uc3m | Universidad Carlos III de Madrid
Departamento de Estadística

Bachelor in Computer Science and Engineering

Statistics Problems

IX Multiple Regression

1. We want to study the relationship between the final grade of our students (variable Grade) and the grades obtained in two activities related to the continuous assessment (P1 and P2). In addition, we want to study if this relationship os different for national or international students (variable Type: 0 for national, 1 for international).

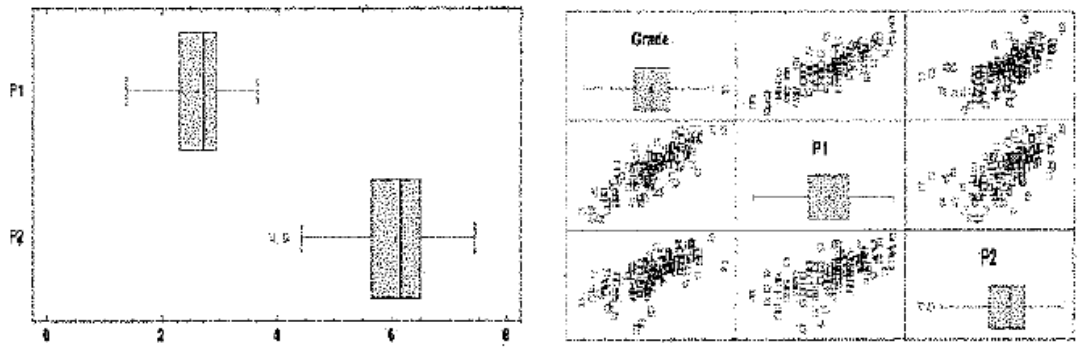

Figure 1.: Box plots for P1 and P2 (left). Dispersion matrix for Grade, P1 and P2 (right)

a) Looking at the box plots, can you tell if there are asymmetries? Say why and which is the asymmetry sign.
b) Looking at the dispersion matrix, can we assume that there is a linear relationship between the variables? Is it possible to use a linear model to represent this relationship? Should the data be transformed? Which should be the sign for the correlation between Grade and P1? Justify your answers.
c) To model the relationship between the response variable Grade, the two explanatory variables P1 and P2 and the influence of being a national or international student the following model is proposed:

Grade =	*1,0374*	*+0,6423*P1*	*+0,1869* P2*	*+0,4256 * Type*
	(p-value=0)	*(p-value=0)*	*(p-value=0)*	*(p-value=0)*

$R^2 = 86,7\%$

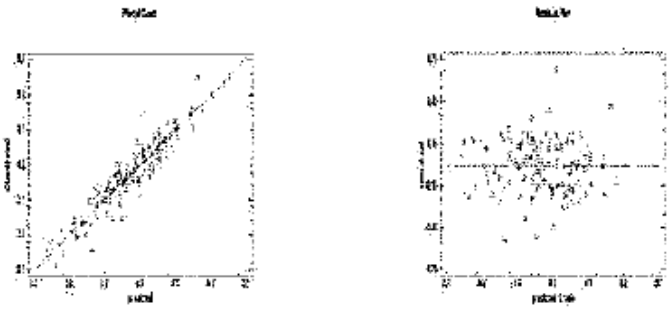

Figure 2. Residual graphs

Considering the p-values for each one of the variables and the residual graphs in figure 2, is this model adequate? Why?

d) According to this model, are there significant differences between national and international students? Explain the meaning of the estimated coefficient for variable Type.

e) Which is the expected final grade for a national student that got P1=4,5 and P2=6

2. In a customer service, the average attention times, X, of 45 employees have been taken, resulting in an average time of 5,5 minutes, with a variance of 4,24 minutes^2.

a) Calculate the coefficient of variation and interpret it.

a. We want to study the relationship of this variable with the months that these employees have been working in the company (Y), obtaining and average of 4,27 months with a variance of 4,93. If S_{xy} = 3,66, calculate the correlation coefficient between X and Y. What can you say about this relationship?

b. Write the equation for the regression line

c. A multiple regression has been performed for variable Y and 3 employee satisfaction ratios proposed by the Quality Department. Looking at the following computer output what can be concluded?

Parameter	Slope	Standard error	T Statistic	p-value
Constant	-0.0077	0.0173	-0.4477	0.6567
X1	0.9994	0.0022	441.27	0.004
X2	0.0008	0.036	0.2421	0.8099
X3	-1.007	0.0060	-164.651	0.000

R-squared = 97.9780 percent
R-squared (adjusted for d.f.) = 96,5431 percent

d) Calculate the mean squared error for the best estimator identified in b)

3.

We would like to predict the average number of yearly failures based on information from three variables, X1 (operating years), X2 (daily usage time, in hours) and X3 (average age of the operator). Two linear regression models are proposed:

MODEL 1

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	3,6034	1,08906	3,30874	0,0162
X1	0,350275	0,0523142	6,69561	0,0005
X2	0,722951	0,234922	-3,07741	0,0217
X3	-0,000989172	0,00777737	-0,127186	0,9029

MODEL 2

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	3,62969	0,991275	3,66163	0,0081
X1	0,350693	0,0484029	7,24529	0,0002
X2	0,732137	0,207239	-3,53281	0,0096

(a) **(0.5 Points)** Is the model 1 adequate? Why? Can we improve this regression?

(b) **(0.5 Points)** How would you interpret the estimated coefficient $\beta_2 = 0.732137$? in model 2?

(c) **(0.5 Points)** A 2 year old machine, with a daily use time of 8 hours, operated by a 25 year-old person is monitored, resulting in an average monthly number of breakdowns of 12.168. What average number of breaksdowns does the model predict? What is the value of the corresponding residual?

(d) **(0.75 Points)** If the residual variance in model 2 is 0.4, calculate the probability that the average number of yearly breakdowns in the above machine will be greater than 11.

4.

The management team of a hypermarket chain is deciding where to build a new market, so it decides to analyze the SALES (euros) of the existing premises by means of a multiple linear regression model, obtaining the following results:

Parameter	Estimate	Standard Error	T-Statistic	P-Value	
CONSTANT	2205.79	4181.57	0.527503	0.602	
METRES	0.872918	0.433204	2.01503	0.0536	Hypermarket surface (m^2)
POPULATION	0.00885157	0.00272513	3.24813	0.003	Municipality population (inhabitants)
SHOPPING.C	-0.0260789	0.0880543	-0.296168	0.7693	Shopping center surface (m^2)
ACCESSES	-1193.81	920.327	-1.29716	0.2052	Num. Accesses to shopping center
PARKING	2.31738	1.69698	1.36559	0.1829	Num. Parking spaces

Simple linear regression	SALES	POPULATION	RESIDUALS
Average	14070.6	310559	
Standard deviation	4839.96	232718	3523.4
Correlation coeff.=0.6856			

(a) **(0.75 Points)** Initially, they only evaluate the possibility of locating the hypermarket in the Municipality A (54000 inhabitants), where there are two shopping centers. The Blue Shopping Center has a surface of 12000 m^2 and 3 accesses, while the Red Shopping Center has 15000 m^2 and only 2 accesses. In which of the two shopping centers will there be a greater volume of sales? Justify your answer.

(b) **(1 Point)** In a second scenario, they propose the possibility of locating the new hypermarket in Municipality B (65000 inhabitants). There are no shopping centers in this municipality. Obtain a new regression model to predict SALES according to the population of the municipality (POPULATION). Using this new model, indicate which of the two municipalities (A or B) is more profitable.

(c) **(1 Point)** Finally, they decide to locate the hypermarket in a municipality with 100000 inhabitants. What is the probability that the SALES do not exceed 9000 Euros? Use the model fitted in the previous section.