Foundations of Econometrics - Part I
Sample Exam 2 Solutions

SampleExam2

December 7, 2025

Question 1

Consider the following model relating a country GDP PC (US$), in log form, with $dem = 1$ if country is defined as a democracy, 0 otherwise, $educP = \%$ of country's population with primary schooling attained, $educS = \%$ of country's population with secondary schooling attained, $educH = \%$ of country's population with higher education attained, and $tradewb =$ Exports plus Imports as a share of GDP.

$$\text{Model}(1) : \ln(GDP\,PC) = \beta_1 + \beta_2 dem + \beta_3 educP + \beta_4 educS + \beta_5 educH + \beta_6 tradewb + \epsilon.$$

This regression model is estimated using data for 2010 from Acemoglu's article, "Democracy does cause growth". Stata's output after estimating by OLS, is:

```
Source       SS           df       MS              Number of obs = 122
                                                   F(5, 116)     = 34.46
Model        182.578494    5       36.5156988      Prob > F      = 0.0000
Residual     122.910661   116      1.05957466      R-squared     = 0.5977
                                                   Adj R-squared = 0.5803
Total        305.489155   121      2.52470376      Root MSE      = 1.0294

lnGDPPC      Coef.        Std. Err.     t      P>|t|    [95% Conf. Interval]
dem          .380598      .2183563      1.74   0.084    -.0518841   .81308
educP        .0267492     .0073504      3.64   0.000    .0121908    .0413076
educS        .0300333     .0061537      4.88   0.000    .0178452    .0422215
educH        .0859244     .0098279      8.74   0.000    .066459     .1053898
tradewb      .004786      .0020264      2.36   0.020    .0007724    .0087996
_cons        4.094374     .4496879      9.10   0.000    3.20371     4.985037
```

Part (a)

(i) Provide the general expression that has been used to calculate the standard error, value 0.2183563 underlined in the output above, in as much detail as possible. Then, substitute the available values provided in the output. Any element missing? If so, properly identify it.

Answer:

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

From our regression output, we have sigma squared, which is the MSE, or the MS from the residual: 1.0596.
We also have beta hat 2, which is 0.380.
So plugging in:

$$\text{se}(\hat{\beta}_2) = \sqrt{1.0596 \cdot [(X'X)^{-1}]_{22}}$$

We do not have X'X for index jj (2,2 in this case). This represents the variance of inflation for the betahat2 coefficient.

(ii) What information is this value meant to provide? Be as specific as you can.

Answer: This value describes how the variance of the coefficient (betahat2) affected by the correlation between betahat2 and the other regressor.

Part (b)

Consider testing whether regressor *dem* is statistically significant using the exact t-test statistic.

(i) Indicate null and alternative hypotheses.

Answer:
The null: $\beta_2 = 0$, Alternate: $\beta_2 \neq 0$

(ii) For this test, indicate the expression of the test statistic and its assumed distribution.

Answer:
The test statistic is defined as: $t = \frac{\hat{\beta}_2 - 0}{\text{se}(\hat{\beta}_2)}$

(iii) List by name all the assumptions we would need to place on the dgp for the t-test statistic to have the distribution you indicated in (ii).

Answer:
A1: Linearity A2: Strict Exogeneity A3: Conditional Homoskedacity A4: Uncorrelatedness of Error Terms

(iv) At what significance level would regressor *dem* be statistically significant? Justify.

Answer: The regressor has a p-value of 0.084. It would then be significant at the 10% significance level, because $\alpha < 0.10$.

Part (c)

(i) Provide the exact expression that defines the p-value $= 0.084$, underlined in the output above.

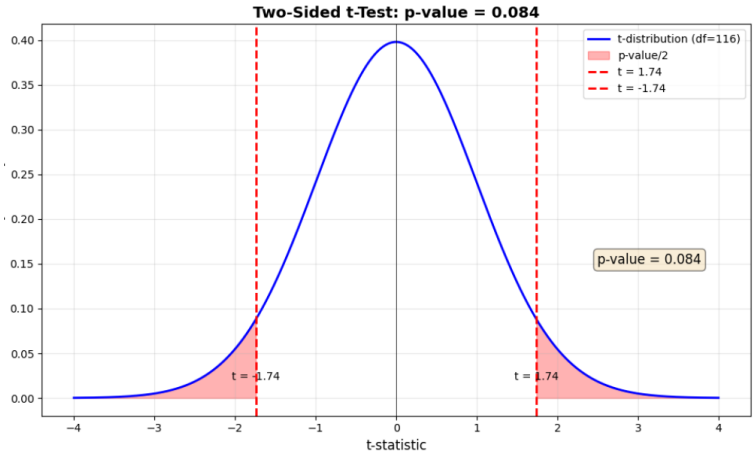
Answer:
The exact expression that defines the p-value is:

$$p\text{-value} = P(|t| > |t_{\text{obs}}| \mid H_0) = 2 \cdot P(t_{116} > 1.74) = 0.084$$

where t_{116} denotes the t-distribution with 116 degrees of freedom ($n - K = 122 - 6 = 116$), and $t_{\text{obs}} = 1.74$ is the observed test statistic.

(ii) Draw this p-value, properly identifying both axes and any relevant value to properly identify this p-value.

Answer:



Part (d)

Consider testing whether *educP*, *educS* and *educH* are jointly significant at $\alpha = 1\%$, using the exact F test statistic with the following assumed distribution:

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

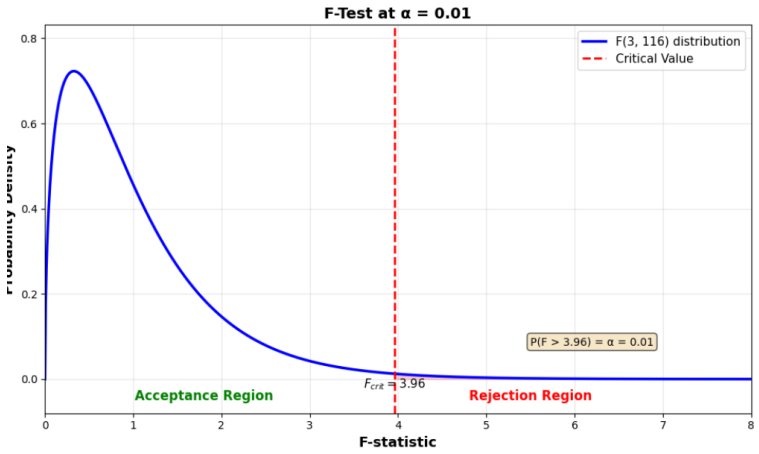
(i) Write down the null and alternative hypotheses associated with this test.

Answer:

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0$$
$$H_1 : \text{not } H_0$$

(ii) Using the information provided below about the F-distribution, draw the acceptance and rejection region associated with this test. Properly label the axes. Include any relevant value to properly identify both regions.

Answer:



(iii) Provide the intuition behind where the acceptance region is located.

Answer:
Under the null hypothesis $H_0 : R\boldsymbol{\beta} = r$, we expect $R\hat{\beta}$ to be close to r. The F-statistic measures the squared Mahalanobis distance between $R\hat{\beta}$ and r, standardized by the residual variance.
Intuition for acceptance region location:

- A **small F-value** (near 0) indicates the restricted model fits almost as well as the unrestricted model, suggesting the restrictions are valid $\rightarrow$ **Fail to reject** H_0
- A **large F-value** indicates the restricted model fits much worse, providing evidence against the restrictions $\rightarrow$ **Reject** H_0

The acceptance region is located at $F \leq F_\alpha(q, n-K)$ because we only reject when the data strongly contradict H_0 (i.e., when F is unusually large).

(iv) Provide the exact expression of the model you would have to estimate to calculate $RSSE$.

Answer:
To calculate $RSSE$, we estimate the restricted model by imposing the null hypothesis $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$:

$$\ln(GDP\,PC) = \beta_1 + \beta_2 dem + \beta_6 tradewb + \epsilon$$

This model excludes $educP$, $educS$, and $educH$. The sum of squared residuals from this restricted regression gives us $RSSE$.

(v) Knowing that the associated F-value = 41.3, what would you conclude? Justify.

Useful information about a $F \sim F(3, 116)$ distribution: $P(F > 2.23) = 0.1$, $P(F > 2.68) = 0.05$, $P(F > 3.96) = 0.01$.

Answer:
Since $F = 41.3 > F_{0.01}(3, 116) = 3.96$, we **reject** H_0 at the 1% significance level.
Conclusion: The variables *educP*, *educS*, and *educH* are **jointly significant**. The extremely large F-value (41.3 vs. critical value of 3.96) indicates very strong evidence that at least one of these education variables has a non-zero effect on $\ln(GDP\,PC)$.
The p-value is far less than 0.01, providing overwhelming evidence against the null hypothesis.

Part (e)

Consider again the test included in question (1b). We wanted to drop classical assumption of disturbances being conditionally normal. Without normality assumption,

(i) Would the distribution of the statistic you used in (1b) change? How?

Answer:
If we drop the assumptions that disturbances are normal, we no longer follow the exact t-stat distribution.
Without normality we move to asymptotics, which uses the normal distribution. We use this for larger samples.
The distribution becomes more conservative, meaning it has less fat tails and smaller critical values.

(ii) Would the acceptance region you drew in (1b)(iii) be affected? Can you tell how? Explain.

Answer: Yes, the acceptance region changes. Under the normal distribution the tails are less fat, so we would have a smaller region to accept in. The acceptance region becomes more conservative.

Part (f)

If we wanted to drop classical assumption of disturbances being homoskedastic, but still wanted to use OLS estimator to estimate model parameters,

(i) Would it affect how R^2 has to be calculated? Justify.

Answer: No. R^2 is a measure of goodness of fit, or in sample predictive power. Homosokedasticity tells us that the variance of errors is constant, so if we drop it the variance of the errors no longer is constant. This doesn't impact the fit of the model but instead the certainity around the estimates.

(ii) Would it affect how the standard errors have to be calculated? Justify.

Answer: Yes, heteroskedaciity would impact how the standard errors are calculated. Under homoskedacitity, the standard errors are calculated assuming they are constant, by design. This would no longer hold.

Part (g)

Acemoglu et al.(2019)'s article "Democracy does cause growth", included the following paragraph: "The estimation of the causal effect of democracy on GDP faces several challenges. Democracies differ from non democracies in unobserved characteristics that can also have an impact on GDP. As a result, cross-country regressions are unlikely to reveal the causal effect of democracy on growth." Rigorously discuss Acemoglu et al.'s argument using the regression presented in this question.

Answer:
Acemoglu's argument is that there is ommitted variable bias - existence of unobserved variables that may impact both democracy and growth. Examples may include political institution quality like corruption. Corruption may both reduce democracy and also impact economic growth. Historical factors may also impact a country's democracy status which in turn impacts economic growth (foriegn intervention stalls democracy and could potentially reduce long run growth).

Question 2

Consider the following dgp:

$$y_i = 2 + 0.5 \cdot x_{i2} + 0.5 \cdot x_{i3} + \epsilon_i \quad \epsilon_i | X \sim \text{i.i.} N(0, 16)$$

$$x_{i2} \sim U[0, 20] \quad x_{i3} = x_{i2} + v_i \quad v_i \sim N(0, \sigma_v^2)$$

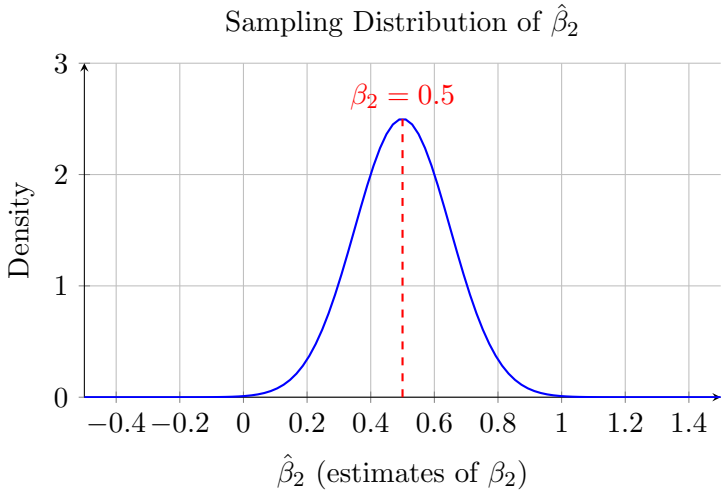
After setting $\sigma_v^2 = 9$ we use this dgp to generate 10,000 samples of 30 observations each ($n = 30$). With each of the generated samples, we estimate by OLS the following regression:

$$(1) \quad y = \beta_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \epsilon,$$

and save the estimates of parameter β_2. Paying careful attention to the dgp and the regression we estimate with each generated sample, answer the following questions:

Part (a)

Consider we summarize the 10,000 OLS estimates of β_2 we got using a density histogram. How would the graph, approximately, look like (shape, location)? Draw the histogram, label both axes, and justify both, the shape and location you selected.



Answer:
The histogram would approximately look like:
Shape: Normal/bell-shaped (symmetric) due to normality of errors.
Location: Centered at $\beta_2 = 0.5$ (the true value), demonstrating unbiasedness: $\mathbb{E}[\hat{\beta}_2] = 0.5$.
Spread: Wider than it would be without collinearity, due to multicollinearity between x_2 and x_3. Since $x_3 = x_2 + v$ with $\sigma_v^2 = 9$, the two regressors are highly correlated, which inflates the variance via the VIF term:

$$\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{SST_2} \cdot \frac{1}{1 - R_2^2}$$

where R_2^2 is large due to collinearity, making $\text{Var}(\hat{\beta}_2)$ large.
Axes:

- x-axis: Values of $\hat{\beta}_2$
- y-axis: Density (probability density function)

Part (b)

Now, we repeat the experiment, but setting $\sigma_v^2 = 1$. How would you expect the histogram you drew in question 2a to change (location, spread, shape)? Rigorously argue using the expression you feel is most appropriate to support your argument.

Answer:
When σ_v^2 decreases from 9 to 1, the histogram changes as follows:
Location: No change. Still centered at $\beta_2 = 0.5$ (OLS remains unbiased).
Shape: No change. Still normal/bell-shaped (errors are still $\epsilon \sim N(0, 16)$).
Spread: Wider (increased variance).
Rigorous argument using variance decomposition:
Recall the variance formula:

$$\text{Var}(\hat{\beta}_2|X) = \frac{\sigma^2}{SST_2} \cdot \frac{1}{1 - R_2^2}$$

where R_2^2 is from regressing x_2 on x_3.
Since $x_3 = x_2 + v$ where $v \sim N(0, \sigma_v^2)$:

- When $\sigma_v^2 = 9$ (original): v has high variance, so x_3 differs substantially from $x_2 \Rightarrow$ **lower correlation** $\Rightarrow$ lower R_2^2
- When $\sigma_v^2 = 1$ (new): v has low variance, so $x_3 \approx x_2 \Rightarrow$ **higher correlation** $\Rightarrow$ higher R_2^2

As R_2^2 increases, the VIF term $\frac{1}{1-R_2^2}$ increases, which **inflates** $\text{Var}(\hat{\beta}_2)$.
Therefore, the histogram becomes **wider** (more spread out), reflecting increased uncertainty due to stronger multicollinearity. The estimator remains unbiased but becomes less precise.

Part (c)

Consider that using the same samples generated in 2a we estimate instead regression:

$$(2) \quad y = \beta_1 + \beta_2 \cdot x_2 + \epsilon.$$

Would you expect the location of the histogram to change? If you do not, explain why. If you do, explain why and how. Rigorously justify in either case.

Answer:
We remove beta3 from the regression but we still have x3 in the dgp. This creates ommitted variable bias because we know that x3 is colinear with x2, and we also say that x3 also effects y. When we drop it, we would expected the location of the histogram to shift, since bias means we are no longer centered around the truth. It will specifically shift to the right because x3 has a positive effect on both x2 and y.

Question 3 (Extra)

For a linear regression model with 4 regressors (const, x_2, x_3, x_4), we want to test: $H_0 : \beta_2 = \beta_3$ versus H_1 : not H_0, using an estimator tilde, $\tilde{\beta}$, (i.e., not the OLS estimator!), which has the following asymptotic distribution:

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

Part (a)

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

Answer:
[Your answer here]

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

Answer: