MTH5131 Actuarial Statistics

Coursework 1 — Solutions

Exercise 1. 1. Describe each of the following terms as it relates to a data set, and give an example of each as it relates to the app providers data:

(a) Cross-sectional data involves recording the values of the variables of interest for each case in the sample at a single moment in time.

In this data set, this relates to the number of messages sent by each user on any particular day.

(b) Longitudinal data involves recording the values of the variables of interest at intervals over time.

In this data set, this relates to the number of messages sent by a particular user on each day over the 5-year period.

Give an example of each of the following types of data analysis that could be carried out using the app providers data:

(a) Examples of descriptive analysis that could be carried out on this data set include

- calculating the mean and standard deviation of the number of messages sent each day by users in each country
- plotting a graph of the total messages sent each day worldwide, to illustrate the overall trend in the number of messages sent over the 5 years
- calculating what proportion of the total messages sent in each year originate in each country

(b) Examples of inferential analysis that could be carried out on this data set include:

- testing the hypothesis that more messages are sent at weekends than on weekdays
- assessing whether there is a significant difference in the rate of growth of the number of messages sent each day by users in different countries over the 5-year period.

(c) Examples of predictive analysis that could be carried out on this data set include

- forecasting which countries will be the major users of the app in 5 years time, and will therefore need the most technical support
- predicting the number of messages sent on the apps busiest day (eg New Years Eve) next year, to ensure that the provider continues to have sufficient capacity

Exercise 2. The key steps in the data analysis process in this scenario are:

1. Develop a well-defined set of objectives that need to be met by the results of the data analysis.

   Here, the objective is to determine whether young drivers are more likely to have an accident in a given year than older drivers.

2. Identify the data items required for the analysis.

   The data items needed would include the number of drivers of each age during the investigation period and the number of accidents they had.

3. Collection of the data from appropriate sources.

   The insurer will have its own internal data from its administration department on the number of policyholders of each age during the investigation period and which of them had accidents.

   The insurer may also be able to source data externally, e.g. from an industry body that collates information from a number of insurers.

4. Processing and formatting the data for analysis, eg inputting into a spreadsheet, database or other model.

   The data will need to be extracted from the administration system and loaded into whichever statistical package is being used for the analysis.

   If different data sets are being combined, they will need to be put into a consistent format and any duplicates (i.e. the same record appearing in different data sets) will need to be removed.

5. Cleaning data, eg addressing unusual, missing or inconsistent values.

   For example, the age of the driver might be missing, or be too low or high to be plausible. These cases will need investigation.

6. Exploratory data analysis,

   which here takes the form of inferential analysis as we are testing the hypothesis that younger drivers are more likely to have an accident than older drivers.

7. Modelling the data

   This may involve fitting a distribution to the annual number of accidents arising from the policyholders in each age group.

8. Communicating the results.

   This will involve describing the data sources used, the model and analyses performed, and the conclusion of the analysis (ie whether young drivers are indeed more likely to have an accident than older drivers), along with any limitations of the analysis.

9. Monitoring the process updating the data and repeating the process if required.

   The car insurer may wish to repeat the process again in a few years time, using the data gathered over that period, to ensure that the conclusions of the original analysis remain valid.

10. Ensuring that any relevant professional guidance and legislation (eg on age discrimination) has been complied with.

Exercise 3. See the script **scattergraphs.R**

Exercise 4. 1. (a) Pearson:

$$S_{cc} = \sum c^2 - \frac{(\sum c)^2}{n} = 6884 - \frac{238^2}{9} = 590.2222,$$

$$S_{c\rho} = \sum_c \rho - \frac{(\sum c)(\sum \rho)}{n} = 983 - \frac{238 \times 33.4}{9} = 99.75556$$

$$S_{\rho\rho} = 140.62 - \frac{33.4^2}{9} = 25.66889$$

$$\Rightarrow r = \frac{S_{c\rho}}{\sqrt{S_{cc} S_{\rho\rho}}} = \frac{99.75556}{\sqrt{590.2222 \times 25.66889}} = 0.81045.$$

(b) Spearman:
The ranks are as follows:

Class	$X1$	$X2$	$X3$	$X4$	$Y1$	$Y2$	$Y3$	$Y4$	$Y5$
Students in class (c)	9	7	4	2	8	6	5	3	1
Average GCSE point score(ρ)	8	6	3	2	9	7	5	4	1
Differences	1	1	1	0	-1	-1	0	-1	0

Hence

$$r_s = 1 - \frac{6 \times 6}{9(9^2 - 1)} = 0.95$$

(c) Kendalll:
Arranging in order of class rank:

Class(c)	$Y5$	$X4$	$Y4$	$X3$	$Y3$	$Y2$	$X2$	$Y1$	$X1$
Students in class (c)	1	2	3	4	5	6	7	8	9
Average GCSE point score(ρ)	1	2	4	3	5	7	6	9	8
#concordantpairs	8	7	5	5	4	2	2	0	0
#discordantpairs	0	0	1	0	0	1	0	1	0

Totalling the rows gives $n_c = 33$, $n_d = 3$. Hence,

$$\tau = \frac{33 - 3}{9(9-1)/2} = 0.83.$$

2. We are testing

$$H_0 : \rho = 0 \quad \text{vs} \quad H_1 : \rho \neq 0$$

Under H_0:

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

The observed value of the test statistic is:

$$\frac{0.81045\sqrt{7}}{\sqrt{1-0.81045^2}} = 3.660.$$

This is greater than 3.499, the upper 0.5% point of the t_7 distribution

Therefore, we have sufficient evidence at the 1% level to reject H_0 . Therefore we conclude that there is a correlation between class size and GCSE results (ie class size does affect GCSE results).

3. There is strong positive correlation between class size and GCSE results (ie bigger classes have better GCSE results).

However, correlation does not necessarily imply causation, ie whilst bigger classes have better results, it is not necessarily the class size that causes the improvement.

Exercise 5. See the script **correlation.R**

Exercise 6.

1. The data matrix is

$$\hat{X} = \begin{pmatrix} 120 & 61 \\ 125 & 60 \\ 125 & 64 \\ 135 & 68 \\ 145 & 72 \end{pmatrix}$$

The sample means vector are easily seen to be 130 and 65. Subtract the sample mean vector from the observations to obtain

$$X = \begin{pmatrix} -10 & -4 \\ -5 & -5 \\ -5 & -1 \\ 5 & 3 \\ 15 & 7 \end{pmatrix}$$

The sample covarianve matrix is

$$S = \frac{1}{5-1} X^T X = \frac{1}{5-1} \begin{pmatrix} 400 & 190 \\ 190 & 100 \end{pmatrix} = \begin{pmatrix} 100.0 & 47.5 \\ 47.5 & 25.0 \end{pmatrix}$$

2. The eigenvalues of S satisfy

$$\det \begin{pmatrix} 100.0 - \lambda & 47.5 \\ 47.5 & 25.0 - \lambda \end{pmatrix} = 0 \Rightarrow (100.0 - \lambda)(25.0 - \lambda) - (47.5)(47.5) = 0$$

$$\Rightarrow \lambda^2 - 125.0\lambda + 243.75 = 0$$

The eigenvalues (to two decimal places) are

$$\frac{125 \pm \sqrt{125^2 - 4 \times 243.75}}{2} = 123.02, 1.98.$$

The largest eigenvalue is 123.02. The principle component is its unit eigenvector:

$$\begin{pmatrix} 100.0 & 47.5 \\ 47.5 & 25.0 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 123.02a \\ 123.02b \end{pmatrix} \Rightarrow 100.0a + 47.5b = 123.02a, 47.5a + 25.0b = 123.02b$$

Thus $47.5b = 23.02a \Rightarrow b = 0.485a$. An eigenvalue is $\begin{pmatrix} 1 \\ 0.485 \end{pmatrix}$. A unit eigenvalue is

$$\frac{1}{\sqrt{1^2 + 0.485^2}} \begin{pmatrix} 1 \\ 0.485 \end{pmatrix} = \begin{pmatrix} 0.900 \\ 0.436 \end{pmatrix}$$

.

Exercise 7. See the script **principal_components.R**