Foundations of Econometrics - Part I

SAMPLE of exam questions

Please use this sample of exam questions as an indication of the style of questions you might encounter in your exam, but not as an indication of the content of the exam. That is, topics covered during the course that do not appear below might be included in your exam.

1. Sachs&Warner(2001) proposed the following regression to test, using data between 1970 and 1990, the so-called *natural resource curse*, which states that a country's natural resource abundance is detrimental for its economic growth.

$$Model(1) \qquad growth7090_i = \beta_1 + \beta_2 sxp_i + \beta_3 lgdp70_i + \beta_4 linv7089_i + \beta_5 open7090_i + \beta_6 rl_i + \epsilon_i$$

where $growth7090$=average economic growth rate from 1970-1990 (as a percentage), sxp=ratio of primary exports on GDP 1970 ($primary\ exports/GDP$), which is used as a proxy of natural resource abundance, $lgdp70$=log of GDP1970, $linv7089$=log of investment 1970-1989, $open7090$=degree of openness of the economy between 1970-1990, rl=rule of law, measure of the quality of country's institutions, set as an index from 0(poorest) to 6(best). The result of estimating this model using OLS estimator (*Stata* default OLS output):

Source	SS	df	MS	Number of obs	=	74
				F(5, 68)	=	39.50
Model	187.586202	5	37.5172404	Prob > F	=	0.0000
Residual	64.5936098	68	.949906026	R-squared	=	0.7439
				Adj R-squared	=	0.7250
Total	252.179812	73	3.45451797	Root MSE	=	.97463

growth7090	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sxp	-8.074527	1.230538	-6.56	0.000	-10.53003	-5.619027
lgdp70	-1.76612	.2087326	-8.46	0.000	-2.182639	-1.349601
linv8089	1.251271	.2898759	4.32	0.000	.6728328	1.829709
open7090	1.542481	.3942322	3.91	0.000	.7558028	2.329159
rl	.32149	.1030098	3.12	0.003	.1159371	.5270429
_cons	12.15513	1.585444	7.67	0.000	8.99143	15.31884

   (a) From the output above, excluding number of observations and degrees of freedom, select 2 different statistics whose calculation does not depend on any assumption regarding the data generating process (dgp). Clearly label the statistics you selected. Be rigorous with the notation. No need to justify.

   (b) From the output above, select 2 different statistics whose calculation does depend on the assumptions placed on the data generating process (dgp). Clearly label the statistics you selected. No need to justify.

   (c) Consider we changed the units of measurement of regressor sxp from the current ratio over 1 to a percentage. From the output above, which statistics would change value? List them by name. Be rigorous with the notation.

   (d) Looking at the output above, we can tell that not all countries in the sample had exactly the same quality of institutions, as measured by regressor rl. Why? Rigorously argue using the expression you feel is most appropriate to support your argument.

   (e) First country in the sample was Algeria. For this country we know that the OLS residual is $\widehat{\epsilon}_1 = 0.791$. ($i$) Specify the units of measurement of this residual. (ii) Interpret the meaning of this residual, explaining what information does it give you about the average economic growth rate from 1970-1990 for this country. Be specific.

   (f) One might say that the coefficient of determination can be interpreted as a measure of the in-sample predictive power of the non-constant regressors. Explain why.

(g) Consider *sxp* is a good measure of a country's natural resource abundance. (*i*) Discuss rigorously up to what point the result of the estimation above can be used by the authors as support for the natural resource curse. (*ii*) Does the high value of the coefficient of determination play any role in the discussion?

(h) Consider the $t-value = -6.6634$ and $p-value = 0.000$, quoted in the first row of the table above. (*i*) Draw this $p-value$ in a graph. Do not forget to clearly label the axes and to identify any relevant element to clearly identify this particular $p-value$. (*ii*) Based on this $p-value$, would you say *sxp* is statistically significant? At what level?

(i) Consider we are interested in testing, at 1% significance level, whether *linv*7089 and *open*7090 are jointly significant, using the following test statistic:

$$F = \frac{(RSSE - SSE)/q}{SSE/(n-K)} \underset{\text{under } H_o}{\sim} F(q, n-K).$$

(*i*) Detail the null and alternative hypotheses associated with this test; (*ii*) detail where would you get the value $RSSE$ needed to calculate the $F-value$. Be specific ; (*iii*) Consider the following information for $F \sim F(2,68)$: $Prob\{F > 4.08\} = 0.01$, $Prob\{F > 2.73\} = 0.05$, $Prob\{F > 2.16\} = 0.1$. If you are told that, for this test, $F-value = 18.5$, what would the conclusion of the test be? (*iv*) Would the associated $p-value$ be smaller or larger than 0.01? Briefly justify.

(j) If all so-called classical assumptions, except for normality, were holding, (*i*) what would the distribution of the test statistic we used in 1i be? (*ii*) Would the failure of normality affect the $p-value$ associated with this test? Discuss rigorously.

(k) Consider that we want to test the normality of disturbances using the Jarque-Bera test statistic:

$$JB \equiv \frac{n}{6}\left(sk^2 + \frac{(kur-3)^2}{4}\right) \overset{a}{\sim} \chi^2(2),$$

where *sk* is the sample coefficient of skewness and *kur* is the sample coefficient of kurtosis. (*i*) Draw the acceptance and rejection regions associated with this test statistic. Clearly label both axes and all relevant elements in your graph. (*ii*) Explain the intuition behind the location of the acceptance region.

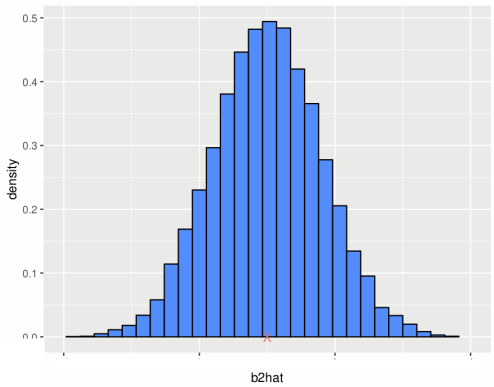
2. Consider the following *dgp*:

$$\begin{aligned} y_i &= 10 + 1\cdot x_{i2} + 1\cdot x_{i3} + \epsilon_i \quad \epsilon_i/X \sim i.i.N(0,81) \\ & x_{i2} \sim U[0,20] \qquad x_{i3} = x_{i2} + v_i \quad v_i \sim i.i.N(0,4) \end{aligned}$$

Using this *dgp*, generated 10,000 samples of 50 observations each ($n = 50$) and with each of the generated samples, we estimated by OLS the following regression:

$$y = \beta_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \epsilon.$$

With the OLS estimates of parameter β_2 the following density histogram was produced:



(a) Could the range of values of the vertical axis have been greater than 1? Justify your answer.

(b) Around which value do you expect the histogram to be centered around? Which property of OLS estimator could this simulation help illustrate and which one it could not? Justify your answer.

(c) Change one element of the dgp that would reduce the uncertainty around the estimation of parameter β_2. Specify which element you selected and justify your choice.

3. Consider article by Enikolopov et al.(2018), "Social media and corruption", *AEJ: Applied Economics*. In this article authors estimate the following regression:

$$(1) \quad Corruption = \beta_1 + \beta_2 ln(gdppc) + \beta_3 Socialnetshare + \epsilon.$$

where $Corruption$=corruption index (higher value indicating higher corruption), $gdppc$=GDP per capita, $Socialnetshare$=share of social network users, and produce the following figure:

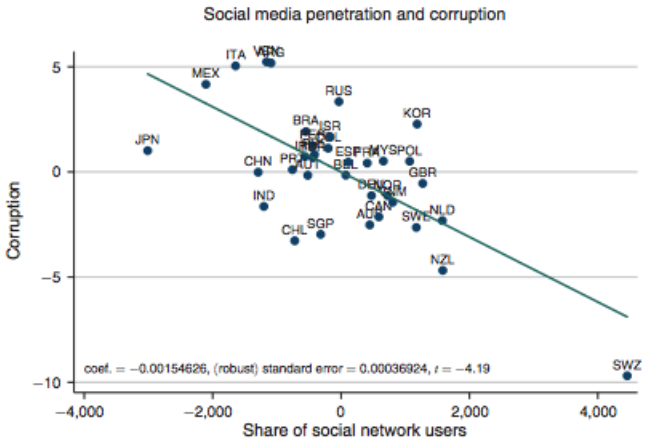


FIGURE 1. SOCIAL MEDIA PENETRATION AND CORRUPTION

Notes: Figure 1 is an added variable plot. GDP per capita is controlled for. Social media penetration is from *ComScore* 2011. Corruption is from Corruption Perception Index 2004–2012. Corruption index is transformed so that higher values mean higher corruption.

where the variable in the horizontal and vertical axis are, respectively, the OLS residuals of regressions:

$$(2) \quad Socialnetshare = \alpha_1^S + \alpha_2^S ln(gdppc) + \epsilon^S,$$
$$(3) \quad Corruption = \alpha_1^C + \alpha_2^C ln(gdppc) + \epsilon^C.$$

(a) Explain why the variables in both, the horizontal axis and the vertical axis, are centered around 0.

(b) Why do the author's state in the figure notes that *GDP per capita is controlled for*? Explain.

(c) In the graph author's quote:

$$coef. = -0.00155626, \text{ (robust) standard error}=0.00036924, \ t = -4.19.$$

(i) Which parameter from which regression do you think the quoted estimate (i.e., $coef. = -0.00155626$) an estimate of? Briefly explain.

(ii) Provide a reason why authors chose to use robust standard errors.

(iii) Which test is the $t - value$ referring to? Provide the null hypothesis.