Foundations of Econometrics - Part I

Sample of exam questions · December 2025

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. Jha&Sarangi(2017) article "Does Social Media Reduce Corruption?" study the effect of internet and social media penetration on corruption using cross-country analysis. Among others, the authors consider the following regression:

$$CI = \beta_1 + \beta_2 fbpen + \beta_3 netpen + \beta_4 prights + \epsilon$$

where CI=country corruption index (the higher CI, the more corrupt the country is), $fbpen$=% of facebook users (as a proxy of social media penetration), $netpen$=internet penetration (%), $prights$=political rights index, index varying from 1(best political rights) to 7 (worse). OLS estimation of this model using 2012 data (*Stata* default output) is:

Source	SS	df	MS	Number of obs	=	177
				F(3, 173)	=	129.52
Model	121.351283	3	40.4504276	Prob > F	=	0.0000
Residual	54.0283574	173	.312302644	R-squared	=	0.6919
				Adj R-squared	=	0.6866
Total	175.37964	176	.996475229	Root MSE	=	.55884

CI	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
fbpen	-.0101358	.004374	-2.32		-.018769	-.0015025
netpen	-.0167164	.0026357	-6.34	0.000	-.0219186	-.0115142
prights	-.1326323	.0245707	-5.40	0.000	-.1811291	-.0841354
_cons	.4844418	.136533	3.55	0.000	.2149569	.7539267

(a) Consider that in the regression above, parameter vector β, including all the regression coefficients, was defined as:

$$\beta = \left[E(xx')\right]^{-1} E(xy),$$

with x being the column vector including all the regressors, including the constant, and y the dependent variable. Keeping this definition in mind: (i) what would be the dimension of element β, $E(xx')$, and $E(xy)$? (ii) Provide one reason why the authors could be interested in estimating vector β as just defined. Justify.

(b) Consider running a regression of the OLS residuals from estimating the model above with respect to a constant, $fbpen$, $netpen$ and $prights$. What would you expect the coefficient of determination of running this regression to be. Justify.

(c) We want to test $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$, using the exact $t-$test statistic.

(i) Provide the expression of the exact $t - statistic$ and its assumed distribution under H_0.

(ii) Perform test using a 1% significance level? Justify. *(Useful information below.)*

(iii) In the output above, the $p - value$ associated with this test is missing. Given your answers to the previous two questions, what can you say about the size of the missing $p - value$? Be as specific as possible.

(iv) Draw the missing $p - value$. Do not forget to label both axes and to identify any relevant element to clearly identify this particular $p - value$.

Needed information regarding $t \sim t(173)$: $Prob\{t > 1.29\} = 0.1$, $Prob\{t > 1.65\} = 0.05$, $Prob\{t > 1.97\} = 0.025$, $Prob\{t > 2.35\} = 0.01$, $Prob\{t > 2.6\} = 0.005$.

(d) List the assumptions, regarding the *dgp* behind the data, that would be needed to justify the use of the exact t-test statistic. Label each assumption and next to each label, provide the corresponding statistical expression that defines the assumption, considering regressors are stochastic. No need to justify.

(e) Consider testing whether $fbpen$ and $netpen$ are jointly significant at 1% significance level, using the exact $F-$test statistic, expressed as:

$$F = \frac{\left(R\hat{\beta} - r\right)' \left[R(X'X)^{-1}R'\right]^{-1} \left(R\hat{\beta} - r\right) \Big/ q}{SSE/\left(n-K\right)} \underset{\text{under } H_o}{\sim} F(q, n-K).$$

(i) Detail the null and alternative hypotheses associated with this test.

(ii) Detail the exact values of the following 4 elements: R, SSE, q and r. No need to justify.

(iii) Draw the acceptance and rejection region associated with this test. Critical value is not available, but label it in the graph, using the notation we used in the course to properly identify it. Label both axes.

(iv) Write the expression that defines the critical value associated with this test.

(v) Provide the intuition of the location of the acceptance region you drew in (iii).

(vi) Knowing that the $F-value = 82.13$, detail what additional information you would need to finish the test, and how you would use it.

(f) If all so-called classical assumptions, except for normality, were holding, (i) what would the distribution of the test statistic used in 1c be? Provide the expression of the test statistic and its distribution (ii) Would the failure of normality affect the size of the $p-value$ you drew in 1c(v)? If so, how? Which one would be larger? Rigorously justify.

(g) Consider the following elements: $\widehat{\beta_2}$, $se(\widehat{\beta_2})$, R^2, and 95% confidence interval for β_2, whose values are reported in the output above. Consider assumption of conditional homoskedasticity did not apply.

(i) Which of the 4 elements listed would require adjustment in calculating its value and which would not? Just state Yes ('adjustment required') or No ('adjustment not required').

(ii) Select one element that <u>would not</u> require adjustment and justify why you selected it.

(iii) Select one element that <u>would</u> require adjustment and justify why you selected it.

(h) (i) Under what condition could we give the OLS estimates of β_2 and β_3 a causal interpretation? Rigorously explain, using the help of a causal path diagram. (ii) Does the value of R^2 play any role in helping us determining whether we can use the estimates of β_2 and β_3 for causal inference? And the size of the sample? Rigorously discuss.

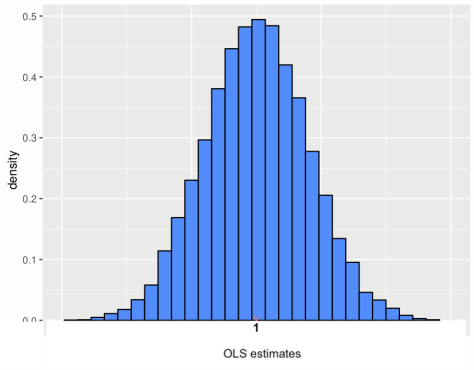
2. *(2.5 points)* Consider the following *dgp*:

$$\begin{aligned} y_i &= 1 + 1 \cdot x_{i2} + 1 \cdot x_{i3} + \epsilon_i \qquad \epsilon_i/X \sim i.i.N(0, 64) \\ & \quad x_{i2} \sim i.i.U[0, 20] \qquad x_{i3} = x_{i2} + v_i \quad v_i \sim i.i.N(0, 4) \end{aligned}$$

Using this *dgp*, we generated 10,000 samples of 50 observations each ($n = 50$), keeping the 50 observations of regressors x_2 and x_3 the same across all samples. 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) *"The histogram above illustrates the concept of the conditional sampling distribution of the OLS estimator under classical assumptions"*. Would you agree with this statement? Rigorously discuss.

(b) Using the described simulation exercise, (i) identify a property of OLS estimator that can be illustrated and (ii) a property that cannot be illustrated. Rigorously justify both, by providing the expression that defines the property and how you can, or cannot, illustrate it with this experiment.

(c) If with each of the generated samples we calculated the 99% confidence interval for parameter β_2, how many of these intervals would you expect not to include a 1? Justify.

(d) Recall the expression of the variance decomposition of OLS estimator under $Gauss - Markov$ assumptions:

$$var(\widehat{\beta}_2/X) = \sigma^2 \cdot \frac{1}{SST_2} \cdot \frac{1}{1 - R_2^2}.$$

(i) Change one of the elements of the *dgp* above to create perfect collinearity. Clearly indicate the element that you changed and how. Using the expression of the variance decomposition provided, rigorously argue why we would not be able to uniquely estimate β_2. Additionally, describe how the histogram provided above would look like.

(ii) Now, go back to the original *dgp* provided and change another element so that with any sample generated, we could perfectly estimate β_2. Again, clearly indicate the element that you changed and how. Using the expression provided, explain why in this case we would be able to perfectly estimate β_2. Additionally, describe how the histogram provided above would look like.

3. A linear regression model with 4 regressors $(const, x_2, x_3, x_4)$ is set. We want to test: $H_0 : \beta_2 = \beta_3$ versus $H_1 : \beta_2 \neq \beta_3$. To perform this test we want to use estimator *tilde*, $\tilde{\beta}$, (i.e., not the OLS estimator!), which has the following asymptotic distribution:

$$\sqrt{n}(\tilde{\beta} - \beta) \overset{a}{\sim} N(0, \Omega).$$

(a) Specify the dimension of the following elements of the asymptotic distribution above: $\tilde{\beta}$, β, 0, Ω. No need to justify.

(b) Derive, step by step, a test statistic to perform the test above using this estimator tilde, for the case where matrix Ω is known.

(c) Justify why the statistic you derived is a proper test statistic.

4. We set the following data generating process (*dgp*):

$$y = 2 + 1 \cdot x_2 + 1 \cdot x_3 + \epsilon \qquad \epsilon / X \sim N(0, 16) \qquad \begin{bmatrix} x_2 \\ x_3 \end{bmatrix} \sim N\left(\begin{bmatrix} 10 \\ 10 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right), \qquad \rho \in [-1, 1]$$

(a) If we set $\rho = -1$, and generate a sample of 50 observations, show that there would exist an infinite number of OLS estimates for parameters β_2 and β_3 in regression:

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

(b) Using the *dgp* above, and setting whatever additional elements are necessary, design a Monte Carlo experiment to illustrate the effects of collinearity on OLS estimator. Outline the main steps of the experiment and what would you be looking at. *(No coding expected at all!)*