The International College of Economics and Finance
Econometrics 2020-2021. First Semester Exam, December 24.
Suggested Solutions
SECTION A. Answer **ALL** questions **1-3** from this section.

Question 1. (17 marks)

A researcher, using a sample of 626 individuals from the National Longitudinal Survey of Youth, is investigating how the probability of a respondent obtaining a bachelor's degree from a four-year college is related to the respondent's score on a test of cognitive ability, $SCORE_i$ ranged from 0 to 100 (22 to 65 in the sample, with mean value 50.2, and most scores lying in the range 40 to 60). $ETHBLACK_i = 1$ and $ETHHISP_i = 1$ for the ethnic black and ethnic hispanic persons correspondingly and zero for the others.

26.7 percent of the respondents earned bachelor's degrees. Defining a variable BA_i to be equal to one if the respondent has a bachelor's degree (or higher degree) and zero otherwise, the researcher fitted the OLS regression (1) (standard errors in parentheses): and logit regressions (2) и (3) (asymptotic standard errors in parentheses):

$$\hat{BA}_i = -0.54 + 0.015 \cdot SCORE_i \qquad R^2 = 0.10 \qquad \textbf{(1)}$$
$$\qquad (0.09) \quad (0.0018)$$

$$\hat{BA}_i = -8.54 + 0.13 \cdot SCORE_i \quad McFadden \;\; R^2 = 0.13 \qquad \textbf{(2)}$$
$$\qquad (0.99) \quad (0.018) \qquad LR\ stat. = 72.49$$

$$\hat{BA}_i = -9.05 + 0.14 \cdot SCORE_i + 1.26 \cdot ETHBLACK_i - 0.34 \cdot ETHHISP_i \quad McFadden \;\; R^2 = 0.14 \qquad \textbf{(3)}$$
$$\qquad (1.09) \quad (0.019) \qquad (0.45) \qquad (0.66) \qquad LR\ stat. = 80.24$$

(a) □ Evaluate marginal effect of $ETHBLACK_i$ for the ethnic black person with $SCORE_i = 70$? (Use the method of direct comparison of two probabilities, explaining why it is most sutable here and presenting and explaining all steps of your calculations).

The change in a variable $ETHBLACK_t$ by the value of its full range is too large to use a differential formula. Therefore, a direct comparison of the two probabilities should be used.

Let us compare two probabilities: for reference ethnic white with the score 70:

$P = \dfrac{1}{1+e^{-(-9.05+0.14 \cdot 70)}} = 0.68$ and $P = \dfrac{1}{1+e^{-(-9.05+0.14 \cdot 70+1.26 \cdot 1)}} = 0.88$, marginal effect is $\Delta = 0.88 - 0.68 = 0.2$.

□ What is the meaning of this marginal effect?

This marginal effect shows the premium that an ethnic black person with score 70 receives over an ethnic white person with the same score.

□ Is this marginal effect significant?

Calculate the z-statistics $z = \dfrac{1.26}{0.45} = 2.8 > 2.57 = z_{crit}^{1\%} \Rightarrow$ significant.

□ Are equations (1), (2) and (3) significant on the whole?

For equation (1) $F = \dfrac{0.1}{(1-0.1)} \cdot 624 = 69.33 >= 6.68 = F_{crit}^{1\%}(1, 624) \implies$ significant.

For equation (2) we use a LR statistic of 72.49 with a critical chi-square value of one degree of freedom for 1% 6.63 or for 0.1% 10.83 $\implies$ significant.

Similarly for equation (3) LR=80.24 with a critical chi-square value with two degrees of freedom for 1% 11.34 or for 0.1% 16.27 $\implies$ significant.

□ Are two variables $ETHBLACK_i$ and $ETHHISP_i$ taken together significant?

In order to perform a test for restrictions, we need to compute the LR-statistics $LR_{restr} = 2(\log(L(3) - \log(L(2))$ for which we need to know the values of logarithmic likelihood functions $\log(L(3), \log(L(2))$ of models (3) and (2). This information is not available, but we have values of LR-statistics for models (3) and (2): $LR(3)=2(\log(L(3)-\log(L(0))$ and $LR(2)=2(\log(L(2)-\log(L(0)))$. Since the value $\log(L(0))$ in both expressions has the same meaning (the value of the logarithmic function for the logit model containing only a constant), the required statistics can be obtained as follows

$$LR_{restr}=2(\log(L(3)-\log(L(2))=2(\log(L(3)-\log(L(0))-2(\log(L(2)-\log(L(0))=$$
$$=LR(3)-LR(2)=80.24-72.49=7.55$$

Since we have two restrictions, we use the critical value of chi square 5% at two degrees of freedom, which is 5.99 - Significant at 5%.

(b) Let again an ethnic black person with score of $SCORE_i=70$ is considered.

□ Are marginal effects of the $SCORE_i$ on the probability to gain bachelor degree different in three models (1), (2) and (3)? (Use derivative method when appropriate).

The marginal effect of $SCORE_i$ for equation (1) is 0.015: an additional score (out of 100) obtained on the exam increases the chances of obtaining a bachelor's degree by 1.5 percentage points (or just 0.015).

To find the marginal effect in models (2) and (3), we use the formula $\frac{dF}{dSCORE}=\frac{dF}{dZ}\cdot\frac{dZ}{dSCORE}=\frac{dF}{dZ}\cdot\beta_2$

Where $F=\frac{1}{1+e^{-z}}$, $\frac{dF}{dZ}=\frac{e^{-z}}{(1+e^{-z})^2}$, $BA_i=\beta_1+\beta_2SCORE_i+u_i$, with the derivative calculated at the specified point $SCORE_i=70$, (in model (3) additionally $ETHBLACK_i=1$, $ETHHISP_i=0$):

For equation (2) we get $\frac{dF}{dZ}=\frac{e^{-(-8.54+0.13\cdot70)}}{(1+e^{-(-8.54+0.13\cdot70)})^2}\cdot0.13=0.03$

For equation (3) it is slightly different $\frac{\partial F}{\partial Z}=\frac{e^{-(-9.05+0.14\cdot70+1.26)}}{(1+e^{-(-9.0504+0.14\cdot70+1.26)})^2}\cdot0.14=0.015$

□ Demonstrate that the maximum marginal effect in the logit model $BA_i=\frac{1}{1+e^{-z}}$, $z=\beta_1+\beta_2SCORE_i+u_i$ is reached at $z=0$. Calculate maximum marginal effect.

As the second part β_i of the expression $f(Z)\cdot\beta_i$ is constant we need to find maximum value of $\frac{dp}{dZ}=f(Z)=\frac{e^{-Z}}{(1+e^{-Z})^2}$. The maximum point can easily be found by taking the derivative of the function and equating it to zero

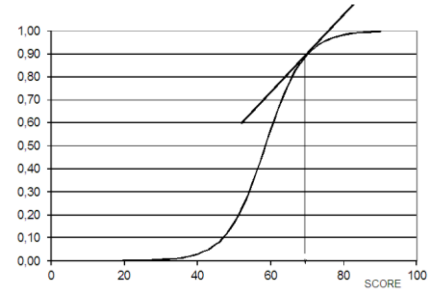
$$\frac{d^2p}{dZ^2}=\left(\frac{e^Z}{(1+e^Z)^2}\right)'=\frac{e^Z(1+e^Z)^2-2(1+e^Z)e^Ze^Z}{(1+e^Z)^4}=\frac{e^Z(1+e^Z)(1+e^Z-2e^Z)}{(1+e^Z)^4}=\frac{e^Z(1+e^Z)(1-e^Z)}{(1+e^Z)^4}=0\Rightarrow Z=0$$

The maximum marginal effect is reached at point $z=0$ and equals $\frac{\partial F}{\partial Z}=\frac{e^{-0}}{(1+e^{-0})^2}\cdot0.14=0.034$

□ Why the value of marginal effect of $SCORE_i$ is far from the maximum for the person under consideration (according to the model (3))? Explain in the context of the considered problem, using graphical illustration.

The maximum marginal effect 0.036 is three times greater than the one calculated by the model 0.012.

The reason is this: the performance of the hypothetical person in question is well above the average of 50.2 and even beyond the highest value in the sample of 65, in which area the marginal effect, equal to the tangent of the slope of the logistic curve, is much smaller than the maximum. In other words, the person in question has practically exhausted the reserves of her/his success determined by $SCORE_i$ growth, so that a further increase gives no more sense.



This is confirmed by the calculated in (a) probability of obtaining a bachelor's degree for the person in question, equal to 0.9.

Question 2 [16 marks]
Consider the following regression model

$$y_i = \beta x_i + u_i .$$

The error term has a zero mean $\mathrm{E}(u_i) = 0$, variance equal to $\mathrm{var}(u_i) = \sigma^2 / x_i^2$, and $\mathrm{E}(u_i u_j) = 0$ if $i \neq j$, $i, j = 1, 2, ..., n$.. You are given a sample of observations $\{(y_i, x_i)\}_{i=1}^{n}$. You may treat x_i as being non-stochastic.

(a) □ Derive the OLS estimator of β taking into account the presence of heteroskedasticity.

To take into account heterocscedasticity one need to apply OLS to the weighted equation (WLS)

$$\frac{y_i}{q_i} = \beta \frac{x_i}{q_i} + \frac{u_i}{q_i} \text{ where } q_i = 1/x_i .$$

Define $y_i^* = \frac{y_i}{q_i} = \frac{y_i}{1/x_i} = x_i y_i$; $x_i^* = \frac{x_i}{q_i} = \frac{x_i}{1/x_i} = x_i^2$; $u_i^* = \frac{u_i}{1/x_i} = u_i x_i$. So WLS equation is $y_i^* = \alpha x_i^* + u_i^*$. The regression without constant can be estimated by $\hat{\beta}^{OLS} = \frac{\sum y_i^* x_i^*}{\sum (x_i^*)^2} = \frac{\sum x_i y_i \cdot x_i^2}{\sum (x_i^2)^2} = \frac{\sum y_i \cdot x_i^3}{\sum x_i^4}$

□ Show that obtained OLS estimator is unbiased.

As $\hat{\beta}^{OLS} = \frac{\sum (\beta x_i + u_i) \cdot x_i^3}{\sum x_i^4} = \frac{\sum \beta x_i^4 + \sum u_i \cdot x_i^3}{\sum x_i^4} = \beta \frac{\sum x_i^4}{\sum x_i^4} + \frac{\sum u_i \cdot x_i^3}{\sum x_i^4} = \beta + \frac{\sum u_i \cdot x_i^3}{\sum x_i^4}$, it is not difficult to prove that the $\hat{\alpha}^{OLS}$ is unbiased: $\mathrm{E}(\hat{\beta}^{OLS}) = \mathrm{E}\left(\beta + \frac{\sum u_i \cdot x_i^3}{\sum x_i^4}\right) = \beta + \frac{\sum \mathrm{E} u_i \cdot x_i^3}{\sum x_i^4} = \beta$.

(b) □ Derive the variance of the OLS estimator of β.

Now evaluate the variance of $\hat{\alpha}$:

$$\mathrm{var}(\hat{\beta}^{OLS}) = \mathrm{var}\left(\frac{\sum u_i \cdot x_i^3}{\sum x_i^4}\right) = \frac{1}{\left(\sum x_i^4\right)^2} \mathrm{var}\left(\sum u_i \cdot x_i^3\right) = \frac{1}{\left(\sum x_i^4\right)^2} \left[\sum x_i^6 \,\mathrm{var}\, u_i \cdot + 2 \sum x_i^3 x_j^3 \,\mathrm{cov}(u_i, u_j)\right] =$$

$$== \frac{\left[\sum x_i^6 \cdot \frac{\sigma^2}{x_i^2}\right]}{\left(\sum x_i^4\right)^2} = \sigma^2 \frac{\sum x_i^4}{\left(\sum x_i^4\right)^2} = \frac{\sigma^2}{\sum x_i^4}$$

□ Demonstrate that OLS estimator of β obtained in (a) is consistent.

As $n \to \infty$ generally speaking $\sum x_i^4 \to \infty$ and so $\mathrm{plim}\,\mathrm{var}(\hat{\beta}^{OLS}) = 0$, so the sufficient condition for the consistency is satisfied ($\hat{\beta}^{OLS}$ is unbiased and $\mathrm{var}(\hat{\beta}^{OLS}) \xrightarrow[n \to \infty]{} 0$).

Question 3. (17 marks)
Consider two equations

$$y_t = \alpha x_t + u_t \qquad (1)$$

$$y_t = \alpha x_t + \beta z_t + u_t \qquad (2)$$

where x_t and z_t are deteministic sequences and u_t follows the conditions $u_t :$ $\mathrm{E}u_t$, $\mathrm{E}u_t^2 = \sigma^2$, $\mathrm{E}u_t u_s = 0, t \neq s$.

(a) □ How parameters α and β can be estimated using ordinary least squares (OLS). Explain. (*application of second order conditions for minima is not expected here and gives no marks*)

The OLS estimator is derived by minimizing the sum of squares of errors, that is, $S = \sum_{t=1}^{T} \hat{u}_t^2 = \sum_{t=1}^{T} (y_t - \hat{\alpha} x_t)^2$.

To achieve the minimum differentiate w.r.t $\hat{\beta}$ and set the derivative equal to 0 and solve, that is,

$\frac{\partial S}{\partial \hat{\alpha}} = -2\sum (y_t - \hat{\alpha} x_t) x_t = 0$ which has the solution $\hat{\alpha} = \frac{\sum_{t=1}^{T} x_t y_t}{\sum_{t=1}^{T} x_t^2}$.

Now for multiple regression.
Let $\hat{\alpha} = a$, $\hat{\beta} = b$.

$$S = \sum_{t=1}^{T} \hat{u}_t^2 = \sum_{t=1}^{T} (y_t - a x_t + b z_t)^2$$

$$\frac{\partial S}{\partial a} = -2\sum (y_t - a x_t - b z_t) x_t = 0$$

$$\frac{\partial S}{\partial a} = -2\sum (y_t - a x_t - b z_t) z_t = 0$$

Or

$$a\sum x_t^2 + b\sum x_t z_t = \sum x_t y_t$$

$$a\sum x_t z_t + b\sum z_t^2 = \sum z_t y_t$$

So using Cramer rule

$$a = \frac{\sum z_t^2 \sum x_t y_t - \sum x_t z_t \sum z_t y_t}{\sum x_t^2 \sum z_t^2 - (\sum x_t z_t)^2}$$

$$b = \frac{\sum x_t^2 \sum z_t y_t - \sum x_t z_t \sum x_t y_t}{\sum x_t^2 \sum z_t^2 - (\sum x_t z_t)^2}$$

□ Demonstate that obtained estimator for β is unbiased providing that (2) is correct specification. (*investigation of efficiency and consistency of estimator is not expected here*).

$$b = \frac{\sum x_t^2 \sum z_t y_t - \sum x_t z_t \sum x_t y_t}{\sum x_t^2 \sum z_t^2 - (\sum x_t z_t)^2} = \frac{\sum x_t^2 \sum z_t (\alpha x_t + \beta z_t + u_t) - \sum x_t z_t \sum x_t (\alpha x_t + \beta z_t + u_t)}{\sum x_t^2 \sum z_t^2 - (\sum x_t z_t)^2}$$

$$= \frac{\sum x_t^2 (\alpha \sum z_t x_t) + \beta \sum z_t z_t + \sum z_t u_t) - \sum x_t z_t (\alpha \sum x_t x_t + \beta \sum x_t z_t + \sum x_t u_t)}{\sum x_t^2 \sum z_t^2 - (\sum x_t z_t)^2} =$$

$$= \beta \frac{\sum x_t^2 \sum z_t z_t - (\sum x_t z_t)^2}{\sum x_t^2 \sum z_t^2 - (\sum x_t z_t)^2} + \frac{\sum x_t^2 \sum z_t u_t - \sum x_t z_t \sum x_t u_t}{\sum x_t^2 \sum z_t^2 - (\sum x_t z_t)^2} = \beta + \frac{\sum x_t^2 \sum z_t u_t - \sum x_t z_t \sum x_t u_t}{\sum x_t^2 \sum z_t^2 - (\sum x_t z_t)^2}$$

Now

$$\mathrm{E}\,b = \beta + \frac{\sum x_t^2 \sum z_t \,\mathrm{E}\,u_t - \sum x_t z_t \sum x_t \,\mathrm{E}\,u_t}{\sum x_t^2 \sum z_t^2 - (\sum x_t z_t)^2} = \beta$$

(b) □ Let (1) be a false model while (2) be a true model. A researcher, using ordinary least squares (*OLS*), estimates α from the false model. Demonstrate that obtained estimator of α is generally biased. Find the bias and indicate the factors determining its sign.

$$\hat{\alpha}_{OLS} = \frac{\sum x_t y_t}{\sum x_t^2} = \frac{\sum x_t(\alpha x_t + \beta z_t + u_t)}{\sum x_t^2} = \alpha + \frac{\beta \sum x_t z_t}{\sum x_t^2} + \frac{\sum x_t u_t}{\sum x_t^2}.$$

$$\mathrm{E}[\hat{\alpha}_{OLS}] = \alpha + \frac{\beta \sum x_t z_t}{\sum x_t^2} + \frac{\sum x_t \,\mathrm{E}\,u_t}{\sum x_t^2} = \alpha + \beta \frac{\sum x_t z_t}{\sum x_t^2}.$$

So $\hat{\alpha}_{OLS}$ is a biased estimator of α except in the case when x_t and z_t are **orthogonal** to each other. Bias is $\beta \frac{\sum x_t z_t}{\sum x_t^2}$. Direction of the bias will depend upon the signs of β and $\sum x_t z_t$. Bias will disappear if x_t and z_t are orthogonal to each other ($\sum x_t z_t = 0$).

□ Let now (2) be a false model while (1) be a true model. A researcher, using ordinary least squares (*OLS*), estimates α from the false model. Examine the properties for this *OLS* estimate of α.

There is no bias, but the estimator $\hat{\alpha}_{OLS}$ is not efficient now, as more parameters are estimated therefore, the number of degrees of freedom is reduced. If correlation between x_t and z_t is r then standard error of $\hat{\alpha}_{OLS}$ is multiplied by $\frac{1}{\sqrt{1-r^2}}$ therefore it is possible that the estimated coefficient will be insignificant, estimated value of the coefficient is unpredictable, and it can even change sign. In the limiting case of perfect correlation of x_t and z_t the estimation is impossible (perfect multicollinearity).

SECTION B.

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

Question 4 [25 marks]

The plot of this question is based on a real story that happened at the thesis defense at ICEF many years ago. A student in her thesis was studying the dependence of final econometrics grade E_i (in %) on average homework grade H_i (also in %) using sample data on performance ICEF students,

$$E_i = \beta_1 + \beta_2 H_i + u_i \quad \textbf{(1)}$$

but she mixed up the variables in the final equation and calculated using OLS a regression

$$\hat{H}_i = \hat{\gamma}_1 + \hat{\gamma}_2 E_i \quad \textbf{(2)}$$

The student had to correct her mistake during the defence. She decided to simply express the variable H_i from equation (1)

$$H_i = -\frac{\beta_1}{\beta_2} + \frac{1}{\beta_2} E_i - \frac{1}{\beta_2} u_i \quad \textbf{(3)}$$

and use the value $\hat{\gamma}_2$ as an estimator of $\frac{1}{\beta_2}$ from which she could obtain an estimate of β_2. To round off the story, it is enough to say that the student managed to successfully defend her diploma. We wish you the same success in answering our questions.

For theoretical discussion let us assume that the values of H_i are assumed to be stochastic (say that they are drawn randomly from a population with variance σ_H^2) and disturbance term u satisfies the following assumptions $\mathrm{E}u = 0$, $\mathrm{E}u^2 = \sigma_u^2$, $\mathrm{E}(u_s u_t) = 0, t \neq s$ and $\mathrm{E}(Hu) = 0$.

(a) Demonstrate that $\hat{\gamma}_2$ is an inconsistent estimator of $\frac{1}{\beta_2}$. Find if it is possible, large sample bias and determine its direction.

$$\hat{\gamma}_2 = \frac{\mathrm{Cov}(E,H)}{\mathrm{Var}(E)} = \frac{\mathrm{Cov}(\beta_1 + \beta_2 H + u, H)}{\mathrm{Var}(\beta_1 + \beta_2 H + u)} = \frac{\beta_2 \mathrm{Var}(H) + \mathrm{Cov}(u,H)}{\beta_2^2 \mathrm{Var}(H) + 2\beta_2 \mathrm{Cov}(u,H) + \mathrm{Var}(u)}$$

$$\mathrm{plim}\,\hat{\gamma}_2 = \frac{\beta_2 \mathrm{var}(H) + \mathrm{cov}(H,u)}{\beta_2^2 \mathrm{var}(H) + 2\beta_2 \mathrm{cov}(H,u) + \mathrm{var}(u)} = \frac{\beta_2 \sigma_H^2}{\beta_2^2 \sigma_H^2 + \sigma_u^2} = \frac{\beta_2}{\beta_2^2}\left(\frac{\sigma_H^2}{\sigma_H^2 + \frac{1}{\beta_2^2}\sigma_u^2}\right) = \frac{1}{\beta_2}\left(\frac{1}{1 + \frac{1}{\beta_2^2}\frac{\sigma_u^2}{\sigma_H^2}}\right)$$

since $\mathrm{cov}(H,u) = \mathrm{E}(Hu) = 0$. The estimator will tend to underestimate $\frac{1}{\beta_2}$ in absolute terms

(b) During the discussion one of the members of defence comittee advised the student to use for estimation $\hat{\gamma}_2$ a third variable A_i (attendence of seminars) that is correlated with E_i but independent of u_i. Demonstrate that if the student had regressed H_i on E_i using A_i as an instrument for E_i, the slope coefficient $\hat{\gamma}_2^{IV}$ would have been a consistent estimator of $\frac{1}{\beta_2}$.

$$\hat{\gamma}_2^{IV} = \frac{\mathrm{Cov}(H,A)}{\mathrm{Cov}(E,A)} = \frac{\mathrm{Cov}(H,A)}{\mathrm{Cov}(\beta_1 + \beta_2 H + u, A)} = \frac{\mathrm{Cov}(H,A)}{\beta_2 \mathrm{Cov}(H,A) + \mathrm{Cov}(u,A)}$$

$$\mathrm{plim}\,\hat{\gamma}_2^{IV} = \frac{\mathrm{cov}(H,A)}{\beta_2 \mathrm{cov}(H,A) + \mathrm{cov}(u,A)} = \frac{\mathrm{cov}(H,A)}{\beta_2 \mathrm{cov}(H,A)} = \frac{1}{\beta_2}$$

(c) One of the students from the public also took part in discussion and suggested to use TSLS approach instead saying that it has many advantages compared to the instrumental variable approach. Demonstate that in case of one instrument A_i this approach has no additional advantages as $\hat{\gamma}_2^{TSLS} = \hat{\gamma}_2^{IV}$.

In TSLS at the first stage E_i is regressed on A_i to get estimation of coefficient α_2 in a regression $E_i = \alpha_1 + \alpha_2 A_i + v_i$: $\alpha_2^{OLS} = \dfrac{\mathrm{Cov}(E, A)}{\mathrm{Var}(A)}$; and to get then estimated values of $\hat{E}_i = \alpha_1 + \alpha_2^{OLS} A_i$. At the second stage the regression $H_i = \gamma_1 + \gamma_2 \hat{E}_i + w_i$ is estimated. The OLS estimator of γ_2 is

$$\hat{\gamma}_2^{\mathrm{OLS}} = \frac{\mathrm{Cov}(H, \hat{E})}{\mathrm{Var}(\hat{E})} = \frac{\mathrm{Cov}(H, \alpha_1 + \alpha_2^{OLS} A)}{\mathrm{Var}(\alpha_1 + \alpha_2^{OLS} A)} = \frac{\alpha_2^{OLS}\,\mathrm{Cov}(H, A)}{(\alpha_2^{OLS})^2\,\mathrm{Var}(A)} = \frac{\mathrm{Cov}(H, A)}{\alpha_2^{OLS}\,\mathrm{Var}(A)}.$$

Substituting $\alpha_2^{OLS} = \dfrac{\mathrm{Cov}(E, A)}{\mathrm{Var}(A)}$ into the last expression we get

$$\hat{\gamma}_2^{\mathrm{TSLS}} = \frac{Cov(H, A)}{\dfrac{Cov(E, A)}{\mathrm{Var}(A)}\mathrm{Var}(A)} = \frac{Cov(H, A)}{Cov(E, A)} = \hat{\gamma}_2^{IV},$$ so $\hat{\gamma}_2^{TSLS} = \hat{\gamma}_2^{IV}$.

So if A_i is valid instrument both approaches (IV and TSLS) to estimation give identical results

Question 5. (25 marks)

The researcher investigates the relationship between ex-spouses for divorced couples with children. The following models jointly determine monthly child support payments S and monthly visitation rights V (we assume that children live with their mothers, so that fathers pay child support):

$$\begin{aligned} S &= \alpha_1 + \alpha_2 V + \alpha_3 I + \varepsilon_1 \\ V &= \beta_1 + \beta_2 S + \varepsilon_2 \end{aligned} . \qquad \textbf{(1)}$$

$$\begin{aligned} S &= \alpha_1 + \alpha_2 V + \alpha_3 I + \alpha_4 D + \varepsilon_1 \\ V &= \beta_1 + \beta_2 S + \beta_3 D + \varepsilon_2 \end{aligned} , \qquad \textbf{(2)}$$

$$\begin{aligned} S &= \alpha_1 + \alpha_2 V + \alpha_3 I + \alpha_4 F + \alpha_5 D + \varepsilon_1 \\ V &= \beta_1 + \beta_2 S + \beta_3 M + \beta_4 D + \varepsilon_2 \end{aligned} \qquad \textbf{(3)}$$

The system of equation (2) includes also variable D (miles currently between the mother and father's residence). System of equations (3) includes two additional dummy variables: F (binary indicator if father remarried), and; M is a binary indicator for whether the woman is remarried. The variables I, D, F and M are considered as exogenous. Disturbance terms of all equations are assumed to be uncorrelated with exogenous variables and with each other.

(a) □ What does it mean that equation is identified (underidentified)? Examine the identification of each structural equation in the system (1) and (2) (using instrumental variables approach) and (3) (using order condition).

The equation is called identified (or exactly identified) if it is possible to get consistent estimates for all its parameters, and underidentified if not.

System (1): The first equation is underidentified as there is no available instruments for endogenous variable *V*. The second equation is exactly identified as there exactly one available instrument *I* for endogenous variable *S*.

System (2): The situation with both equations does not change as additional exogenous variable *dist* is added at the same time into both equations and so it cannot be used as additional instrument neither in the first nor in the second equation as it is present there on its own rights.

System (3): Use order condition. Let $G = 2$ denote the number of equations in the system of simultaneous equations and the number of endogenous variables. The first equation exclude exactly $G-1=1$ variable (M) and so it is likely exactly identified. The second equation excludes $2 > G-1=1$ variables (F and I), and so it is likely overidentified.

□ What are consequences of using OLS to estimate second equation of the system (1)? Give some arguments to your opinion? (*Mathematical proof of insconsistensy is not expected here and gives no mark*).

If we express all endogenous variables through exogenous variable and random terms (this is called the reduced form of system (1)),

$$S = \alpha_1 + \alpha_2 V + \alpha_3 I + \varepsilon_1$$

$$V = \beta_1 + \beta_2 S + \varepsilon_2$$

$$S = \alpha_1 + \alpha_2(\beta_1 + \beta_2 S + \varepsilon_2) + \alpha_3 I + \varepsilon_1$$

$$V = \beta_1 + \beta_2(\alpha_1 + \alpha_2 V + \alpha_3 I + \varepsilon_1) + \varepsilon_2$$

$$(1-\alpha_2\beta_2)S = \alpha_1 + \alpha_2\beta_1 + \alpha_3 I + \varepsilon_1 + \alpha_2\varepsilon_2$$

$$(1-\alpha_2\beta_2)V = \beta_1 + \alpha_1\beta_2 + \alpha_3\beta_2 I + \beta_2\varepsilon_1 + \varepsilon_2$$

$$S = \frac{1}{(1-\alpha_2\beta_2)}(\alpha_1 + \alpha_2\beta_1 + \alpha_3 I + \varepsilon_1 + \alpha_2\varepsilon_2)$$

$$V = \frac{1}{(1-\alpha_2\beta_2)}(\beta_1 + \alpha_1\beta_2 + \alpha_3\beta_2 I + \beta_2\varepsilon_1 + \varepsilon_2)$$

we can see that both endogenous variables include the same random terms, which means that the explanatory variable of each equation in structural form system

$$S = \alpha_1 + \alpha_2 V + \alpha_3 I + \varepsilon_1$$
$$V = \beta_1 + \beta_2 S + \varepsilon_2 .$$

correlates with the random term of equation, which is a violation of the Gauss-Markov condition.

(b) □ What is the difference between instrumental variable (instrument) and proxy variable? What are conditions to use some variable as an instrument for estimation?

Proxy variables are used instead of missed and unavailable explanatory variables that are important in explanation of fluctuation of the dependent variable. The aim of introducing proxy variable is to prevent omitted variable bias. The instrumental variable is introduced instead of explanatory variable to prevent or mitigate (slack) the violation of Gauss-Markov condition when stochastic explanatory variable is correlated with the disturbance term of the equation.

Using instrument is based on the following assumptions: a variable is a valid instrument if it correlates with instrumented explanatory variable but not correlated with disturbance term of the equation for dependent variable.

□ Why can't IV method be used to estimate the coefficient α_2 in (2)?

There are no available instruments to estimate α_2 in (2). Both exogenous variables are already present in the first equation of system (2).

□ Prove that IV estimator of β_2, in system (1) based on using instrument I is consistent.

$$\hat{\beta}_2^{\mathrm{IV}} = \frac{\mathrm{Cov}(V,I)}{\mathrm{Cov}(S,I)} = \frac{\mathrm{Cov}(\beta_1+\beta_2 S+\varepsilon_2, I)}{\mathrm{Cov}(S,I)} = \frac{\mathrm{Cov}(\beta_1, I)+\beta_2\,\mathrm{Cov}(S,I)+\mathrm{Cov}(I,\varepsilon_2)}{\mathrm{Cov}(S,I)} = \beta_2 + \frac{\mathrm{Cov}(I,\varepsilon_2)}{\mathrm{Cov}(S,I)}$$

$$\mathrm{plim}\,\hat{\beta}_2^{\mathrm{IV}} = \beta_2 + \frac{\mathrm{plim}\,\mathrm{Cov}(I,\varepsilon_2)}{\mathrm{plim}\,\mathrm{Cov}(S,I)} = \beta_2 + \frac{\mathrm{cov}(I,\varepsilon_2)}{\mathrm{cov}(S,I)} = \beta_2 + \frac{0}{\mathrm{cov}(S,I)} = \beta_2$$ (I and ε_2 are uncorrelated).

(c) □ How to use TSLS method to estimate coefficient β_2 in (1). Prove that the TSLS method gives the same coefficient estimate of β_2 in (1) as the IV method.

In TSLS at the first stage S is regressed on I to get estimation of coefficient γ_2 in a regression $S_i = \gamma_1 + \gamma_2 I_i + v_i$: $\gamma_2^{OLS} = \dfrac{\mathrm{Cov}(S,I)}{\mathrm{Var}(I)}$; and to get then estimated values of $\hat{S}_i = \gamma_1 + \gamma_2^{OLS} I_i$. At the second stage the regression $V_i = \beta_1 + \beta_2 \hat{S}_i + w_i$ is estimated. The OLS estimator of β_2 is

$$\beta_2^{\mathrm{TSLS}} = \frac{\mathrm{Cov}(V,\hat{S})}{\mathrm{Var}(\hat{S})} = \frac{\mathrm{Cov}(V,\gamma_1+\gamma_2^{OLS} I)}{\mathrm{Var}(\gamma_1+\gamma_2^{OLS} I)} = \frac{\gamma_2^{OLS}\,\mathrm{Cov}(V,I)}{(\gamma_2^{OLS})^2\,\mathrm{Var}(I)} = \frac{\mathrm{Cov}(V,I)}{\gamma_2^{OLS}\,\mathrm{Var}(I)} .$$

Substituting $\gamma_2^{OLS} = \dfrac{Cov(S,I)}{\mathrm{Var}(I)}$ into the last expression we get $\hat{\beta}_2^{\mathrm{TSLS}} = \dfrac{Cov(V,I)}{\dfrac{Cov(S,I)}{\mathrm{Var}(I)}\mathrm{Var}(I)} = \dfrac{Cov(V,I)}{Cov(S,I)} = \hat{\beta}_2^{IV}$,

so $\hat{\beta}_2^{\mathrm{TSLS}} = \hat{\beta}_2^{IV}$.

□ Is the same true for estimation the coefficient β_2 in (3)? Which method is preferable here and why?

No, they are not the same here. The point is that there are two different instruments I and F for the estimation of β_2 in (3), which give different numerical values, which causes a conflict of estimates. Using TSLS allows us to resolve this conflict by obtaining the best in terms of efficiency compromise between these estimates.