Multivariate Linear Analysis Exam – 19333982 – Jack Cleary

Q1 (a)
We should standardise the predictor variables in both of these cases to ensure that no one variable dominates due to having a large variance. This is necessary because the variables are all measured in different units and so we must standardise across all predictors.

(b)
Cross-validation is a technique whereby we split the data into three sets; training, test and validation; and use each section for a different purpose. The training set firstly, is used to teach the model how to classify unlabelled points. The test set is then treated as an unlabelled set of points and these are used to find the best value of k for classification. The validation set lastly, is used to estimate the classification error when using the optimal k-value identified by the test set.
Based on the graphic in the Appendix, the optimal number k of nearest neighbours is 17.
I would interpret this as a very high misclassification rate, as this means that the model will only correctly diagnose heart disease about 65% of the time. This is not a very high figure, especially due to the weight of a diagnosis like this in a realistic sense. A false positive diagnosis would be very emotionally draining, potentially life changing and a false negative diagnosis could cost a patient their life. Overall, I would say that this model would not be an useful one in predicting heart disease.

(c)(i)
Firstly, on examining the coefficients for each variable we can see that Age is not useful to our result as its p-value is extremely insignificant. In contrast, MaxHeartRate seems to have a major impact on the result, with a large negative coefficient and a negligible p-value. This indicates that this variable is significant in our prediction. The difference between the Null deviance and residual deviance indicates that this model does not perform extremely well as there is not a major difference between the two. This means that a model without any variables would perform almost as well. Lastly, the AIC value is quite high which would indicate again that the model does not perform very well, although this is slightly improved with glm2. I would predict that this model could be much improved if variables were standardised.

(ii)
The output from a logistic regression model can be used for classification by forming an equation in all of the variables using the coefficients outputted. The solution to this equation is the log odds of a patient having heart disease. Using the output from glm1, I compute that the probability of the referenced patient having heart disease is -1.8074, which corresponds to a probability of about 85% that the patient has heart disease.

(iii)
There are a number of ways of comparing regression models. Firstly, we could compare the difference between the Null deviance and residual deviance for each model. In this case, glm1 has a difference of 59.64 whereas glm2 only has a difference of 59.63. Another thing

to compare is the Alkaike Information Criterion (AIC), which is an estimate of the prediction error in the model. The goal is to minimize the AIC and in this case, glm2 has a slightly lower AIC value. Lastly, we could regard the p-values of the variables, if they are below 0.05 then we can say that the results are statistically significant at a 95% confidence interval. In both models, we have variables with p-values greater than 0.05, indicating again that neither model performs exceptionally well. Overall however, I would say that glm2 is a slightly superior model.

(d)
Discriminant analysis is a supervised learning method used to reveal structure in a dataset. These methods assume the use of a distribution over the data, and hence the uncertainty around a structure in the data can be quantified. Where LDA and QDA are used it is assumed that there are a definite number k of groups within the dataset, and that there is a subset of the data that is labelled, i.e. it is known to which group they belong. Hence, we use both a training and test set, similar to KNN. We do not need a validation set however as the test set will be assigned with probabilities of uncertainty as opposed to assignments. In this case, I would expect LDA to perform very similarly to logistic regression but probably slightly worse as this method would rely on the data having a fixed distribution whereas logistic regression does not. Also, the linear relationships between variables do not seem very strong and hence a linear discriminant analysis would probably not perform very well

Q2(a)
Principal Component analysis performs well on highly correlated data and so we can expect this to be a good method to use on this data. The scatter plots show fairly high correlations between variables, in particular birth rate, death rate and life expectancy. This analysis also makes it easier to identify clusters, or sub-groups within the data. The covariance matrix also shows high levels of variance between variables, and this analysis aims to retain as much variance as possible when it is used, hence it would be appropriate here.

(b)
PCA is effectively a method for re-expressing data, with as much variance as possible, in terms of a few linear combinations of the original variables in order to reveal internal structures. It can be used as either a dimension reduction technique or to identify relationships between variables. The aim is to describe the correlated variables X_1, X_2,….X_m, using a set of uncorrelated variables Y_1, Y_2,….Y_p, where p will be less than m and each of the Y variables are a linear combination of X_1, X_2,….X_m. The first 'principal component', Y_1, will be selected as the most important principal component, that describes more variation in the original dataset than any other possible combination of X_1, X_2,…. X_m. The next principal component, Y_2, will be selected to account for as much of the leftover variation as possible, as long as it is not correlated with Y_1. And this process continues until all variance is accounted for.
This data should have been standardised before performing the PCA analysis, as this would prevent unnecessary weight being given to the variables of greatest variance. These independent variables are all measured in different units and for example, inflation rates

have tiny values compared to the other variables which means it would barely contribute to the results. Hence, standardisation would have been appropriate here.

(c)
The first part of our output shows the importance of each component, or how much variance is explained by each. As we see from the bottom row, which shows the cumulative proportion of variance, 58.18% is explained by PC1, 76.78% by PC1 and 2, and 92.98% by PC1, 2 and 3. Choosing the optimal number of principal components can be subjective, however a general rule of thumb is to account for a minimum of 80% of the variance. This would mean choosing to use the first three principal components. We could potentially also choose to use only the first two as a fairly large portion of the variance is still explained and our dimensions are reduced further, however in a realistic sense I think that three groupings would make more sense in this case as this also reflects the real world groupings of first, second and third world countries. The second table shows the loadings of each variable in each PC, so for example, we would expect observations that have lower birth rates and death rates, and higher Life expectancy and GDP-per-capita to have high PC1 values. Similarly, countries with very low inflation rates would have much higher PC2 values.

(d)(i)
In multidimensional scaling, we look to reduce the dimensions in our dataset while preserving the similarity/ dissimilarity between points. In simple terms, if two points are very close to each other in 5 dimensions, we want them to be very close in 2 dimensions. The metric least squares scaling method looks to find a configuration of points which will minimize a loss function S. This function maps the difference between the original dissimilarities and the new dissimilarities after dimension reduction. T-SNE is a slight variation of this as instead of using distances between points, it converts these to probabilities. Effectively, instead of minimizing the distance between two points in a lower dimension, it will maximise the probability that they are each other's nearest neighbour. These differ from principal component analysis as PCA starts by calculating correlations among samples, whereas multidimensional scaling methods start by calculating pairwise distances among samples.

(ii)
Procrustes method compares two dimension reduction techniques by taking their outputs and matching them to one another. For example, if there are two outputs X and Y, it will match the *i*th point in X to the *i*th point in Y by dilation, rotation, reflection and translation. It will then minimize a sum of squared distances known as Procrustes sum of squares in order to measure the match between the two configurations.
The Procrustes output showing the Sammon scaling method vs PCA shows a great degree of similarity between the two methods. This is evident from the very blue short arrows protruding from each point. In contrast, the comparison of t-SNE and PCA is not nearly as accurate as we can see much longer blue arrows which shows that the points do not agree after matching.