**PILLA SATISH**

**STUDENT NO-210472095**

**Newcastle University**

**MAS8404 Summative Assignment**

**Cluster Analysis**

The ISLR gene expression dataset (Ch10Ex11.csv) consists of 40 tissue samples with measurements on 1,000 genes. The first 20 samples are from healthy patients, while the second 20 are from a diseased group.

Reading the dataset

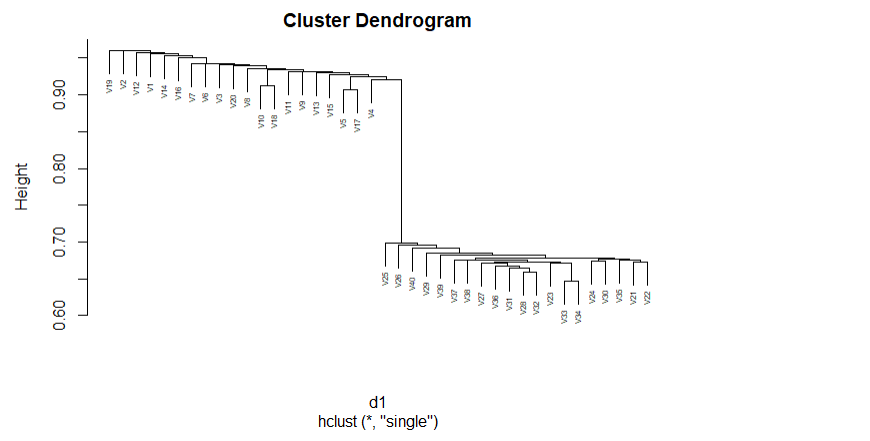


While checking the data we observed that the gene expression data is stored in wrong way around the rows representing genes and columns representing tissue samples. We need to do transpose to get the data matrix.



**Hierarchical clustering**

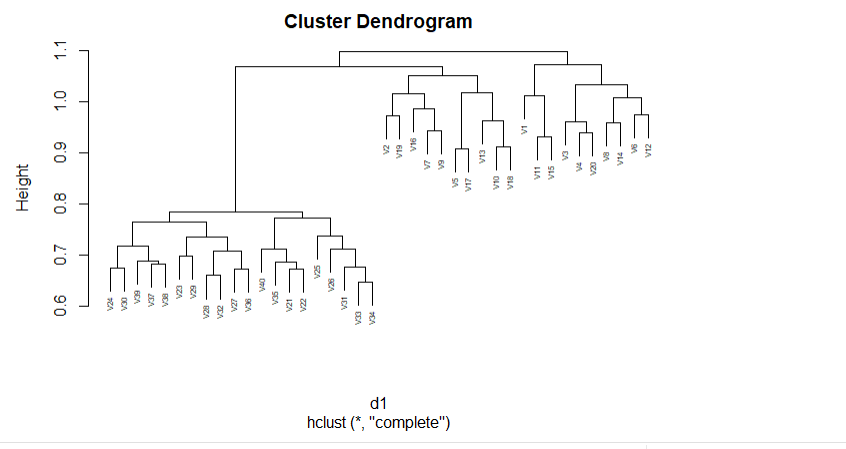
1a. Apply hierarchical clustering with single-linkage using correlation-based distance.



**Fig 1: Dendrogram from applying agglomerative hierarchical clustering with single-linkage using correlation-based distance to ISLR gene expression data**

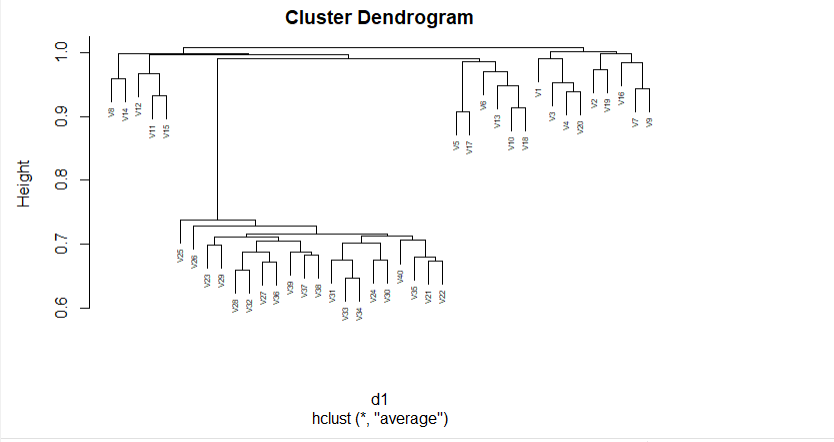
By observing the dendrogram we can clearly see that the nearby observations are fused one-by-one leading to long, thin cluster. We can say that gene is not separating the samples is two groups. When we cut the tree at height 0.80 it is producing k=21 clusters if divided at height 0.95 it produces K=7 clusters.

1b. Complete linkage and average linkage



**Fig 2: Dendrogram from applying agglomerative hierarchical clustering with complete-linkage using correlation-based distance to ISLR gene expression data**

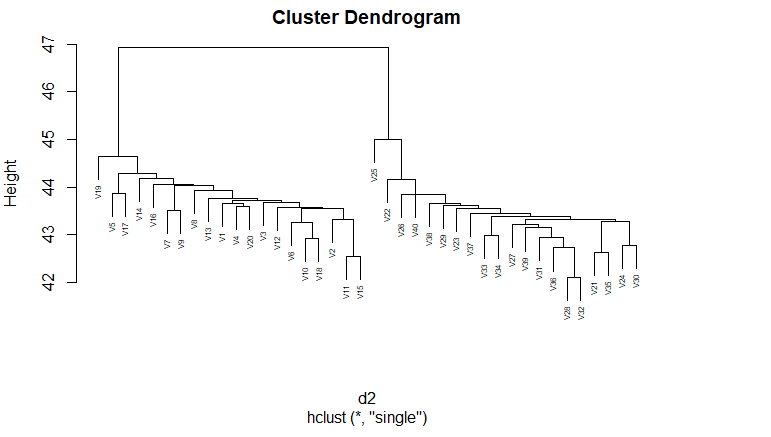
The dendrogram of complete linkage balanced, When the dendrogram is cut a height of 1.05 we can see at least 5 well separated clusters.



**Fig 3: Dendrogram from applying agglomerative hierarchical clustering with average-linkage using correlation-based distance to ISLR gene expression data**

The dendrogram for average-linkage gives k=21 when the dendrogram is cut a height of 0.9. For all the three linkage when the dendrogram is cut a height of 0.9 it is producing k=21 clusters.

1c. Single-linkage using Euclidean distance

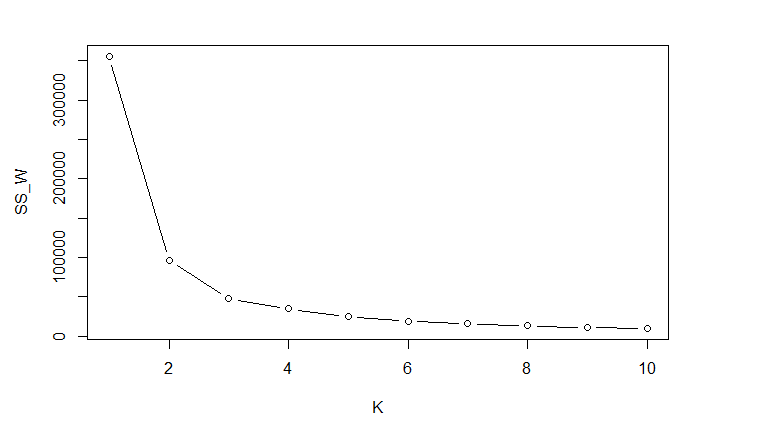


**Fig 4: Dendrogram from applying agglomerative hierarchical clustering with Single-linkage using Euclidean distance to ISLR gene expression data.**

From the above dendrogram we can say that when the dendrogram is cut at height=45.5, it gives exactly two clusters. By comparing the above, we can say that numbers of clusters produced depends on the distance metric used, when we used correlation-based distance we are getting higher cluster values(k=21) when Euclidean distance is used we are getting k=2 clusters. From above we say that the using Euclidean distance gives us right number of clusters and it’s is best method to determine the cluster for ISLR gene expression data.

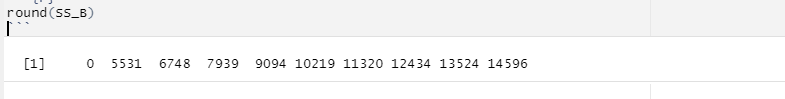
**K-means clustering**

2a. Applying the K-means algorithm for a range of values of K



**Fig 5: SSw against K for K-means clustering of the US ISLR gene expression dataset**

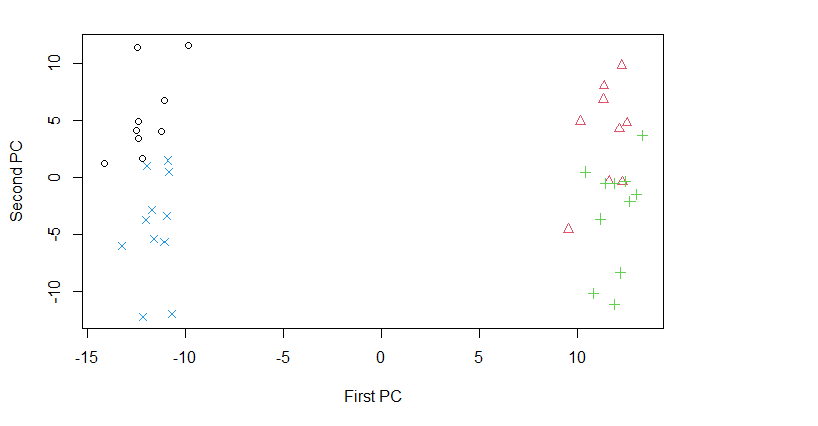
From the above plot we can see there is an elbow at K=2, which suggest it might be appropriate to use 2 cluster.



We can also see that difference between-cluster sum of squares (SSb) values after K=2 are approximately equal.

Due to large number of variables p=1000 makes it very difficult to get any kind of feel for whether the clustering procedure as found any interesting structure.

2b. visual display of the four clusters in a two-dimensional plot.



**Fig 6: Cluster solutions in the ISLR gene expression data in the space of two PCs. Clusters were identified by K-means clustering with K=4**

From the above we see that in the low-dimensional representation of the ISLR gene expression data, the cluster appear to be reasonably well separated when K=4.

**Linear Regression**

3b. Test error from the full least squares model



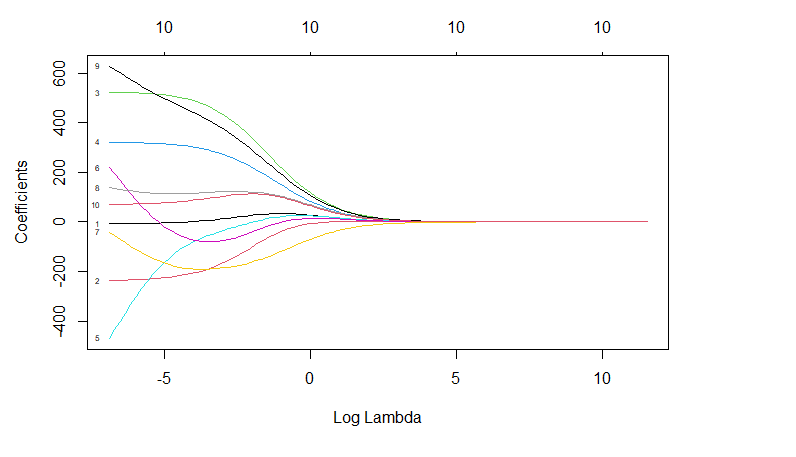
3c. Test error from fitting a least squares model to the training data using the 6 predictors



3di. Optimal Lambda value and test error

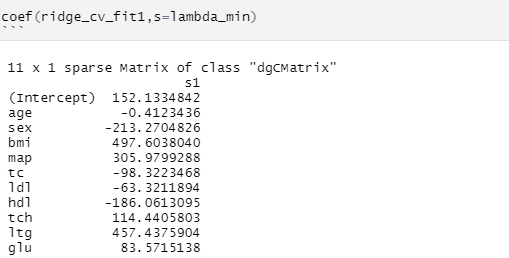


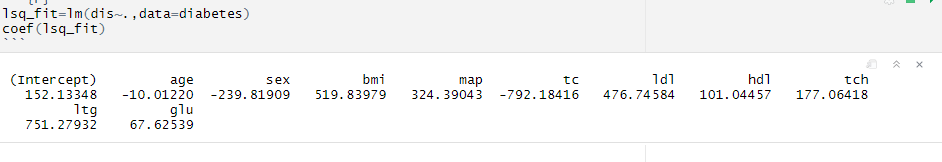
3dii. Plot showing how the estimates of the regression coefficients change as the tuning parameter is increased



**Fig 7: Graphical illustration of the effect of varying the tuning parameter when applying ridge regression to the diabetes dataset.**

Each line represents the regression coefficient for a different variable. We can see from the plot that by the time log lambda is around 4(i.e. lambda is around e^5=57), all regression coefficients are essentially equal to zero. Ridge regression includes all the variables in the model. Ridge minimizes the residual sum of squares plus a shrinkage penalty of lambda multiplied by the sum of squares of the coefficients. As lambda increases, the coefficients approach zero.





**Fig 8: estimated regression coefficients for ridge regression and estimated coefficients in the full model fitted by least squares**.

From above we can see that regression coefficients (age, sex, tc, hdl) for the ridge regression are tending towards zero, while in least square method (age, sex, tc) are tending towards zero.

3e. Comparing the three test errors

Test error on the least square using 6 variables is best method as it produces least test error value among other two methods. Using only 6 variables we are able to build a model which helps in predicting the response variable with low test error among 3 methods.