# The International College of Economics and Finance
# Econometrics 2018-2019. First Semester Exam, December 27.
## Part 2. (2 hours). Answer all questions (1,2,3) from section A and one (4 or 5) - from section B.

***IMPORTANT: Start answering each question on the form with the desired question number (ask for extra paper if necessary). Structure your answers in accordance with the structure of the questions. Testing hypotheses always state clearly null and alternative hypotheses provide critical value used for test, mentioning degrees of freedom and the significance level chosen for the test.***

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

**1. [15 marks]** A researcher has data on the hourly earnings, $E$, measured in U.S. \$, tenure with current employer, $T$, in years, and age, $AGE$, also in years for 112 male and female graduates of a large university (with no further education) and 28 university drop-outs. Defining a dummy variable $D$ to be equal to 0 in the case of a graduate and 1 in the case of a drop-out, and a slope dummy $DT$ as the product of $D$ and $T$, the researcher runs the following regressions (standard errors in parentheses; $RSS$ = sum of squares of residuals):

$$\text{Graduates only} \quad \underset{(1.70)\ \ (0.14)}{\hat{E}_i = 14.82 + 0.70T_i} \qquad R^2 = 0.19 \quad RSS = 15{,}220 \qquad \text{(eq.1)}$$

$$\text{Drop-outs only} \quad \underset{(0.94)\ (0.16)}{\hat{E}_i = 7.41 + 0.47T_i} \qquad R^2 = 0.24 \quad RSS = 294 \qquad \text{(eq.2)}$$

$$\text{Combined sample} \quad \underset{(1.39)\ \ (0.12)}{\hat{E}_i = 12.26 + 0.81T_i} \qquad R^2 = 0.24 \quad RSS = 17{,}022 \qquad \text{(eq.3)}$$

$$\text{Combined sample} \quad \underset{(1.52)\ \ (0.12)\quad\ (2.34)}{\hat{E}_i = 14.94 + 0.70T_i - 8.46D_i} \qquad R^2 = 0.31 \quad RSS = 15{,}535 \qquad \text{(eq.4)}$$

$$\text{Combined sample} \quad \underset{(1.55)\ \ (0.12)\quad\ (3.35)\quad (0.53)}{\hat{E}_i = 14.82 + 0.70T_i - 7.41D_i - 0.23DT_i} \qquad R^2 = 0.31 \quad RSS = 15{,}514 \qquad \text{(eq.5)}$$

**(a)** Give an economic interpretation to all coefficients of the last equation.
Derive from the last equation two regression equations stating the relationship between $T$ and $E$ separately for graduates and for drop-outs.

The constant estimates the earnings of a high school graduate at the beginning of her/his tenure with a current employer as \$14.82 per hour

The coefficient of *T* estimates the increase in the hourly rate to be \$0.70 per year of the current tenure The coefficient of *D* estimates that drop-outs earn \$7.41 less per hour, for any fixed level of *T* (or holding T constant)
The coefficient of the slope dummy estimates that the hourly earnings of drop-outs increase by \$0.23 less per year for drop-outs.
Mathematically, the relationship between *T* and *E* corresponding to eq.5 is:

$$E_i = \beta_1 + \beta_2 T_i + \beta_3 D_i + \beta_4 DT_i + u_i$$

For graduates $D = 0$ so $\quad E_i = \beta_1 + \beta_2 T_i + u_i$

For drop-outs $D = 1$ so $\quad E_i = \beta_1 + \beta_2 T_i + \beta_3 1 + \beta_4 1 \cdot T_i + u_i =$

$$E_i = (\beta_1 + \beta_3) + (\beta_2 + \beta_4)T_i + u_i$$

According to this we get

*Graduates* $\quad \hat{y} = 14.82 + 0.70T$

*Drop-outs* $\quad \hat{y} = (14.82 - 7.41) + (0.70 - 0.23)T = 7.41 + 0.47T$

**(b)** Investigate whether the earnings function for drop-outs differs from that for graduates using a Chow test. State the null hypothesis; indicate the test used, degrees of freedom, and critical values. Explain the logic of the Chow test.

The null hypothesis that the parameters are the same for graduates and drop-outs.
Writing $RSS_G$, $RSS_D$ and $RSS_P$ for the residual sums of squares in the equations (eq.1), (eq.2) and (eq.3), the $F$ statistic for the Chow test is

$$F(k+1, n-2k-2) = \frac{RSS_P - [RSS_G + RSS_D]/k}{(RSS_G + RSS_D)/(n-2k)} .$$

where $k$ is the number of parameters in the regression.
Since $k$ is 2 and $n$ is 140 in this case, the $F$ statistic is

$$F(2, 140) = \frac{(17{,}022 - [15{,}220 + 294])/2}{(15{,}220 + 294)/140 - 2 \cdot 2} = 6.61 .$$

The critical value of $F$ at the 1% significance level is about 4.79.. Hence the null hypothesis is rejected and one concludes that the earnings functions for graduates and drop-outs are significantly different.

**(c)** Investigate the same problem testing the explanatory power of the dummy variables. Run the $t$-tests and the appropriate $F$-test, indicating clearly a pair of hypotheses. Compare the findings of the two test approaches. Discuss the results obtained from different approaches.

**t-tests: The *t* ratios for *D* and *DT*** are -7.41/3.35 = -2.21 and -0.23/0.53 = -0.43, respectively. The first is significant at 5% level (critical value for df=120 is 1.98), while the second is not.
**F test of the joint explanatory power of the dummy variables***:* Writing the coefficients of $D$ and $DT$ as $\gamma_1$ and $\gamma_2$, respectively, the null hypothesis is $H_0 : \gamma_1 = \gamma_2 = 0$ and the alternative hypothesis is that one or both coefficients is non-zero.
Writing the residual sums of squares in the third and fifth equations as $RSS_3$ and $RSS_5$, the $F$ statistic is

$$F(2; 1{,}064) = \frac{(RSS_3 - RSS_5)/2}{RSS_5/136} = \frac{(17{,}022 - 15{,}514)/2}{15{,}514/136} = 6.61$$

This of course is the same $F$ statistic as in the case of the Chow test, and so the conclusion is the same, the null hypothesis being rejected.The Chow test and the $F$ test on the dummy variables as a group are equivalent.
If the null hypothesis is rejected, the dummy variable approach may identify which coefficient is significantly different via $t$ tests on the individual coefficients. In this case, only one $t$ test is significant, possibly as a consequence of multicollinearity.

**2. [15 marks]** A researcher has data on school enrolment, $N$, and annual fixed expenditure, $EXP$, measured in thousands of Korean won, for the sample of 100 schools in South Korea and estimated cost function of the quadratic form

$$\widehat{EXP} = 17{,}999 + 1{,}060N - 1.27N^2 \qquad R^2 = 0.74 \qquad (1)$$
$$(12{,}908)\ (133)\qquad (0.29)$$

Suspecting that the regression was subject to heteroscedasticity, the researcher runs the regression twice more, first with the 43 schools with lowest enrolments, then with the 31 schools with the highest enrolments The residual sum of squares ($RSS$) in the two regressions are 8,416,000 and 65,525,000 respectively.

The researcher defines a new variable, $EXPSTUD$, expenditure per student, as $EXPSTUD = EXP/N$, and runs the regression

$$\widehat{EXPSTUD} = 1{,}316 - 1.92N \qquad R^2 = 0.63 \qquad (2)$$
$$(33)\quad (0.15)$$

He then runs regressions with the 43 smallest schools and the 31 largest schools and the residual sums of squares are 967,000 and 698,000 respectively.

Her colleague suggests that it might be a good idea to include the reciprocal of $N$ in the regression. The researcher does this and obtains the following result:

$$\underset{}{\hat{EXPSTUD}} = 1{,}078 - \underset{(0.26)}{1.31}N + \underset{(5{,}863)}{16{,}374}NRECIP \qquad R^2 = 0.66 \qquad (3)$$
$$\qquad (91)$$

where $NRECIP = 1/N$.

He again runs regressions with the 43 smallest schools and the 31 largest schools and the residual sums of squares are 948,000 and 677,000.

**(a)** Explain why the researcher should be prepared for the presence of heteroscedasticity in the regression under consideration.

Explain what is meant by heteroscedasticity and describe the consequences of its presence in a regression model.

It is natural to assume that schools with a large number of pupils incur large costs of education, therefore, the spread of values of the costs of large schools is also naturally expected to be large.

This causes differences in the variance of the disturbance term for large and small schools. This phenomenon of the presence of differences in the magnitude of the variance of a random term is called heteroscedasticity.

The usual consequences of heteroscedasticity are inefficient estimates of regression coefficients (while maintaining their unbiasedness), increase in the standard error of the coefficients, as they are incorrectly calculated and, accordingly, the invalidity of the significance tests. At the same time, estimates of standard errors of regression coefficients are usually underestimated.

**(b)** Describe the Goldfeld-Quandt test for heteroscedasticity and explain why under certain conditions it may detect heteroscedasticity.

What assumptions must be made to correctly apply this test?

Why the researcher may believe that these assumptions are fulfilled?

Perform a Goldfeld-Quandt test for heteroscedasticity on each of the regressions.

Goldfeld-Quandt test uses F-statistic in the form $F = \dfrac{\dfrac{RSS(bigger)}{(volume\ of\ \ the\ sample\ with\ bigger\ \ RSS) - k}}{\dfrac{RSS(smaller)}{(volume\ of\ \ the\ sample\ with\ smaller\ \ RSS) - k}}$, where $k$ is the number of parameters in the regression under consideration

The test is based on the assumption that the standard deviation of the value of the disturbance term is proportional to the number of students in the school.

The fact that large schools have a large variance of the disturbance term provides the probable validity of this assumption. This makes it possible to use the Goldfeld-Quandt special heteroscedasticity test.

For regression (1) $F = \dfrac{\dfrac{65{,}525}{31-3}}{\dfrac{8{,}416}{43-3}} = 11.12$ while $F_{crit}^{1\%}(43{,}31) \approx F_{crit}^{1\%}(40{,}30) = 2.30$, so null hypothesis of no heteroscedasticity is rejected. For regression (2) $F = \dfrac{\dfrac{967}{43-2}}{\dfrac{698}{31-2}} = 0.98$ while $F_{crit}^{5\%}(43{,}31) \approx F_{crit}^{5\%}(40{,}30) = 1.79$, so null hypothesis of no heteroscedasticity is not rejected.

The same for regression (3) $F = \dfrac{\dfrac{948}{43-3}}{\dfrac{677}{31-3}} = 0.98$ while $F_{crit}^{5\%}(43{,}31) \approx F_{crit}^{5\%}(40{,}30) = 1.79$, so null hypothesis of no heteroscedasticity is not rejected.

Hence the null hypothesis of homoscedasticity is rejected at the 1% level for regression (1) but is not rejected at the 5% level for regressions (2) and (3).

**(c)** Explain why the researcher ran the second regression.

Give an economic interpretation of regression (3) and explain why it may be preferable to regression (2).

What significance, if any, should be attached to the fact that the coefficient of *N* is smaller in regression (3) and its standard error is larger?

Expenditure per student may have less extreme variation than size of school and hence a regression using it as the dependent variable might be free of the heteroscedasticity found in regression (1). Technically, if the standard deviation of the disturbance term were proportional to *N*, as was assumed in the Goldfeld-Quandt test, scaling through by *N* will result in a model with a homoscedastic disturbance term.

According regression (3) the annual marginal cost per child starts at $1,078 per year and falls by $1.31 per child.

After scaling through, the model should be as in regression (3), the intercept in the original equation becoming the coefficient of *NRECIP*. Model (3) is preferable to model (2) because it includes the missing *NRECIP* term. This is important from an economic viewpoint, since it may be interpreted as an estimate of the overhead cost of running a school, in this case $16,374 per year, and it is important statistically since the *t* statistic is 2.79.

The coefficient of *N* in regression (2) is subject to omitted variable bias equal to $\alpha \mathrm{Cov}(N, NRECIP)/\mathrm{Var}(N)$. The covariance is of course negative, and a cab be expected to be positive, so the bias is downwards, which explains the (numerically) larger coefficient in regression (2). A comparison of the standard errors is invalid since regression (2) is subject to omitted variable bias.

**3. [15 marks]** A researcher has data from the World Bank *World Development Report 2000* on *F*, average fertility (average number of children born to each woman during her life), *M*, under-five mortality (number of children, per 100, dying before reaching the age of 5), and *S*, average years of female schooling, for a sample of 54 countries. She hypothesizes that fertility is negatively related to schooling and positively related to mortality, and that mortality is negatively related to schooling:

$$F = \beta_1 + \beta_2 S + \beta_3 M + u \qquad \text{(eq1)}$$

$$M = \alpha_1 + \alpha_2 S + v \qquad \text{(eq 2)}$$

where *u* and *v* are disturbance terms that may be assumed to be distributed independently of each other. $S$ may be assumed to be exogenous. $M$ is assumed to be distributed independently of $u$ and $S$ is assumed to be distributed independently of $v$.

**(a)** Analyze the structure of the system of equations, and derive the reduced form equations for *F* and *M*.
Explain what would be the most appropriate method to fit equation (1).
Explain what would be the most appropriate method to fit equation (2).

There is no circular dependence in this system of equations. The second regression has no endogenous variable in its right side. So there is no problem in its estimation. The explanatory variable $M$ in (eq1) is certainly stochastics and endogenous but it is assumed to be distributed independently of $u$, so GMC are not violated.

(2) is the reduced form equation for *M*.
Substituting for *M* in (1), we have

$$F = (\beta_1 + \alpha_1\beta_3) + (\beta_2 + \alpha_2\beta_3)S + u + \beta_3 v.$$

Since *M* does not depend on *u*, OLS may be used to fit (1).
There are no endogenous explanatory variables in (2), so again OLS may be used.

**(b)** The researcher decides to fit (eq1) using ordinary least squares, and she decides also to perform a simple regression of *F* on *S*, again using ordinary least squares, with the following results (standard errors in parentheses):

$$\underset{(1.86)}{\hat{F} = 4.20} - \underset{(0.041)}{0.18S} + \underset{(0.019)}{0.039M} \qquad R^2 = 0.31 \qquad \text{(eq1*)}$$

$$\underset{(1.91)}{\hat{F} = 4.42} - \underset{(0.042)}{0.17S} \qquad R^2 = 0.25 \qquad \text{(eq2*)}$$

Give detailed explanations why the coefficient of *S* differs in the two equations.
Is indirect effect of schooling $S$ via mortality $M$ on fertility $F$ significant and should it be taken into account? (answer this question using the comparison of $R^2$ and $R^2_{adj}$ and doing appropriate tests).
Explain whether one may validly perform *t* tests on the coefficients of (eq2*).

In (eq3), the coefficient is an estimate of the direct effect of *S* on fertility, controlling for *M*. In (eq4) it is an estimate of the total effect, taking account of the indirect effect via *M* (female education reduces mortality, and a reduction in mortality leads to a reduction in fertility).

$R^2_{adj} = R^2 - (1-R^2)\dfrac{k-1}{n-k}$ where $k$ is the number of estimated parameters, $n$ is the number of obsevations.

For (eq2*) $R^2_{adj} = 0.25 - (1-0.25)\dfrac{1}{52} = 0.236$

For (eq1*) $R^2_{adj} = 0.31 - (1-0.31)\dfrac{2}{51} = 0.283$ what is greater than in (eq1*) so the t-statistics for coefficient of *M* should be greater than 1, in fact it is $t = \dfrac{0.039}{0.019} = 2.05$ what is grater than $t^{5\%}_{crit}(51) \approx t^{5\%}_{crit}(50) = 2.01$, so the coefficient is significant and so indirect effect of schooling $S$ via mortality $M$ on fertility $F$ should be taken into account.

Slight difference in the values of coefficient's estimate can be explained by omitted variable bias in (eq1*).

**(c)** At a seminar someone hypothesizes that female schooling $S$ may be negatively influenced by fertility $F$, especially in the poorer developing countries in the sample, so $S$ cannot be assumed to be exogenous and so some endogeneity is present in the model. To investigate this, the researcher adds the following equation to the model:

$$S = \delta_1 + \delta_2 F + \delta_3 G + w \qquad \text{(eq3)}$$

where $G$ is GNP per capita and $w$ is a disturbance term which is assumed to be distributed independently of $F$ and $G$ where $G$ is assumed exogenous. She estimates this equation, then memorizes its residuals in variable $E$ and adds it to the equation (1), getting the following results

$$\underset{(2.14)}{\hat{F} = 18.58} - \underset{(0.047)}{0.49S} + \underset{(0.013)}{0.003M} + \underset{(0.06)}{0.49E} \quad R^2 = 0.70 \qquad \text{(eq4*)}$$

What test run the researcher? What are null hypothesis and assumptions of the test? What are your conclusions?

If the researcher asks your advice what method would you recommend her to estimate equation (1)?

Discuss whether *G* is likely to be a valid instrument.

The researcher run Durbin-Wu-Hausman test in the Davidson and MacKinnon version. The test is based on the null hypothesis that there is no endogeneity. In this case, the least squares method gives the best estimates of the coefficients. However, in the case of endogeneity, the estimates will no longer be correct, and tools must be used (for example, using TSLS), which will give correct estimates.

As the coefficient of residuals $E$ of the auxiliary equation where the variable *S* was regressed using instrument *G*, is significant ($t = \dfrac{0.49}{0.06} = 8.17 > t_{crit}^{1\%}(50) = 2.009$) the null hypothesis of no endogeneity is rejected. So the assumption of of *S* being exogenous is not correct, and this variable should be assumed endogenous.

One should be recommended to use TSLS for estimation of equation (1), using *G* as the instrument.

*G* should be a valid instrument since it is highly correlated with *S*, it may reasonably be considered to be exogenous and therefore uncorrelated with the disturbance term in (5.4), and it does not appear in the equation in its own right (though perhaps it should).

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

**4. [30 marks]** A researcher investigating the shadow economy (illegal economy) using international cross-section data for 40 countries hypothesizes that consumer expenditure on shadow goods and services, *q*, is related to total consumer expenditure, *z*, by the relationship

$$q = \alpha + \beta z + v$$

where $v$ is a disturbance term which satisfies the Gauss-Markov conditions. Both variables $q$ and $z$ are measured with error, and from the meaning of variables of this model $q$ is part of $z$, so any error in the estimation of *q* affects the estimate of *z* **by the same amount**. Hence

$$y_i = q_i + w_i$$

and

$$x_i = z_i + w_i$$

where $y_i$ is the estimated value of $q_i$, $x_i$ is the estimated value of $z_i$, and $w_i$ is the measurement error **affecting both variables** in observation *i*. It is assumed that the expected value of *w* is zero and that *v* and *w* are distributed independently of *z* and of each other. Note since shadow expenditure is a component of total consumer expenditure, $\beta$ will lie between 0 and 1.

**(a)** Derive an expression for the large-sample bias in the estimate of $\beta$ when Ordinary Least Squares is used to regress $y_i$ on $x_i$, and determine its sign if this is possible.

**a)** The relationship between the observed variables is

$$(y - w) = \alpha + \beta(x - w) + v$$

so

$$y = \alpha + \beta x + v + (1 - \beta)w = \alpha + \beta x + u$$

where

$$u = v + (1-\beta)w$$

If *y* is regressed on *x* using OLS,

$$\begin{aligned} b_{OLS} &= \frac{\operatorname{Cov}(x,y)}{\operatorname{Var}(x)} = \frac{\operatorname{Cov}(x,[\alpha+\beta x+u])}{\operatorname{Var}(x)} \\ &= \frac{\operatorname{Cov}(x,\alpha)+\operatorname{Cov}(x,\beta x)+\operatorname{Cov}(x,u)}{\operatorname{Var}(x)} = \\ &= \beta + \frac{\operatorname{Cov}(x,u)}{\operatorname{Var}(x)} \end{aligned}$$

It is not possible to obtain an expression for the expected value of $b_{\text{OLS}}$ because both the numerator and the denominator of the error term are functions of *w*, so we will investigate its plim instead.

$$\begin{aligned} \operatorname{plim}\operatorname{Cov}(x,u) &= \operatorname{plim}\operatorname{Cov}([z+w],[v+(1-\beta)w]) \\ &= \operatorname{plim}\operatorname{Cov}(z,v) + (1-\beta)\operatorname{plim}\operatorname{Cov}(z,w) + \\ &+ \operatorname{plim}\operatorname{Cov}(w,v) + (1-\beta)\operatorname{plim}\operatorname{Var}(w) \end{aligned}$$

The first three plims are zero because *v* and *w* are distributed independently of *z* and each other, so

$$\operatorname{plim}\operatorname{Cov}(x,u) = (1-\beta)\sigma_w^2$$

Next

$$\begin{aligned} &\operatorname{plim}\operatorname{Var}(x) = \operatorname{plim}\operatorname{Var}(z+w) = \\ &= \operatorname{plim}\operatorname{Var}(z) + \operatorname{plim}\operatorname{Var}(w) + 2\operatorname{plim}\operatorname{Cov}(z,w) = \\ &= \sigma_z^2 + \sigma_w^2 \end{aligned}$$

Hence

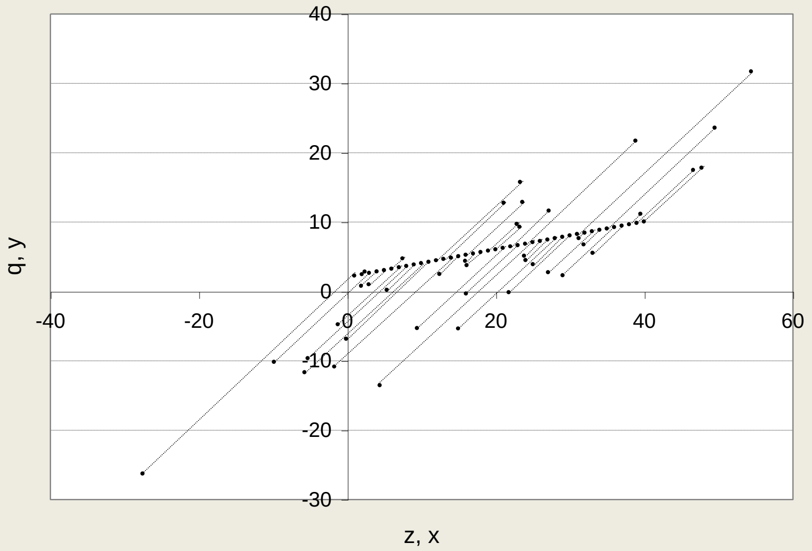
$$\operatorname{plim} b_{OLS} = \beta + (1-\beta)\frac{\sigma_w^2}{\sigma_z^2+\sigma_w^2}$$

Since $\beta$ will lie between 0 and 1, the bias will be positive.

**(b)** Show that large sample value of the slope of the regression line is the weighted average of the true slope, $\beta$, and unity: $\operatorname{plim} b_{OLS} = C_1 \cdot \beta + C_2 \cdot 1$; $C_1 + C_2 = 1$. Find $C_1$ and $C_2$. Explain the possible meaning of this relation formally, economically and graphically.

From the result above $\operatorname{plim} b_{OLS} = \beta + (1-\beta)\frac{\sigma_w^2}{\sigma_z^2+\sigma_w^2}$ we get immediately $\operatorname{plim} b_{OLS} = \frac{\sigma_z^2}{\sigma_z^2+\sigma_w^2}\beta + \frac{\sigma_w^2}{\sigma_z^2+\sigma_w^2}$. Hence the slope of the regression line will be a compromise between the true slope, $\beta$, and unity. Thus plim $b_{\text{OLS}}$ is a weighted average of $\beta$ and unity, the weights being the variances of *z* and *w*.

The graph below shows plots the points $(q,z)$ and $(y,x)$ for the first sample, with each $(q,z)$ point linked to the corresponding $(y,x)$ point.



The diagram shows how the measurement error causes the observations to be displaced along 45° lines (slope of these lines is unity). The greater is the variance of the error term $w_i$ the more displacement is observed along these lines with the slope 1, and so the greater if the weight of 1.

**(c)** The researcher is worried of the fact that consumer expenditure on the shadow economy were systematically underestimated, so the expected value of the measurement error being not zero but rather negative?
He is also worried of the fact that the analysis could be affected by positive correlation of *w* with *z*, as countries with large *z* tend to have larger measurement errors *w*. Comment.

The results would not be affected by the expected value of the measurement error term. The analysis of the large-sample bias in fact use no assumption concerning *E*(*w*).

The second modification is in fact essential for analysis. We would now have

$$\operatorname{plim}\operatorname{Cov}(x,u) = \operatorname{plim}\operatorname{Cov}([z+w],[v+(1-\beta)w]) =$$
$$= \operatorname{plim}\operatorname{Cov}(z,v) + (1-\beta)\operatorname{plim}\operatorname{Cov}(z,w) + \operatorname{plim}\operatorname{Cov}(w,v) + (1-\beta)\operatorname{plim}\operatorname{Var}(w)$$

Hence

$$\operatorname{plim} b_{OLS} = \beta + (1-\beta)\frac{\sigma_w^2 + \sigma_{w,z}}{\sigma_z^2 + \sigma_w^2 + 2\sigma_{w,z}}$$

where $\sigma_{w,z}$ is $\operatorname{plim}\operatorname{Cov}(w,z)$. The bias could either increase or decrease, depending on the sign and the value of $\beta$.

**(d)** Trying to overcome consequences of bias caused by measurement errors the researcher decided to use disposable personal income, *I* as an instrument for total consumer expenditure, *z*, assuming that *I* correlates with *z* but not correlates with *v* and *w*. Comment providing necessary proofs, taking into account that consumer expenditures on shadow goods and services, *q*, still are under measurement errors *w* .

The idea is quite reasonable, as instrumental variable *I* allows to get unbiased estimator for $\beta$. Taking in mind that both $q$ and $z$ are unobservable we get

$$b_{IV} = \frac{\operatorname{Cov}(I,y)}{\operatorname{Cov}(I,x)} = \frac{\operatorname{Cov}(I,q+w)}{\operatorname{Cov}(I,z+w)} = \frac{\operatorname{Cov}(I,\alpha+\beta z+v+w)}{\operatorname{Cov}(I,z+w)} =$$
$$= \frac{\operatorname{Cov}(I,\alpha)+\operatorname{Cov}(I,\beta z)+\operatorname{Cov}(I,v+w)}{\operatorname{Cov}(I,z+w)} = \frac{0+\beta\cdot\operatorname{Cov}(I,z)+\operatorname{Cov}(I,v)+\operatorname{Cov}(I,w)}{\operatorname{Cov}(I,z)+\operatorname{Cov}(I,w)}$$

Taking probability limit we get

$$\operatorname{plim} b_{IV} = \frac{\beta \cdot \operatorname{plim}\operatorname{Cov}(I,z) + \operatorname{plim}\operatorname{Cov}(I,v) + \operatorname{plim}\operatorname{Cov}(I,w)}{\operatorname{plim}\operatorname{Cov}(I,z) + \operatorname{plim}\operatorname{Cov}(I,w)} =$$

$$= \frac{\beta \cdot \sigma_{I,z} + \sigma_{I,v} + \sigma_{I,w}}{\sigma_{I,z} + \sigma_{I,w}} = \frac{\beta \cdot \sigma_{I,z} + 0 + 0}{\sigma_{I,z} + 0} = \beta \cdot \frac{\sigma_{I,z}}{\sigma_{I,z}} = \beta$$

We should be aware that efficiency of this estimator could be low. Remaining errors *w* in the measurement *q* make no effect on consistency of estimator, but increase the variance of disturbance term and so increase standard errors of estimator.

**5. [30 marks]** The following simultaneous equations model is considered:

$$Y = \beta_1 + \beta_2 X + u \qquad (1)$$

$$X = \alpha_2 Y + v \qquad (2)$$

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

**(a)** Derive reduced form system of equations for the system above.
What information gives the reduced form system on the properties of possible estimators for the coefficients of equations (1) and (2)?
What can be said on the identification of these equations?

Find reduced form equation (1)

$$Y = \beta_1 + \beta_2 X + u \;\Rightarrow\; Y = \beta_1 + \beta_2(\alpha_2 Y + v) + u \;\Rightarrow\; Y = \beta_1 + \alpha_2\beta_2 Y + u + \beta_2 v \;\Rightarrow$$

$$\Rightarrow\; (1-\alpha_2\beta_2)Y = \beta_1 + u + \beta_2 v \;\Rightarrow\; Y = \frac{1}{1-\alpha_2\beta_2}(\beta_1 + u + \beta_2 v)$$

Find reduced form equation (2)

$$X = \alpha_2 Y + v \;\Rightarrow\; X = \alpha_2(\beta_1 + \beta_2 X + u) + v \;\Rightarrow\; X = \alpha_2\beta_1 + \alpha_2\beta_2 X + v + \alpha_2 u \;\Rightarrow$$

$$\Rightarrow\; (1-\alpha_2\beta_2)X = \alpha_2\beta_1 + v + \alpha_2 u \;\Rightarrow\; X = \frac{1}{1-\alpha_2\beta_2}(\alpha_2\beta_1 + v + \alpha_2 u)$$

Equation $Y = \frac{1}{1-\alpha_2\beta_2}(\beta_1 + u + \beta_2 v)$ shows that disturbance term $v$ is a part of $Y$, so in equation $X = \alpha_2 Y + v$ (2) the explanatory variable $Y$ certainly correlates with $v$, and so GMC are violated so the OLS estimate of the equation (2) is not BLUE.

The same situation with equation $Y = \beta_1 + \beta_2 X + u$ (1), as from $X = \frac{1}{1-\alpha_2\beta_2}(\alpha_2\beta_1 + v + \alpha_2 u)$ follows that disturbance term $u$ is a part of explanatory variable $X$.

The situation with identification of these equations is not simple: from the first glance there is no explanatory variable at all. But analyzing the structure of these equations we can notice that equation (1) includes constant term $\beta_1$ while the equation (2) does not. This constant term can be considered here as coefficient of artificial variable $Z$ identically equal to one: $Y = \beta_1 Z + \beta_2 X + u$. So from order condition follows that $G-1=2-1=1$, while the number of missed variables in the right side of the second equation is also 1 (we consider here $Z$ as the missed variable), and so second equation is identified, while first equation has no missed variable and so is underidentified.

**(b)** Show that OLS estimator $\hat{\alpha}_2^{OLS}$ of $\alpha_2$ is inconsistent and find large sample bias.

As the second equation $X = \alpha_2 Y + v$ does not include intercept OLS estimator of $\alpha_2$ is $\hat{\alpha}_2^{OLS} = \dfrac{\sum X_i Y_i}{\sum Y_i^2}$.

Using reduced form equation for $X = \dfrac{1}{1-\alpha_2\beta_2}(\alpha_2\beta_1 + v + \alpha_2 u)$ and $Y = \dfrac{1}{1-\alpha_2\beta_2}(\beta_1 + u + \beta_2 v)$ find

$$\hat{\alpha}_2^{OLS} = \frac{\dfrac{1}{(1-\alpha_2\beta_2)^2}\sum(\alpha_2\beta_1 + v_i + \alpha_2 u_i)(\beta_1 + u_i + \beta_2 v_i)}{\dfrac{1}{(1-\alpha_2\beta_2)^2}\sum(\beta_1 + u_i + \beta_2 v_i)^2} =$$

$$= \frac{\sum(\alpha_2\beta_1(\beta_1 + u_i + \beta_2 v_i) + v_i(\beta_1 + u_i + \beta_2 v_i) + \alpha_2 u_i(\beta_1 + u_i + \beta_2 v_i))}{\sum(\beta_1^2 + u_i^2 + \beta_2^2 v_i^2 + 2\beta_1 u_i + 2\beta_2 u_i v_i + 2\beta_1\beta_2 v_i)} =$$

$$= \frac{\frac{1}{n}\sum((\alpha_2\beta_1\beta_1 + \alpha_2\beta_1 u_i + \alpha_2\beta_1\beta_2 v_i) + (v_i\beta_1 + v_i u_i + \beta_2 v_i^2) + (\alpha_2 u_i\beta_1 + \alpha_2 u_i^2 + \alpha_2\beta_2 u_i v_i))}{\frac{1}{n}\sum(\beta_1^2 + u_i^2 + \beta_2^2 v_i^2 + 2\beta_1 u_i + 2\beta_2 u_i v_i + 2\beta_1\beta_2 v_i)}$$

Summing up and simplifying

$$= \frac{\alpha_2\beta_1^2 + 2\alpha_2\beta_1\frac{1}{n}\sum u_i + (\alpha_2\beta_1\beta_2 + \beta_1)\frac{1}{n}\sum v_i + (1+\alpha_2\beta_2)\frac{1}{n}\sum u_i v_i + \beta_2\frac{1}{n}\sum v_i^2 + \alpha_2\sum u_i^2}{\beta_1^2 + \frac{1}{n}\sum u_i^2 + \beta_2^2\frac{1}{n}\sum v_i^2 + 2\beta_1\frac{1}{n}\sum u_i + 2\beta_2\frac{1}{n}\sum u_i v_i + 2\beta_1\beta_2\frac{1}{n}\sum v_i}$$

Now taking probability limits $\operatorname{plim}\hat{\alpha}_2^{OLS}$ we have the following recurring situations

$\operatorname{plim}\frac{1}{n}\sum u_i = Eu = 0$, $\operatorname{plim}\frac{1}{n}\sum v_i = Ev = 0$, $\operatorname{plim}\frac{1}{n}\sum u_i v_i = Euv = 0$, $\operatorname{plim}\frac{1}{n}\sum u_i^2 = \sigma_u^2$, $\operatorname{plim}\frac{1}{n}\sum v_i^2 = \sigma_v^2$

So finally we get $\operatorname{plim}\hat{\alpha}_2^{OLS} = \dfrac{\alpha_2\beta_1^2 + \alpha_2\sigma_u^2 + \beta_2\sigma_v^2}{\beta_1^2 + \sigma_u^2 + \beta_2^2\sigma_v^2} \neq \alpha_2$

**(c)** For the estimation of parameter $\alpha_2$ of equation (2) a researcher suggests the following estimator $\hat{\alpha}_2 = \dfrac{\bar{X}}{\bar{Y}}$ (where $\bar{X}$ and $\bar{Y}$ are the sample means of *X* and *Y*). The researcher believes that it is consistent. Investigate this suggestion.

Consider suggested estimator $\hat{\alpha}_2 = \dfrac{\bar{X}}{\bar{Y}} \Rightarrow \hat{\alpha}_2 = \dfrac{\dfrac{1}{1-\alpha_2\beta_2}(\alpha_2\beta_1 + \bar{v} + \alpha_2\bar{u})}{\dfrac{1}{1-\alpha_2\beta_2}(\beta_1 + \bar{u} + \beta_2\bar{v})} = \dfrac{\alpha_2\beta_1 + \bar{v} + \alpha_2\bar{u}}{\beta_1 + \bar{u} + \beta_2\bar{v}}$

As both numerator and denominator are stochastic it is impossible to take expectations, so consider probability limit of the estimator

$\operatorname{plim}\hat{\alpha}_2 = \operatorname{plim}\dfrac{\alpha_2\beta_1 + \bar{v} + \alpha_2\bar{u}}{\beta_1 + \bar{u} + \beta_2\bar{v}} = \dfrac{\alpha_2\beta_1 + \operatorname{plim}(\bar{v}) + \alpha_2\operatorname{plim}(\bar{u})}{\beta_1 + \operatorname{plim}(\bar{u}) + \beta_2\operatorname{plim}(\bar{v})} = \dfrac{\alpha_2\beta_1 + Ev + \alpha_2 Eu}{\beta_1 + Eu + \beta_2 Ev} = \dfrac{\alpha_2\beta_1}{\beta_1} = \alpha_2$, as

$\operatorname{plim}(\bar{u}) = Eu = 0$ and $\operatorname{plim}(\bar{v}) = Ev = 0$. So $\operatorname{plim}\hat{\alpha}_2 = \alpha_2$, and estimator $\hat{\alpha}_2 = \dfrac{\bar{X}}{\bar{Y}}$ is consistent.

**(d)** What is ILS estimation? Prove that $\overline{X}$ and $\overline{Y}$ are OLS estimators for $\pi_1$ and $\pi_2$ in degenerate regressions: $Y = \pi_1 + \omega_1$ (3) and $X = \pi_2 + \omega_2$ (4). Derive from here that ILS estimator $\hat{\alpha}_2^{ILS}$ for $\alpha_2$ is $\hat{\alpha}_2^{ILS} = \frac{\hat{\pi}_2^{OLS}}{\hat{\pi}_1^{OLS}} = \frac{\overline{X}}{\overline{Y}}$. Is it consistent?

The ILS method consists in the evaluation of an auxiliary system of equations and the subsequent derivation of estimates algebraically.
Let's write reduced form equations in a short form

$$Y = \frac{1}{1-\alpha_2\beta_2}(\beta_1 + u + \beta_2 v)$$

$$X = \frac{1}{1-\alpha_2\beta_2}(\alpha_2\beta_1 + v + \alpha_2 u)$$

In brief

$$Y = \pi_1 + \omega_1 \qquad (3)$$

$$X = \pi_2 + \omega_2 \qquad (4)$$

where

$$\pi_1 = \frac{\beta_1}{1-\alpha_2\beta_2}$$

$$\pi_2 = \frac{\alpha_2\beta_1}{1-\alpha_2\beta_2}$$

$$\pi_1 = \frac{\beta_1}{1-\alpha_2\beta_2}$$

$$\alpha_2 = \frac{\frac{\alpha_2\beta_1}{1-\alpha_2\beta_2}}{\frac{\beta_1}{1-\alpha_2\beta_2}} = \frac{\pi_2}{\pi_1}$$

The values $\overline{Y}$ and $\overline{X}$ are OLS estimators for $\pi_1$ and $\pi_2$ in degenerate equations (3) and (4):
Really, for example for $Y = \pi_1 + \omega_1$: $S = \sum (Y_i - \hat{\pi}_1^{OLS})^2 \to \min \Rightarrow$ FOC: $-2\sum (Y_i - \hat{\pi}_1^{OLS}) = 0 \Rightarrow$
$\Rightarrow n\hat{\pi}_1^{OLS} = \sum Y_i \Rightarrow \hat{\pi}_1^{OLS} = \frac{1}{n}\sum Y_i = \overline{Y}$, so $\hat{\pi}_1^{OLS} = \overline{Y}$. The same for $\hat{\pi}_2^{OLS} = \overline{X}$.

Now ILS estimator of $\alpha_2$ can be obtained as $\hat{\alpha}_2^{ILS} = \frac{\hat{\pi}_2^{OLS}}{\hat{\pi}_1^{OLS}} = \frac{\overline{X}}{\overline{Y}}$.
This estimator coincides with the suggested estimator in **(c)** so it is consistent.