Econometrics – 2022-2023. Midterm exam. 2022 December 29.
Part 2. Free Response Questions (1 hour 30 minutes)
SUGGESTED SOLUTIONS

SECTION A

Answer **ALL** questions from this section (questions **1-2**).

Question 1. (25 marks). A student is investigating factors that affect schoolchildren's consumption of unhealthy food at fast food restaurants (McDonalds etc.) Y_i (say, the average number of hamburgers consumed per month in 2021), assuming that age X_i may influence Y_i, and wants to understand if there any difference in this dependence for boys and girls. She introduces a dummy variable D_i equal to 1 for boys and 0 for girls. Using a sample of 17 boys and 13 girls (total 30 observations), she runs first simple regression Y_i on X_i

$$\hat{Y}_i = -0{,}56 + 0.24X_i \qquad R^2 = 0.17$$
$$(0.53)\;(0.10) \qquad\qquad (1)$$

Assuming that boys eat more frequently at a fast food restaurant, she defines a slope dummy variable $(XD)_i$ as the product of X_i and D_i and fits the regression (standard errors in parentheses):

$$\hat{Y}_i = -1{,}43 + 0.19X_i + 0.52D_i + 0.78(XD)_i \qquad R^2 = 0.36$$
$$(0.36)\;(0.07) \quad (0.33) \quad (0.42) \qquad\qquad (2)$$

(a) □ What is the meaning of coefficients of regression (2)?

The meaning of all coefficinets of equation

$$\hat{Y}_i = -1{,}43 + 0.19X_i + 0.52D_i + 0.78(XD)_i$$

is clear from separate regressions

Girls

$$\hat{Y}_i = -1.43 + 0.19X_i$$

Boys

$$\hat{Y}_i = -0.91 + 0.97X_i$$

where $-0.91 = -1.43 + 0.52$ and $0.97 = 0.19 + 0.78$

[3 marks]

□ Is there any difference in the influence of X_i on Y_i between boys and girls? How to test the significance of this difference?

t-statistics for dummy variables are $t = \frac{0.52}{0.33} = 1.58$ and $t = \frac{0.78}{0.42} = 1.86$ while critical value of $t(crit.,\ 5\%,\ df = 30 - 4 = 26) = 2.056$, so differences in both coefficients are insignificant.

□ Can the answer to the previous question be obtained using the Chow test? What additional information is needed for this and how can it be obtained?

To evaluate the significance both dummy variables taken together one could use F-test for a group of dummies. Comparing R-squared for two regression we get:

$$F(2{,}26) = \frac{(0.36 - 0.17)/2}{(1 - 0.36/26} = 3.86$$

The critical value of F(2,26) at the 5% significance level is 3.37. Hence the null hypothesis that the coefficients are the same for boys and girls is rejected.

[6 marks]

□ Can the answer to the previous question be obtained using the Chow test? What additional information is needed for this and how can it be obtained?

The regression $e_i = \beta_1 + \beta_2 y_i + u_i$ should be estimated using data for the whole sample both for boys and girls, obtaining the sum of squared residuals RSS_{toral}, then estimation is performed for boys (SSR_{boys}) and girls (SSR_{girls}) separately, so these three values of SSR should be obtained. Now Chow test can be evaluated

$$F = \frac{(SSR_{total} - SSR_{boys} - SSR_{girls})/2}{(SSR_{boys} + SSR_{girls})/(30 - 2 \cdot 2)}$$

Chow test is equivalent to F-test for full set of dummies.

(b) When the student showed her results to the supervisor, the latter advised the student to evaluate also a simplified regression of the form

$$Y_i = \beta_0 + \beta_1 D_i + u_i \qquad (3)$$

which does not take into account the effect of age X_i. Unfortunately, the student did not have a computer with her to recalculate the coefficients of this equation. The supervisor noted that, it is sufficient to know the average number of consumed hamburgers per month for girls $\overline{Y}_0$ and boys $\overline{Y}_1$, as it can be shown that $\beta_1 = \overline{Y}_1 - \overline{Y}_0$ and $\beta_0 = \overline{Y}_0$. The student was confused.

□ Help the student to show that these statements are true for the regression (3).

For any value of D_i, regression

$$Y_i = \beta_0 + \beta_1 D_i + u_i \qquad (3)$$

is degenerate, having only a constant on the right side, which is different for girls and boys
Let for simplicity n_0 first observations for D_i are 0, and next n_1 observations are 1, $n_0 + n_1 = n$.
For girls $Y_i = \beta_0 + u_i$,
using OLS $\sum(Y_i - \hat{\beta}_0)^2 \to min$, $-2(\sum Y_i - n_0\hat{\beta}_0) = 0$, $\hat{\beta}_0 = \overline{Y}_0$
For boys $Y_i = \beta_0 + \beta_1 + u_i$,
$\sum(Y_i - \hat{\beta}_0 - \hat{\beta}_1)^2 \to min, -2(\sum Y_i - n_1\hat{\beta}_0 - n_1\hat{\beta}_1) = 0, \hat{\beta}_1 = \overline{Y}_1 - \overline{Y}_0$
This elegant solution was suggested by some students at the exam.
The original solution is more compicated but also possible
As we know

$$\beta_1 = \frac{\sum D_i Y_i - n\overline{D}\overline{Y}}{\sum D_i^2 - n(\overline{D})^2}$$

$$\beta_0 = \overline{Y} - \beta_1 \overline{D}$$

We have $\overline{D} = \frac{\sum D_i}{n} = \frac{n_1}{n}$. Next $\sum D_i Y_i = \sum_{i=n_0+1}^{n} Y_i = n_1 \bar{Y}_1$, also $\sum D_i^2 = \sum D_i = n_1$
Using formula for weighted average we have

$$\bar{Y} = \frac{n_0}{n}\bar{Y}_0 + \frac{n_1}{n}\bar{Y}_1$$

So

$$\beta_1 = \frac{\sum D_i Y_i - n\overline{D}\overline{Y}}{\sum D_i^2 - n(\overline{D})^2} = \frac{n_1\bar{Y}_1 - n \cdot \frac{n_1}{n} \cdot \left(\frac{n_0}{n}\bar{Y}_0 + \frac{n_1}{n}\bar{Y}_1\right)}{n_1 - n \cdot \left(\frac{n_1}{n}\right)^2} = \frac{n_1\bar{Y}_1 - \frac{n_0 n_1}{n}\bar{Y}_0 - \frac{n_1 n_1}{n}\bar{Y}_1}{n_1 - \frac{n_1 n_1}{n}}$$

$$= \frac{n n_1 \bar{Y}_1 - n_0 n_1 \bar{Y}_0 - n_1 n_1 \bar{Y}_1}{n_1(n - n_1)} = \frac{(n_0 + n_1) n_1 \bar{Y}_1 - n_0 n_1 \bar{Y}_0 - n_1 n_1 \bar{Y}_1}{n_0 n_1}$$

$$= \frac{n_0 n_1 \bar{Y}_1 + n_1 n_1 \bar{Y}_1 - n_0 n_1 \bar{Y}_0 - n_1 n_1 \bar{Y}_1}{n_0 n_1} = \frac{n_0 n_1 \bar{Y}_1 - n_0 n_1 \bar{Y}_0}{n_0 n_1} = \bar{Y}_1 - \bar{Y}_0$$

Now $\beta_0 = \bar{Y} - \beta_1 \bar{D} = \frac{n_0}{n}\bar{Y}_0 + \frac{n_1}{n}\bar{Y}_1 - (\bar{Y}_1 - \bar{Y}_0)\frac{n_1}{n} =$

$$= \frac{n_0}{n}\bar{Y}_0 + \frac{n_1}{n}\bar{Y}_1 - \frac{n_1}{n}\bar{Y}_1 + \frac{n_1}{n}\bar{Y}_0 = \frac{n_0}{n}\bar{Y}_0 + \frac{n_1}{n}\bar{Y}_0 = \frac{n}{n}\bar{Y}_0 = \bar{Y}_0$$

[10 marks]

□ Provide the intuition behind these statements

The intuition here comes from simplified example. Let all girls eat the same number of hamburger, say $Y_i = \bar{Y}_0 = a$, and let all boys eat also the same number of hamburger say $Y_i = \bar{Y}_1 = b$. The meaning of the constant $\hat{\beta}_0$ in the regression $Y_i = \hat{\beta}_0 + \hat{\beta}_1 D_i$ is the consumption of hamburgers for girls $\bar{Y}_0 = a$ (when $X_i = 0$). The meaning of $\hat{\beta}_1$ if the premium in consumption for hamburger for boys, so it should be $b - a = \bar{Y}_1 = \bar{Y}_0$. Any similar explanation based on the meaning of the coefficients in the regression with dummies are welcome.

[2 marks]

Question 2. (25 marks) A student tries to find how the expenditure on education E_i (in billions of dollars) relates to Y_i - GDP (in billions of dollars) and P_i - population of each country (in millions) having data on 34 developed and developing countries with high, medium and low aggregate income for the year 2020. She runs a regression model (1) (e_i here and below are the residuals of the regression equation)

$$\underset{(3.40)}{E_i = -4.52} + \underset{(0.004)}{0.043Y_i} + e_i \; R^2 = 0.75, i = 1, \ldots, 34, \quad \textbf{(1)}$$

,

(a) □ Why the student may fear of the presence of heteroscedasticity? Explain on the basis of your understanding of heteroscedasticity.

One of the Gauss-Markov conditions on disturbance term u_i is $Var(u_i) = const$. Violation of this condition $Var(u_i) \neq Var(u_j)$ for some i, j is called heteroscedasticity.
In cross section analysis like in our case the samle includes developed and developing countries significantly different in their GDP and population, so it is naturally to expect the presence of heteroscedasticity: big countries usually have greater deviations (residuals) in their education expenditures.
[2 marks]

□ How heteroscedasticity could influence the regression results?

Under heteroscedasticity the estimates being unbiased become inefficient, significance tests are invalid. Using the option of White heteroscedasticity-consistent standard errors allows to estimate them correctly.
[2 marks]

□ How heteroscedasticity can be detected using graphs (what graphs)?

The heteroscedasticity can be visually observed on the scatter diagram as different range of observation scattering for different values of explanatory variable.
It can be observed also on residual graph after sorting observation by explanatory variable: deviations of residuals from zero are generally greater at the certain parts of the graph (for example residuals become greater with the increase of explanatory variable).
[2 marks]

□ The student arranges countries by Y_i and runs two regressions in specification (1) first for the 10 countries with the highest Y_i values, getting $SSR_1 = 5795.4$, and then for the 14 countries with the lowest Y_i, getting $SSR_2 = 41.7$. Help the student conduct an appropriate heteroscedasticity test using this data.

This data allows to conduct Goldfeld-Quandt test $F = \frac{RSS_1}{RSS_2} = \frac{5795.4/8}{41.7/12} = 208.47$ while $F^{1\%}_{crit}(8.12) = 4.50$ so null of no heteroscedasticity is rejected.
[3 marks]

□ The supervisor advised the student to use per capita values E_i/P_i and Y_i/P_iinstead of absolute values E_i and Y_i. Following his advice, the student again arranges the countries by Y_i/P_i and running corresponding regressions for this data obtains $SSR_1 = 0.19$ for the 10 countries with the highest values Y_i/P_i, and $SSR_2 = 0.33$ for the 14 countries with the lowest Y_i/P_i. What was the idea behind the adviser's advice? Help the student conduct the test to assess the usefulness of the advisor's advice.
While variations of education expenditures in different countries are significantly different in absolute value the differences in per capita values will be substantially smaller, which can help eliminate heteroscedasticity. Now $F = \frac{RSS_2}{RSS_1} = \frac{0.33/12}{0.19/8} = 1.16$ while $F^{5\%}_{crit}12.8) = 3.28$ so null of no heteroscedasticity is not rejected.
[3 marks]

(b) Next, the student calculates multiple regression E_i on Y_i and P_i (2).

$$\underset{(0.94)}{E_i = -1.57} - \underset{(0.0027)}{0.0056\, Y_i} + \underset{(0.044)}{0.88\, P_i} + e_i \,, R^2 = 0.98, , \quad \textbf{(2)}$$

□ Compare the coefficients for the variable Y_i in equations (1) and (2): How has the meaning of the coefficient, its value and its significance changed. Which value seems more reasonable to you and why?

0.043 in (1) is the marginal effect of income on education expenditure, while -0.0056 is the partial marginal effect under population being fixed.
Important variable P_i is omitted from equation (1), which probably leads to a bias in the marginal income effect estimate in (1) and also to the fact that tests for significance become invalid. This explanation is confirmed by the significance of the included variable P_i at 1%. The negative partial marginal effect in (2) is significant at the 5% level. According to equation (2), the main factor in the growth of spending on education is the population of the country.

[3 marks]

□ To test equation (2) for heteroscedasticity, the student uses the Breusch-Pagan test, obtaining the value $R^2 = 0.38$ for it. Help the student complete the test, describing the test procedure and its result.

In order to perform the Breusch-Pagan test, the student memorized the residuals e_i of equation (2) and evaluated the auxiliary regression

$$\hat{e}_i^2 = b_0 + b_1 Y_i + b_2 P_i \quad R^2 = 0.38.$$

Test statistic here is $n \cdot R^2$ that has chi-square distribution with the degrees of freedom equal to the number of variables in the auxillary equation. Here $n \cdot R^2 = 34 \cdot 0.38 = 12.92$ while $\chi^2(1\%,\ 2) = 9.21$, so we reject null of no heteroscedasticity. Alternatvely it is possible also to use F-test $\frac{R^2/2}{(1-R^2)/(34-3)} = \frac{0.38/2}{(1-0.38)/(34-3)}$ =9.5, while $F\left(1\%, 2,\ 31\right) = 5.39$ so the conclusion is the same

[4 marks]

On the advice of a friend, a student calculates multiple regression in logarithms (3)

$$\underset{(0.34)}{\ln E_i = 9.63} - \underset{(0.09)}{0.37}\ \ln Y_i + \underset{(0.09)}{1.37}\ \ln P_i + e_i\ ,\ R^2 = 0.95, , \qquad \textbf{(3)}$$

and obtains the value of $R^2 = 0.16$ for the Breusch-Pagan test and the value of $R^2 = 0.36$ for the White test with cross terms

□ Why might using logarithms help get rid of heteroscedasticity? Did this technique help, judging by the results of the Breusch-Pagan and White tests? Help the student to complete both tests. Why are the results of the two tests different? Which one can be trusted more and why?

The logarithm of the dependent variable significantly reduces its values, and thereby reduces the differences between its values for different observations, which can lead to getting rid of heteroscedasticity. In our case, for Breusch-Pagan test $n \cdot R^2 = 34 \cdot 0.16 = 5.44$ while $\chi^2(5\%,\ 2) = 5.99$ so null of homoscedasticity is not rejected. Alternatvely it is possible also to use F-test $\frac{0.16/2}{(1-0.16)/(34-3)}$ =2.95, while $F\left(5\%, 2,\ 31\right) = 3.32$ so the conclusion is the same
But for White's test (which has the same test statistic) $n \cdot R^2 = 34 \cdot 0.36 = 12.24$ under $\chi^2(5\%,\ 5) = 11.07$ which indicates the presence of heteroscedasticity. Alternatvely it is possible also to use F-test $\frac{0.36/5}{(1-0.36)/(34-5)}$ =3.26, while $F\left(5\%, 5,\ 29\right) = 2.55$ so the conclusion is the same
We use 5 degrees of freedom in the test, since White's auxiliary equation with cross-term contains 5 variables

$$\hat{e}_i^2 = c_0 + c_1 Y_i + c_2 Y_i^2 + c_3 P_i + c_4 P_i + c_5 Y_i P_i \quad R^2 = 0.36.$$

This test is more accurate, as it allows you to identify more complex forms of heteroscedasticity.

[3 marks]

□ What would you advise a student to get rid of heteroscedasticity in equation (2)?

The weighted least squares (WLS) method should be used, for this the entire equation (1) should be divided by P_i which, as we saw earlier, can be considered as the main heteroscedasticity factor.

[3 marks]

SECTION B.

Answer **ONE** question from this section (**3 OR 4**).

Question 3. (25 marks) A student of ICEF's econometrics course, using data from a sample of 100 students received from a teacher, wants to know which factors determine the Y_i score (out of 100 points) on the winter exam in econometrics. Since econometrics is essentially based on statistics, it is natural to assume that one of these factors can be students' knowledge of statistics Z_i

$$. Y_i = \beta_1 + \beta_2 Z_i + u_i \qquad (1)$$

Direct measurement of Z_i is not possible, the available variable is the S_i – score (also out of 100 points) obtained in the second-year exam in statistics. Since the students were nervous during this exam what could positively or negatively affect the result of the exam, the student assumes that this introduces the measurement error $S_i = Z_i + w_i$, where w_i can be considered as distributed independently of Z_i and u_i with zero expectation $\mathrm{E}w_i = 0$ and constant variance σ_w^2.

She gets the following equation using OLS

$$\hat{Y}_i = -8{,}26 + 0.80 S_i \qquad R^2 = 0.42$$
$$(6.00) \; (0.09) \qquad (2)$$

(a) □ What are the consequences of the presence of measurement errors in regressor when estimating the coefficient β_2 by the OLS method?

Let $S = Z + w$. Substituting for $Z = S - w$ in $Y = \beta_1 + \beta_2 Z + u$ we have

$Y = \beta_1 + \beta_2(S - w) + u = \beta_1 + \beta_2 S + r$ where $r = u - \beta_2 w$.

$$b_2 = \beta_2 + \frac{\widehat{\mathrm{Cov}}(S, u)}{\widehat{\mathrm{Var}}(S)} = \beta_2 + \frac{\widehat{\mathrm{Cov}}([Z + w], [u - \beta_2 w])}{\widehat{\mathrm{Var}}(S)}$$

It is impossible to use expectations as S in denominator is stochastic

We will use plim instead

$$\mathrm{plim}\, b_2 = \beta_2 + \frac{\mathrm{plim}\, \widehat{\mathrm{Cov}}(Z, u) - \beta_2 \mathrm{plim}\, \widehat{\mathrm{Cov}}(Z, w) + \mathrm{plim}\, \widehat{\mathrm{Cov}}(w, u) - \beta_2 \mathrm{plim}\, \widehat{\mathrm{Cov}}(w, w)}{\mathrm{plim}\, \widehat{\mathrm{Var}}(S)} =$$

$$= \beta_2 + \frac{\mathrm{Cov}(Z, u) - \beta_2 \mathrm{Cov}(Z, w) + \mathrm{Cov}(w, u) - \beta_2 \mathrm{Cov}(w, w)}{\mathrm{Var}(S)}$$

$$= \beta_2 + \frac{\sigma_{Z,u} - \beta_2 \sigma_{Z,w} + \sigma_{u,w} - \beta_2 \sigma_w^2}{\sigma_S^2}$$

The first three terms in the numerator are 0 and so

$$\mathrm{plim}\, b_2 = \beta_2 + \frac{-\beta_2 \sigma_w^2}{\sigma_S^2} = \beta_2 - \frac{\beta_2 \sigma_w^2}{\sigma_Z^2 + \sigma_w^2}$$

Taking into account that, for meaningful reasons, the coefficient β_2 must be positive, we obtain a downward bias.

[6 marks]

□ A friend of the student pointed out that the statistics exam was assessed very harshly with student grades lowered, so it is more correct to assume that $\mathrm{E}w_i = \mu_{w_i} < 0$. What additional consequences for the estimation of β_2 by the OLS method will this assumption lead to?

If $\mathrm{E}(w)$ is not equal to 0, b_2 is not affected as $\mathrm{E}(w)$ is not used in the derivation of large sample bias.

b_2 remains inconsistent as in the standard case $\mathrm{E}(w) = 0$.

[2 marks]

(b) □ What are the consequences of measurement errors when estimating the coefficient β_1 by the OLS method assuming $\mathrm{E}(w) = 0$?

The OLS estimator of the intercept is affected in both cases, but like the slope coefficient, it was inconsistent anyway.

$$b_1 = \bar{Y} - b_2 \bar{S} = \beta_1 + \beta_2 \bar{S} + \bar{r} - b_2 \bar{S} = \beta_1 + \beta_2 \bar{S} + \bar{u} - \beta_2 \bar{w} - b_2 \bar{S}$$

Hence $\mathrm{plim}\, b_1 = \beta_1 + (\beta_2 - \mathrm{plim}\, b_2)\mathrm{plim}\bar{S} + \mathrm{plim}\, \bar{u} - \beta_2 \mathrm{plim}\, \bar{w}$

In the standard case $\mathrm{E}w_i = 0$ this would reduce to

$$\text{plim } b_1 = \beta_1 + (\beta_2 - \text{plim} b_2)\text{plim}\bar{S} = \beta_1 + \beta_2 \frac{\sigma_w^2}{\sigma_z^2 + \sigma_w^2}\bar{Z}$$

or

$$\beta_1 + \beta_2 \frac{\sigma_w^2}{\sigma_z^2 + \sigma_w^2} E(Z)$$

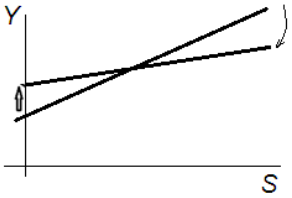
if Z is stochastic.
We see that the intercept in this case experiences an upward bias.

[4 marks]

□ Illustrate graphically the result obtained in the previous question connected with estimation of β_1.

Taking into account that the regression graph is located entirely in the first quadrant (both estimates S and Y are positive), and the midpoint of the regression $(\bar{S}, \bar{Y})$ is also located there, (the regression line always passes through this point) we obtain that when the slope coefficient is shifted downward, the intercept moves upward.

[3 marks]



(c) □ Since the student also has data on grades in other subjects M_i (mathematics), B_i (banking), L_i (linear algebra) etc. (the student assumes that these variables are not subject to measurement errors), she builds a regression of S_i on all these variables, remembers the residuals of this auxiliary regression E_i and includes them in her equation

$$\hat{Y}_i = -20{,}69 + 1.00S_i - 0.47E_i \qquad R^2 = 0.46$$
$$(7.64) \; (0.12) \quad (0.19) \qquad\qquad\qquad\qquad\qquad (3)$$

Comment on the aim and the logic of the performed procedure and also on the obtained results.

This is Darbin-Wu-Hausman test for the endogeneity due to the measurement errors in the version proposed by Davidson and MacKinnon. If the OLS estimates are consistent, then the coefficient of E_i should not be significantly different from zero. So the significance of the residuals could be considered as a sign of possible endogeneity of suspect explanatory variable. This is just our case as

$$t_{E_i} = \frac{|-0.47|}{0.19} = 2.47 > 1.96$$

[5 marks]

□ In fact during winter exam in econometrics the students were also nervous, which introduced also measurement error v_i $(E\,v_i = 0,\ \sigma_{v_i} = \sigma_v^2)$ into the variable Y_i. How will this assumption affect the properties of the estimate β_2 in equation (2)?

If the students were nervous also taking econometric examinations (so the variable Y_i contains the measurement error term q_i

$$Y_i + q_i = \beta_1 + \beta_2 Z_i + u_i$$

this term simply adds to the disturnance term of equation

$$Y_i = \beta_1 + \beta_2 Z_i + (u_i - q_i)$$

So in general case the estimators of β_1 and β_2 become less efficient. No additional bias observed.

[3 marks]

□ Will your conclusions change if you take into account that as econometrics exam was just before the New Year the the graders were instructed to consider all controversial cases in favor of the students, so we can assume that $\mathrm{E}\,u_i = a > 0$ (*no rigorous derivation expected*)?

Obviously, in this case, the estimate of the coefficient β_2 will not change, while the estimate of $\beta_1 2$ will acquire a positive bias equal to a, which will contribute to the rise in the pre-New Year mood of the students.

[2 marks]

Question 4. (25 marks) After graduating from university a student joined a consulting firm dealing with the promotion of candidates and the organization of elections for various positions by voting on the Internet. Her first assignment was to advise potential candidate **A** for the position of head of the student organization. Candidat **A** is young and inexpierenced person who can afford to spend only $2000 on advertising but people liked the way he looks and ranked his attractiveness 5. His competitor is Candidate **B**. Being quite experienced and rich **B** plans to spend $5000 on advertising, having rank of attractiveness 2. Each candidate is granted 3 free appearances on local television, additional appearances must be paid for by them at a cost of $799 per performance.
The student got data on similar situations from past election results. The data includes 200 observations on the following indicators:
V – dependent variable - number of votes cast for the candidate.
E – dependent binary variable which takes value 1 if the candidate was elected.
AD – the amount of money (in thousands of US dollars) spent promoting the candidate.
TV – number of times candidate appears on TV special events (debates, speaches and so on).
APP – people's estimate of the personal appeal of a candidate out of 5.

Using this data the student estimates different models (standard errors or their counterparts in parentheses):
For the probit model, $f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}$ *(the standard normal probability density function).*

$$\underset{(17.84)}{\hat{V}_i = -41.60} + \underset{(2.08)}{25.15\ AD_i} + \underset{(2.75)}{32.64\ TV_i} + \underset{(4.72)}{21.60\ APP_i} \quad R^2 = 0.66, \qquad \textbf{(1-OLS)}$$

$$\underset{(0.13)}{\hat{E}_i = -0.74} + \underset{(0.016)}{0.14\ AD_i} + \underset{(0.02)}{0.18\ TV_i} + \underset{(0.04)}{0.11\ APP_i} \quad R^2 = 0.51, \qquad \textbf{(2-OLS)}$$

$$\underset{(0.94)}{\hat{E}_i = -5.64} + \underset{(0.13)}{0.64\ AD_i} + \underset{(0.12)}{0.75\ TV_i} + \underset{(0.19)}{0.54\ APP_i} \quad McFadden\ R^2 = 0.49, \qquad \textbf{(3-Probit)}$$

(a) □ Explain the meaning of regression (1). Compare the chances of candidates based on model (1), assuming that a higher expected number of votes can be considered as an indicator of success.

Conventional OLS regression: for example, the coefficient of AD_i equal 25.15 shows that every additional thousand of dollars spent on advertising (other factors being equal) increases the number of votes by 25.15 on average. The other coefficients are interpreted in the same way.
Substituting data into equation we get expected values of votes for each candidate
For Candidate A:

$$V_i = -41.6 + 25.15 \cdot 2 + 32.64 \cdot 3 + 21.61 \cdot 5 = 214.7$$

Candidate B

$$V_i = -41.6 + 25.15 \cdot 5 + 32.64 \cdot 3 + 21.61 \cdot 2 = 225.3$$

So the chances of B are slightly higher.
[2 marks]

□ Explain the meaning of regression (2) and its coefficients.

This is Linear Probability Model LPM: for example, the coefficient of AD_i equal 0.14 shows that every additional appearance on TV increases the chances to be elected by 14 p.p. keeping other variavles constant and so on.
[2 marks]

□ Explain the logic of the model (3) descibing the mechanism of obtaining regression results?

Model (ii) is LPM, estimated by OLS, while model (iii) is a binary choice probit model where the probability of the success is determined by the cumulative normal distribution function, $F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}} dx$, where $z_i = \beta_0 + \beta_1 AD_i + \beta_2 TV_i + \beta_2 APP_i + u_i$. It is estimated by maximum likelihood techniques.
[4 marks]

(b) □ According to model (3) what are the chances for each candidate to be elected? Compare with the results of the model (2). Which model can be trusted more? *Recall briefly the data on both candidates*:
A – $AD = 2, \ \ TV = 3, \ APP = 5$, B – $AD = 5, \ \ TV = 3, \ APP = 2$.
Model (3)

For A

$z_i = -5.64 + 0.64 \cdot 2 + 0.75 \cdot 3 + 0.54 \cdot 5 = 0.59 \ \Rightarrow$ normal tables $P(SUCCESS) = \Phi(0.59) = 0.72 \Rightarrow$ **72%**

For B

$z_i = -5.64 + 0.64 \cdot 5 + 0.75 \cdot 3 + 0.54 \cdot 2 = 0.89 \ \Rightarrow$ normal tables $P(SUCCESS) = \Phi(0.89) = 0.81 \Rightarrow$ **81%**

Model (2) gives a little different results

Candidate A

$$E_i = -0.74 + 0.14 \cdot 2 + 0.18 \cdot 3 + 0.11 \cdot 5 = 0.72$$

Candidate B

$$E_i = -0.74 + 0.14 \cdot 5 + 0.18 \cdot 3 + 0.11 \cdot 2 = 0.63$$

As we can see candidate A is slightly behind in the probability of success.
Model (2) underestimates the probability of success, and, in addition, it has all the disadvantages of LPM model: the constant marginal effect, heteroscedasticity, the lack of normality in the distribution of residuals, and possible going beyond the boundaries of the interval [0; 1]. So the model (3) can be trusted more.

[8 marks]

(c) □ Based on the analysis of the marginal effects of advertising and TV appearances according to model (3), help the student to advise Candidate A on how he should reallocate funds between advertising and TV appearances in order to close the gap with Candidate B or overtake him (*recall that one additional TV appearance costs 799 dollars, candidates has no free funds*). Show with calculations that your advice can really bring success to A.

To evaluate marginal effect we need to use chain rule for evaluation derivative of $P(SUCCESS) = p$ with respect to X_i $\frac{\partial p}{\partial X_i} = \frac{dp}{dZ} \cdot \frac{\partial Z}{\partial X_i} = f(Z) \cdot \beta_i$, where $f(Z) = \frac{dp}{dZ} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2}$

All effects in equation (3) are significant (using z-statistics)

From (b) for Candidate A:

$z_i = -5.64 + 0.64 \cdot 2 + 0.75 \cdot 3 + 0.54 \cdot 5 = 0.59$ but now we substitute this to the normal probability density function $f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(0.59)^2} = 0.335$

and then mutiply it by AD coefficient

$\frac{\partial p}{\partial AD_i} = f(Z) \cdot \beta_2 = 0.335 \cdot 0.64 = 0.21 \Rightarrow$ **21** p.p.

This is marginal effect of \$1000 spend on advert.

One appearance on TV give the following effect

$\frac{\partial p}{\partial TV_i} = f(Z) \cdot \beta_2 = 0.335 \cdot 0.75 = 0.25 \Rightarrow$ **25** p.p.

(this is a much higher value than gives LPM –- 14 и 18 p.p correspondingly).

One additional performance costs 799 dollars, less than a thousand, and the effect will be 24.4 points. Thus it makes sense for A to use some of the advertising money to buy additional TV appearances. One can buy 2 additional performances. As a result, there will be a $2000 = 2 \cdot 799 = 402$ dollars left for advertising, and the new vector of parameters for candidate A will take the form $AD = 0.402, \ \ TV = 5, \ APP = 5$

Let's calculate what effect it will actually give. Recall the previous estimates of the probability of success. From (b) $P(SUCCESS) = 0.81$ for A и $P(SUCCESS) = 0.81$ for B. Now For A

$z_i = -5.64 + 0.64 \cdot 0.402 + 0.75 \cdot 5 + 0.54 \cdot 5 = 1.12 \ \ \Rightarrow$ normal tables $P(SUCCESS) = \Phi(1.12) = 0.868 \Rightarrow$ **86.8%** which exceeds the probability of success of candidate B. However, if Candidate B does the same, then with his huge sum of \$10,000, by buying several TV appearances at once, he will certainly secure a victory. Therefore, try to keep the scheme found by the student a secret from B.

[9 marks]