37459 Multivariate Statistics - Assignment 1

- This assignment is worth 35% of the marks for the subject.
- Late submissions are penalised by reducing the maximum attainable mark by 2% per day, or part thereof. e.g. after one day, it will be worth a maximum of 33 %, and e.g. after 18 days it will score zero.
- For each question, you are expected to perform (only) the analysis listed. Write up your analysis as a short report. Any graphics that you use to support your discussion should also be included.
- The R code that you use to produce the analysis must also be submitted with your assignment as an appendix. Your code must show all working to get full marks; it is not sufficient just to give the answer.

Question 1

(10 marks) Consider the oxygen consumption data given in Table 6.12 of Johnson and Wichern. (Note: the only thing we are using from chapter 6 here is this data; the assignment should be done by using methods from earlier chapters, not methods from chapter 6 for example.) Look at the *female* observations only. It is recommended that you convert the data into .csv format before importing into R.

a) Examine the univariate normality of the variables x_2 and x_4 using a $Q-Q$ plot.

b) Examine the multivariate normality of the variables x_2 and x_4 using a chi-square plot of ordered squared generalised distances.

c) What are the conditions under which we can make meaningful inferences from the chi-square plot?

d) Find the best Box–Cox transformation to make each of these variables near-normal.

e) Show the $Q-Q$ plot of each transformed variable. Comment.

Question 2

(12 marks) Using the flea beetle data for the *concinna* species, available from the course website with this assignment, answer the following questions.

a) Find and plot the 95% confidence ellipse for the population means μ_1 and μ_2.

b) Suppose that it is known that $\mu_1 = 125$mm and $\mu_2 = 14$ units for the *heikertingeri* species. Are these plausible values for the *concinna* species? Explain.

c) Construct the simultaneous 95% T^2-intervals for μ_1 and μ_2.

d) Construct the simultaneous 95% Bonferroni intervals for μ_1 and μ_2.

e) Compare these intervals. What are the advantages and disadvantages of each type of interval?

f) Use $Q - Q$ plots and scatter plots to investigate whether a bivariate normal distribution is appropriate. What is your conclusion?

Question 3

(22 marks) You are given that $Y = (Y_1, Y_2, Y_3)^\top$ has a multivariate normal distribution, with zero mean $(0, 0, 0)^\top$, and with exact and true covariance matrix

$$\Sigma = \frac{1}{5630} \begin{pmatrix} 575 & -60 & 10 \\ -60 & 300 & -50 \\ 10 & -50 & 196 \end{pmatrix}.$$

a) Find Σ^{-1}, the inverse of the covariance matrix.

b) Are there any subsets of the variables that are conditionally independent?

c) Give a few lines of R code to draw random samples of Y.

d) Using your R code, simulate a few thousand points, to generate a sample of data from this distribution. Visualise the distribution: Create three pairs of 2D plots, and one 3D plot.

e) Find the eigenvalues and the eigenvectors of the covariance matrix.

f) Consider a Principal Component Analysis. What are the principal components?

g) If you had to represent the data using two principal components, then which two would they be? What percentage of the variance is explained by your two principal components?

h) Perform a linear regression to find the coefficient β_1 in

$$Y_2 = \beta_1 Y_1 + \epsilon_2$$

where $\hat{Y}_2 = \beta_1 Y_1$ is the linear least squares predictor of Y_2 based on Y_1, and where $\epsilon_2 = Y_2 - \hat{Y}_2$ is the prediction error.

i) Perform a linear regression to find the coefficients β_1 and β_2 in

$$Y_3 = \beta_1 Y_1 + \beta_2 Y_2 + \epsilon_3$$

where $\hat{Y}_3 = \beta_1 Y_1 + \beta_2 Y_2$ is the linear least squares predictor of Y_3 based on Y_1 and Y_2, and where $\epsilon_3 = Y_3 - \hat{Y}_3$ is the prediction error.

j) Estimate $\mathrm{Var}(Y_2|Y_1)$.

k) Estimate $\mathrm{Var}(Y_3|Y_1, Y_2)$.

For (h)-(k), you should find two different ways (one way is via exact formulas, and the other way is via simulation in R) to compute the same answer. Then one way can be a check on the other.

Question 4

(16 marks) Use the 'diabetes data set,' available in `R` through the `LARS` package. In this data, $p = 10$ variables are observed in $n = 442$ patients. For example, a snippet of this data appears in Table 7.2 of the book (which is freely available online):

"Computer Age Statistical Inference: Algorithms, Evidence and Data Science"
by Efron and Hastie

The data in the book and the data in the R package have some minor differences, which you might note. It might be better to use the data from the book. You might like to consider issues of inference after model selection, discussed in chapter 20 of the book, and Table 20.1; however those topics from chapter 20 are NOT required for this assignment.

Consider *linear regression* to predict the response y from the ten covariates. Do not use PCA.

Use this data, and `R`, to demonstrate your knowledge of linear regression. Base your work on chapter 7 of the textbook by Johnson and Wichern (mainly *sections 7.1 to 7.6, and 7.8*). Examples of things you could try are:

a) Check for collinearity between the predictor variables.

b) Estimate the coefficients β in a linear predictor, using Result 7.1 of that chapter.

c) Show how to use the coefficient of determination R^2 in section 7.3.

d) What is the residual standard error? Is there a relationship between the response and the predictors?

e) What is your final model?

f) Examine the residuals, using the techniques of Section 7.6.

g) Examine the one-at-a-time confidence intervals and the simultaneous confidence intervals for the coefficients. Do these intervals include 0?

h) Are you able to make any comments on the issue of choosing a 'good' subset of predictors?