The International College of Economics and Finance
Econometrics – 2021-2022. Midterm exam. 2021 December 23.

Part 2. Free Response Questions (1 hour 30 minutes)

SECTION A

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

Question 1. (25 marks)
The superviser set the task to his student to investigate the effects of background characteristics and admission assessment scores on the performance of students in the final examinations in their university. The student estimated the following equation using Ordinary Least Squares:

$$\hat{FS} = 53.89 + 0.03L + 0.05M + 0.06INT - 0.04PR + 0.17MALE + 0.06PR * MALE \quad n = 325,\ R^2 = 0.06 \quad \textbf{(1)}$$

where FS is the average finals score (the outcome), L and M are the pre-admission language and math test scores respectively, INT is the pre-admission interview score (all scores are measured in % of 100), PR indicates whether the student attended a private school ($1 = \text{yes},\ 0 = \text{no}$), and $MALE$ indicates whether the student is male ($1 = \text{yes},\ 0 = \text{no}$).

(a) [13 marks] □ Using dummy variables allows you to write equations separately for each category. Using equation **(1)** as a basis, write an equation for the following category: girls in public schools (denote this equation **(2)**). Using the same principle, write an equation for another category: young men in private schools (equation **(3)**).

Girls in public schools: $\hat{FS} = 53.89 + 0.03L + 0.05M + 0.06INT$ **(2)**

Young men in private schools: $\hat{FS} = 54.08 + 0.03L + 0.05M + 0.06INT$ **(3)**

□ What is gender performance gap measured as the difference between male and female average final scores? How to test whether gender has a significant impact on students' finals performance? Indicate what information you would need to enable you to implement this test.

The gender performance gap is measured as the difference between male and female average final scores FS
Let $FS = \theta_0 + \theta_1 L + \theta_2 M + \theta_3 INT + \theta_4 PR + \theta_5 MALE + \theta_6 PR * MALE + u$ (notations from b). For state schools it is $\hat{\theta}_5 = 0.17$ For private schools $\hat{\theta}_5 + \hat{\theta}_6 = 0.17 + 0.06 = 0.23$. To test for significance use F-test $H_0 : \theta_5 = 0$ and $\theta_6 = 0$ vs. $H_1 : \theta_5 \neq 0$ or $\theta_6 \neq 0$. We need the R_R^2 of a regression of FS on a constant, L, M, I and PR (the restricted model) $FS = \theta_0 + \theta_1 L + \theta_2 M + \theta_3 INT + \theta_4 PR + v$. Compute the test statistic:

$$F = \frac{R_{ur}^2 - R_r^2}{1 - R_{ur}^2} \times \frac{n-7}{2} = \frac{0.06 - R_r^2}{1 - 0.06} \times \frac{325 - 7}{2}.$$

- Under Gauss-Markov assumptions plus normality, $F \sim F_{2,318}$. At chosen signicance level, reject H_0 if $F > F(cr., 2, 318)$.

□ Is Equation **(1)** overall statistically significant?

$H_0 : \theta_1 = 0, \theta_2 = 0, ..., \theta_6 = 0$ against 'at least one of these coeffitients is not zero'.

$$F = \frac{R^2 / 6}{(1 - R^2)/(325 - 7)} = \frac{0.06/6}{0.94/318} = 3.38 > F_{cr}^{1\%}(6, 318) = 2.86.\ \text{Reject.}$$

(b) [12 marks] □ Suppose students who did not attend a private school, attended a state school (dummy variable ST is defined as being equal 1 if a student attended state school, and 0 otherwise.

Let $\hat{\theta}_s$ indicate the estimated parameters from the initial regression

$$\hat{FS} = \hat{\theta}_0 + \hat{\theta}_1 L + \hat{\theta}_2 M + \hat{\theta}_3 INT + \hat{\theta}_4 PR + \hat{\theta}_5 MALE + \hat{\theta}_6 PR * MALE$$

Now let instead the regression be estimated using OLS

$$\hat{FS} = \hat{\beta}_0 + \hat{\beta}_1 L + \hat{\beta}_2 M + \hat{\beta}_3 INT + \hat{\beta}_4 ST + \hat{\beta}_5 MALE + \hat{\beta}_6 ST * MALE$$

where ST indicates whether the student attended a state school ($1 = \text{yes},\ 0 = \text{no}$).

Express the parameters $\hat{\beta}_r$ in terms of $\hat{\theta}_s$ and, if possible, find their numerical values,

$$\hat{FS} = \hat{\theta}_0 + \hat{\theta}_1 L + \hat{\theta}_2 M + \hat{\theta}_3 INT + \hat{\theta}_4 PR + \hat{\theta}_5 MALE + \hat{\theta}_6 PR \times MALE_i =$$

$$= \hat{\theta}_0 + \hat{\theta}_1 L + \hat{\theta}_2 M + \hat{\theta}_3 INT + \hat{\theta}_4 (1 - ST) + \hat{\theta}_5 MALE + \hat{\theta}_6 (1 - ST) \times MALE =$$

$$= (\hat{\theta}_0 + \hat{\theta}_4) + \hat{\theta}_1 L + \hat{\theta}_2 M + \hat{\theta}_3 INT - \hat{\theta}_4 ST_i + (\hat{\theta}_5 + \hat{\theta}_6) MALE - \hat{\theta}_6 ST \times MALE\ ,$$

$$\hat{\beta}_0 = \hat{\theta}_0 + \hat{\theta}_4 = 53.89 - 0.04 = 53.85\,.$$

$$\hat{\beta}_4 = -\hat{\theta}_4 = -(-0.04) = 0.04\,.$$

$$\hat{\beta}_6 = -\hat{\theta}_6 = -0.06\,.$$

$$\hat{\beta}_5 = \hat{\theta}_5 + \hat{\theta}_6 = 0.17 + 0.06 = 0.23\,.$$

$\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ and $\hat{\beta}_4$ remain the same from (1). So the final new equation is

$$\hat{FS} = 53.85 + 0.03L + 0.05M + 0.06INT + 0.04ST + 0.23MALE - 0.06ST * MALE$$

□ Discuss any problem you may have in estimating the model if all males in your sample have attended a private school prior to enter university. What can you do to mitigate this problem?

Solution:

If all males attended an independent school, then $MALE = MALE \times PR$, so the variables will be perfectly collinear and he would not be able to obtain the OLS estimates from (1). In this case, we could either drop the interaction term from the regression or obtain more data such that some men that have not attended independent school will also be in the sample.

Question 2. (25 marks) An ICEF student, worried about the incidence of COVID-19, decided to independently figure out the factors that could affect the chances of contracting COVID. Using the Internet, he interviewed his classmates and received 40 responses, which allowed him to generate data on the following variables: ***COVID*** – covid disease during the last academic year (BINARY, 1 if yes), ***VAC*** – vaccination (DUMMY, 1 if vaccinated), OUT - time spent with friends outside of ICEF facilities (clubs, restaurants, in hours per week), ***TR*** – time spent on public transport (in hours per week), ***M*** (DUMMY, 1 if male), ***FIT*** – playing sports in fitness centers (DUMMY , 1 if yes), ***DORM*** – accommodation in HSE dormitory (DUMMY, 1 if yes). Based on the data obtained, he calculated three models (dependent variable ***COVID***)

$$\begin{aligned}&COVID_i = 0.25 - 0.44VAC_i + 0.006OUT_i + 0.024TR_i + 0.054M_i + 0.024FIT_i + 0.28DORM_i + e_i\\ &\textbf{(OLS)}\quad (0.15)\ (0.13)\qquad (0.006)\qquad (0.014)\qquad (0.12)\qquad (0.14)\qquad (0.13)\ \ R^2 = 0.6\end{aligned} \qquad \textbf{(1)}$$

$$\begin{aligned}&COVID_i = -3.38 - 4.99VAC_i + 0.05OUT_i + 0.48TR_i + 0.27M_i - 2.45FIT_i + 3.08DORM_i + e_i\\ &\textbf{(LOGIT)}\ \ (1.74)\ \ (2.25)\qquad (0.06)\qquad (0.24)\qquad (1.32)\quad (2.35)\qquad (1.56)\ McFaddenR^2 = 0.63\end{aligned} \qquad \textbf{(2)}$$

$$\begin{aligned}&COVID_i = -0.66 - 4.05VAC_i + 0.74M_i + 1.16FIT_i + 2.67DORM_i + e_i\\ &\textbf{(LOGIT)}\ \ (0.77)\ \ (1.45)\qquad (1.07)\quad (1.49)\qquad (1.19)\ \ McFaddenR^2 = 0.51\end{aligned} \qquad \textbf{(3)}$$

(a) [13 marks] □ Can we draw some practical conclusions from the analysis of coefficients for ***TR***, ***OUT***, ***M*** and ***FIT*** in regression **(1)**? Does model (2) support them?

0.024 is the marginal effect of TR on probability to get ill (oth. var. unchanged), and so other coefficients. All of them are insignificant $t_{cr}^{5\%}(33) \approx 2.037$ what prevents to draw any meaningful conscusions. Eq.(2) supports this except that the coefficient of TR is significant using z-statistic $0.48/0.24 = 2 > 1.96 = z_{cr}^{5\%}$.

□ Comment on the coefficients of ***VAC*** and ***DORM*** variables in equation **(1)**. What practical conclusions can a student draw based on the analysis of these coefficients, given that he lives in a dormitory and has not yet been vaccinated. Does model **(2)** support the conclusions drawn from model **(1)**?

-0.44 means that being vaccinated reduces probability to get ill by 44 p.p. and this effect is significant at 1% $(0.44/0.13 = 3.38 > t_{cr}^{1\%}(33) \approx 2.738)$. Living in dorm increases the prob. to get ill by 28 p.p. (sign. only at 5%). Model (2) generally supports this (we take imto account only signs of coefficients and their significance using z-statistic). So certainly student should think of moving from dorm and vaccination.

□ In model **(2)** (Logit) McFadden R-squared is slightly higher than R-squared in model **(1)** and higher than in the model **(3)**. What conclusions can be drawn from here?

No conclusion can be drawn from this as R-squared and McFadden R-squared are based on different principles and McFadden R-squared has nothing to do with the percentage of explained variance.

□ Is it possible to evaluate the significance of the ***OUT*** and ***TR*** variables taken together on the basis of the information on McFadden R-squared in **(2)** and **(3)**, if it is known additionally that Restricted log likelihood for the logit model is -26.92? What conclusions follow from here?

$McFaddenR^2(2) = 1 - \dfrac{\log L(2)}{\log L_0} = 0.63$, $McFaddenR^2(3) = 1 - \dfrac{\log L(3)}{\log L_0} = 0.51$ where $\log L_0 = -26.92$. So

$\log L(2) = 0.37 * \log L_0 = 0.37 * (-26.92) = -9.96$

$\log L(3) = 0.48 * \log L_0 = 0.49 * (-26.92) = -13.19$

Now $LR = 2(\log L(2) - \log L(3)) = 2(-9.96 + 13.19) = 6.46 > 5.99 = \chi_{crit}^2(2,\ 5\%)$, so null is rejected and the student is recommended to move out of dorm and/or be vaccinated.

□ Are equations (1), (2), (3) significant?

For (1) $F = R^2/6/(1-R^2) \cdot 33 = 0.6/6/0.4 \cdot 33 = 8.25 > 3.37 = F_{cr}^{1\%}(6,33)$.

$LR(2) = 2(\log L(2) - \log L_0) = 2(-9.96 - (-26.92)) = 33.92 > 16.812 = \chi_{crit}^2(6,\ 1\%)$

$LR(3) = 2(\log L(3) - \log L_0) = 2(-13.19 - (-26.92)) = 27.46 > 16.812 = \chi^2_{crit}(6,\ 1\%)$

b) [12 marks] □ The student plans to move from the dorm to a rented apartment to reduce chance of infection. Help the student calculate by how much it will help to reduce probability of being indected using any method. Use the following information: the student is not vaccinated, is engaged in fitness, uses public transport an average 14 hours a week, and likes to hang out with friends an average of 20 hours a week.

$z = -3.38 - 4.49 \cdot 0 + 0.05 \cdot 20 + 0.48 \cdot 14 + 0.27 \cdot 1 - 2.45 \cdot 1 + 3.08 \cdot 1 = 5.24$, $P = 1/(1 + e^{-5.24}) = 0.995$.

$z = -3.38 - 4.49 \cdot 0 + 0.05 \cdot 20 + 0.48 \cdot 14 + 0.27 \cdot 1 - 2.45 \cdot 1 + 3.08 \cdot 0 = 2.16$, $P = 1/(1 + e^{-2.16}) = 0.897$.

$\Delta P = 0.995 - 0.897 = 0.098$. As $DORM$ is a dummy this approach is more appropriate.

Alternative approach: $P = e^{-5.24}/(1 + e^{-5.24})^2 \cdot 3.08 = 0.016$ or $P = e^{-2.16}/(1 + e^{-2.16})^2 \cdot 3.08 = 0.28$

□ Will this effect change if he gets vaccinated?

$z = -3.38 - 4.49 \cdot 1 + 0.05 \cdot 20 + 0.48 \cdot 14 + 0.27 \cdot 1 - 2.45 \cdot 1 + 3.08 \cdot 1 = 0.25$, $P = 1/(1 + e^{-0.25}) = 0.56$.

$z = -3.38 - 4.49 \cdot 1 + 0.05 \cdot 20 + 0.48 \cdot 14 + 0.27 \cdot 1 - 2.45 \cdot 1 + 3.08 \cdot 0 = -2.83$, $P = 1/(1 + e^{2.83}) = 0.056$.

$\Delta P = 0.56 - 0.056 = 0.5$. Again as $DORM$ is a dummy this approach is more appropriate.

Alternative approach: $P = e^{-0.25}/(1 + e^{-0.25})^2 \cdot 3.08 = 0.76$ or $P = e^{2.83}/(1 + e^{2.83})^2 \cdot 3.08 = 0.16$.

SECTION B.

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

Question 3. (25 marks) A student doing an internship at a large analytical company received an assignment to investigate the relationship of GDP per capita in 2020 (***GDPPC***20 – with an average value 20953.12) on the GDP per capita in 2019 (***GDPPC***19 – with an average value 21803.18, both in dollars) using data on 170 countries (World Bank data, several outliers – small island states – removed).

(a) [8 marks]. First she runs simple regression model (standard errors in parenthesis)

$$GDPPC20_i = -127.13 + 0.967 GDPPC19_i + e_i \quad R^2 = 0.998 \qquad \textbf{(1)}$$

$$(112.42) \quad (0.004)$$

□ What is the economic meaning of the constructed equation? Can you trust the discovered pattern?

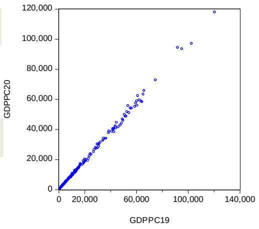
0.967 is marginal effect of $GDPPC19$ on $GDPPC20$, it is significant but possibly biased as only one varible is included into equation.

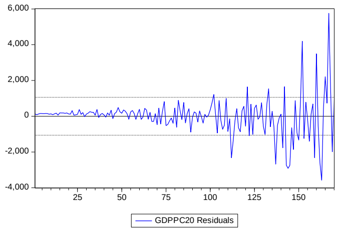
□ Why should the student be wary of possible heteroscedasticity?

Large countries are characterized by a higher level of GDP and, accordingly, by larger deviations from the general trend (larger residuals).

□ Explain to the student how heteroscedasticity can be found using simple graphical tools: scatter diagram and residual plot in the context of the problem under consideration. How to get these pattern from the data.

Here on the graph are typical picture of heteroscedasticity on a scatter diagram and residual plot (to reveal the picture on the residual plot all observations first should be arranged in ascending order of $GDPPC19$.





□ Assume heteroscedasticity is detected. How this change the conclusions from estimated equation **(1)**?

The estimates of the coefficients remain unbiased, but are no longer efficient, this is manifested in the growth of standard errors of estimates, but the standard formula for them is now incorrect, and therefore we usually observe the opposite picture of the decrease in standard errors. Therefore, the tests become incorrect, in particular, the significance of the coefficients is usually overestimated.

(b) [8 marks]. The student performs White's test by calculating White's auxiliary equation

$$RESID^2{}_i = -3627554.6 + 54.87 GDPPC19_i + 0.00028 GDPPC19_i^2 + e_i \quad R^2 = 0.265 \qquad \textbf{(2)}$$

$$(380354.0) \ (25.78) \qquad\qquad (0.0003)$$

□ Help the student get the most out of this equation. Is heteroscedasticity detected?

$n \cdot R^2 = 170 \cdot 0.265 = 45.05 > \chi^2_{cr,1\%}(2) = 9.21$ – the null of no heteroscedasticity is rejected. t-statistic $54.87/25.78 = 2.128 > t_{cr}^{5\%}(150) = 1.976$ so $GDPPC19$ may well turn out to be a factor of heteroscedasticity.

□ The student run the regression in specification **(1)** using two different samples: 1st – 60 developing countries with lowest values of $GDPPC19_i$ (RSS = 929275.8), and 2nd – 35 developed countries with greatest values of $GDPPC19_i$ (RSS =1.34E+08). Continue Goldfeld-Quandt test and interpret its results

Assuming $\sigma_u = k \cdot GDPPC19$ evaluate $F = (RSS(big)/(n_1 - k))/(RSS(small)/(n_2 - k)) =$
$= (1.23 \cdot 10^8 / 33)/(929275.8/58) = 3.62 \cdot 10^6 > F_{cr}^{1\%}(33, 58) \approx 1.98$ – reject null of no heteroscedasticity.

(c) [9 marks]. Calculation with WLS (weights $w_i = 1/GDPPC19_i$) gives

$$GDPPC20_i = 20.8 + 0.960 GDPPC19_i + e_i \quad R^2 = 0.998$$

(11.52) (0.00336) **(3)**

[10.02] [0.00340]

(standard errors in parenthesis, White heteroscedasticity consistent standard errors in square brackets)

□ Explain how this equation was obtained? Why can we be sure that in this equation the problem of heteroscedasticity is eliminated or mitigated? What is the difference between regular and heteroscedasticity consistent standard errors (some math expected).

The equation $GDPPC20_i = \beta_1 + \beta_2 GDPPC19_i + u_i$ was devided by $GDPPC19_i$ so equation $\frac{GDPPC20_i}{GDPPC19_i} = \frac{\beta_1}{GDPPC19_i} + \beta_2 + \frac{u_i}{GDPPC19_i}$ was obtained. This allows you to exclude or reduce the risk of heteroscedasticity. Let us show this. Let $X_i = GDPPC19_i$. Then under assumption $\sigma(u_i) = X_i \cdot \sigma_u$ after dividing all members of equation 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 heteroscedasticity consistent (HEC) standard errors in White form $\sigma^2_{\hat{\beta}_2} = \frac{\sum_{i=1}^{n} x_i^2 \sigma_i^2}{\left(\sum_{j=1}^{n} x_j^2\right)^2}$ instead of regular standard errors $\sigma^2_{\hat{\beta}_2} = \frac{\sigma_u^2}{\sum \left(X_i - \bar{X}\right)^2}$ allows to ensure significance tests be correct.

□ R^2 for the auxiliary equation of White's test similar to (2) is 0.006861. Comment on this.

$n \cdot R^2 = 170 \cdot 0.006861 = 1.7 < \chi^2_{cr,5\%}(2) = 5.991$ – the null of no heteroscedasticity is not rejected.

□ Finally the student constructs a logarithmic regression (OLS) ($R^2 = 0.0169$ for White test))

$$\log GDPPC20_i = 0.006 + 0.9953 \log GDPPC19_i + e_i \quad R^2 = 0.999$$

(0.021) (0.0023) **(4)**

What is the interpretation of the slope coefficient. Obtain from here the value of the corresponding marginal effect and compare it with the magnitude of the marginal effect calculated using equations **(1)** and **(3)**. Which one can be more trusted? Comment on your findings.

0.9953 is the elasticity of $GDPPC20$ on $GDPPC19$ $e = \frac{dY}{dX} \cdot \frac{X}{Y} \Rightarrow \frac{dY}{dX} = e \cdot \frac{Y}{X}$ so the marginal effect is now not constant. A reasonable estimate of the constant marginal effect corresponding to linear equation (1) can be obtained using information on the averages $\overline{GDPPC20} = 20953.12$ and $\overline{GDPPC19} = 21803.18$: $\frac{dY}{dX} = e \cdot \frac{\bar{Y}}{\bar{X}} = 0.9953 \cdot \frac{20953.12}{21803.18} = 0.9566$. Both equations (4) and (3) give less estimate of the marginal effect so probably in equation (1) it was overestimeted (0.965).

Question 4. (25 marks) The student explores the relationship between gas prices and gas demand in Europe in connection with the completion of the Nord Stream-2 gas pipeline. In year *t*, aggregate demand for a gas in Europe, Q_{Dt}, is related to its price, P_t, and also to the aggregate income, Y_t, which is supposed to be exogenous:

$$Q_{Dt} = \beta_1 + \beta_2 P_t + \beta_3 Y_t + u_{Dt} \qquad (1)$$

Aggregate supply of gas in year *t*, Q_{St}, is also a function of P_t:

$$Q_{St} = \alpha_1 + \alpha_2 P_t + u_{St} \qquad (2)$$

u_{Dt} and u_{St} are disturbance terms that satisfy the Gauss–Markov conditions and are distributed independently of each other. The market clears in each year, so that $Q_{Dt} = Q_{St}$.

For the purposes of this question, any problems associated with nonstationary time series may be ignored.

(a) [9 marks] □ Derive reduced form equations and explain why OLS would not be appropriate for estimation parameters of both equations..

To derive the reduced form equation for *P* it is sufficient to deduct equation (2) from equation (1) taking into account equality $Q_{Dt} = Q_{St}$

$$Q_{Dt} = \beta_1 + \beta_2 P_t + \beta_3 Y_t + u_{Dt} \qquad (1)$$

$$Q_{St} = \alpha_1 + \alpha_2 P_t + u_{St} \qquad (2)$$

$$Q_{Dt} - Q_{St} = (\beta_1 - \alpha_1) + (\beta_2 - \alpha_2)P_t + \beta_3 Y_t + u_{Dt} - u_{St},$$

so the reduced form equation for *P* is $P_t = \dfrac{\beta_1 - \alpha_1}{\alpha_2 - \beta_2} + \dfrac{\beta_3 Y_t}{\alpha_2 - \beta_2} + \dfrac{u_{Dt} - u_{St}}{\alpha_2 - \beta_2}$

To get the reduced form equation for Q it is convenient to multiply equation (1) by α_2, equation (2) – by β_2 and then subtract them taking into account that $Q_{Dt} = Q_{St} = Q_t$

$$\alpha_2 Q_t = \alpha_2\beta_1 + \alpha_2\beta_2 P_t + \alpha_2\beta_3 Y_t + \alpha_2 u_{Dt} \qquad (1')$$

$$\beta_2 Q_t = \alpha_1\beta_2 + \alpha_2\beta_2 P_t + \beta_2 u_{St} \qquad (2')$$

$$(\alpha_2 - \beta_2)Q_t = (\alpha_2\beta_1 - \alpha_1\beta_2) + \alpha_2\beta_3 Y_t + (\alpha_2 u_{Dt} - \beta_2 u_{St}) \qquad (1') - (2')$$

So $$Q_t = \frac{\alpha_2\beta_1 - \alpha_1\beta_2}{\alpha_2 - \beta_2} + \frac{\alpha_2\beta_3}{\alpha_2 - \beta_2} Y_t + \frac{\alpha_2 u_{Dt} - \beta_2 u_{St}}{\alpha_2 - \beta_2}\ .$$

One can see from reduced form system that *P* is partly determined by both u_{Dt} and u_{St}, the Gauss–Markov condition that the explanatory variables are distributed independently of the disturbance term is violated for both equations and as a consequence OLS yields inconsistent estimates.

□ Use reduced form equations to find large sample bias when OLS is applied for estimation of α_2 in equation **(2)**.

OLS estimator of α_2 is $\hat{\alpha}_2^{IV} = \dfrac{\mathrm{Cov}(Q_t, P_t)}{\mathrm{Var}(P_t)}$

At the first sage we use equation $Q_{St} = \alpha_1 + \alpha_2 P_t + u_{St}$

$$\hat{\alpha}_2^{IV} = \frac{\mathrm{Cov}(Q_t, P_t)}{\mathrm{Var}(P_t)} = \frac{\mathrm{Cov}(\alpha_1 + \alpha_2 P_t + u_{St}, P_t)}{\mathrm{Var}(P_t)} = \alpha_2 \frac{\mathrm{Var}(P_t)}{\mathrm{Var}(P_t)} + \frac{\mathrm{Cov}(u_{St}, P_t)}{\mathrm{Var}(P_t)} = \alpha_2 + \frac{\mathrm{Cov}(u_{St}, P_t)}{\mathrm{Var}(P_t)}$$

Now use reduced system equation $P_t = \dfrac{1}{\alpha_2 - \beta_2}(\beta_1 - \alpha_1 + \beta_3 Y_t + u_{Dt} - u_{St})$

$$\hat{\alpha}_2^{IV} = \alpha_2 + \frac{\dfrac{1}{\alpha_2 - \beta_2}\mathrm{Cov}(u_{St}, \beta_1 - \alpha_1 + \beta_3 Y_t + u_{Dt} - u_{St})}{\dfrac{1}{(\alpha_2 - \beta_2)^2}\mathrm{Var}(\beta_1 - \alpha_1 + \beta_3 Y_t + u_{Dt} - u_{St})} =$$

$$= \alpha_2 + (\alpha_2 - \beta_2)\frac{\beta_3\,\mathrm{Cov}(u_{St}, Y_t) + \mathrm{Cov}(u_{St}, u_{Dt}) - \mathrm{Cov}(u_{St}, u_{St})}{\beta_3^2\,\mathrm{Var}(Y_t) + \mathrm{Var}(u_{Dt}) + \mathrm{Var}(u_{St}) + 2\,\mathrm{Cov}(Y_t, u_{Dt}) - 2\,\mathrm{Cov}(Y_t, u_{St}) - 2\,\mathrm{Cov}(u_{St}, u_{Dt})}$$

$$\text{plim}\,\hat{\alpha}_2^{IV} = \alpha_2 + (\alpha_2 - \beta_2)\frac{\beta_3\sigma_{u_{St},Y_t} + \sigma_{u_{St},u_{Dt}} - \sigma^2_{u_{St}}}{\beta_3^2\sigma^2_{Y_t} + \sigma^2_{u_{Dt}} + \sigma^2_{u_{St}} + 2\sigma_{Y_t,u_{Dt}} - 2\sigma_{Y_t,u_{st}} - 2\sigma_{u_{St},u_{Dt}}} =$$

Using assumptions and exogeneity of Y_t $\sigma_{u_{St},Y_t} = 0$, $\sigma_{u_{Dt},Y_t} = 0$, $\sigma_{u_{Dt},u_{St}} = 0$ so we get

$$\text{plim}\,\hat{\alpha}_2^{IV} = \alpha_2 - (\alpha_2 - \beta_2)\frac{\sigma^2_{u_{St}}}{\beta_3^2\sigma^2_{Y_t} + \sigma^2_{u_{Dt}} + \sigma^2_{u_{St}}}$$ so the estimator is biased.

(b) [8 marks] □ What is meant by identification of equations? Use any method to tell whether each equation is identified, underidentified or overidentified. Give some explanations of the method chosen.

The equation is said to be identified if there is a method to get consistent estimates of its parameters.

For the second part of the question two approaches can be used

Method of finding instruments

$$Q_{Dt} = \beta_1 + \beta_2 P_t + \beta_3 Y_t + u_{Dt} \qquad (1)$$
$$Q_{St} = \alpha_1 + \alpha_2 P_t + u_{St} \qquad (2)$$

Both the first and second equations contain an endogenous variable P_t on the right. However, there is an instrument Y_t for consistent estimation of the coefficient (an exogenous variable not contained in the second equation, so it is exactly identified.

For consistent estimation of the coefficient β_2 in the first equation, there are no free instruments (the only exogenous variable Y_t is already contained in this equation – so it is underoidentified.

Order condition method

Key parameter of the system of equations $Q_{Dt} = \beta_1 + \beta_2 P_t + \beta_3 Y_t + u_{Dt}$ (1)

$$Q_{St} = \alpha_1 + \alpha_2 P_t + u_{St} \qquad (2)$$

Is G–1=2–1=1, where G – the number of equations in the system (and at the same time the number of endogenous variables). Then in the second equation one variable Y_t is missed, so that the identification condition 1 = 1 is fulfilled.

The first equation has no missed variables at all - it contains all the variables, so 1> 0 and the equation is not identifiable.

□ Obtain IV estimator of α_2 in equation **(2)** using appropriate instrument. Prove that this estimator is consistent.

IV estimator $\hat{\alpha}_2^{IV} = \dfrac{\text{Cov}(Q_t, Y_t)}{\text{Cov}(P_t, Y_t)}$. Use structural equation $Q_t = \alpha_1 + \alpha_2 P_t + u_{St}$ to transform it

$$\hat{\alpha}_2^{IV} = \frac{\text{Cov}(Q_t, Y_t)}{\text{Cov}(P_t, Y_t)} = \frac{\text{Cov}(\alpha_1 + \alpha_2 P_t + u_{St}, Y_t)}{\text{Cov}(P_t, Y_t)} = \frac{\text{Cov}(\alpha_1, Y_t) + \alpha_2\,\text{Cov}(P_t, Y_t) + \text{Cov}(u_{St}, Y_t)}{\text{Cov}(P_t, Y_t)} = \alpha_2 + \frac{\text{Cov}(u_{St}, Y_t)}{\text{Cov}(P_t, Y_t)}$$

and then take plim

$$\text{plim}\,\hat{\alpha}_2^{IV} = \alpha_2 + \frac{\text{plim}\text{Cov}(u_{St}, Y_t)}{\text{plim}\text{Cov}(P_t, Y_t)} = \alpha_2 + \frac{\text{cov}(u_{St}, Y_t)}{\text{plim}\text{Cov}(P_t, Y_t)} = \alpha_2 + \frac{0}{\text{plim}\text{Cov}(P_t, Y_t)}$$ since the exogenous variable Y_t does not correlate with the random term u_{St}

So $\text{plim}\,\hat{\alpha}_2^{IV} = \alpha_2$ and $\hat{\alpha}_2^{IV}$ is consistent

□ Obtain TSLS estimator of α_2 in equation **(2)** using the same instrument.

To estimate α_2 at the first stage of TSLS we regress endogenous explanatory variable P on available instrument Y: $\hat{P} = \hat{\gamma}_1 + \hat{\gamma}_2 Y$

and then at the second stage use forecasted value of $\hat{P}$ as an explanatory variable in OLS instead of endogenous variable P: $\hat{\alpha}_2^{TSLS} = \dfrac{\text{Cov}(Q, \hat{P})}{\text{Var}(\hat{P})}$.

(c) [8 marks] Prove that IV (Instrumental Variables) and TSLS (Two Stage Least Squares) approaches used for estimation of coefficient α_2 in equation **(2)** give the identical results.

Let us show that it is possible to use at the second stage the IV estimator $\frac{\mathrm{Cov}(\hat{P},Q)}{\mathrm{Cov}(\hat{P},P)}$. It is sufficient to show that $\mathrm{Cov}(\hat{P},P)=\mathrm{Cov}(\hat{P},\hat{P}+e)=\mathrm{Cov}(\hat{P},\hat{P})+\mathrm{Cov}(\hat{P},e)=\mathrm{Var}(\hat{P})+0=\mathrm{Var}(\hat{P})$ according well known property of OLS $\mathrm{Cov}(\hat{P},e)=0$.

Now

$$\hat{\alpha}_2^{TSLS}=\frac{\mathrm{Cov}(\hat{P},Q)}{\mathrm{Var}(\hat{P})}=\frac{\mathrm{Cov}(\hat{P},Q)}{\mathrm{Cov}(\hat{P},P)}=\frac{\mathrm{Cov}(P+e,Q)}{\mathrm{Cov}(P+e,P)}=\frac{\mathrm{Cov}(\hat{\gamma}_1+\hat{\gamma}_2Y,Q)}{\mathrm{Cov}(\hat{\gamma}_1+\hat{\gamma}_2Y,P)}=\frac{\mathrm{Cov}(\hat{\gamma}_1,Q)+\hat{\gamma}_2\,\mathrm{Cov}(Y,Q)}{\mathrm{Cov}(\hat{\gamma}_1,P)+\hat{\gamma}_2\,\mathrm{Cov}(Y,P)}=$$

$$\frac{0+\hat{\gamma}_2\,\mathrm{Cov}(Y,Q)}{0+\hat{\gamma}_2\,\mathrm{Cov}(Y,P)}=\frac{\hat{\gamma}_2\,\mathrm{Cov}(Y,Q)}{\hat{\gamma}_2\,\mathrm{Cov}(Y,P)}=\frac{\mathrm{Cov}(Y,Q)}{\mathrm{Cov}(Y,P)}=\hat{\alpha}_2^{IV}$$

So, $\hat{\alpha}_2^{IV}=\hat{\alpha}_2^{TSLS}$.