Barcelona School of Economics

Assignment 4

Foundations of Econometrics

Group 7

October 22, 2025

Question 1

Class slide 3(33) (Unit 3) illustrated, via simulation, the effects of collinearity. The script used to generate the sample is included in file `data33.R`.

Part (a)

(i) Estimate the regression model included in the slide, presenting OLS estimates and the 95% confidence intervals for each parameter; Include the output in your answer.

Answer:

Table 1: Regression Results with 95% CI

	Estimate	Std..Error	Lower	Upper
(Intercept)	7.9365	1.6426	4.5863	11.2867
x2	0.5953	0.0747	0.4429	0.7476
x3	0.2996	0.5264	-0.7740	1.3733
x4	0.7818	0.5277	-0.2945	1.8582

(ii) Using `confidenceEllipse()` function, or equivalent, draw the 95% confidence region for parameters β_3, β_4. Include also in the drawing the confidence intervals for each parameter.

Answer:

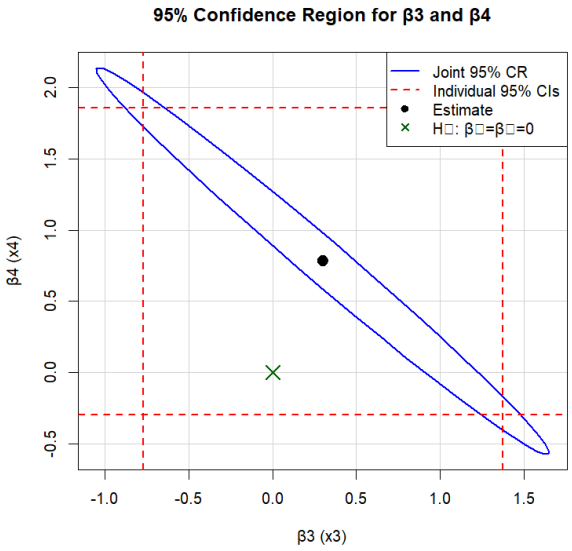

Figure 1: 95% Confidence Region for β_3 and β_4

(iii) Describe what the confidence region you just drew provides.

Answer:

The ellipse we drew illustrates the joint confiendence region, while the red lines show the individual confidence intervals. It represents every combination of the two parameters that falls in the 95% confidence region. It also illsutrates the correlation between the two parameters, as the confidence region is elongated along a diagonal line. The negative slope provides insight into the negative correlation between the two parameters.

(iv) Use the figure of confidence intervals and confidence region to show the difference between testing statistical significance of regressors separately or jointly, and explain why this is so relevant under the presence of collinear regressors.

Answer:

The figure illustrates that joint testing can show us when collinear variables are jointly significant, even if they are not individually significant.

Individual tests: Testing $H_0 : \beta_3 = 0$ and $H_0 : \beta_4 = 0$ separately uses the rectangular region formed by the individual 95% confidence intervals (red dashed lines). From the regression output, both intervals include zero: x_3 CI $= [-0.7740, 1.3733]$ and x_4 CI $= [-0.2945, 1.8582]$. Since the rectangle contains the origin $(0, 0)$, we fail to reject both null hypotheses individually—neither x_3 nor x_4 appears statistically significant.

Joint test: Testing $H_0 : \beta_3 = \beta_4 = 0$ jointly uses the 95% confidence ellipse. The ellipse is much smaller than the rectangle due to the negative correlation between $\hat{\beta}_3$ and $\hat{\beta}_4$. If the origin $(0, 0)$ falls outside the ellipse, we reject the joint null hypothesis, meaning x_3 and x_4 are jointly significant even though neither is individually significant.

Collinearity: When regressors are collinear, their coefficients are negatively correlated because if one increases, the other must decrease to fit the same data. This creates the tilted ellipse. The individual tests ignore this correlation and use the wider rectangular region, making them overly conservative. The joint test correctly accounts for correlation, revealing that while we cannot precisely determine which variable drives the effect, we can confidently say that together they have a significant impact on y. This demonstrates why collinearity makes individual t-tests unreliable while joint F-tests remain valid.

latex

Part (b)

Modify the script used to estimate now the same regression with data generated from the same dgp but now using a sample of 3500 observations.

(i) Surprised with how the estimates have changed? Rigorously justify.

Answer:

Not surprised. The estimates have converged closer to the true parameter values ($\beta_3 = \beta_4 = 0.5$, shown by purple dotted lines in Figure **??**). For $n = 35$, the estimates were $\hat{\beta}_3 = 0.2996$ and $\hat{\beta}_4 = 0.7818$, showing considerable sampling variability. With $n = 3500$, the estimates are much closer to 0.5 for both parameters.

By the Law of Large Numbers, as sample size increases, the OLS estimators converge in probability to their true values. This is consistency: $\text{plim}(\hat{\beta}_j) = \beta_j$ as $n \to \infty$. The larger sample provides more information about the true relationship, reducing the influence of random sampling variation. The collinearity between x_3 and x_4 still exists (since $x_4 = x_3 + \text{noise}$ in both samples), but with more observations, the estimator can better distinguish their individual effects on y.

(ii) Surprised of the change of the 95% confidence intervals? Rigorously justify.

Answer:

Not surprised. The confidence intervals have become much narrower with the larger sample. For $n = 35$: β_3 CI = [-0.7740, 1.3733] with width = 2.15, and β_4 CI = [-0.2945, 1.8582] with width = 2.15. With $n = 3500$, the confidence intervals should be approximately $\sqrt{3500/35} = 10$ times narrower, giving widths around 0.215. This follows from the asymptotic distribution of OLS estimators.

(iii) Surprised of the change of the 95% confidence region for parameters β_3, β_4? Rigorously justify.

Answer:

Not surprised. The confidence ellipse has shrunk dramatically while maintaining its elongated shape oriented along the same diagonal. The area of the confidence ellipse is proportional to $|\text{Var}(\hat{\beta}_3, \hat{\beta}_4)|$, which is proportional to $1/n$. Therefore, increasing sample size from 35 to 3500 (a factor of 100) reduces the ellipse area by a factor of 100. The ellipse remains tilted because the negative correlation between $\hat{\beta}_3$ and $\hat{\beta}_4$ persists—it's determined by the correlation between x_3 and x_4 in the DGP, which doesn't change with sample size. However, the smaller ellipse means:

- Higher precision in joint estimation of β_3 and β_4
- The point $(0.5, 0.5)$ (true values) is much more likely to be contained in the ellipse
- If the origin $(0, 0)$ was near the boundary for $n = 35$, it may now fall outside the ellipse for $n = 3500$, making the joint test reject $H_0 : \beta_3 = \beta_4 = 0$ even though individual tests might still fail to reject

(iv) Using the variance decomposition expression for $\text{var}(\hat{\beta}_3 \mid X)$, or $\text{var}(\hat{\beta}_4 \mid X)$, discuss why increasing n can explain the changes observed. Be specific.

Answer:

The variance of $\hat{\beta}_3$ conditional on X is:

$$\text{Var}(\hat{\beta}_3|X) = \sigma^2[(X'X)^{-1}]_{33} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_{i3} - \bar{x}_3)^2(1 - R_3^2)}$$

where R_3^2 is the R-squared from regressing x_3 on all other regressors (including x_4).

This can be decomposed as:

$$\text{Var}(\hat{\beta}_3|X) = \sigma^2 \cdot \frac{1}{\sum_{i=1}^{n}(x_{i3} - \bar{x}_3)^2} \cdot \frac{1}{1 - R_3^2} = \sigma^2 \cdot \frac{1}{n \cdot \text{Var}(x_3)} \cdot \text{VIF}_3$$

Three components:

(a) σ^2: Error variance (constant across samples if same DGP)

(b) $\frac{1}{n \cdot \text{Var}(x_3)}$: Decreases linearly with n

(c) $\text{VIF}_3 = \frac{1}{1-R_3^2}$: Variance Inflation Factor due to collinearity (constant for same DGP)

Effect of increasing n from 35 to 3500:

- The term $\frac{1}{n}$ decreases by factor of 100
- Therefore $\text{Var}(\hat{\beta}_3|X)$ decreases by factor of 100
- Standard errors decrease by factor of $\sqrt{100} = 10$
- Confidence intervals become 10 times narrower
- The confidence ellipse shrinks by factor of 100 in area

Part (c)

Now, go back to the original script generating 35 observations, and modify the script so that now $x_{i3} + 2x_{i4} = 0$.

(i) Run the script again. Include the output in your answer.

Answer:

Table 2: Regression Results with Perfect Collinearity

	Estimate	Std.Error	Lower	Upper
(Intercept)	10.1838	2.2007	5.7012	14.6665
x2_c	0.4697	0.0999	0.2661	0.6733
x3_c	0.1985	0.1015	-0.0083	0.4052
x4_c				

(ii) How many estimates did you get an estimate for β_3? And for β_4? You should be able to show, using the proper derivation, that in fact you got an infinite number of estimates for β_3 and β_4.

Answer:

From the regression output, we obtained **one estimate** for β_3 (the one R reported) and **zero estimates** for β_4 (R returned NA). However, mathematically, there are **infinitely many** valid estimates for both β_3 and β_4.

Derivation:

The OLS estimator solves the normal equations:

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

When $x_3 + 2x_4 = 0$ (perfect collinearity), the matrix $X'X$ is singular (non-invertible) because its columns are linearly dependent. This means the normal equations have either no solution or infinitely many solutions. Since the least squares problem always has a solution (we can always minimize SSR), it must have infinitely many solutions.

To show infinite solutions:

Suppose $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$ is a solution. The fitted values are:

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_4 x_{i4}$$

Now consider an alternative set of coefficients:

$$\beta_3^* = \hat{\beta}_3 + t, \quad \beta_4^* = \hat{\beta}_4 + 2t$$

for any constant $t \in \mathbb{R}$. The fitted values become:

$$\begin{aligned}
\hat{y}_i^* &= \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + (\hat{\beta}_3 + t) x_{i3} + (\hat{\beta}_4 + 2t) x_{i4} \\
&= \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_4 x_{i4} + t x_{i3} + 2t x_{i4} \\
&= \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_4 x_{i4} + t(x_{i3} + 2x_{i4}) \\
&= \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_4 x_{i4} + t \cdot 0 \\
&= \hat{y}_i
\end{aligned}$$

Since $x_{i3} + 2x_{i4} = 0$ for all i, the fitted values (and thus the SSR) are identical for any value of t. Therefore, $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3 + t, \hat{\beta}_4 + 2t)$ is also a valid OLS solution for **any** $t \in \mathbb{R}$.

Conclusion: There are infinitely many combinations of (β_3, β_4) that minimize the sum of squared residuals. We cannot uniquely identify β_3 and β_4 separately—only their weighted combination $\beta_3 x_3 + \beta_4 x_4$ is identified. This is why R drops one variable and returns NA: it's R's way of acknowledging that the parameters cannot be uniquely estimated.

Question 2

Data file `microsoft.csv` includes monthly data from May 1986 to April 2013 on RP_{msft} (excess return of Microsoft stock), $RP_{s\&p}$ (excess return on the S&P500 portfolio), $Dprod$ (variation of Industrial production), $Dinflation$ (change in inflation rate), $Dterm$ (change in interest rate) and $m1$ (an indicator variable that takes value 1 if t is the month of January and 0 otherwise). The following regression is set to measure the reaction of the excess return of Microsoft stocks to changes in macroeconomic variables:

$$RP_{msft,t} = \beta_1 + \beta_2 RP_{s\&p,t} + \beta_3 Dprod_t + \beta_4 Dinflation_t + \beta_5 Dterm_t + \beta_6 m1_t + \epsilon_t$$

(a) Estimate the model above by OLS. Present the complete output (estimates, standard errors, p-values) as your answer.

Answer:

Dep. Variable:	RPmsoft	**R-squared:**	0.213
Model:	OLS	**Adj. R-squared:**	0.201
Method:	Least Squares	**F-statistic:**	17.26
Date:	Wed, 22 Oct 2025	**Prob (F-statistic):**	4.08e-15
Time:	12:54:20	**Log-Likelihood:**	-1276.7
No. Observations:	324	**AIC:**	2565.
Df Residuals:	318	**BIC:**	2588.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	-0.9291	0.760	-1.223	0.222	-2.424	0.566
RPsandp	1.3232	0.152	8.678	0.000	1.023	1.623
Dprod	-1.5216	1.283	-1.186	0.237	-4.046	1.003
Dinflation	0.4716	2.351	0.201	0.841	-4.154	5.097
Dterm	4.1587	2.487	1.672	0.095	-0.735	9.052
m1	5.4352	2.869	1.894	0.059	-0.210	11.081

Omnibus:	203.965	**Durbin-Watson:**	2.141
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1809.211
Skew:	-2.541	**Prob(JB):**	0.00
Kurtosis:	13.401	**Cond. No.**	21.2

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

(b) The January effect states that on average, every else equal, the returns (or excess returns) are larger in the month of January than the rest of the months.

Test, at $\alpha = 1\%$, the presence of the January effect using the exact t-test statistic. Would you say the data supports the presence of this effect?

Answer:

We test for the presence of the January effect by examining whether the coefficient on the January dummy variable (m_1) is statistically significant.

Hypotheses:

$$H_0 : \beta_6 = 0 \quad \text{(No January effect)}$$
$$H_1 : \beta_6 \neq 0 \quad \text{(January effect exists)}$$

We use a two-tailed t-test at significance level $\alpha = 0.01$.

Test Statistic:

$$t = \frac{\hat{\beta}_6}{\text{SE}(\hat{\beta}_6)} = \frac{5.44}{2.87} = 1.89$$

With 318 degrees of freedom, the critical values are ± 2.59. So, reject H_0 if $|t| > 2.59$ or equivalently if $p < 0.01$. The t-statistic is 1.89 with a p-value of 0.059. Since $|1.89| < 2.59$ and $p = 0.059 > 0.01$, we fail to reject the null hypothesis. Therefore, we do not have sufficient evidence at the 1% significance level to conclude that there is a January effect in Microsoft stock returns.

(c) Aside from normality, list the assumptions needed to justify the use of the t test statistic. Justify your answer.

Answer:

(d) For the test you performed in questions (2b), you were asked to use the t test statistic. Using this test requires, among others, for disturbances to be normally distributed. One of the available tests of normality of the distribution of a given random variable is the Jarque-Bera test. Under the null hypothesis of normality, the Jarque-Bera (JB) test statistic is:

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

where sk is the sample coefficient of skewness of the variable and kur is its sample coefficient of kurtosis. Using a significance level of 1%, draw (by hand absolutely fine) the distribution of JB under H_0 and the corresponding acceptance and rejection regions. Provide an intuition for the location of the acceptance region.

Answer:

Hypotheses:

$$H_0 : \text{Residuals are normally distributed}$$
$$H_1 : \text{Residuals are not normally distributed}$$

Test Statistic: Under H_0, the Jarque-Bera statistic follows a $\chi^2(2)$ distribution:

$$JB = \frac{n}{6}\left[sk^2 + \frac{(kur-3)^2}{4}\right] \sim \chi^2(2)$$

where sk is the sample skewness and kur is the sample kurtosis.

Results:

- Sample size: $n = 324$
- Skewness: $sk = -2.54$
- Kurtosis: $kur = 13.40$
- $JB = 1809.21$
- Critical value: $\chi^2_{0.01}(2) = 9.21$
- p-value: < 0.0001

Decision: Since $JB = 1809.21 > 9.21$, we strongly reject the null hypothesis at the 1% significance level.

(e) Now, we want to test for the presence of normality in our disturbances using the JB test. Ideally, to test normality of disturbances, JB test should be applied to a sample of disturbances, but, given that they are unobservable, the JB test is usually applied to our OLS residuals. Explain how, if all the assumptions regarding the dgp for consistency of OLS estimator are met, it would be justified for $\hat{\epsilon}_t$'s to take the place of ϵ_t's to perform the test.

Answer:

(f) Perform the JB test on the OLS residuals. What do you conclude? Comment.

Answer:

(g) Repeat the test performed in (2b) using the asymptotic T-test statistic. Use a 1% significance level.

Answer:

(h) Is the use of asymptotic tests justified in this case? Rigorously argue.

Answer:

(i) Consider the following statement: "Using the exact t test statistic leads to slightly more conservative inference, because we get larger acceptance regions and larger p-values than if we used the asymptotic version." Do you agree? Rigorously argue.

Answer: