MTH5131 Actuarial Statistics

Coursework 1

This coursework is not to be turned in. You may ask questions about the coursework in tutorial or by email.

Exercise 1. The data analysis department of a mobile phone messaging app provider has gathered data on the number of messages sent by each user of the app on each day over the past 5 years. The geographical location of each user (by country) is also known.

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

   (b) longitudinal

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

   (a) descriptive

   (b) inferential

   (c) predictive

Exercise 2. A car insurer wishes to investigate whether young drivers (aged 17-25) are more likely to have an accident in a given year than older drivers. Describe the steps that would be followed in the analysis of data for this investigation.

Exercise 3. This exercise is to be carried out in R.

A new computerised ultrasound scanning technique has enabled doctors to monitor the weights of unborn babies. This data is contained in the text file: ‘baby weights,

1. Load the data frame and store it in the data frame BABY.
2. Obtain a scattergraph of the data.
3. Comment on the linear relationship.

The numbers of new AIDS cases recorded in the US in successive years during the early part of the AIDS epidemic are contained in the CSV file: AIDS.

1. (a) Load the data frame and store it in the data frame AIDS.

   (b) Plot a scattergraph and comment on the linear relationship.

2. (a) Create a new data frame $\mathrm{AIDS2}$ which contains the log of the number of cases.

   (b) Obtain a scattergraph of the logged data.

   (c) Comment on the linear relationship.

Exercise 4. This example is to be carried out by hand.

A schoolteacher is investigating the claim that class size does not affect GCSE results. His observations of nine GCSE classes are as follows.

Class	$X1$	$X2$	$X3$	$X4$	$Y1$	$Y2$	$Y3$	$Y4$	$Y5$
Students in class (c)	35	32	27	21	34	30	28	24	7
Average GCSE point score(ρ)	5.9	4.1	2.4	1.7	6.3	5.3	3.5	2.6	1.6

You can easily verify that

$$\sum c = 238,\ \sum c^2 = 6884,\ \sum \rho = 33.4, \sum \rho^2 = 149.62,\ \sum c\rho = 983$$

1. Calculate Pearsons, Spearmans and Kendalls correlation coefficients.

2. Use Pearsons correlation coefficient to test whether or not the data agrees with the claim that class size does not affect GCSE results.

3. Following his investigation, the teacher concludes, bigger class sizes improve GCSE results. Comment on this statement.

Exercise 5. This exercise is to be carried out in R.

The built in data set Iris contains measurements (in cm) of the variables sepal length, sepal width, petal length and petal width, respectively, for 50 flowers from each of 3 species (Iris setosa, versicolor, and virginica) of iris.

1. (a) Extract the four measurements for the setosa species only and store them in the 50×4 data frame, SDF.
   (b) Use plot to obtain a scattergraph of each pair of measurements for the setosa species.
   (c) Comment on the relationship between Petal Width and the other measurements.

2. For this and the remaining parts of the exercise, use the data frame SDF.

   Calculate the following correlation coefficients between all four pairs of variables:

   (a) Pearson
   (b) Spearman
   (c) Kendall

3. Obtain the Spearman correlation coefficient between the Sepal Length and the Petal Length only.

4. (a) Carry out the following test for Pearsons correlation coefficient $H_0 : \rho = 0$ vs $H1 : \rho > 0$ between Sepal Length and Petal Length.
   (b) Extract the statistic and the degrees of freedom for the test.
   (c) Use the statistic to obtain the p-value for the test.

5. Test whether the true value of Kendalls correlation coefficient, τ, is zero between Sepal Width and Petal Length.

Exercise 6. This exercise is to be carried out by hand.

The following table lists the weights and heights of five boys:

Boy	1	2	3	4	5
Weight(lb)	120	125	125	135	145
Height(in)	61	60	64	68	72

1. Center the matrix and construct the sample covariance matrix.

2. Find the principle component that explains most of the variation in the data.

Exercise 7. This exercise is to be carried out in R.

This question uses the setosa iris data from question 4.

1. Obtain a scaled matrix of the 50 observations of 4 variables which have zero mean and store in the matrix object X.

2. Derive the eigenvectors of $X^T X$ and store them in the matrix object W.

3. Obtain $P = XW$, the principal components decomposition P of X.

4. (a) Obtain the diagonal matrix $S = P^T P$.

   (b) Calculate what percentage each of the variances in matrix S are of the total.

   (c) State which principal component(s) should be dropped to simplify the dimensionality.

5. (a) Obtain the matrix P using R's built in PCA function.

   (b) Obtain the percentages in part 4(b) from this PCA function.

   (c) Draw a scree diagram using plot of the result and hence state which principal component(s) should be dropped to simplify the dimensionality.

6. (a) Obtain a new matrix P_1 which has only the first two principle components and vectors of zeroes for the removed components.

   (b) Obtain the reduced data set X_1 using $X_1 = P_1 W^T$.

   (c) Plot X_1 and compare to the original data.