Foundations of Econometrics - Part I
Sample Exam 3 Solutions

SampleExam3

December 7, 2025

Question 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.

$$\text{Model(1)} : \text{growth7090}_i = \beta_1 + \beta_2 sxp_i + \beta_3 lgdp70_i + \beta_4 linv7089_i + \beta_5 open7090_i + \beta_6 rl_i + \epsilon_i$$

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
```

Part (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.

Answer: Assumptions on the dfp include linearity, homoskedacitiy, and uncorrelatedness of error terms.
As a results, two statistics whos calculation do not depend on these assumptions are

$$\hat{\beta}$$

and R^2.
Betahat is calculated as

$$\hat{\beta} = (X'X)^{-1}X'y$$

While R^2 is:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Part (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.

Answer: Standard error of the coefficients and the confidence interval.

$$\text{se}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 \cdot [(X'X)^{-1}]_{jj}}$$

$$\hat{\beta}_j \pm t_{\alpha/2,n-K} \cdot \text{se}(\hat{\beta}_j)$$

Part (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.

Answer: Would change: Coefficient of betahat2, standard error of betahat2, and confidence interval of betahat2
Wouldn't change $\text{R}^2, p-value, t-statistic or any of the coffiecients regression output for other regressors$

Part (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.

Answer: If all countries had the same value for the regressor rl, we would not be able to estimate this beta value.
The standard error would become zero, since there is no variation in the observation. As a result the ViF goes to infintiy:

$$VIF_2 = \frac{1}{1 - R_2^2} = \frac{1}{1-1} = \frac{1}{0} = \infty$$

Also from the prior definition of betahat2, the X matrix becomes sinuglar (as rl is constant across all observations), so we can no longer invert it and calculate betahat2. The regressor would be dropped from the output if this was the case.

Part (e)

First country in the sample was Algeria. For this country we know that the OLS residual is $\hat{\epsilon}_1 = 0.791$.

(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.

Answer: The units of measurement is the average grwoth rate from 1970-90 as a percentage.
The residual is positive, meaning that the model underestimates the average gdp growth rate.
The actual model is 0.791 above the estimate, as given by the residual.

Part (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.

Answer: Yes, R^2 is only used for how much variation in dependent variable is explained by the model. It is not used for projections.
The SST accounts for the constant, while the R^2 explains explanatory power gained by adding additional regressors.

Part (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?

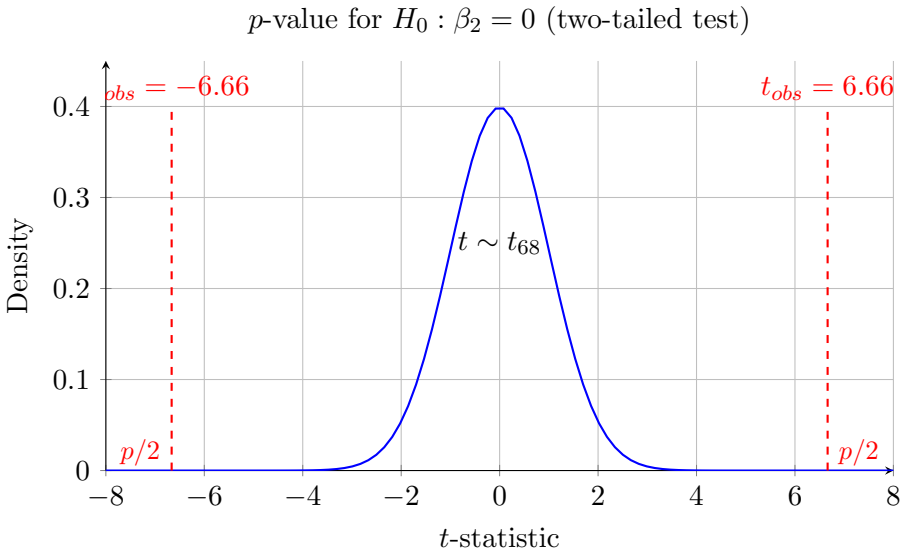
Answer: From the output we can tell there is a negative relationship between exports and economic growth and it is statistically significant. However, by construction the regression output does not prove causality.
R^2 simply measures the explanatory power of the model or goodness of fit, it does not extend into proving causality. It measures the linear relationship between regressors and residuals.

Part (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?



The shaded red areas represent the p-value $= 0.000$, which is:

$$p\text{-value} = P(|t| > 6.66 \mid H_0) = 2 \cdot P(t_{68} > 6.66) \approx 0.000$$

where t_{68} denotes the t-distribution with 68 degrees of freedom ($n - K = 74 - 6 = 68$).

Answer: Yes, it is significant at all levels.

Part (i)

Consider we are interested in testing, at 1% significance level, whether $linv7089$ and $open7090$ are jointly significant, using the following test statistic:

$$F = \frac{(RSSE - SSE)/q}{SSE/(n-K)} \sim_{\text{under } H_0} F(q, n-K).$$

(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 $F \sim F(2, 68)$: $P(F > 4.08) = 0.01$, $P(F > 2.73) = 0.05$, $P(F > 2.16) = 0.1$. 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.

Answer:

Part (j)

If all so-called classical assumptions, except for normality, were holding,

(i) What would the distribution of the test statistic we used in 1(i) be?

(ii) Would the failure of normality affect the p-value associated with this test? Discuss rigorously.

Answer:

Part (k)

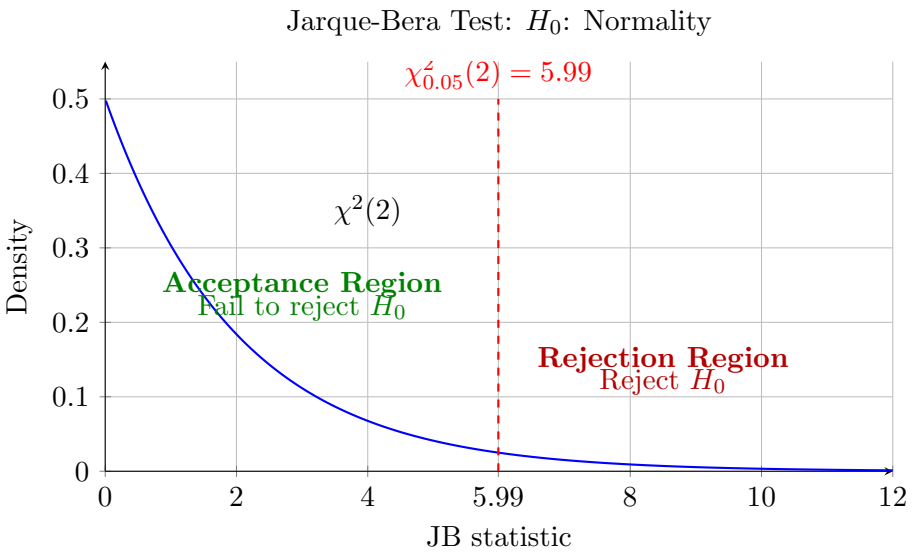
Consider that we want to test the normality of disturbances using the Jarque-Bera test statistic:

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

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.



Note: Using $\alpha = 0.05$, the critical value is $\chi^2_{0.05}(2) = 5.99$.

Answer:
Part (i): See graph above.
Part (ii): Intuition for acceptance region location
The acceptance region is located at **low values of JB** (near 0) because:
Under H_0: Errors are normally distributed

- Normal distribution has: $sk = 0$ (symmetric) and $kur = 3$
- Therefore: $JB = \frac{n}{6}\left(0^2 + \frac{(3-3)^2}{4}\right) = 0$
- Small deviations from normality $\Rightarrow$ JB close to 0

Under H_1: Errors are not normally distributed

- If $sk \neq 0$ (asymmetric) or $kur \neq 3$ (fat/thin tails)
- Then: sk^2 and/or $(kur - 3)^2$ become large
- Therefore: JB becomes large

Conclusion: Small JB values are consistent with normality (accept H_0), while large JB values indicate significant departures from normality (reject H_0). The test is **one-sided** (right-tailed) because we only reject when JB is unusually large, i.e., when $JB > \chi^2_\alpha(2)$.

Question 2

Consider the following dgp:

$$y_i = 10 + 1 \cdot x_{i2} + 1 \cdot x_{i3} + \epsilon_i \quad \epsilon_i | X \sim \text{i.i.} N(0, 81)$$
$$x_{i2} \sim U[0, 20] \quad x_{i3} = x_{i2} + v_i \quad v_i \sim \text{i.i.} N(0, 4)$$

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:

Part (a)

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

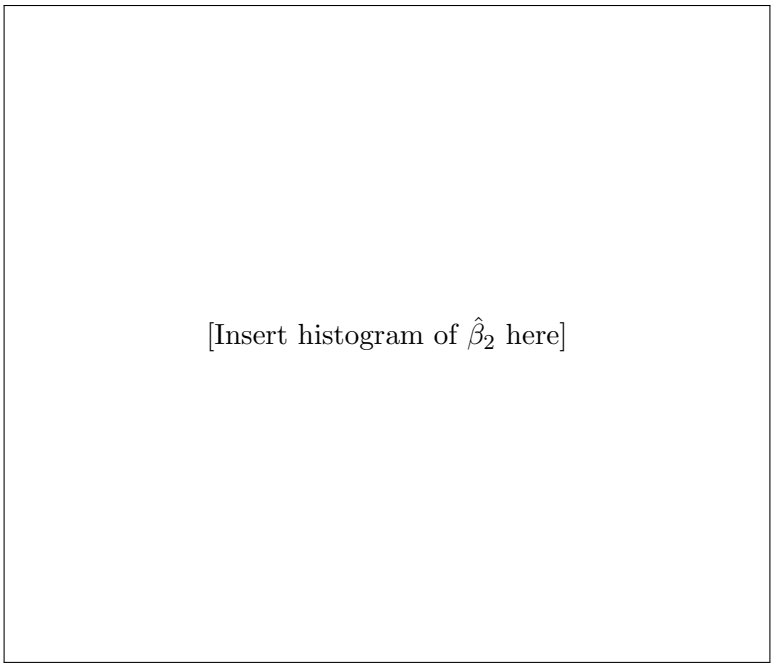


Figure 1: Density histogram of OLS estimates of β_2

Answer:
Yes, the range of values on the vertical axis can be greater than 1.
Key distinction:
The vertical axis shows **density** (probability per unit), not probability itself. There is an important difference:

- The **density** (height of the curve) can exceed 1
- The **total area under the curve** must equal 1 (representing total probability)

Intuition:
Think of density as "concentration" of probability. A histogram or density curve shows how probability is distributed, not the probability itself.
- If a distribution is very **narrow/concentrated**, the curve must be **tall** to ensure the total area still equals 1 - If a distribution is very **wide/spread out**, the curve will be **short**
It's like squeezing a fixed amount of clay: squeeze it narrower and it becomes taller; spread it wider and it becomes shorter. The total amount (area) stays the same, but the height changes.
In this context:
The sampling distribution of $\hat{\beta}_2$ could be quite narrow (low variance) if: - The sample size is large ($n = 50$) - The error variance is moderate ($\sigma^2 = 81$) - There's limited multicollinearity between regressors
A narrow distribution would produce tall peaks that could easily exceed 1 on the vertical axis.
Conclusion: Density is not probability—it's "probability per unit width." Therefore, density values can and often do exceed 1, especially for concentrated distributions.

Part (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.

Answer: We expect it to be centered around the true value of beta2, which from the dgp is one. OLS properties of unbiasedness illustrate this, since it centers around the true value. But it cannot estimate consistency, since n is fixed at 50 and consistency requires n moving to infinity.

Part (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.

Answer: To reduce the uncertainty around β_2, we could condense the distribution of the error terms ei—X, or remove the colinearity given by xi3 (remove xi2 from the model)

Question 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:

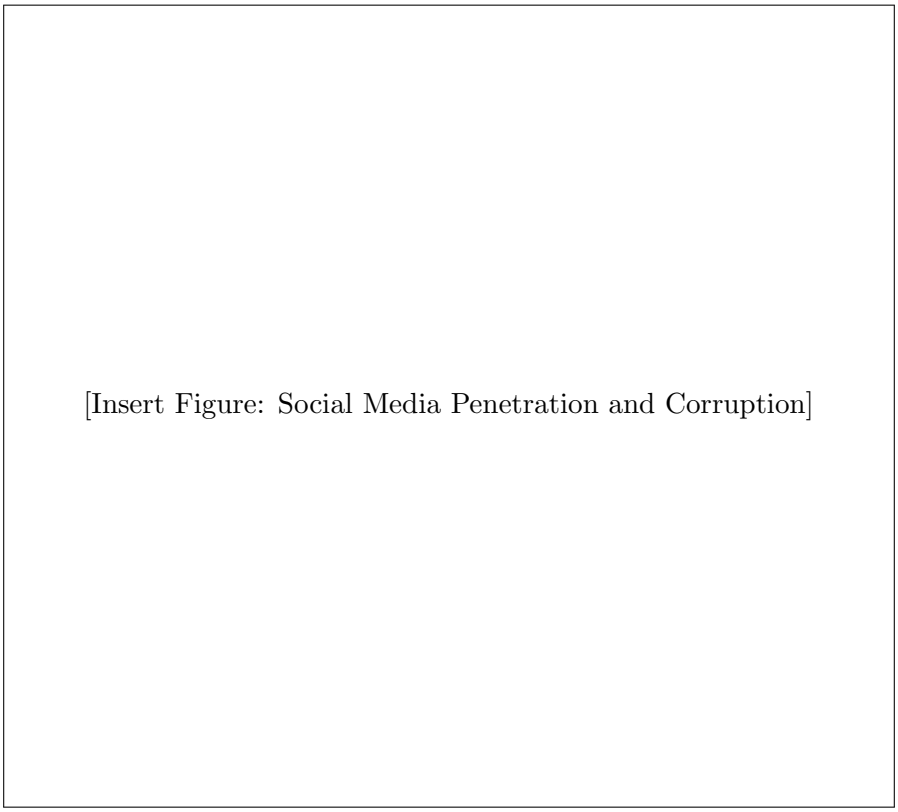


Figure 2: Social media penetration and corruption (from Enikolopov et al. 2018)

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

$$(2) \quad Socialnetshare = \alpha_1^S + \alpha_2^S ln(gdppc) + \epsilon^S, \tag{1}$$

$$(3) \quad Corruption = \alpha_1^C + \alpha_2^C ln(gdppc) + \epsilon^C. \tag{2}$$

Part (a)

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

Answer: Because this is a plot of OLS residuals, then both the x and y axis should be centered around 0 as OLS by construction centers residuals around 0.

Part (b)

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

Answer: GDP per capita is controlled for so they can isolate the effect of social media - GDP per capita may influence both share of social network users and corruption.

Part (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.

Answer: The quote estimate is the coefficeint for socialnetshare, because the slope from figure 1 is negative.
They used robust standard errors because the data is cross-country, so it controls for heteroskedasticity.
The t-test is a test for the slope coefficeint. The null is that beta3 is zero, where the alternate is that it is non-zero.