Chapter IX: Regression

PROBLEMS

Proposed Problems

1. True or false? Consider the general linear model (multiple regression)

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki} + e_i$$

Which of the following propositions are required assumptions for the validity of the model?

   a) The regressors $X_1, \cdots, X_K$ follow a normal distribution.
   b) The variance of $X_1, \cdots, X_K$ must be constant.
   c) The response variable Y conditional on the observed values X must be constant.
   d) The term of error e should be homoskedastic.
   e) The effects of the variables that are not included in the regressors have a joint effect that can be modeled as normal with mean zero.
   f) Parameters $\beta_0, \beta_1, \cdots, \beta_K$ must be positive.

2. True or false? Consider the following multiple regression model:

$$Y = 10 + 0.5\, X_1 - 3X_2 + e$$

with $e \sim N(0, \sigma^2 = 1)$.

   a) If $X_1 = 0$ and $X_2 = 0, Y \sim N(10,1)$.
   b) If $X_1 = 1$ and $X_2 = 1, Y = 12$.
   c) If $X_1 = 1$ and $X_2 = 1, Y \sim N(7.5,1)$.
   d) If $X_1 = 1$ and $X_2 = 1, \hat{Y} = 7.5$.
   e) If $X_1 = 1$ and $X_2 = 1, Y = 7.5$.
   f) If $X_1 = 1$ and $X_2 = 1, E[Y|X_1, X_2] = 10$.

3. True or false? Consider the following multiple regression model:

$$Y = 10 + 0.5\, X_1 - 3X_2 + e$$

with $e \sim N(0, \sigma^2 = 1)$.

   a) If X_1 increases by one unit and X_2 remains constant, the mean of Y increases by 10 + 0.5 = 10.5 units
   b) If X_1 increases by one unit and X_2 remains constant, the mean of Y increases by 0.5 units
   c) If X_2 increases by one unit and X_1 remains constant, Y decreases on average by 3 units

4. The file *FrenosITV.sf3* has some properties of a sample of vehicles. We want to know if the age of the car and its power help to predict the effectiveness of braking. Build a multiple regression model that predicts the efficiency as a function of the kilometers traveled by car (variable KM) and its power (variable "*Potencia*").

5. Let denote by PM10 the small sized solid or liquid particles that are dispersed in the atmosphere. High concentrations of PM10 may be harmful to health. The file *ConcentraPM10.sf3* contains a sample of 500 hourly observations of PM10 concentration together with other variables. These variables are:
   - "*Coches*" (cars) = number of cars that passed in front of the concentration meter during the hour of measurement
   - "*Temperatura*" (temperature) = Air temperature measured at a height of 2m from ground
   - *"Viento"* (wind) = wind speed (m / s)
   - "*Hora*" (hour) = day time

   It calls for:

   a) Build a model to explain the concentration of PM10 as a function of wind speed. It is assumed that the greater the wind speed, the cleaner the air is as PM10 particles get dispersed. The effect can be nonlinear and maybe if could be necessary to aplly transformations to the variables.

b) Build a model to explain the concentration of PM10 as a function of car traffic, temperature and wind speed. Analyze the residuals. The model only has to contain significant variables.
c) Some analysts believe that sunlight reduces the concentration of PM10 by destroying some particles by biochemist effect. Other analysts say that the fact that during the night there is less PM10 is just due to the effect of less car traffic. Knowing that the daylight hours in the place where data were taken range from 7AM to 6PM, analyze which of two groups of analysts (or both) is right.

6. Using the file *AlumnosIndustriales.sf3* contrasts the following statements:
   a) Boys tend to carry more money than girls
   b) Students who live farther away (takes longer to get to college) have more money with them
   c) A boy and girl who are just as high, will have, on average, the same shoe size
7. We want to determine what is the fastest computer language to sort a list of 1,000,000 integers, belonging to the set [0, 100], by using the same sorting *QuickSort* algorithm. The computer languages we want to compare are the C and Java. We repeat the experiment 200 times and record the execution time of each experiment (in milliseconds) in the file *CyJava.sf3*. Which is the fastest?

   SOLUTION:

   Do a regression with the variable Y = "*Tiempo"* (Time) and X = "*Lenguaje"* (Language: 0 = Java, 1 = C) and check that the C language is significantly faster.
8. The file *FrenosITV.sf3* has data of a set of cars that come to the ITV station of Leganes. The variable "*eficacia*" (effectiveness) takes values between 0 and 100 and measures the effectiveness of braking, so that better efficiency means better braking, in the sense that the pressure to brake is stronger. The "*ABS"* variable has the value 1 if the vehicle has ABS and 0 if it has not ABS system. Do vehicles with ABS have a braking more effective?
9. The file *IndiceMC.sf3* contains the Body Mass Index (variable "*Indice Masa Corporal*") of a set of students, together with other information like their sex (variable "*sexo*", it is equal to 1 for males and 0 for females).

   Is the average Body Mass Index of boys equal to the one of the girls?

   SOLUTION:

   Yes, using "*sexo*" as explicative variable, it does result significant, therefore the average Body Mass Index when "*sexo"* = 0 is different from the average Body Mass Index when "*sexo"* = 1.