Athletes Data Analysis

Written sections (other sections in AthletesAnalysis.R file)

The **athletes** data were collected at the Australian Institute of Sport by Richard Telford and Ross Cunningham, see Cook and Weisberg (1999). For the 100 female and 102 male athletes the variables are listed in Table 4.3 of Lecture 2 and in Chapter 4 of Koch and Pope which accompanies the Lecture 2 slides.

The character f in the column Sex indicates 'female' and an m in this column indicates 'male'.

[Question 1: 20 marks]

Denote by ais.q1 the raw athletes data without the variable Sex and without the variable LBM, lean body mass.

(a) [2 marks]

i. What is the number of observations and the number of variables in ais.q1?
ii. Show a parallel coordinate plot of the male and female athletes either as two separate plots or in different colours in the same plot.
iii. Point out the main differences in the variables between the male and female observations.

1.
    a.
        i. 202 observations, 10 variables
        ii.
        iii. Most variables have a greater ranger for males, especially Ferr, Ht and Wt. Males also tend to generally have higher values for most variables with the exception of PctBfat and SSF. For SSF especially, females seem to have a larger range and higher values than males

(b) [4 marks]

i. Carry out a principal component analysis of the raw data ais.q1 and present in the form of a graph of the eigenvalues against the index, and of the cumulative contribution to total variance against the index. (Hint. Recall that cumulative contribution to total variance is expressed as a ratio.)
ii. What is the largest eigenvalue of the covariance matrix of the data ais.q1?
iii. What is the smallest number of principal components required to yield at

least 95% of the cumulative contribution to total variance?

iv. How many principal components represent a good approximation to the raw data? Give a reason for your choice. (Hint. A kink is not the answer we want.)

    b.
        i.
        ii. 2300.305
        iii. 3
        iv. 3 components as this is the number that account for at least 95% of the variation in the data. This is also the point on the graph where the plot stops decreasing quickly relative to the fast decrease from k=1 to 2 and k=2 to 3

(c) [6 marks]

Consider the principal component data obtained in part (b) for the raw data ais.q1.

i. Show score plots of
   • PC1 against PC2 of the raw data ais.q1 and • PC2 against PC3 of the raw data ais.q1 with the first-named component displayed on the x-axis. Use different colours or plotting symbols for male and female athletes in your score plots.
ii. Scale the raw data ais.q1 and call its scaled version ais.q1.sc. Calculate the first three principal component scores PC1, PC2 and PC3 for the scaled data ais.q1.sc. Show score plots of
   • PC1 against PC2 of the scaled data ais.q1.sc and • PC2 against PC3 of the scaled data ais.q1.sc with the first-named component displayed on the x-axis. Use different colours or plotting symbols for male and female athletes in your score plots.
iii. List the variable(s) that contribute(s) most in absolute value to the first and second PC-directions separately for the raw and scaled data.
iv. Briefly comment on the graphs and calculations of part c(i) and c(ii) and highlight differences between the score plots and contributions to the PCs.

c.

i.

ii.

iii.

iv. For PC1/PC2, the scaled plot is much more separated by gender than that of the raw data plot. In terms of variable contribution, the raw data is heavily weighted by a single variable (Ferr for PC1, SSF for PC2) in each case, while that of the scaled data is much more spread between variables, making this data better for PCA. For PC2/PC3, the scaled plots are more spread out indicating that they capture more variation of the data when compared to the raw data plots

(d) [4 marks]

Consider the principal component data obtained from the raw data ais.q1 in part (b) as predictors in principal component regression. Take the variable LBM as response.

i. Write down a mathematical expression for the linear regression model based on the components of the principal component data. (Hint. Note that the response variable is not centred.)
ii. For k = 1, . . . , 10, calculate the regression coefficients and the residual standard deviation based on the first k principal component vectors as predictors. Show a graph of the residual standard deviations against the number of principal components k. How many PCs would you chose? Give a reason for your choice.

d.

i. $Y = \beta_0 + \beta^1 PC1 + \beta^2 PC2 + \cdots + \beta^{10} PC10$

ii. 3 would be a good choice of components to use as this is the point when the residual standard deviation stops decreasing quickly relative to previous points

4. (e) [4 marks]
   Consider (standard) linear regression (SLR) with multivariate predictors, and let PCR denote the linear regression approach with the PC components as predictors.
   i. Without doing any calculations, briefly describe two or three of the main differences between the two approaches. (Hint. You may want to think of forward selection of variables for SLR, and adding of PCs as in part (d).)
   ii. Regard the analysis in part (d). Consider and comment on the coefficient estimates you obtained for the PCs as you increase the number of components in the model. What do you notice? Give a reason for this behaviour.

e.

   i.
      1. Dimension reduction – because PCR is derived from PCA, it will generally require less variables to be used in the regression than SLR while keeping most of the variability of the original predictors. This helps reduce model complexity which is an advantage
      2. Multicollinearity – if there is some degree of multicollinearity present in the data, PCR will usually avoid this problem through of its use of linear combinations. SLR will be prone to multicollinearity however in this case
      3. Feature selection – in SLR we use forward selection and know exactly which factors affect the response variable and the order of their significance. PCR is not a feature selection method and can make it harder to determine which variable is affecting what without looking further into the PC's
   ii. The coefficient estimates show to how much each PC affects the response variable LBM. We can see that in absolute value, PC3 (-0.75) has the greatest effect meaning it accounts for the most change in LBM of all the PC's. 3 PC's is also the number that was chosen from the residual standard deviation graph, so it aligns with this choice

[Question 2: 20 marks]
Write ais.q2 for the raw athletes data without the variable Sex. Define the two subsets ais.phys and ais.sero of ais.q2 which consist of all observations and the following variables ais.phys: consists of the variables ”PctBfat” ”BMI” ”Ht” ”LBM” ”SSF” ”Wt”, ais.sero: consists of the variables ”Ferr” ”Hc” ”Hg” ”RCC” ”WCC”.
Write X[1] for the data matrix which is the transpose of ais.phys, and write X[2] for the data matrix which is the transpose of ais.sero. Put
□X[1]□ T □ □ Xnew = [2] and Xnew = ais.phys ais.sero .
(a) [3 marks]

i. Calculate the correlation matrix of Xnew and display the part of the correlation matrix (including the variables names) which corresponds to the 'between correlation matrix' of the X[1] and the X[2] data.
ii. Which variable of X[1] and which variable of X[2] give rise to the strongest correlation coefficient in absolute value? What is this value?

iii. Show a scatterplot of the two variables you identified in part (a)ii with the variable of X[1] on the x-axis.

2.

a.

i.

ii. LBM from X[1] and Hg from X[2] give the strongest correlation of 0.6109861

iii.

2. (b) [5 marks]
Carry out a canonical correlation analysis for Xnew with the two parts X[1] and X[2] which are defined at the beginning of this question. Determine the pairs of canonical correlation scores (U•k , V•k) for k ≤ 5.

i. What do you think the two parts of the data represent and why should there be correlation between the two parts? (Hint. Give an interpretation in terms of their actual meaning, eg, by making use of Table 4.3 of the lectures slides or Chapter 4.)

ii. List the strength of the correlation of the pairs (U•k , V•k) in decreasing order of stregth.

iii. Show separate score plots, one for each pair of canonical correlation scores.

b.

i. The two parts seem to represent physical attributes of athletes in X[1] and attributes to do with blood measurements of athletes in X[2]

ii.

iii.

2. (c) [6 marks]

i. Comment on the strength of the correlations in part (b) and compare these and the CC score plots to the scatterplot and correlation coefficient obtained in part (a).

ii. What do you notice about the scatterplot of the first pair of canonical correlation scores? Can you suggest an interpretation of this scatterplot? How does, what you find, relate to the other scatterplots in part (a)iii and (b)ii? Comment.

iii. Without doing any calculations, discuss the following and give a reason for your answer. Do the CC scores differ depending on
• whether the data are first scaled and then split into the two parts, or
• whether the X[1] and the X[2] parts are scaled separately?

c.

i. For any two variates, the best possible pair of between variate correlations is the first pair of canonical correlations due to its use of linear combinations. It is not possibly to select a more correlated pair from the between correlation matrix of part (a). The most correlated pair in CCA has a correlation score of 0.74955341 while that of part (a) is 0.6109861.

ii. The CC scatterplot of the first pair shows a relatively strong positive correlation, i.e., as X tends to increase Y tends to increase also. The scatterplot of part (a) is more stretched out and random in comparison. As well at this the remaining pairs of CC scores are also more spread out and less indicative of a relationship than CC1. This is expected, as their CC values are all less than half of the first CC pair

iii. Scaling has no effect on CC scores as all scores will be identical in strength and correlation regardless of whether the data are scaled initially, split first then scaled, or not scaled at all. The only changes are usually small changes in the pairs of eigenvectors (p's and q's). For their correctness, we separate the data into X[1] and X[2] and scale them at this point, then apply CCA on the two scaled sets

2. (d) [6 marks]
   The pairs of CC scores vary in the strength of their correlation. To find out whether the data could arise from a model with zero correlation between some of the pairs of CC scores we conduct a hypothesis test.

i. Write down hypotheses for testing whether the population correlation could be zero, state an approximate test statistic for these hypotheses, list what distributional assumptions are required of the data, and how the test statistic is used to make a decision.

ii. For k ≤ 4 test the hypotheses that the correlation could be zero. For each k, list the value of the test statistic, the p-value of the test and the degree of freedom. Make a decision and give a reason for your decision.

iii. Without doing any calculations indicate how one may check whether the data satisfy the assumptions of the null hypotheses. Explain implications of testing if the data do not satisfy the distributional assumptions of test set-up.

d.

i. The hypotheses are:

$$H_0^k: \gamma_{k+1} = 0 \quad vs \quad H_1^k: \gamma_{k+1} > 0$$

We keep testing all k from k=1 (which tests the second pair) or start with the last k until we accept the null hypothesis, as at this point there is zero correlation in the model. The test statistic, -2logLR, for which small values favour the null hypothesis can be approximated to the T(k) statistic which we actually use in testing since it is easier to calculate. To make a decision, we receive the T(k) and chi-square critical value from our test for each k. If T(k) > crit value, we reject the null hypothesis and conclude that there is correlation for that specific k pair. If starting from k=1 we repeat the test for k=2 and so on until we accept the null hypothesis. If starting from the largest k, if T(k) > crit value we conclude that all pairs could have some (non-zero) correlation. If T(k) < crit value in this case, we work backwards until T(k) > crit value and conclude that up to that k, all pairs are correlated. The data is assumed to be normal throughout testing.

ii. At a 2% significance level, for k=4 we can see that the test statistic T(k) is greater than the critical value. We therefore reject the null hypothesis and conclude that all CC pairs could be correlated

iii. The null hypothesis and assumes that there is some correlation between the first CC pair (otherwise doing CCA is pointless). To check this, we can simply look at the correlations of X[1] variables with X[2] variables (the between correlation matrix) and if any have an absolute value of above, say 0.5, we know that there is at least some correlation between these variables. If the data does not satisfy the normality assumptions, it can lead to inaccuracies in the analysis. It is still possible to test non-normal data however, and generally, if the sample size of the data is not too small and not too close to the critical value, testing can be done with some degree of accuracy

3. [Question 3: 20 marks]
Consider ais.phys and ais.sero as in Question 2. Write ais.q3 for the data which consist of the columns of ais.phys followed by the columns of ais.sero.
Scale the data ais.q3 and call the scaled data ais.q3.sc. Split ais.q3.sc into two parts in the following way:
the first part, called ais.phys.sc, contains the (scaled) variables ”PctBfat” ”BMI” ”Ht” ”LBM” ”SSF” ”Wt”;
the second part, called ais.sero.sc, contains the (scaled) variables ”Ferr” ”Hc” ”Hg” ”RCC” ”WCC”.
(a) [1 mark]
List the first entry of ais.phys.sc and the first entry of ais.sero.sc. (Hint. The first entry of ais.phys.sc should be almost three times larger (in absolute value) than the first entry of ais.sero.sc.)

a. ais.phys.sc: 1.008521772
   ais.sero.sc: -0.355279947

(b) [6 marks]
Conduct separate cluster analyses of the two datasets ais.q3.sc and ais.sero.sc defined in (a) using the k-means approach.

i. Using kmeans with nstart=25, calculate the cluster membership of each observation for k = 2,3 and 4. Determine the number of observations in each cluster and list them in an appropriate table.
ii. For each of the two datasets of part (a) and for each k = 2,3 and 4, use the cluster allocation obtained in (b)i and present this information in a table of the form shown in Table 1 which shows how the clusters align with the two genders.

Table 1: Number of observations in k clusters C1, . . . Ck by Sex C1 ... ... Ck
female n1 ... ... nk male m1 ... ... mk
iii. Comment on the results of these analyses and the information presented in the corresponding tables.

b.
   i.
   ii.
   iii. For both subsets of data, the observations seem to be split relatively well between clusters with Phys being split more evenly than Sero for k=3 and k=4. Looking at how the clusters align with gender show that they do well

at clustering by gender, most notably in the k=4 case for Phys and to a lesser extent the k=2 case for Sero

(c) [4 marks]
Consider the two datasets ais.phys.sc and ais.sero.sc defined in part (a). Separately classify each of the two datasets using linear discriminant analysis based on lda with uniform priors (equal proportions in each class). Use Sex as response variable.
i. Show your results in the format of Table 1, where C1 and C2 now refer to classes the observations have been assigned to by the rule.
ii. Comment on the results of the analyses and shown in the tables.

c.
   i.
   ii. The analysis has performed much better for Phys. For Phys, there is only 1 misclassification as an M misclassified for F. The error rate is 0.5%. For Sero, 9 F and 5 M are misclassified, and the error rate is 6.93%

(d) [4 marks]
Repeat the calculations of part (c) using logistic regression based on glm. Present your results similarly to those of (c)i and provide comments corresponding to (c)ii.

d. The analysis has performed slightly better for Phys. 2 M and 1 F are misclassified giving it an error rate of 1.49%. For Sero, 6 M and 9 F are misclassified giving an error rate of 4.95%

(e) [5 marks]
   i. Compare the results of the analyses in parts (b)–(d) with reference to the different methods of analysis. Do not repeat comments previously made in parts (b)–(d).
   ii. Compare classification based on linear discriminant analysis and logistic regression by highlighting some of the differences and advantages of the two approaches.

e. Comparing k-means, LDA and logistic regression
To compare all 3 methods, we can use the 2-cluster case of k-means to create a table similar to those created for the other methods. Then, we can rank each by accuracy

Error rate

	Phys %	Sero %	total
K-means	6.93	11.88	18.81
LDA	0.50	6.93	7.43
Logistic regression	1.49	4.95	6.44

The most accurate is LDA overall, with LDA being most accurate for Phys, and logistic regression being most accurate for Sero. K-means ha a much higher level of error overall but this is not surprising as it is an unsupervised method and so does not consider class labels. Both LDA and logistic regression are supervised forms of classification which make them more useful for this type of analysis

- LR essentially predicts whether something is true or false and can be seen as both regression and classification
- LDA focuses on maximising the separability between classes by creating a new 1D axis from a 2D graph of both classes. It projects the data onto the new axis buy maximising the distance between clusters and minimising the distance within classes
- LR is based on maximum likelihood estimation while LDA is based on least squares estimation
- LR is not as sensitive to outliers as LDA
- LR is usually preferred over LDA as it is a more robust method