Questions 10:
Multilinear Regression

Roman Belavkin

Middlesex University

Question 1

Suppose that a multiple regression model using $m = 6$ input variables should be created based on some data. What is the minimum number of cases the dataset should have?

Answer: *The dataset should have at least* $m + 2 = 8$ *cases. This is because the mean-square regression requires several points not laying in one hyperplane. A linear function of* m *variables describes a hyperplane in a* $m+1$ *dimensional space. Thus,* $m + 1$ *points always lay in one hyperplane, and we need one extra point not laying on this hyperplane. Hence, a minimum* $m + 2$ *cases are required.*

Question 2

Consider the following data:

Monthly Income (£K)	Monthly Expenses (£K)	Home Owner	Credit Score
2	1	0	3
1	2	0	1
6	2	1	5
3	1	1	4
3	2	0	2

Suppose you are building a linear mean-square model based on this data to predict the customers' credit scores. Let us denote the three input variables (the first three columns of the table) as x_1, x_2 and x_3, while the output variable as y. Answer the following questions:

a) Does this dataset contain a sufficient number of cases to build the model?

b) Compute the centre of gravity $(E\{x_1\}, E\{x_2\}, E\{x_3\}, E\{y\})$.

c) The regression coefficients for this model are $b_1 = 0,69$, $b_2 = -1,31$ and $b_3 = 0,56$. Using the coefficients and the centre of gravity of the data above, write the linear function $f(x_1, x_2, x_3)$.

d) Using the model, compute the credit score for the following customer:

Monthly Income (£K)	Monthly Expenses (£K)	Home Owner	Credit Score
5	3	1	?

e) Explain what does this model represent.

Answer:

a) *The dataset must have at least $m + 2$ different cases, where m is the number of input variables. So, the minimum number of cases is 5. Hence, the dataset contains the required number (the table has five rows).*

b) *The centre of gravity is:*

$$E\{x_1\} = 3\,, \quad E\{x_2\} = 1,6\,, \quad E\{x_3\} = 0,4\,, \quad E\{y\} = 3$$

c) *The we can use the following form of the linear mean-square model:*

$$y = E\{y\} + b_1(x_1 - E\{x_1\}) + b_2(x_2 - E\{x_2\}) + b_3(x_3 - E\{x_3\})$$

The values of the regression coefficients are already given, and the centre of gravity was computed above. Thus, the equation of the linear mean-square model is:

$$\begin{aligned} f(x_1, x_2, x_3) &= E\{y\} + b_1(x_1 - E\{x_1\}) + b_2(x_2 - E\{x_2\}) + b_3(x_3 - E\{x_3\}) \\ &= 3 + 0,69(x_1 - 3) - 1,31(x_2 - 1,6) + 0,56(x_3 - 0,4) \end{aligned}$$

d) *In this case, $x_1 = 5$, $x_2 = 3$ and $x_3 = 1$. The credit score is*

$$y = 3 + 0,69(5 - 3) - 1,31(3 - 1,6) + 0,56(1 - 0,4) = 2,88$$

e) *The model represents multiple linear dependencies between the first three variables and the Credit Score, as it is observed in the data. This model can be used to predict Credit Score for new customers.*

Question 3

How to test whether a model is good in forecasting?

Answer: *Usually, the dataset is split into two parts. One is used for creating the model, and another for testing the model. Because the model has not 'seen' the cases from the testing part of the data, they can be used to estimate how good the model is for forecasting. The errors between the testing data and the model's predictions can estimate how reliable is the forecast.*

Question 4

Compute correlation for the following set of data:

Monthly Expenses (£K)	Home Owner
1	0
2	1
1	1
2	0

What does the value of the correlation mean in this case?

Answer: *The correlation is defined as*

$$Corr(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}}$$

We can compute it as follows:

a) *Find the means:*

$$\begin{aligned} E\{x\} &= \frac{1+2+1+2}{4} = \frac{3}{2} \\ E\{y\} &= \frac{0+1+1+0}{4} = \frac{1}{2} \end{aligned}$$

b) *Find the covariance*

$$\begin{aligned} Cov(x,y) &= \frac{1}{4}[(1-3/2)(0-1/2) + (2-3/2)(1-1/2) \\ &+ (1-3/2)(1-1/2) + (2-3/2)(0-1/2)] = 0 \end{aligned}$$

c) *Find the variances*

$$\begin{aligned} Var(x) &= \frac{(1-3/2)^2 + (2-3/2)^2 + (1-3/2)^2 + (2-3/2)^2}{4} = \frac{1}{4} \\ Var(y) &= \frac{(0-1/2)^2 + (1-1/2)^2 + (1-1/2)^2 + (0-1/2)^2}{4} = \frac{1}{4} \end{aligned}$$

d) *The correlation is*

$$Corr(x, y) = \frac{0}{\sqrt{1/41/4}} = 0$$

Zero correlation means that the variables are uncorrelated, and, therefore, they are not related linearly. This, however, does **not** *mean that the variables are independent. There still can be some non-linear dependency between the variables.*