TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

Quantitative Methods III - Exercises

1. The *revenue* (X, in million euros) and the number of *employees* (Y) were recorded for 5 companies. The joint distribution is as follows:

X	Y
10	4
13	4
15	6
18	8
20	12

   (a) Estimate the parameters of the regression line (with revenue as a function of the number of employees).

   The model to be estimated is the revenue, x_i, as a function of employees, y_i, i.e.:

   $$x_i = \beta_0 + \beta_1 y_i + u_i$$

   The estimated coefficients are:

   $$\hat{\beta}_0 = 7.73, \quad \hat{\beta}_1 = 1.1$$

   (b) What if we wanted to estimate a function for the number of employees instead?

   The estimated coefficients are:

   $$\hat{\beta}_0 = -5.1083, \quad \hat{\beta}_1 = 0.7834$$

   (c) Using the function from the point (b), predict how many employees would be needed for a company to reach 30 million euros in revenue.

   The predicted number of employees is:

   $$\hat{y}_{30} = 18.3937$$

   (d) Compute the appropriate goodness-of-fit measures for the estimated regression line in the point (b).

The commonly used goodness-of-fit measures are the R^2 and a measure of residual standard deviation, the SER (Residuals Standard Error).
The formula for R^2, where $\hat{u}_i$ denotes the residuals, is:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

The Residuals Standard Error, SER, is given by:

$$SER = \sqrt{\frac{\sum_{i=1}^{n}\hat{u}_i^2}{n-2}}$$

$$R^2 = 0.8604, \quad SER = 1.444$$

(e) Given that the standard error of the slope is 0.1822, is it possible that the slope could be 1.5 for the regression line of point (b)?

To determine whether the true slope β_1 could be 1.5, we perform two approaches:

- constructing a confidence interval,
- conducting a hypothesis test.

The 95% confidence interval for the true slope β_1 is given by:

$$\hat{\beta}_1 \pm t_{\alpha/2} \cdot SE \tag{1}$$

where:

- $\hat{\beta}_1 = 0.7834$ (estimated slope),
- $SE = 0.1822$ (standard error of the slope),
- $t_{\alpha/2}$ is the critical value from the t-distribution with $df = n - 2 = 3$ degrees of freedom.

For $df = 3$ and a 95% confidence level, the critical t-value is approximately $t_{0.025,3} \approx 3.182$. Thus, the confidence interval is:

$$0.7834 \pm (3.182 \times 0.1822) = (0.203, 1.364) \tag{2}$$

Since 1.5 is outside this interval, it is unlikely that the true slope is 1.5.
We conduct a t-test for:

$$H_0 : \beta_1 = 1.5$$
$$H_A : \beta_1 \neq 1.5$$

The test statistic is computed as:

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE} = \frac{0.7834 - 1.5}{0.1822} = \frac{-0.7166}{0.1822} \approx -3.93 \tag{3}$$

For $df = 3$, the critical t-value at 95% confidence is approximately 3.182. Since $|t| = 3.93 > 3.182$, we reject H_0.
Since both the confidence interval and the hypothesis test indicate that $\beta_1 = 1.5$ is unlikely, we conclude that it is not plausible that the true slope is 1.5.

2. Suppose you want to estimate the causal effect of the regressors x_{1i} and x_{2i} on y_i.

   (a) Write the regression model and the formula for the estimation of the unknown regression parameters.

   Model: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$.
   OLS Estimators

   $$\hat{\beta}_{OLS} = (X'X)^{-1}(X'Y)$$

   (b) Estimate β_1, the coefficient associated with x_{1i}, and interpret its value, knowing that:

   $$(X'X)^{-1} = \begin{bmatrix} 5 & -2 & 1 \\ -2 & 3 & -1 \\ 1 & -1 & 2 \end{bmatrix} \qquad and \qquad (X'Y) = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$$

   The OLS vector is:

   $$\hat{\beta}_{OLS} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 5 & -2 & 1 \\ -2 & 3 & -1 \\ 1 & -1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 12 \\ -1 \\ 3 \end{bmatrix}$$

   (c) Find the predicted value of y when $x_{1i} = 2$ and $x_{2i} = 5$.

   Since the estimated model is $\hat{y}_i = 12 - x_{1i} + 3x_{2i}$, the predicted value is: $\hat{y}_i = 12 - 1 \times 2 + 3 \times 5 = 25$

   (d) What are the main goodness-of-fit indices used to evaluate multiple linear regression models?

   The goodness-of-fit indices generally used to evaluate multiple linear regression models are the SER and the $\overline{R}^2$ (Adjusted R^2).

3. Always in relation to the regression model of exercise 2, the following table reports the observed and predicted values of Y_i, together with the first 4 residuals, $\hat{u}_i$:

i	Y_i	$\hat{Y}_i$	$\hat{u}_i$	$\hat{u}_i^2$
1	15	13.5	1.5	2.25
2	5	5	...	0
3	15	17.5	-2.5	6.25
4	5	4.3	0.7	0.49
5	10	9.7	...	0.09

(a) Complete the table with the residual for observation 5 using the space below for calculations.

The sum of the residuals is always 0, therefore:
$\hat{u}_5 = -\sum_{i=1}^{4} \hat{u}_i = -(1.5 + 0 - 2.5 + 0.7) = 0.3$

(b) Find the SER of the estimated model, after filling in the last column of the previous table.

The formula for the SER (Residuals Standard Error) is:

$$SER = \sqrt{\frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k}} = \sqrt{\frac{SSR}{n-k}}$$

where $\hat{u}_i$ are the residuals, and k is the number of regressors included in the model (including the intercept). The SSR is obtained based on the sum of squared residuals, computed previously.
In this case, $N = 5$ and $k = 3$, so:

$$SER = \sqrt{\frac{9.08}{5-3}} = 2.131$$

4. Suppose you have the following regression model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + u_i$$

Based on a sample of $N = 206$ observations, the following results are obtained:

$$\hat{y}_i = \underset{(0.02)}{-12} - \underset{(0.32)}{1.8}\, x_{1i} + \underset{(1.23)}{7.1}\, x_{2i} + \underset{(2.50)}{3.4}\, x_{3i} - \underset{(1.91)}{2.9}\, x_{4i}$$

(a) Estimate the confidence interval for the parameter β_1, setting $\alpha = 5\%$. Is it statistically significant?

The required confidence interval:

$$95\%CI = [\hat{\beta}_1 \pm z_{\alpha/2} \times SE(\hat{\beta}_1)] = [-1.8 \pm 1.96 \times 0.32] = [-2.427; -1.173]$$

Since the null value is not included in the confidence interval, we conclude that the estimated coefficient $\hat{\beta}_1$ is statistically significant.

(b) Verify if x_{3i} has a significant effect on y_i.

i. $H_0 : \beta_3 = 0$ vs. $H_1 : \beta_3 \neq 0$
ii. Test Statistic:
$z = \frac{3.4}{2.5} = 1.36$
iii. Decision Rule: Reject H_0 if $|1.36| > Z_{1-\frac{\alpha}{2}}$

Setting $\alpha = 0.10$, we conclude in favor of the null hypothesis, leading to the conclusion that the variable x_{3i} does not have a statistically significant causal effect on the variable y_i (the conclusion holds, even more so, at any other significance level lower than 10%)

(c) Verify if β_3 and β_4 are **both** statistically significant at $\alpha = 5\%$ using the Bonferroni method.

The Bonferroni test allows for testing joint hypotheses on q coefficients starting from the t statistics related to individual hypotheses but correcting the critical value as follows:

$$z_{\alpha/2} = z_{1-\frac{\alpha/q}{2}}$$

Since H_0 is a joint hypothesis on 2 coefficients, we have that $q = 2$. Having set $\alpha = 5\%$, the adjusted critical value becomes:

$$z_{\alpha/2} = z_{1-\frac{\alpha/q}{2}} = z_{1-\frac{0.05/2}{2}} = z_{0.9875} = 2.24$$

H_0 is rejected if at least one of the individual test statistics is, in absolute value, greater than the critical value.
The test statistics of the two coefficients are $z_{\beta_3} = 1.36$ and $z_{\beta_4} = 1.526$. Since in neither case is the absolute value greater than 2.24, the null hypothesis $H_0 : \beta_3 = \beta_4 = 0$ is not rejected.

(d) In a multiple linear regression model, the $\overline{R}^2$ (Adjusted R^2):

- ☐ cannot be negative
- ☐ always decreases by adding a regressor
- ☐ is the square of the linear correlation coefficient
- ■ is never greater than R^2 of the regression
- ☐ none of the above

5. On the time series y_t of T =198 observations, two different models are estimated. The results are reported in the following Table.

   (a) Complete the table of the estimated models (fill in where the '...' are).

   In the column relating to the Significance Level LS write:

   $-$ when the coefficient is not significant;

   $***$ when the coefficient is significant at the 1% level;

   $**$ when the coefficient is significant at the 5% level;

   $*$ when the coefficient is significant at the 10% level.

Model	Model (1) ADL(...,...)				Model (2) ADL(...,...)			
Coefficients	Estimate	Std. Error	t-value	LS	Estimate	Std. Error	t-value	LS
Intercept	-0.0044	0.0161	-0.273	$-$	-0.0028	0.0158	-0.177	$-$
y_{t-1}	-1.3789	0.1049	-13.145	$***$	-1.3282	0.0945	-14.055	$***$
y_{t-2}	-0.1343	0.1080	-1.245	$-$	-0.0723	0.0712	-1.015	$-$
y_{t-3}	0.0083	0.0716	0.116	$-$				
x_{t-1}	0.2379	0.1051	2.264	$**$	0.2033	0.0932	2.181	$**$
x_{t-2}	0.0932	0.1061	0.878	$-$				
k	5				3			
$R^2_{adj.}$	0.6026				0.6057			
F	59.54				100.3			
$ln(SSR(p)/T)$	-3.0202				-3.0129			

Model (1) is an ADL(3,2) since it includes 3 lags of y_t and 2 of x_t. Similarly, Model (2) is an ADL(2,1). It follows that k, the number of estimated parameters excluding the constant, is 5 for Model (1) and 3 for Model (2). The missing values can be derived by inverting the t-ratio formula appropriately.

As for the t-ratios to be found, consider, as an example, the case of the variable y_{t-1} in the ADL(3,2) model. In this case, the t-ratio is:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{-1.3789}{0.1049} = -13.145$$

Since the absolute value of the t-ratio is greater than the critical value at the 1% level of a Student's t-distribution $t_{198-5-1,0.005} = 2.602$, we conclude that the associated p-value will be less than 1%, thus indicating the associated significance code (***). In Model (1): x_{t-1}: $t = \frac{0.2379}{0.1051} = 2.264$. With $198 - 5 - 1 = 192$ df, the critical value for a two-tailed test at the 5% level is 1.98. Thus, it is significant at the 5% level (**).

y_{t-2}: $t = \frac{-0.1343}{0.1080} = -1.245$. Not significant.

y_{t-3}: $t = \frac{0.0083}{0.0716} = 0.116$. Not significant.

x_{t-2}: $SE = \frac{0.0932}{0.878} = 0.1061$. Not significant.

In Model (2): The y-intercept: $t = \frac{-0.0028}{0.0158} = -0.177$. Not significant.

y_{t-1}: $SE = \frac{-1.3282}{-14.055} = 0.0945$. Significant at the 1% (***).

y_{t-2}: $t = \frac{-0.0723}{0.0712} = -1.015$. Not significant.

x_{t-1}: $t = \frac{0.2033}{0.0932} = 2.181$. Significant at the 5% (**).

   (b) Calculate the BIC of Model (1).

For Model (1) we have that:

$$BIC = ln\left(\frac{SSR}{T}\right) + \frac{k \times ln(T)}{T} = -3.0202 + \frac{6 \times ln(198)}{198} = -2.856$$

(c) Which model would you choose knowing that Model (2) has an AIC of -1.57 and a BIC of -2.906?

Model (2) would be preferred since the BIC associated with it is smaller than that of Model (1). This is in line with the fact that Model (2) is also the more robust ($R^2_{adj.} = 61\%$) and significant ($F = 100.3$).

6. Based on the estimates of the following $AR(4)$ for y_t:

$$\hat{y}_t = \underset{(0.14)}{7.9} + \underset{(0.07)}{0.76}y_{t-1} + \underset{(0.09)}{0.13}y_{t-2} + \underset{(0.001)}{0.018}y_{t-3} - \underset{(0.05)}{0.21}y_{t-4}$$

with the values of y_t observed between February 2011 and June 2011 reported in the following table.

Date	2011:M6	2011:M5	2011:M4	2011:M3	2011:M2
y_t	5.6	35.2	7.6	23.3	14.8

(a) Derive the one-step-ahead forecast of the variable y_t in July 2011.

The one-period-ahead forecast will be equal to:

$$\hat{y}_{T+1|T} = \hat{\beta}_0 + \hat{\beta}_1 y_t + \hat{\beta}_2 y_{t-1} + \hat{\beta}_3 y_{t-2} + \hat{\beta}_4 y_{t-3}$$

Given that $T + 1 =$ 2011:M7, we can rewrite the model by making the observation time explicit:

$$\hat{y}_{2011:M7|2011:M6} = \hat{\beta}_0 + \hat{\beta}_1 y_{2011:M6} + \hat{\beta}_2 y_{2011:M5} + \hat{\beta}_3 y_{2011:M4} + \hat{\beta}_4 y_{2011:M3}$$

Hence, the one-period-ahead forecast will be:
$\hat{y}_{2011:M7|2011:M6} = 7.9 + (0.76 \cdot 5.6) + (0.13 \cdot 35.2) + (0.018 \cdot 7.6) + (-0.21 \cdot 23.3) = 7.3998$

(b) Calculate the forecast error associated with the forecast produced above, knowing that $y_{2011:M7}$ = 7.9

$\tilde{u}_{2011:M7} = y_{2011:M7}$ - $\hat{y}_{2011:M7} = 7.9 - 7.4 = 0.5$

(c) Can we verify through a test if the $AR(4)$ process is stationary? If yes, calculate an appropriate test and comment on its outcome.

It is possible to apply an Augmented Dickey-Fuller Test for the non-stationarity of a series (verification of the existence of a unit root).
$H_0 : y_t$ is not stationary (= has a unit root) vs
$H_1 : y_t$ is stationary (= does not have a unit root)
In other terms, y_t is not stationary $H_0 : \hat{\beta}_1$=1 vs $H_1 : \hat{\beta}_1$<1 (left-tailed one-sided test)

Let's calculate the test statistic:

$$t = \frac{\hat{\beta}_1 - 1}{SE(\hat{\beta}_1)} = \frac{0.76 - 1}{0.07} = -3.429$$

The table of critical values for an ADF test with only the intercept tells us that at a significance level of 5 percent, the series will be stationary ($-3.429 < -2.86$).

(d) At a significance level of 1%, can the same conclusions be drawn? Argue your answer.

The critical value for an ADF test with only the intercept and a significance level of 1% is -3.43.
At this significance level, we cannot conclude to reject the null hypothesis and consequently that the series is stationary.