Questions 9:
Linear Models

Roman Belavkin

Middlesex University

Question 1

What is a model and its error, and how can they be defined in terms of information?

Answer: *A model is a simplified representation of another object. Because the model contains fewer details and fewer number of states, it is less complex, and, therefore, it is less uncertain than the object it represents. It means also that the model can be described using less information than the original object. The greater certainty makes the model more predictable, which can be used to study and predict the real object. However, the loss of information leads to errors, which are the differences between the model's predictions and the behaviour of the real object.*

Question 2

What are the linear models, and why are they called 'linear'?

Answer: *Linear models are represented by linear functions. In the simplest case, the linear function of one variable is* $f(x) = a + bx$*, and it represents a line on a plane.*

Question 3

What are the data-driven models, and what does the data represent?

Answer: *Data-driven models are based on data that represents the object to be modelled. Usually, these models use large datasets from datawarehouses. The data in this case is the reality that the model has to simplify. More precisely, the data represents measurements of the object of interest or the part of reality that is to be modelled.*

Question 4

Why plotting the data on a chart can be useful?

Answer: *Charts visualise the data and can be used to check for some visible relationships in the data before a model is created. Their use, however, is limited to two or three dimensions.*

Question 5

How can the errors of the model be measured? Why is it important for the choice of model?

Answer: *The error is the difference between the data (y) and the model's prediction of this data ($f(x)$). The difference is measured by the cost function $c(y, f(x))$. There can be different cost functions:*

- *Boolean (or binary) cost*

$$c(y, f(x)) = \begin{cases} 0, & \textit{if success (e.g. } y = f(x)\textit{)} \\ 1, & \textit{otherwise} \end{cases}$$

- *Absolute deviation*

$$c(y, f(x)) = |y - f(x)|$$

- *Squared (or quadratic) deviation*

$$c(y, f(x)) = |y - f(x)|^2$$

Which cost function is used is important, because it influences the choice of the optimal model. For the same data, the mean absolute error can be minimised by one model, and the mean square error by another.

Question 6

Compute the linear mean-square model for the following set of data:

Monthly Income (£K)	Home Owner
2	0
1	0
6	1

Answer: *Let us denote Monthly Income by x and Home Ownership by y. The computations can be done as follows:*

a) *Compute expected values $E\{x\}$ and $E\{y\}$*

$$E\{x\} = \frac{2+1+6}{3} = 3\,, \quad E\{y\} = \frac{0+0+1}{3} = 1/3$$

b) *Find covariance Cov(x,y)*

$$Cov(x,y) = \frac{(2-3)(0-1/3)+(1-3)(0-1/3)+(6-3)(1-1/3)}{3} = 1$$

c) *Find variance Var(x)*

$$Var(x) = \frac{(2-3)^2+(1-3)^2+(6-3)^2}{3} = 14/3$$

d) *Find the slope*

$$b = \frac{Cov(x,y)}{Var(x)} = \frac{3}{14} = 0,21$$

e) *Find the intercept*

$$a = E\{y\} - bE\{x\} = -0,31$$

The resulting linear model is

$$f(x) = -0,31 + 0,21x$$

Question 7

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.*