.1 + 4.8E_i + 1.7IQ_i \quad RSS = 1622106 \qquad \textbf{(4)}$$

$$\hat{S}_i = -14.4 + 6.4E_i + 1.8IQ_i \quad RSS = 954905 \qquad \textbf{(5)}$$

□ Is it possible, based on the information contained in equations (4) and (5), and involving additionally necessary information, to evaluate the effectiveness of training?

From (1),(4),(5): $F = \dfrac{(2735345 - (1622106 + 954905))/3}{(1622106 + 954905)} \cdot (2000 - 6) = 40.8$, $F_{crit}^{1\%}(3, 1000) = 3.80 \Rightarrow$ signif.

□ Compare the assessment of the effectiveness of training with the calculations and conclusions in (a).

Results of Chow test and F-test for full group of dummies in (a) coincide as well as critical values, so their conclusions are similar.

□ In all of the equations considered above, a value RSS was specified, but no information about the R^2 values was available. Suppose, on the contrary, that all equations would have had a R^2 value but no information on RSS. Which of the questions in items (a) to (b) could still be answered and which would not be available for analysis?

All sample and two subsamples in Chow test in (b) have different number of observations and so different TSS, so R^2 of three equations are not comparable as $R^2 = 1 - \frac{RSS}{TSS}$. All equations used in (a) are based on the same sample and so on the same TSS, so all tests based on R^2 remain valid.

(c) [8 marks] The researcher decided to additionally take into account the factor M of the store location (M equal to one for sellers working in Moscow stores and zero for the region), for which he estimated the following equations.

$$\hat{S}_i = -24.3 + 5.3E_i + 1.7IQ_i + 18.9T_i + 3.3M_i \quad RSS = 2574433 \qquad \textbf{(6)}$$

$$\hat{S}_i = -23.2 + 5.3E_i + 1.7IQ_i + 16.5T_i + 1.9M_i + 4.6T_i * M_{i_i} \quad RSS = 2572170 \qquad \textbf{(7)}$$

□ Is the location factor significant for equation (6)?

From (2) and (6): $F = \frac{(2579858 - 2574433)}{2574433} \cdot (2000 - 5) = 4.2$, $F_{crit}^{5\%}(1, 1000) = 3.85 \Rightarrow M$ is significant

□ Give an interpretation of the coefficients of the variables T, M and $T*M$ in equation (7).

For manager from region increase in sales by participating in a training is 16.5 on average, and for manager from Moscow stores it is even more: 16.5+4.6=21.1 (for any equal combination of other factors).

□ Is a group of variables and training and store location T, M and $T*M$ significant in equation (7)

Comparing equations using F-test we get

From eq. (1) and (7): $F = \frac{(2735345 - 2572170)/3}{2572170/(2000 - 6)} = 42.2$, $F_{crit}^{1\%}(3, 1000) = 3.80 \Rightarrow$ significant

(d) [7 marks]□ Is it possible to answer the question of the joint significance of the factors of the training and the location using the Chow test? (no calculations are expected here).

Divide 2000 managers into 4 groups: untrained from region $\overline{T}\overline{M}$, untrained from Moscow $\overline{T}M$, and so on, run 4 regressions in specification (1), obtaining $RSS(\overline{T}\overline{M})$ and so on in addition to $RSS(1)$ for equation (1), and then run Chow test $F = \frac{(RSS(1) - (RSS(\overline{T}\overline{M}) + RSS(\overline{T}M) + RSS(T\overline{M}) + RSS(TM))/(3*3)}{(RSS(\overline{T}\overline{M}) + RSS(\overline{T}M) + RSS(T\overline{M}) + RSS(TM)/(2000 - 3*4)}$

□ Will the calculation of this test answer the question of the significance of a group of variables T, M and $T*M$?

This test is equivalent to F-test with dummy variables M and T only if all slope dummies are included in the equation, but equation (7) does not contain slope dummies, so it cannot answer this question.

(More detailed explanations in a, b, c, d are welcome).

Question 5. [30 marks] The researcher studies the factors that affect the volume of paid services per capita V_i in 82 regions of Russia (in rubles). He suggests that this indicator may depend primarily on average per capita monthly income in rubles I_i (from 14000 to 70000 rubles depending on the region), as well as on the level of unemployment in percent U_i for each region. In addition, the researcher suggests that the situation with paid services in the central (near Moscow) and northwestern regions of Russia (near the city of St. Petersburg) may differ from the rest of the country, so he introduces a dummy variable R_i equal to 1 for central and northwestern regions, and equal to 0 for other regions of Russia.

(a) [7 marks] To assess the impact of income on paid services, the researcher first runs a simple linear regression

$$\hat{V}_i = -2448.2 + 2.05 I_i \qquad R^2 = 0.78 \qquad \textbf{(1)}$$
$$(3546.8)\ \ (0.12)$$

The researcher is afraid that the equation may not be of sufficient quality due to possible heteroscedasticity.

□ What is heteroscedasticity? Explain how heteroscedasticity can arise here. What characteristics of the equation can heteroscedasticity influence and how?

Let $V_i = \beta_1 + \beta_2 I_i + u_i$. Heteroscedasticity: $\text{var}(u_i) \neq \text{var}(u_j)$ for some $i \neq j$ violates $\text{var}(u_i) \neq \sigma_u^2$ for all i, and so estimators of β_1, β_2 are no longer efficient, remaining unbiased. The income varies by region, so we can expect that the income variation in the low-income group will be less than in the high-income group. Standard errors are calculated incorrectly $\Rightarrow$ standard tests are invalid.

□ The researcher then rank all regions in order of increasing per capita income, and then regresses first for the 20 regions with the lowest income (getting RSS value equal $4.81 \cdot 10^8$), and then for the 30 regions with the highest income (getting RSS value equal $5.87 \cdot 10^9$). How can this information be used to check the data for heteroscedasticity? Carry out the necessary calculations, explaining your actions, and make the conclusion.

Goldfeld-Quandt test (G-Q test). The statistics $F = \dfrac{RSS_2/(n_2 - k)}{RSS_1/(n_1 - k)}$ $(RSS_2(n_2) > RSS_1(n_1)$ $\sim \text{F}(n_2 - k, n_1 - k)$.

$F = \dfrac{5.87 \cdot 10^9/(30-2)}{4.81 \cdot 10^8/(20-2)} = 7.84$, $F_{crit}^{1\%}(28,18) = 2.98$ $\Rightarrow$ significant $\Rightarrow$ heteroscedasticity.

(b) [7 marks] The researcher runs two regressions

$$\hat{V}_i = 3817.85 + 1.97 I_i - 683.3 U_i - 4877.5 R_i \qquad R^2 = 0.80 \qquad \textbf{(2)}$$
$$(5365.2)\ \ (0.13)\ \ (360.6)\ \ (2491.6)$$

$$\hat{V}_i = -27725.7 + 3.75 I_i - 2.26 \cdot 10^{-5} I_i^2 - 348.9 U_i - 6323.1 R_i \qquad R^2 = 0.82 \qquad \textbf{(3)}$$
$$(10858.6)\ \ (0.56)\ \ (6.89 \cdot 10^{-6})\ \ (354.8)\ \ (2389.7)$$

After applying the Ramsey test to the equations (2) and (3) and comparing results, he chooses equation (3) as more reasonable.

□ What is the Ramsey test? Can you explain how it's done? Give additional arguments for equation (3).

Ramsey Test: the square of residuals is included additionally in equation. Significant t-test for it indicates on the need to change the specification, introducing some nonlinearity (as in (4)).

Another approach: $|t| = \dfrac{|-2.26 \cdot 10^{-5}|}{6.89 \cdot 10^{-6}} = 3.28$ $\Rightarrow$ significant.

□ Analyze equations (2) and (3) from an economic point of view, based on interpretations of the coefficients and their significance (you are not required to interpret coefficients or conduct significance tests).

Non-linear nature of income impact allows to reflect the decrease in the growth of consumption of paid services with the growth of income of the population. The impact of unemployment is negligible

(c) [8 marks] After conducting White's test with all the cross-terms for equation (3), the researcher obtained the value of the determination coefficient $R^2 = 0.57$ for the auxiliary equation and concluded that heteroscedasticity is present.

□ Explain the mathematics of the White's test: how is the auxiliary equation constructed, what regressors it includes, what are its advantages and disadvantages?

Regress squares of residuals on variables of the model, their squares and cross-terms to obtain R^2. $\chi^2 = n \cdot R^2$ have χ^2-distribution with the d.f.=number of variables in the auxiliary equation.
White test is more flexible than G-Q test but it requires large sample.
Here we have 4 explaining variables plus 2 squares plus all sorts of cross-terms, so 12 variables in total:

□ Explain the logic of the researcher: how he comes to the conclusion that there is heteroscedasticity.

$\chi^2 = nR^2 = 82 \cdot 0.57 = 46.7$ while $\chi^2_{crit,1\%}(10) = 23.2$ $\Rightarrow$ significant. (Note: critical value for df=12 can be found using extrapolation).

(d) [8 marks] In an effort to get rid of heteroscedasticity, the researcher runs the following equations (both equations demonstrate an absence of heteroscedasticity)

$$\underset{(0.58)\ (8.71\cdot10^{-6})\ (9728.1)\ (215.7)\ (2062.8)}{(\hat{V}/I_i) = 2.99 - 1.19\cdot10^{-5} I_i - 16117.4\cdot(1/I_i) - 292.6(U_i/I_i) - 6249.5(R_i/I_i)} \quad R^2 = 0.22 \qquad \textbf{(4)}$$

$$\underset{(0.17)\ (8.82\cdot10^{-6})\ (1.09\cdot10^{-10})\ (0.006)\ (0.04)}{\log\hat{V}_i = 9.2 + 8.42\cdot10^{-5} I_i - 6.95\cdot10^{-10} I_i^2 - 0.02U_i - 0.11R_i} \quad R^2 = 0.85 \qquad \textbf{(5)}$$

□ Explain why each of the specifications for equations (5) and (6) was able to eliminate heteroscedasticity.

If $\sigma(u_i) = X_i \cdot \sigma_u$ then after dividing by X_i we get new disturbance term $\frac{u_i}{X_i} = \sigma_u$ and $\sigma(\frac{u_i}{X_i}) = \frac{\sigma(u_i)}{X_i} = \frac{X_i\sigma_u}{X_i} = \sigma_u$ $\Rightarrow$ no heteroscedasticity. Using log makes makes the values and so the fluctuations of a random term smaller which reduces the variance of the disturnabce term.

□ In equation (4) the value R^2 is less than in equation (3), and the income factor became negative and insignificant? Is this an indication that the resulting equation is of poor statistical quality?

When divided by *I* the variable *I* becomes a constant, now the value of R^2 shows the contribution of other factors to the explanatory power of the equation. To restore and correctly interpret equation (6), it must be multiplied by *I* again.

□ Give interpretation to the coefficient of variable R in equation (5).

In areas other than the Central and North-Western regions, the volume of paid services per capita is 0.11*100=11% higher than in those regions.

□ Consider again equation (5). Suppose the average per capita income of a certain region is 50 thousands rubles. How will its increase in additional one thousand rubles affect the volume of paid services ?

Let $\log(V) = aI + bI^2$ $\Rightarrow$ $\frac{dV}{V} = (a + 2bI)dI$ or $\frac{dV}{V}\cdot100\% = (a + 2bI)dI * 100$. For $I = 50000$, $dI = 1000$, $\Rightarrow$ increase in paid services 100*1000*(8.42/100000-6.95/10000000000*2*50000)=1.47 percent.
(More detailed explanations in a, b, c, d are welcome).