The International College of Economics and Finance
Econometrics – 2020-2021.
Midterm exam. 2020 October 22.
Part 2. Free Response Questions
(1 hour 50 minutes + 10 min reading time)

SECTION A

Answer **ALL** questions from this section (questions **1-3**).

Question 1. (17 marks)

Two students developed different models for their course papers: student (a) was convinced that the coefficient of variable Z_i is equal to 1, while student (b) believed that this coefficient is the opposite of the coefficient for the variable X_i

$$Y_i = \beta_1 + \beta_2 X_i + Z_i + u_i \qquad \textbf{(a)}$$
$$Y_i = \beta_1 + \beta_2 X_i - \beta_2 Z_i + v_i \qquad \textbf{(b)}$$

They think that in fact both models will give the same estimates of β_2, but both faced unexpected difficulties when trying to evaluate these models (what difficulties?). So they decided instead to calculate sample variances and covariances on the base of the same data (n observations): $\mathrm{Var}(Y)=4$, $\mathrm{Var}(X)=3$, $\mathrm{Var}(Z)=5$, $\mathrm{Cov}(Y,X)=6$, $\mathrm{Cov}(Y,Z)=1$, $\mathrm{Cov}(X,Z)=2$.

(a) □ Help the students to find the least squares estimates of β_2 for their models, indicating all necessary steps.

Both regressions are simple linear regression models with some restrictions, so in both cases one should use conventional estimator of the type $\hat{\beta} = \frac{\mathrm{Cov}(Y,X)}{\mathrm{Var}(X)}$. But the transformations of the data needed to reduce problem to simple regression model, are different, hence estimators will be different.

For specification (a) one can rewrite the model in the form $Y_i - Z_i = \beta_1 + \beta_2 X_i + u_i$ and then estimate the model $Y_i^* = \beta_1 + \beta_1 X_i + u_i$, where $Y_i^* = Y_i - Z_i$.

For specification (b) one can rewrite the model in the form $Y_i = \beta_1 + \beta_2 (X_i - Z_i) + v_i$ and then estimate the model $Y_i = \beta_1 + \beta_2 X_i^* + v_i$ where $X_i^* = X_i - Z_i$.

□ Are these estimates really the same as students think?

For (a) we get $\hat{\beta}_2 = \frac{\mathrm{Cov}(Y-Z,X)}{\mathrm{Var}(X)}$. For (b) we get $\hat{\beta}_2 = \frac{\mathrm{Cov}(Y,X-Z)}{\mathrm{Var}(X-Z)}$. To compare

For (a) we get $\hat{\beta}_2 = \frac{\mathrm{Cov}(Y-Z,X)}{\mathrm{Var}(X)} = \frac{\mathrm{Cov}(Y,X)-\mathrm{Cov}(Z,X)}{\mathrm{Var}(X)} = \frac{6-2}{3} = \frac{4}{3}$

For (b) we get $\hat{\beta}_2 = \frac{\mathrm{Cov}(Y,X-Z)}{\mathrm{Var}(X-Z)} = \frac{\mathrm{Cov}(Y,X)-\mathrm{Cov}(Y,Z)}{\mathrm{Var}(X)+\mathrm{Var}(Z)-2\mathrm{Cov}(Z,X)} = \frac{6-1}{3+5-2\cdot 2} = \frac{5}{4}$.

So the estimates are different.

(b) The scientific advisor told the students that both their models are restricted versions of the more general model

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + w_i \qquad \textbf{(c)}.$$

□ What are restrictions in each case?

The restriction for the model (a) is $\gamma = 1$; for (b) it is $\gamma = -\beta$.

□ How to test the restriction for the model (a)? Indicate necessary steps.

Run model (c) and test the restriction $\gamma = 1$ using t-test.

Alternatively

1) Run model (c) and find RSS_c.

2) Run transformed model (a) $Y_i - Z_i = \beta_1 + \beta_2 X_i + u_i$ and find RSS_a.

3) Do F-test $F = \frac{RSS_a - RSS_c}{RSS_c/(n-3)}$. If $F > F_{crit}$ reject $\gamma = 1$.

□ What model should be chosen if both restrictions are invalid? What model (a), (b) or (c) should be chosen if only one restiction is valid?

If both restrictions are invalid choose model (c).

Let for example restriction $\gamma = 1$ is invalid while restriction $\gamma = -\beta$ is valid. In pair (a-c) we choose (c), then in pair (b-c) we choose (b). So final choice is (b).

□ Let both restrictions be valid. Which model out of (a) and (b) is preferable and why? Use numerical data above to choose between (a) and (b).

Any of restricted regressions can be estimated, but the one with smaller variance is preferable. For the model (a) $Y_i - Z_i = \beta_1 + \beta_2 X_i + u_i$ (if it is valid) $\sigma^2_{\beta_2} = \frac{\sigma^2_u}{n\,\mathrm{Var}(X)}$; For specification (b) $Y_i = \beta_1 + \beta_2 (X_i - Z_i) + v_i$ $\sigma^2_{\beta_2} = \frac{\sigma^2_u}{n\,\mathrm{Var}(X - Z)}$. In our case (assuming equal variances of disturbance terms σ^2_u - we have no information on them) $\mathrm{Var}(X) = 3 < \mathrm{Var}(X) + \mathrm{Var}(Z) - 2\mathrm{Cov}(X,Z) = 3 + 5 - 2 \cdot 2 = 4$, so model (b) is preferable being more efficient.

Question 2. (16 marks)

(a) □ What is R^2? What are its main properties and usage? What are its advantages and disadvantages in econometric analysis.

Definition $R^2 = \frac{Explained\ Sum\ of\ Squares}{Total\ Sum\ of\ Squares} = \frac{ESS}{TSS}$. The definition implies its meaning as a percentage of the explained variance of the dependent variable. One of the main indicators of the quality of regression. The property $TSS = ESS + RSS$ is true only for simple and multiple regression with the intercept. From here for simple and multiple regression with the intercept $R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$ and so $R^2 \in [0; 1]$.

For regression without intercept, the formula $TSS = ESS + RSS$ is wrong, so the determination coefficient can even become negative and loses its meaning as a percentage of the explained variance of the dependent variable, and in fact useless.

The main drawback of R^2 is that it always increases with the adding of a new variable, which does not allow us to judge the value of including a new variable and its significance.

□ Why some disadvantages of R^2 can be overcomed by using instead adjusted $R^2_{adj} = \overline{R}^2 = 1 - \frac{RSS/(n-k)}{TSS/(n-1)}$?

Consider $\overline{R}^2 = R^2 - \frac{k-1}{n-k}(1 - R^2)$. For any number of observations the increase in the number of included variables increases R^2, but at the same time increases the number of estimated parameters, so the term $\frac{k-1}{n-k}$ increases and $(1 - R^2)$ decreases - the result for $\overline{R}^2$ is uncertain, and depends on the quality and value of added variables.

□ What are its main properties and usage? What are disadvantages of $\overline{R}^2$ and how they can be overcomed?

$\overline{R}^2$ is used in practice for preliminary estimation of reasonability of inclusion of additional variable or group of variables into the equation. In this case, we can be guided by the known property: it increases when and only when the t-statistics at the included variable is greater than 1. So if $\overline{R}^2$ has decreased, the included variable is insignificant. In case of growth of $\overline{R}^2$, the question of utility of inclusion and significance of the new variable remains open, because 1 is not a critical value of the t-distribution for the required significance levels. An unambiguous answer can only give an F-test for inclusion of a variable (a group of variables).

(b) □ Prove the following properties of $\overline{R}^2$ (based on the definition of $\overline{R}^2$ in (a) above):

1) $\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k}$, **2)** $\overline{R}^2 = R^2 - \frac{k-1}{n-k}(1 - R^2)$.

$$R^2_{adj} = 1 - \frac{RSS \cdot (n-1)}{TSS \cdot (n-k)} = 1 - \frac{(TSS - ESS) \cdot (n-1)}{TSS \cdot (n-k)} = 1 - \frac{(TSS/TSS - ESS/TSS) \cdot (n-1)}{(n-k)} = 1 - \frac{(n-1)}{(n-k)}(1-R^2)$$

$$\overline{R}^2 = 1 - \frac{n-1}{n-k}(1-R^2) = \frac{n-k+R^2 n - R^2 + 1 - n}{n-k} = \frac{n-k+R^2 n - R^2 + 1 - n}{n-k} = \frac{-k+R^2 n - R^2 + 1}{n-k} =$$

$$= \frac{R^2 n - R^2 k + R^2 k - R^2 + 1 - k}{n-k} = R^2 \frac{n-k}{n-k} + \frac{R^2(k-1) - (k-1)}{n-k} = R^2 - \frac{(k-1) - R^2(k-1)}{n-k} = R^2 - \frac{k-1}{n-k}(1-R^2)$$

□ Investigate in what limits $\overline{R}^2$ can vary.

$\overline{R}^2 = R^2 - \frac{k-1}{n-k}(1-R^2) < R^2 \le 1$ as $R^2 \in [0;1]$.

To search for the lower boundary it is more convenient to use the formula $\overline{R}^2 = 1 - (1-R^2)\frac{n-1}{n-k}$.

Obviously, there is no lower limit, because with a sufficiently large number of observations n and a sufficiently large number of parameters k, the expression $\frac{n-1}{n-k}$ can be made as large as you want. At the same time, if $k-1$ independent variables have a weak relation to the dependent variable in the whole, then $R^2 \approx 0$ and the the expression $\overline{R}^2 = 1 - (1-R^2)\frac{n-1}{n-k}$ can become any number in the range $(-\infty; 1)$.

Question 3. (17 marks) The researcher investigates the relation between GDP – variable Y_t and money supply (M1) – variable M_t for Kingdom of Manama in millions of dirhams for a period 1994-2018 (in the beginnig of the year). She tries different models to fit the real data (all coefficients are significant)

$$\hat{Y}_t = 0.337 + 0.08 \cdot M_t \qquad R^2 = 0.89, \quad RSS = 7.0 \qquad (1)$$

$$\hat{Y}_t = -34.02 + 6.1 \cdot \log(M_t) \qquad R^2 = 0.85, \quad RSS = 8.43 \qquad \mathbf{(2)}$$

$$\log(\hat{Y}_t) = 0.77 + 0.0014 \cdot M_t \qquad R^2 = 0.83, \quad RSS = 0.36 \qquad \mathbf{(3)}$$

(a) □ Give interpretation to the models and their coefficients. Are they significant?

Correct interpretation of (1) is: an increase of money supply by one million of dirhams leads to the growth of GDP by 80 thousands of dirhams.
Correct interpretation of (2) is: an increase of money supply by one percent leads to the growth of GDP by 61 thousands of dirhams on average according to the regression.
Correct interpretation of (3) is: an increase of money supply by one million of dirhams leads to the growth of GDP by 0.14 percent on average according to the regression.
All equations are significant based on F-test . For example for equation (1) $F = \frac{R^2}{1-R^2}(n-k) = \frac{0.89}{1-0.89} \cdot 23 = 186$ - significant at any significance level.

□ Which regressions from (1) - (3) can be compared directly (without additional transformations) in their quality? Compare them and choose better model providing appropriate explanations. Can this comparision be done on the basis of their R^2? Explain.

The regressions (1) and (2) are comparable directly as they have the same dependent variable, and judging by RSS regression (1) is better. The comparison here can be also done on the basis of their R^2, as these equations have the same dependent variable Y_t and so have the same TSS so to choose equation based on smaller RSS is equivalent to choose them based on greater $R^2 = 1 - \frac{RSS}{TSS}$.

The regressions (1) and (3) cannot be compared directly (why?). The researcher used Zarembka scaling for the linear regression. After this transformation new variable $YZ = \frac{Y}{geometric\ mean} = \frac{Y}{(Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n)^{1/n}}$ was regressed on M_t with the following results

$$\hat{YZ}_t = 0.05 + 0.001 \cdot M_t \qquad R^2 = 0.89, \quad RSS = 0.18 \qquad (4)$$

□ Do Box-Cox test to compare regressions (1) and (3) and choose the better one on the basis of their RSS. Can this comparision be done on the basis of their R^2?

The dimension of the dependent variable in (1) and (2) is millions of dirhams while dependent variable in (3) is measured in their logarithms, so their *RSS*'s also have different dimension so they are not comparable directly. On the other hand regressions (1) and (4) are comparable, and (4) is better.

According to Box-Cox test we have to evaluate χ-square statistic and compare it with critical value of χ-square distribution with 1 degree of freedom: $\chi^2 = (25/2)\,|\ln(0.36/0.18)| = 8.66 > 6.63 = \chi^2_{crit}(1\%, df = 1)$. Thus linear specification with lower RSS (after Zarembka transformation) is significantly better. As it differs only by constant term from equation (1) we choose regression (1) as best specification.

(b) Demonstrate that Zarembka transformation provides approximate comparability of RSS values for the model with the natural logarithm on the left and for the model in which the dependent variable is subject to the Zarembka scaling.

Accoding Zarembka transformation, dependent variable Y is replaced by transformed one $Y^* = \frac{Y}{geometric\ mean} = \frac{Y}{(Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n)^{\frac{1}{n}}}$. The values of Y^* in regression $Y^* = \beta_1^{(1)} + \beta_2^{(1)} X$ are directly comparable with $\ln Y = \beta_1^{(2)} + \beta_2^{(2)} X$, and the discrepancy can be estimated for significance using Box-Cox test.

One of the possible approaches to show that $\frac{Y}{geometric\ mean} = \frac{Y}{\overline{Y}_{geom}}$ is comparable with $\ln Y$ is the following. Let ΔY be a deviation of Y from its geometric mean $\overline{Y}_{geom}$: $\Delta Y = Y - \overline{Y}_{geom}$. We assume that this deviation is 'small enough' relatively to Y and so $\overline{Y}_{geom}$. Then $\frac{Y}{\overline{Y}_{geom}} = \frac{\overline{Y}_{geom} + \Delta Y}{\overline{Y}_{geom}} = 1 + \alpha$, so it is only by constant equal 1 (which is absorbed by the constant term of regression) differs from α. On the other hand dependent variable $\ln \frac{Y}{\overline{Y}_{geom}} = \ln Y - \ln \overline{Y}_{geom}$ only by constant $-\ln \overline{Y}_{geom}$ (which is also absorbed by the constant term of regression) differs from $\ln Y$. But the representation $\ln\left(\frac{Y}{\overline{Y}_{geom}}\right) = \ln\left(\frac{\overline{Y}_{geom} + \Delta Y}{\overline{Y}_{geom}}\right) = \ln(1 + \alpha) \approx \alpha$ shows that both $\ln Y$ and $\frac{Y}{\overline{Y}_{geom}}$ could be represented by α and some constant, so they are directly comparable, and so residual sums of squares RSS of both regressions are also comparable.

SECTION B.

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

Problem 4. [25 marks]. Consider the following model

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad i = 1, \ldots, n$$

together with the fitted model

$$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$$

Variables X_2, X_3 are assumed to be fixed, assumptions of model A are satisfied.

(a) Show that OLS estimator of b_3 is given by expression

$$b_3 = \frac{\mathrm{Cov}(X_3, Y)\mathrm{Var}(X_2) - \mathrm{Cov}(X_2, Y)\mathrm{Cov}(X_2, X_3)}{\mathrm{Var}(X_2)\mathrm{Var}(X_3) - \left[\mathrm{Cov}(X_2, X_3)\right]^2}$$

Let $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$ and $\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$

The residuals: $e_i = Y_i - \hat{Y}_i = Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i}$.

OLS: $$RSS = S = \sum e_i^2 = \sum (Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i})^2 \underset{b_1, b_2, b_3}{\longrightarrow} \min$$

F.O.C.:

$$\begin{cases} \frac{\partial S}{\partial b_1} = -2\sum (Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i}) = 0 \\ \frac{\partial S}{\partial b_2} = -2\sum X_{2i}(Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i}) = 0 \\ \frac{\partial S}{\partial b_3} = -2\sum X_{3i}(Y_i - b_1 - b_2 X_{2i} - b_3 X_{3i}) = 0 \end{cases}$$

From the first equation $b_1 = \bar{Y} - b_2 \bar{X}_2 - b_3 \bar{X}_3$

Using this expression, we can transform the other two equations as follows:

$$\frac{\partial S}{\partial b_2} = -2\sum X_{2i}(Y_i - \bar{Y} + b_2 \bar{X}_2 + b_3 \bar{X}_3 - b_2 X_{2i} - b_3 X_{3i}) = 0$$

$$\frac{\partial S}{\partial b_3} = -2\sum X_{3i}(Y_i - \bar{Y} + b_2 \bar{X}_2 + b_3 \bar{X}_3 - b_2 X_{2i} - b_3 X_{3i}) = 0 \qquad (*)$$

Dividing by (-2) and rearranging

$$\begin{cases} b_2 \sum X_{2i}(\bar{X}_2 - X_{2i}) + b_3 \sum X_{2i}(\bar{X}_3 - X_{3i}) + \sum X_{2i}(Y_i - \bar{Y}) = 0 \\ b_2 \sum X_{3i}(\bar{X}_2 - X_{2i}) + b_3 \sum X_{3i}(\bar{X}_3 - X_{3i}) + \sum X_{3i}(Y_i - \bar{Y}) = 0 \end{cases} \qquad (**)$$

$$\begin{cases} b_2 \sum X_{2i}(X_{2i} - \bar{X}_2) + b_3 \sum X_{2i}(X_{3i} - \bar{X}_3) = \sum X_{2i}(Y_i - \bar{Y}) \\ b_2 \sum X_{3i}(X_{2i} - \bar{X}_2) + b_3 \sum X_{3i}(X_{3i} - \bar{X}_3) = \sum X_{3i}(Y_i - \bar{Y}) \end{cases} \qquad (**)$$

$$\sum X_{2i}(X_{2i} - \bar{X}_2) = \sum X_{2i}^2 - n\bar{X}_2^2 = n\mathrm{Var}(X_2)$$

$$\sum X_{2i}(X_{3i} - \bar{X}_3) = \sum X_{2i}(X_{3i} - \bar{X}_3) - \bar{X}_2 \sum (X_{3i} - \bar{X}_3) = \sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3) = n\mathrm{Cov}(X_2, X_3)$$

as $\sum (X_{3i} - \bar{X}_3) = 0$

$$\sum X_{2i}(Y_i - \bar{Y}) = \sum X_{2i}(Y_i - \bar{Y}) - \bar{X}_2 \sum (Y_i - \bar{Y}) = \sum (X_{2i} - \bar{X}_2)(Y_i - \bar{Y}) = n\mathrm{Cov}(X_2, Y)$$ (the same).

The same applies for the second equation in (**):

$$\sum X_{3i}(X_{2i} - \bar{X}_2) = n\mathrm{Cov}(X_2, X_3)$$

$$\sum X_{3i}(X_{3i} - \bar{X}_3) = n\mathrm{Var}(X_3)$$

$$\sum X_{3i}(Y_i - \bar{Y}) = n\mathrm{Cov}(X_3, Y)$$

Hence, (**) is equivalent (after dividing by *n*) to:

$$\begin{cases} b_2 \mathrm{Var}(X_2) + b_3 \mathrm{Cov}(X_2, X_3) = \mathrm{Cov}(X_2, Y) \\ b_2 \mathrm{Cov}(X_2, X_3) + b_3 \mathrm{Var}(X_3) = \mathrm{Cov}(X_3, Y) \end{cases}$$

Now obtain solution using Cramer's rule:

$$b_2 = \frac{\mathrm{Cov}(X_2, Y)\mathrm{Var}(X_3) - \mathrm{Cov}(X_3, Y)\mathrm{Cov}(X_2, X_3)}{\mathrm{Var}(X_2)\mathrm{Var}(X_3) - \left[\mathrm{Cov}(X_2, X_3)\right]^2}$$

$$b_3 = \frac{\mathrm{Cov}(X_3, Y)\mathrm{Var}(X_2) - \mathrm{Cov}(X_2, Y)\mathrm{Cov}(X_2, X_3)}{\mathrm{Var}(X_2)\mathrm{Var}(X_3) - \left[\mathrm{Cov}(X_2, X_3)\right]^2}$$

(b) Derive the formula of decomposition of the coefficient's estimator into fixed and random components.

$$b_3 = \beta_3 + \left(\frac{\mathrm{Cov}(X_3, u)\mathrm{Var}(X_2) - \mathrm{Cov}(X_2, u)\mathrm{Cov}(X_3, X_2)}{\Delta} \right) \text{ where } \Delta = \mathrm{Var}(X_2)\mathrm{Var}(X_3) - \left[\mathrm{Cov}(X_2, X_3)\right]^2$$

$$b_3 = \frac{1}{\Delta}(\mathrm{Cov}(X_3,Y)\mathrm{Var}(X_2) - \mathrm{Cov}(X_2,Y)\mathrm{Cov}(X_3,X_2)) =$$
$$= \frac{1}{\Delta}((\mathrm{Cov}(X_3,\beta_1+\beta_2X_2+\beta_3X_3+u)\mathrm{Var}(X_2) - \mathrm{Cov}(X_2,\beta_1+\beta_2X_2+\beta_3X_3+u)\mathrm{Cov}(X_3,X_2)) =$$
$$= \frac{1}{\Delta}((\beta_2\mathrm{Cov}(X_3,X_2)+\beta_3\mathrm{Var}(X_3)+\mathrm{Cov}(X_3,u))\mathrm{Var}(X_2) -$$
$$- \beta_2\mathrm{Var}(X_2) - \beta_3\mathrm{Cov}(X_3,X_2) - \mathrm{Cov}(X_2,u))\mathrm{Cov}(X_3,X_2)) =$$
$$= \frac{1}{\Delta}(\beta_2\mathrm{Cov}(X_3,X_2)\mathrm{Var}(X_2)+\beta_3\mathrm{Var}(X_3)\mathrm{Var}(X_2)+\mathrm{Cov}(X_3,u)\mathrm{Var}(X_2) -$$
$$- \beta_2\mathrm{Var}(X_2)\mathrm{Cov}(X_3,X_2) - \beta_3\mathrm{Cov}(X_3,X_2)\mathrm{Cov}(X_3,X_2) - \mathrm{Cov}(X_2,u)\mathrm{Cov}(X_3,X_2)) =$$
$$= \frac{1}{\Delta}(\beta_3\{\mathrm{Var}(X_3)\mathrm{Var}(X_2) - [\mathrm{Cov}(X_3,X_2)]^2\} + \beta_2\{\mathrm{Cov}(X_3,X_2)\mathrm{Var}(X_2) - \mathrm{Var}(X_2)\mathrm{Cov}(X_3,X_2)\} +$$
$$+ \{\mathrm{Cov}(X_3,u)\mathrm{Var}(X_2) - \mathrm{Cov}(X_2,u)\mathrm{Cov}(X_3,X_2)\}) =$$
$$= \beta_3\frac{\Delta}{\Delta} + \beta_2\frac{1}{\Delta}\{\mathrm{Cov}(X_3,X_2)\mathrm{Var}(X_2) - \mathrm{Var}(X_2)\mathrm{Cov}(X_3,X_2)\} +$$
$$+ \frac{1}{\Delta}\{\mathrm{Cov}(X_3,u)\mathrm{Var}(X_2) - \mathrm{Cov}(X_2,u)\mathrm{Cov}(X_3,X_2)\}) =$$

So $E(b_3) = \beta_3 + 0 + \frac{1}{\Delta}E\{\mathrm{Cov}(X_3,u)\mathrm{Var}(X_2) - \mathrm{Cov}(X_2,u)\mathrm{Cov}(X_3,X_2)\}$

(c) □ Prove that OLS estimator b_3 of the coefficient β_3 in multiple regression is unbiased.

As it is shown in (b) $b_3 = \beta_3 + \left(\frac{\mathrm{Cov}(X_3,u)\mathrm{Var}(X_2) - \mathrm{Cov}(X_2,u)\mathrm{Cov}(X_3,X_2)}{\Delta}\right)$ where

$$E(b_3) = \beta_3 + \frac{\mathrm{Var}(X_2)E(\mathrm{Cov}(X_3,u)) - \mathrm{Cov}(X_3,X_2)E(\mathrm{Cov}(X_2,u))}{\Delta} =$$
$$= \beta_3 + \frac{\mathrm{Var}(X_2)(\mathrm{Cov}(X_3,Eu)) - \mathrm{Cov}(X_3,X_2)(\mathrm{Cov}(X_2,Eu))}{\Delta} = \beta_3 + 0 = \beta_3$$

Question 5. [25 marks] A lazy student found an interesting paper with econometric models describing how $COVID$ - aggregate anti COVID expendidtures, depend on GDP, aggregate gross national product, and POP, total population, for a sample of 70 countries in the second quarter of 2020. $COVID$ and GDP are both measured in US$ billion. POP is measured in million. *RSS* – Residual Sum of Squares. He decided to use this article in his course paper pretending that he got all equations himself using original data (what in fact was not true). He wrote the equations on a paper:

$$\log\frac{\hat{COVID}}{POP} = -3.74 + 1.27\log\frac{GDP}{POP} \qquad R^2 = 0.90 \quad RSS = 15.45$$

(1)

$$\log\hat{COVID} = -3.60 + 1.27\log GDP - 0.33\log POP \quad R^2 = 0.95 \quad RSS = 13.90 \qquad \textbf{(2)}$$

$$\log\frac{\hat{COVID}}{POP} = -3.60 + 1.27\log\frac{GDP}{POP} - 0.06\log POP \quad R^2 = 0.91 \quad RSS = 13.90 \qquad \textbf{(3)}$$

But when he wrote his coursework some details seemed a little strange to him and he began to doubt that he had correctly rewritten the equations on paper. He asked your advice.

(a) □ The student now believes that by mistake he repeated the same coefficient 1.27 in equations (1), (2) and (3), as well as he repeated the intercept -3.60 in equations (2) and (3), but he does not remember the correct values. Are these coincedences really happened by mistake?

Theoretical equations and their transformations

First

$$\log\frac{COVID}{POP} = \beta_1 + \beta_2\log\frac{GDP}{POP} + u \quad \Leftrightarrow \quad \log COVID = \beta_1 + \beta_2\log GDP + (1-\beta_2)\log POP + u$$

Second

$$\log COVID = \beta_1 + \beta_2 \log GDP + \beta_3 \log POP + u$$

Third

$$\log \frac{COVID}{POP} = \beta_1 + \beta_2 \log \frac{GDP}{POP} + \beta_3 \log POP + u \;\Leftrightarrow\; \log COVID = \beta_1 + \beta_2 \log GDP + (1 - \beta_2 + \beta_3) \log POP + u$$

From the equations in the transformed form it is clearly seen that under fixed *POP*, all the first slope coefficients express the same - the elasticity of *COVID* by *GDP*, and therefore after the evaluation of these equations should get there the same numbers, so it is no coincidence that there is 1.27 everywhere.
The same with the coincidence of intercepts in equations (2) and (3) - both transformed equations are multiple regression equations with the same set of variables that are estimated from the same data.

□ It seems strange to him that both coefficietns of the variable $\log POP$ in equations (2) and (3) are negative but different in absolute value. Help the student to understand these.
If we fix *GDP*, on the contrary, and increase the *POP* value, there will be a shortage of money, which will reduce the expenditures on *COVID*

The coefficient $(1 - \beta_2 + \beta_3)$ of the second equation, as we already know, must be equal to the coefficient β_3' of the third equation: $1 - \beta_2 + \beta_3 = \beta_3'$. From here we get for the third original equation

$$\log \frac{COVID}{POP} = \beta_1 + \beta_2 \log \frac{GNP}{POP} + \beta_3 \log POP + u$$

different expression for the coefficient $\beta_3 = \beta_3' + \beta_2 - 1$.

□ The scietific advisor asked the student to add standard errors to all coefficients in equations. Having no records the student decided to write arbitrary numbers as standard errors, but he is afraid that some of them must match, and some not, and if he makes a mistake, then his supervisor immediately recognizes the cheat. Can't you help the student (giving necessary explanations)?
From the same theoretical equations, taking into account the already drawn conclusions about the coincidence of coefficients, it follows that the errors of the same coefficients 1.27 and -3.60 of the second and third equations must coincide.

(b) □ *RSS'* are the same in models (2) and (3) while their R^2 are different? Explain.
The same is applicable to the *RSS* of second and third equations. But in the initial equations they have different dependent variables (and therefore different TSS), so they must have different R^2.

□ In equations (1) and (3) both R^2 and *RSS'* are different. Explain.
The first and third equations have the same dependent variables and the same set of independent variables, as seen from the equivalent theoretical equations (1) and (3), but the first equation includes a restriction (the third coefficient is zero), while the third equation has no restriction, so they must have different RSS and R^2.

(c) □ The supervisor made it clear to the student that equation (1) is a restricted version of equation (2) and asked the student to find the restriction and test it using an F test, and on this basis to choose the best equation. Can you help?
The first equation $\log \frac{COVID}{POP} = \beta_1 + \beta_2 \log \frac{GDP}{POP} + u$ can be rewritten as

$$\log COVID = \beta_1 + \beta_2 \log GDP + (1 - \beta_2) \log POP + u$$

This is a restricted version of the more general specification

$$\log COVID = \beta_1 + \beta_2 \log GDP + \beta_3 \log POP + v$$

with the restriction $\beta_3 = 1 - \beta_2$.
Using values of RSS for equations (1) and (2) we get

$F(1,67) = \frac{(14.26 - 13.90)/1}{13.90/67} = 1.74$. The null hypothesis is $H_0 : \beta_3 = 1 - \beta_2$. The critical value of $F(1,67)$ at the 5 percent significance level is about 3.99. Hence we do not reject the restriction and choose restricted equation (1).

□ Show the student that the previous F-test in (c) may well be replaced by the t-test. Is there any advantage to this approach?

We take the unrestricted equation

$$\log COVID = \beta_1 + \beta_2 \log GDP + \beta_3 \log POP + v$$

and we transform it a little so that the restriction (we already know it) comes out as a coefficient for the variable:

$$\log \frac{COVID}{POP} = \beta_1 + \beta_2 \log \frac{GDP}{POP} + (\beta_2 + \beta_3 - 1) \log POP + v$$

Now we can now test the restriction using a t-test. Its advantage is that it can be executed also as a one-way test under certain assumptions, which may help to make result significant.