Estadística II
Exercises Chapter 4
Academic Year 2010/11

---

1. The following table shows information regarding the sales in 1998 of daily press in number of daily copies sold for every thousand inhabitants in 8 autonomous Spanish regions. The sales are related to the economic production based on the Gross domestic product (GDP) per inhabitant in thousands of euros (Source: INE. Anuario Estadístico).

GDP	8,3	9,7	10,7	11,7	12,4	15,4	16,3	17,2
Copies	57,4	106,8	104,4	131,9	144,6	146,4	177,4	186,9

   a) Estimate with Least Squares a model of simple regression to explain the number of sold copies in terms of GDP.

   b) Construct a confidence interval at 95 % for the slope of the regression line and test the hypothesis that the slope obtains the value zero. Can be verified that the number of sold copies depends linearly on GDP?

   c) What would be the the level of sales predicted for a region with GDP 15000 euros per inhabitant?

   d) If for a region the value of GDP increases by 2500 euros, how are the daily press sales expected to change?

2. One of the administrators of a company argues that the use of the internet is the main cause of the cost in the telephone invoice. In order to corroborate this affirmation, data are taken about the monthly telephone costs in euros and the duration of the connection in minutes in different departments.

quantity of the telephone invoice	55	100	118	120	142
Connection time	200	500	700	800	1000

   a) Calculate the correlation coefficient between the two variables. Is there any kind of linear relation between them?

   b) Estimate a linear regression model that permits estimate the monthly cost of the telephone invoice as a function of the connection time.

   c) According to this linear relation, what would be the level of the telephone invoice in a department that doesn't connect to the internet? Calculate a confidence interval at 95 % for this estimation.

   d) Taking into account the linear relation, what would be the the estimated telephone expense in a department if the connection time to the internet was 2000 minutes? Does it seem to be acceptable such a prediction?

3. In order to study the linear relation between the prices of the cars and the number of sold items, information was gathered during the last month in a certain region. The obtained results were the following ones:

Price (thousands of euros)	Sold quantity (units per month)
7,5	450
9	425
10,5	400
12	350
14	325
16	300
18	290
20,5	280
23,5	260
27	200

a) Calculate and interpret the correlation coefficient between the two variables.

b) A company situated in the region for the next month schedules to increase the price of its more sold model by 500 euros. If we suppose that the linear relation between the two variables holds for the data of last month, how would this fact affect the sales of the above mentioned model?

c) If we express the price in euros and the sold quantities in 10^2 units, which would be the linear model that expresses the sales as a function of the price? And the correlation coefficient?

4. A gas station has collected information about its daily collection for a week and the number of clients that went to the gas station each day:

Collection (10^3 euros)	1,5	10	8	3	5	15	2
Number of clients (10^2)	3	6	5	3,5	4	8	3,2

a) Find a linear model that expresses the collection of the gas station as a function of the number of clients.

b) Determine what would be the mean predicted collection for the days that reach 720 clients at the gas station. Obtain a confidence interval at 95 % for this prediction.

c) Determine what would be the predicted collection for one day that reach 720 clients at the gas station. Obtain a confidence interval at 95 % for this prediction.

d) Obtain the regression line if we express the collection in euros and the number of clients in units.

5. A distributor of drugstore and cleaning products, distributes his products among the trades of the sector of all towns of a region. Among the products that he distributes, has selected a sample of 7 of them, that are the ones that considers more important, due to their demand on the part of the merchants and of the public in general. Of these 7 products there is information corresponding to last April regarding the unit price of each product in euros and the volume of sales corresponding to this product in thousands of euros, that are the following:

	Preice	Volume
Bleach	0,65	2,50
Perfumed ammonia	0,85	1,68
Swab green fiber	1,25	4,23
Concentrated detergent	2,6	5,69
Conditioner	2,1	5,17
Dishwasher detergent	2,5	5,50

a) Construct the regression of the volume of sales with respect to the price per unit for April.

b) Test the hypothesis that the volume of sales depends linearly on the price of the products.

c) Assuming that the prices during May don't vary with respect to the previous month (April), obtain a prediction for the volume of the sales in May of a product with unit price of 1 euro. Give a measure of the reliability of this prediction through a confidence interval at 95 %.

d) The distributor is convinced that the volume of the sales would remain the same if in May increases by 5 cents the price of every product. Obtain the equation of the regression line in this case and a prediction for the sales in May of a product that its price per unit is 1 euro (including the ascent of 5 cents).

6. An important application of regression analysis in accounting is to estimate the costs as a function of the volume of sales. From a sample of 8 pairs of data, using the method of least squares, it is obtained the regression line of the costs with respect to the volume of sales, giving the following residuals:

e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8
10	-3	2	1	1	-1	-2	

a) Calculate the residual number 8.

b) Calculate the residual variance.

c) Calculate a confidence interval at 95 % for the variance of the errors.

d) If the estimated linear regression model is $y = 2 + 5x$ and the variance of the observations of the independent variable is $s_X^2 = 100$, analyze if the linear relation between the two variables is significant.

e) What is the mean of the observations of the dependent variable, $\bar{y}$, if the mean of the observations of the independent is $\bar{x} = 3$.

7. We wish to carry out a study among the staff of a company in order to estimate the number of daily errors committed by an employee (variable Y) as a function of the number of years hired in the company (variable X). For this purpose, the work of 8 employees selected at random was checked and the following information was obtained:

$$\sum_{i=1}^{10} x_i^2 = 296; \qquad \sum_{i=1}^{10} x_i = 50; \qquad \sum_{i=1}^{10} y_i^2 = 1036; \qquad \sum_{i=1}^{10} y_i = 86; \qquad \sum_{i=1}^{10} x_i y_i = 331$$

a) Calculate and interpret the correlation coefficient between the two variables.

b) Estimate the corresponding regression line indicating the meaning of each coefficient.

c) Estimate the mean quantity of errors that will commit an employee who works 4 years in this company.

8. A company wants to investigate the relation between the number of days per year (variable Y) that the employees are absent without permission and the distance in kilometers from their home to their work (variable X). Analyzing a sample of employees the following results were obtained:

$$cov(x, y) = 5{,}4; \qquad cor\,(x, y) = 0{,}7838; \qquad \bar{x} = 25; \qquad s_x^2 = 12$$

Moreover, for a distance of 20 kilometers from home to work, it has been estimated that the mean number of days of absence without permission was 4.

a) Estimate a linear regression model.

b) In average, what will be the number of days without permission that an employee will be absent, if he lives 18 kilometers far away from his work?

c) Answer to the previous questions expressing the distance in meters.

9. From a sample of paired data, $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, the following results have been obtained:

$$cor\,(x, y) = 0{,}9; \qquad \bar{x} = 5; \qquad s_x^2 = 1{,}44; \qquad s_y^2 = 4{,}41$$

a) Determine the regression line of Y on X.

b) What will be the regression line if $cor\,(x, y) = 0$?

c) Answer to the previous questions if the value of the observations of the response variable, y_i, augments 2 units.