Athletes Data Analysis

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.
      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.
      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.
      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
   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

2.

a.

i.

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

iii.

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.

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

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.

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

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.

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. 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. Comparing k-means, LDA and logistic regression

i. 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

ii.

- 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