Multivariate Data Analysis : Milk Transportation Data

Dilshad Imon (191040)
Muskan Kaur (191080)
Saptarshi Roy (191128)
Souvik Paul (191152)

Department of Mathematics and Statistics,
Indian Institute of Technology, Kanpur

Abstract

In this project we have used the Milk transportation Data which consists of three variables Fuel cost, Repair cost and Capital cost. These all costs are due to transportation of milk from farms to diary plants and during transportation only Gasoline or Diesel is used as fuel.Here we considered two populations - one due to Gasoline and other due to Diesel. firstly we have checked the Normality of two populations and accordingly we took a suitable transformation of our variables. We did principal component analysis to reduce the number of variables if possible and it was possible for gasoline data. Then we have used other techniques to study these populations. We checked if there any outlier using hat matrix and we replace them with respective column means. We calculated confidence region for each of the variables in each population to see if there any significant difference. We performed Box's M test to see if there any homogeneity between two populations. accordingly we used QDA and LDA and QDA came better and we found the possible reasons. We also performed KNN classification, Logistic Regression and observed the error rate. Finally we used Lachenbruch's Holdout procedure to check the performance.

Contents

1 Introduction

Here we have the transportation data where the variables are given as Fuel(Gasoline or Diesel) used for transportation and three types of cost - Fuel cost(X_1), repair cost(X_2) and capital cost(X_3) measured on cent/mile. In the data under the variable Fuel, '1' denotes Gasoline and '2' denotes Diesel. Some typical obseravations are shown below :

	Fuel Type	Fuel Cost	Repair Cost	Capital Cost
33	1	9.18	9.18	9.49
34	1	12.49	4.67	11.94
35	1	17.32	6.86	4.44
36	2	8.50	12.26	9.11
37	2	7.42	5.13	17.15
38	2	10.28	3.32	11.23
39	2	10.16	14.72	5.99

Figure 1: Data

Here we have data of the form $(\underline{X}^T, Y) = (X_1, X_2, X_3)$ with $n_1 = 36$ observations for Gasoline and $n_2 = 23$ observations for Diesel.
Here our objective is to check :

- Can we consider the linear combinations (< 3) of the observed variables to explain the variability in the data ?
- Can we construct some confidence intervals for some functions of class-specific mean vectors ?
- Can we construct some rule for discriminating the Fuel based on the observed costs ?

To get the answer , we can start with :

- Checking the class-specific data on the costs are multivariate normal or not. If not we will take transformation and will again check.
- Checking both the covariance matrices are same or not (for Gasoline and Diesel data)

- If Covariances are same , we will check if mean vectors of these two populations are same or not.

2 Data Visualisation

Here we will see the correlation scatter plot of two population Gasoline and Diesel:

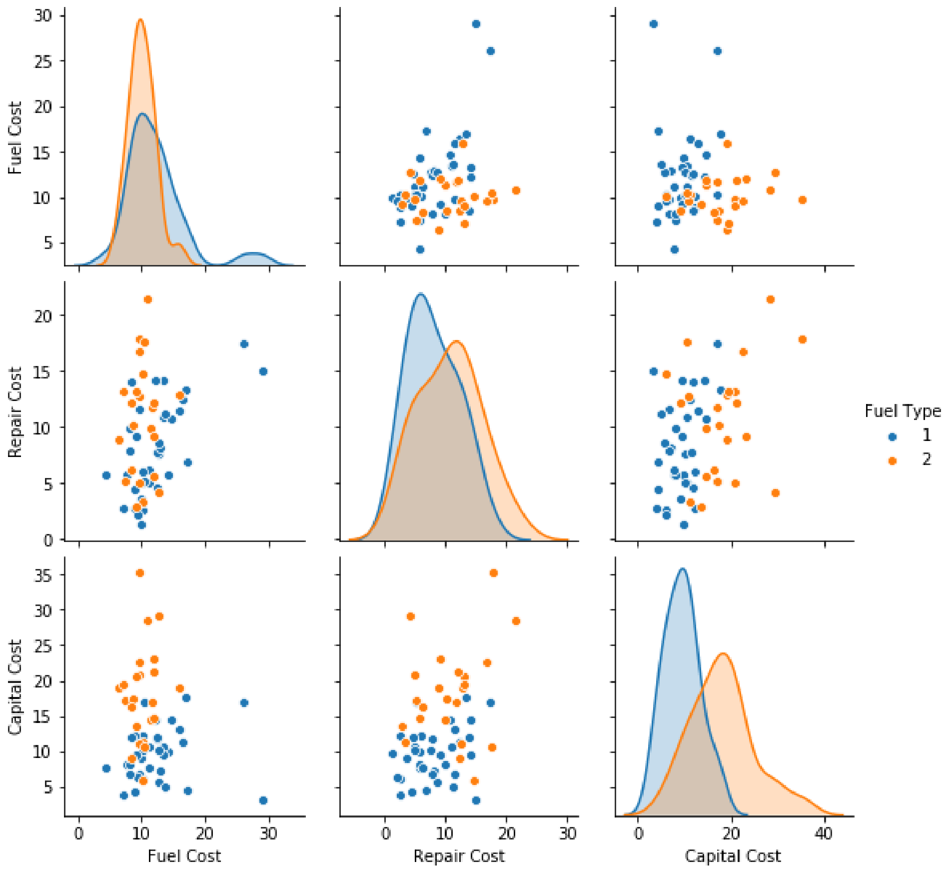

Figure 2: Correlation Structure

From the graph we can see that marginal density of each of the variables are not look like Normal, especially for gasoline data. Graphs also tells that there may be some outliers.

Now we will plot 3d diagram of **Fuel Cost**, **Repair Cost** and **Capital Cost** ,grouped over **Fuel Type**

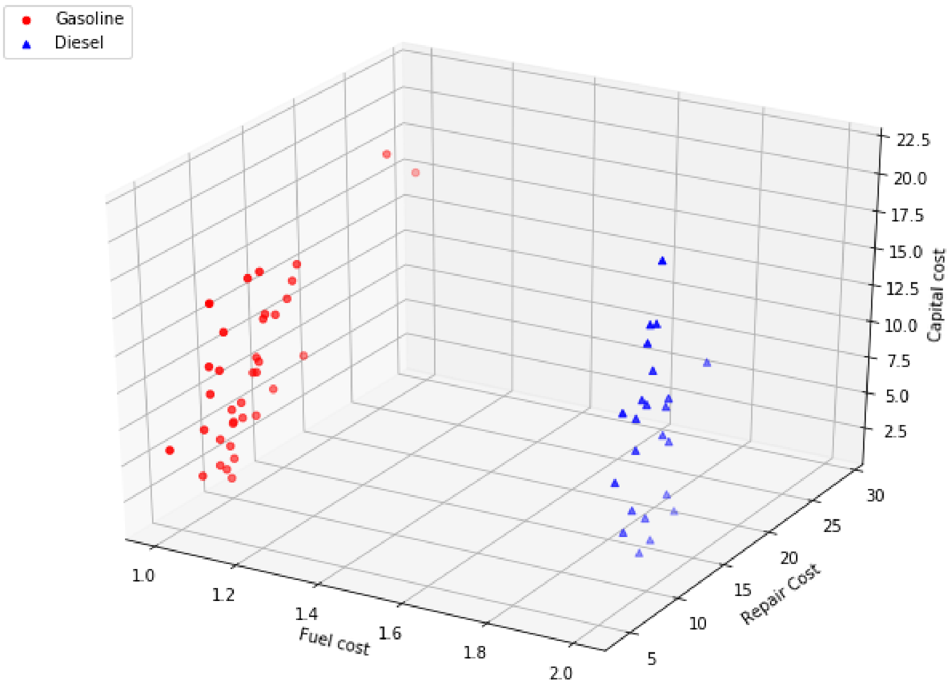

Figure 3: Scatter Plot: Red-Gasoline, Blue-Diesel

From the 3d plot we can see that there is clear distinction between two population. This implies that **Fuel type** influences the other costs. An discriminant analysis may reveal the properties of these populations. Also in both cases we can see some outliers.

3 Checking for Normality

In the later part of the analysis, we may require Principle Component Analysis(PCA), Discriminant Analysis, Confidence interval for Mean Vector and parellel analysis on the dataset (all these techniques are to reveal the differences between two populations). Although PCA does not require the normality of the dataset but the other procedures stated above require normality assumption on the dataset. So we will be checking the normality individually for diesel and gasoline using QQ Plot, Shapiro Wilk's test and Mardia's test.

3.1 Using QQ Plot

Let us draw the QQ plots of the individual variables for Diesel and Gasoline seperately.

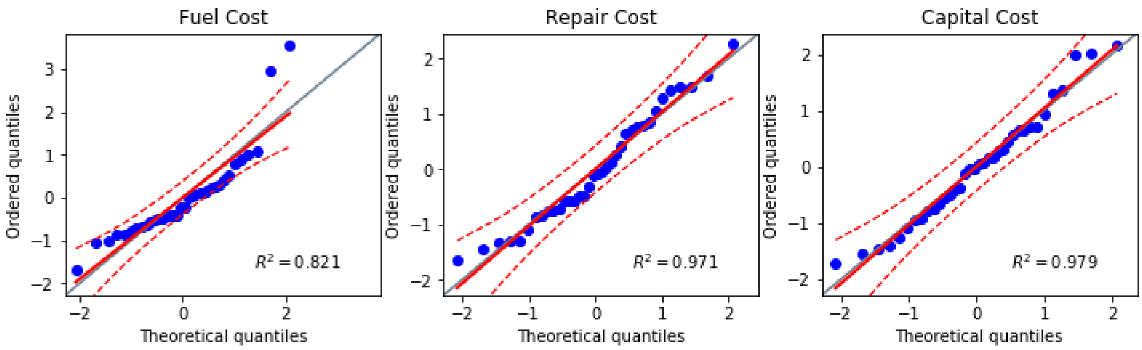

Figure 4: Q-Q Plot for Gasoline Data

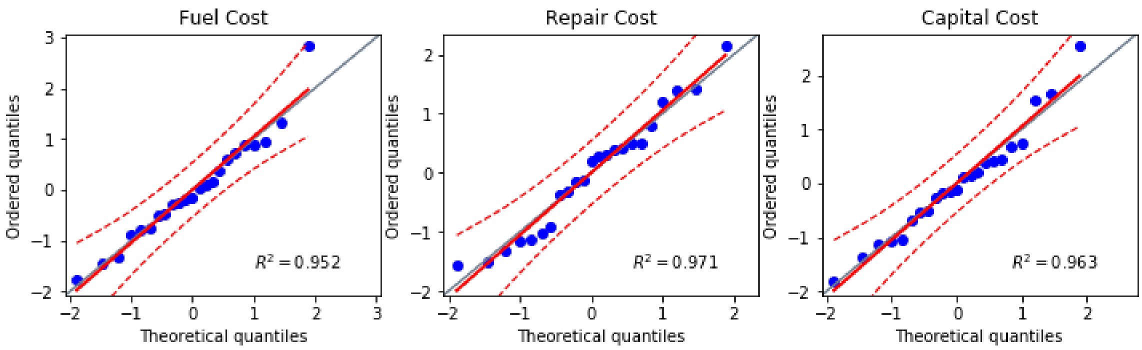

Figure 5: Q-Q Plot for Diesel Data

From the plots we can conclude that our Gasoline data is not normal but Diesel data seems normal.

3.2 Using Shapiro Wilk's Test

The Shapiro-Wilk test is a test of normality in a dataset. It was published in the year 1965 by Samuel Sanford Shapiro and Martin Wilk. It basically tests whether the sample observations have come from a normally distributed population or not i.e. it tests,
H_o: The sample arises from a normal population against $H_1 : H_o^c$
The suitable test statistic for the above testing procedure is given by,

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Where, $x_{(i)}$ is the i^{th} order statistic ,$\bar{x}$ is the sample mean. The coefficients $\mathbf{a} = (a_1, a_2, \ldots, a_n)^T isgivenby, (a_1, a_2, \ldots, a_n) = \frac{(m^T V^{-1})}{C}$ Where $C = (m^T V^{-2} m)^{1/2}$ and the vector $\mathbf{m} = (m_1, m_2, \ldots, m_n)^T$ is made of the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution; finally V is the covariance matrix of those normal order statistics.

The null-hypothesis of this test is that the population is normally distributed. Thus, on the one hand, if the p-value is less than the chosen alpha level, then the null hypothesis is rejected and there is evidence that the data tested are not normally distributed. On the other hand, if the p-value is greater than the chosen alpha level, then the null hypothesis that the data came from a normally distributed population can not be rejected (e.g., for an alpha level of 0.05, a data set with a p-value of less than 0.05 rejects the null hypothesis that the data are from a normally distributed population). Like most statistical significance tests, if the sample size is sufficiently large this test may detect even trivial departures from the null hypothesis.

The table for the Shapiro-Wilk's test statistic and the corresponding p-values is given below.

Fuel Used	**Data**	**Shapiro Wilk's Test Statistic**	**P-value**
Gasoline	Fuel Cost	0.83672	9.555e-05
	Repair Cost	0.96282	0.2623
	Capital Cost	0.97099	0.4532
	Multivariate Data	0.94245	0.009902
Diesel	Fuel Cost	0.96232	0.5117
	Repair Cost	0.96177	0.5
	Capital Cost	0.96872	0.6583
	Multivariate Data	0.96557	0.7312

Thus we conclude from here that the p-values suggest that null hypothesis of normality is accepted for diesel whereas it is rejected for gasoline.

3.3 Using Mardia's test

Another test for multivariate normality in a dataset was introduced by Prof K V Mardia. Basically it checks whether the multivariate skewness and kurtosis are consistent with a multivariate normal distribution. For a size n sample $X_1, X_2, ..., X_n$ with each being a $k \times 1$ vector, let us define,

$$\text{skew} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ (X_i - \bar{X})^T S^{-1} (X_j - \bar{X}) \right\}^3$$

$$\text{Kurt} = \frac{1}{n}\sum_{i=1}^{n}\left\{(X_i - \bar{X})^T S^{-1}(X_i - \bar{X})\right\}^2$$

where $S = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T$.Actually, we will use the sample versions of skew and kurt, which are obtained by multiplying skew as described above by $\left(\frac{n}{n-1}\right)^3$ and kurt by $\left(\frac{n}{n-1}\right)^2$. Under the null hypothesis we know that,

$$\frac{n}{6} skew \sim \chi^2_{\frac{k(k+1)(k+2)}{6}} \qquad \frac{nc}{6} Kurt \sim \chi^2_{\frac{k(k+1)(k+2)}{6}}$$

where $c = \frac{(n+1)(n+3)(k+1)}{n(n+1)(k+1)-6}$. And the results of Mardia Test is as follows,

Fuel Used	**Test**	**Mardia's Test Statistic**	**P-value**
Gasoline	Skewness	37.9072	3.93898e-05
	kurtosis	2.77972	0.00544058
Diesel	Skewness	7.292359	0.6975862
	Kutosis	-0.430022	0.667197

Thus the Mardia's test for multivariate normality suggests that we should accept the assumption of normality for diesel data whereas to reject it for gasoline data.

4 Making Data Normal

As discussed in the previous section Gasoline data is normally distributed but Diesel data is normally dis-tributed, we will perform Box-Cox transformation on Gasoline data to make the data Normally distributed and use the same transformation on Diesel data for sake of comparison.

4.1 Box-Cox Transformation

The value of λ that maximises the multivariate normal likelihood for gasoline data is $\lambda = (0.0644923, 0.6983734, 0.6457150)$ Since each column is cost column and transforming these columns with different λ doesn't make any sense as the resulting variable have no longer the same unit, it will create difficulty in comparison. So, we will consider the transformation as, $\lambda = (0.5, 0.5, 0.5)$ We will again perform Shapiro-Wilk test, Mardia test and QQ plot to check multivariate normality. Let us draw the QQ plots of the individual variables for the cars which use Gasoline and diesel as fuel.

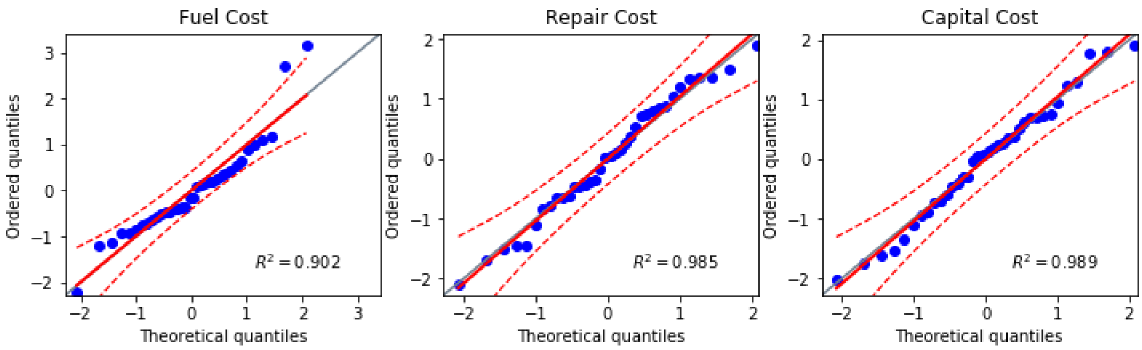

Figure 6: Q-Q Plot for Gasoline Data(After Transformation)

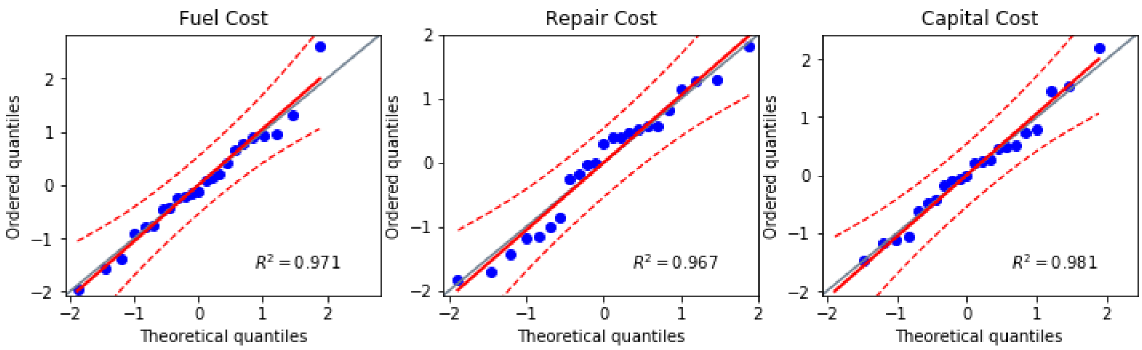

Figure 7: Q-Q Plot for Diesel Data(After transformation)

Below we give the results for Shapiro-Wilk's test and Mardi's tests.QQ plots and p-values of Shapiro-Wilk tests and Mardia tests suggest that we should accept the assumption of normality of both gasoline and Diesel data.

Fuel Used	**Data**	**Shapiro Wilk's Test Statistic**	**P-value**
Gasoline	Fuel Cost	0.91708	0.01035
	Repair Cost	0.97726	0.6525
	Capital Cost	0.98092	0.7761
	Multivariate Data	0.96421	0.2505
Diesel	Fuel Cost	0.97936	0.8950
	Repair Cost	0.95683	0.4024
	Capital Cost	0.98687	0.9851
	Multivariate Data	0.97142	0.8936

Table 1: Results for Shapiro Wilk's Test

Fuel Used	**Test**	**Mardia's Test Statistic**	**P-value**
Gasoline	Skewness	24.91345	0.00551
	kurtosis	1.487878	0.1367568
Diesel	Skewness	5.18073	0.878782
	Kutosis	-0.92976	0.352495

Table 2: Results for Mardia's Test

4.2 Detection and removal of outliers

Since we are aware of the fact that the data contains leverage points our primary interest is now to detect the leverage points and do some remedial steps. To detect the leverage points we will make use of Hat Matrix. Let us denote the coefficient matrix of the i^{th} population by,

$$Z_i = \begin{pmatrix} Y_{i1}^T \\ Y_{i2}^T \\ . \\ . \\ . \\ Y_{in_i}^T \end{pmatrix} \quad i = 1, 2$$

and Hat Matrix is defined as follows,$H_i = Z_i(Z_i^T Z_i)^{-1} Z_i^T \quad i = 1, 2$. We will consider the cut off of the leverage points as $\frac{2p}{n_i}$, with $n_1 = 36$ and $n_2 = 23$.The plot of hat values for two population is given below:

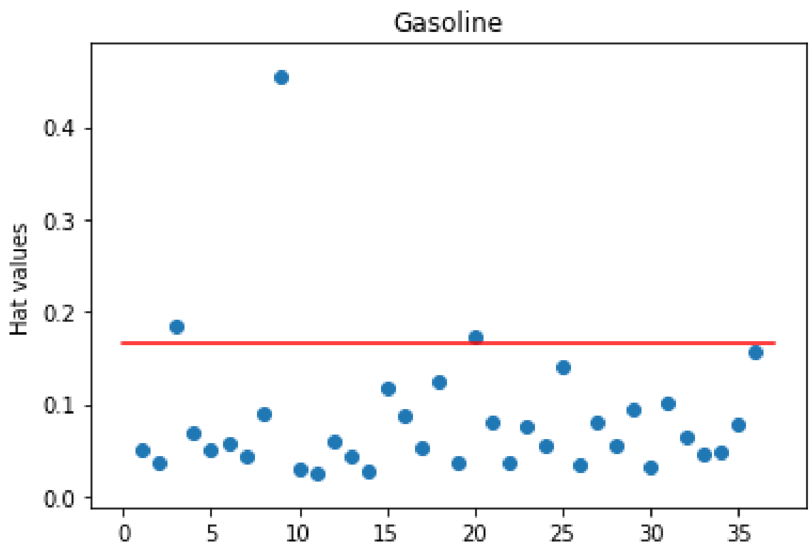

Figure 8: Hat values for Gasoline Data

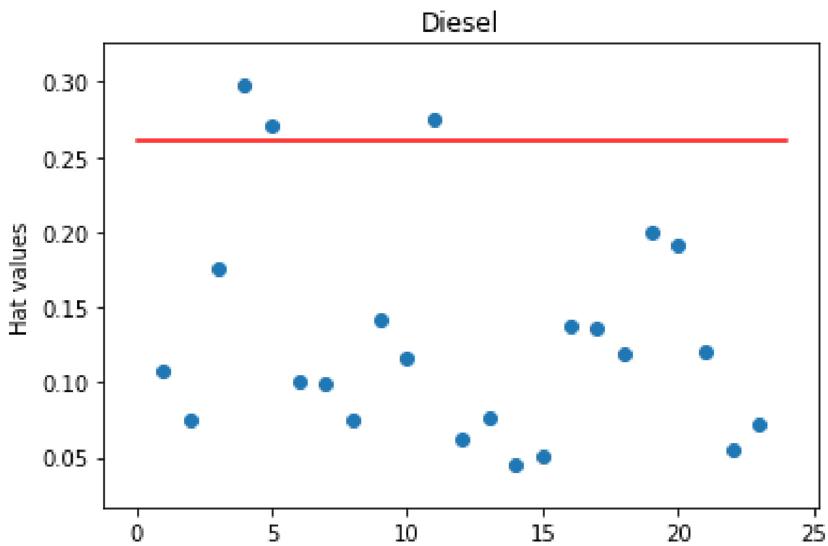

Figure 9: Hat values for Diesel Data

From the plot, we consider those points as outliers and replace 9^{th} observation from Gasoline data and $4^{th}, 5^{th}, 11^{th}$ (originally, $40^{th}, 41^{th}, 47^{th}$) observations from Diesel data by respective column means(after discarding those points).
After doing this lastly we again check from the normality of the data. The results obtained are given below in the form of tables.

Fuel Used	**Data**	**Shapiro Wilk's Test Statistic**	**P-value**
Gasoline	Fuel Cost	0.9544	0.1435
	Repair Cost	0.9821	0.8126
	Capital Cost	0.9793	0.7234
	Multivariate Data	0.9642	0.2505
Diesel	Fuel Cost	0.9641	0.5506
	Repair Cost	0.9606	0.4755
	Capital Cost	0.9690	0.6654
	Multivariate Data	0.97142	0.8936

Table 3: Results for Shapiro Wilk's Test

Fuel Used	**Test**	**Mardia's Test Statistic**	**P-value**
Gasoline	Skewness	9.18957	0.51421
	kurtosis	0.14516	0.0.88458
Diesel	Skewness	8.674548	0.563243
	Kutosis	-0.370027	0.71136

Table 4: Results for Mardia's Test

The p-values of Shapiro-Wilk tests and Mardia test accept the assumption of normality of both the gasoline and Diesel data. Hence we are now ready for doing further analysis.

5 Principal Component Analysis

Let us now move to Principal Component Analysis to have idea of the linear combination of the variables that explains the variability of the data. If we see the contributions of of the variables to the PC are different is two population then we can have idea of the relationship of the variables.

5.1 PCA for Gasoline Data

The sample covariance matrix and correlation matrix for Gasoline data set is given by,

$$S_1 = \begin{pmatrix} 1.1409654 & 0.8393898 & 0.4938155 \\ & 2.1955563 & 0.6527992 \\ & & 1.3383613 \end{pmatrix} \qquad R_1 = \begin{pmatrix} 1 & 0.5303410 & 0.3996150 \\ & 1 & 0.3808208 \\ & & 1 \end{pmatrix}$$

We see that the variability of the (transformed) variables Fuel cost, Repair cost, Capital cost are not same. So,we will work with correlation matrix.
The EValue-EVector pairs$(\hat{\lambda}, \hat{\mathbf{e}})$ of R_1 are

$$\left(1.8774604, \begin{pmatrix} 0.6020909 \\ 0.5949829 \\ 0.53243020 \end{pmatrix}\right), \left(0.6536039, \begin{pmatrix} -0.3328173 \\ -0.4191258 \\ 0.84472848 \end{pmatrix}\right), \left(0.4689357, \begin{pmatrix} -0.7257542 \\ 0.6858053 \\ 0.05433116 \end{pmatrix}\right)$$

The following table gives the contribution of the variables to the principal components,

	PC_1	PC_2	PC_3
Fuel Cost	0.6020909	-0.3328173	-0.72575422
Repair Cost	0.5949829	-0.4191258	0.68580532
Capital Cost	0.5324302	0.8447285	0.05433116

Table 5: Contribution of the variables to the principle components

Let us now see the Scree plot and proportion of explained variability in the following table,

	Eigenvalue	Percentage of variance
PC_1	1.8774604	65.38201
PC_2	0.6536039	21.38680
PC_3	0.4689357	15.63119

Table 6: Percentage of variance of Principle components

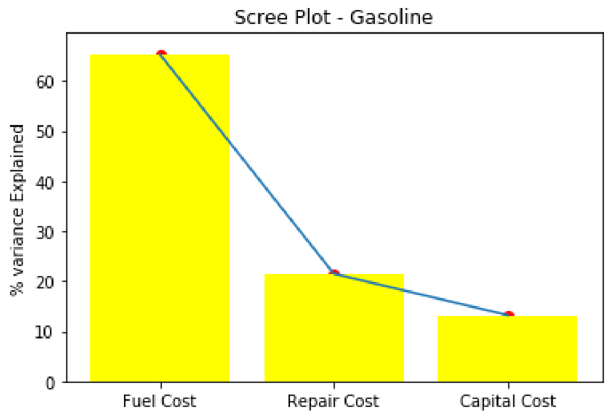

Figure 10: Scree Plot for Gasoline Data

Seeing the Scree plot and the table of percentage of variance of PC , we can drop the PC_3 since the first two PC has about 86.4% variability.The following table represents the correlation between the variables and principal components.

	PC_1	PC_2	PC_3
Fuel Cost	0.8249876	-0.2690688	-0.49698834
Repair Cost	0.8152483	-0.3388455	0.46963177
Capital Cost	0.7295383	0.6829273	0.03720537

Table 7: Correlation between the variables and the principle components

Correlation Circle Plot: It uses coordinates as the correlation between variables and the first two PC's having highest variance. Features with positive correlation will be grouped together. totally uncorrelated features are orthogonal to each other. Features with a negative correlation will be plotted on the oppsing quadrants of this plot.

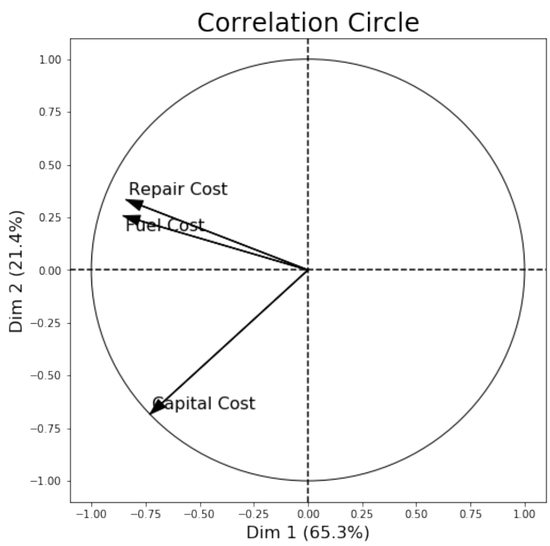

Figure 11: Correlation Circle for Gasoline Data

i.e.Arrows at 90 degree and 180 degree at each other shows zero correlation and negative correlation respectively.

- Fuel cost and Repair cost are highly correlated and these variables have more or less zero correlation with Capital cost.
- Further, note that the percentage values shown on the x and y axis denote how much of the variance in the original data set is explained by each proncipal componenet.

5.2 PCA for Diesel Data

The sample covariance matrix and correlation matrix for Gasoline data set is given by,

$$S_2 = \begin{pmatrix} 0.38752239 & 0.1641008 & 0.0.08090999 \\ & 2.1551809 & 0.0.50386795 \\ & & 1.22220283 \end{pmatrix} \qquad R_2 = \begin{pmatrix} 1 & 0.1795645 & 0.1175661 \\ & 1 & 0.3104583 \\ & & 1 \end{pmatrix}$$

We see that the variability of the (transformed) variables Fuel cost, Repair cost, Capital cost are not same. So,we will work with correlation matrix.
The EV-EV pairs$(\hat{\lambda}, \hat{\mathbf{e}})$ of R_2 are

$$\left(1.4169940, \begin{pmatrix} 0.4520790 \\ 0.6496940 \\ 0.0.6111648 \end{pmatrix}\right), \left(0.9012142, \begin{pmatrix} 0.8748847 \\ -0.1894780 \\ -0.4457296 \end{pmatrix}\right), \left(0.6817919, \begin{pmatrix} 0.1737856 \\ -0.7382037 \\ 0.6540663 \end{pmatrix}\right)$$

The following table gives the contribution of the variables to the principal components,

	PC_1	PC_2	PC_3
Fuel Cost	0.4520790	0.8748847	0.1737856
Repair Cost	0.6496940	-0.1894780	-0.7362037
Capital Cost	0.6111648	-0.4457296	0.6540663

Table 8: Contribution of the variables to the principle components

Let us now see the Scree plot and proportion of explained variability in the following table,

	Eigenvalue	Percentage of variance
PC_1	1.4169940	41.53313
PC_2	0.9012142	30.91047
PC_3	0.6817919	27.72640

Table 9: Percentage of variance of Principle components

Seeing the Scree plot and the table of percentage of variance of PC , it is difficult drop the PC_3 since the first two PC has about only 72% variability and there is no formation of elbow shape in scree plot.

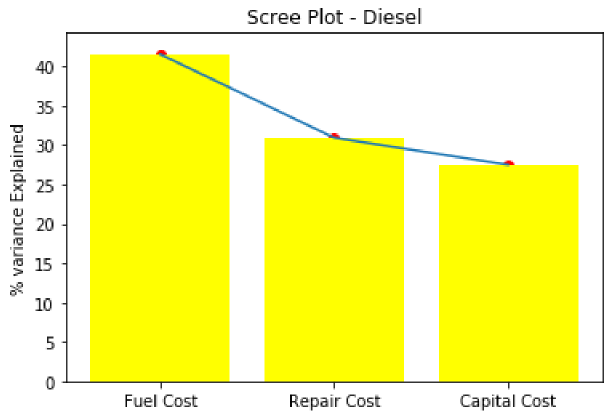

Figure 12: scree Plot for Diesel Data

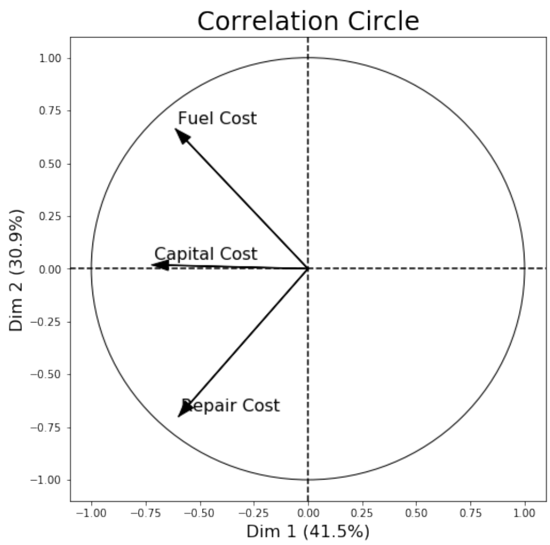

Figure 13: Correlation Circle for Diesel Data

- Fuel cost and Capital cost are moderately correlated and repair cost and Capital cost are also moderately correlated. Fuel Cost has more or less zero correlation with Repair cost.

5.3 Findings

- It is clear from the above analysis that PC_3 can be dropped from the Gasoline data (only 15% variance explained) but that is not possible for the Diesel data (27% variance explained). So, dimension reduction for Diesel data is not possible this way.
- For car using Gasoline as fuel, repair cost has high correlation with fuel cost whereas for car using Diesel as fuel, Repair cost has moderate correlation with Capital cost as well as Fuel cost has moderate correlation with Capital cost.
- The Principal Components for these populations is not similar, so we have to

analysis for each population separately.

6 Confidence Interval for Mean

Define

$$\mu_i := \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \\ \mu_{i3} \end{pmatrix}$$

denoting the mean vector, containing the 3 (transformed) variables, for the i^{th} class of fuel (1 and 2 denoting Gasoline and Diesel respectively). Having already established normality for our data, we would now like to construct appropriate confidence regions for both μ_1 and μ_2.

6.1 Gasoline Mean Vector

We use three different methods to find the confidence region for the mean vector of interest, μ_1, which are given below. We use 95% confidence for all purposes.

- **Individual Confidence Intervals**: $[4.406381, 5.129207] \times [2.926083, 3.928781] \times [3.752178, 4.535039]$
- **Bonferroni's Confidence Interval**: $[4.320138, 5.215449] \times [2.806448, 4.048415] \times [3.658773, 4.628444]$
- **Simultaneous Confidence Interval**: $[4.227800, 5.307788] \times [2.678357, 4.176506] \times [3.558765, 4.728452]$

In the figure below, we have plotted the three confidence regions alongside the confidence ellipsoid.

From the plot, it is evident that the Simultaneous Confidence Interval provides the largest confidence region and the Individual Confidence Intervals provide the largest. It is also important to note that the Simultaneous Confidence Interval is actually the projection of the confidence ellipsoid on the respective axes.

6.2 Diesel Mean Vector

In a similar manner we find confidence regions for the mean vector, μ_1, (using 95% confidence) and plot them along with the confidence ellipsoid. Our observations are given below.

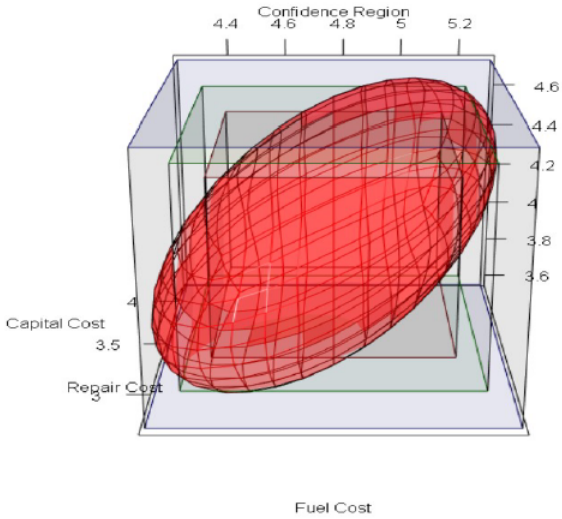

Figure 14: 95% Confidence Ellipsoid for Gasoline

- **Individual Confidence Intervals**: $[4.022195, 4.560585] \times [3.680056, 4.949723] \times [5.792567, 6.748703]$
- **Bonferroni's Confidence Interval**: $[3.955043, 4.627737] \times [3.521693, 5.108086] \times [5.673310, 6.867960]$
- **Simultaneous Confidence Interval**: $[3.876331, 4.706449] \times [3.336069, 5.293710] \times [5.533524, 7.007746]$

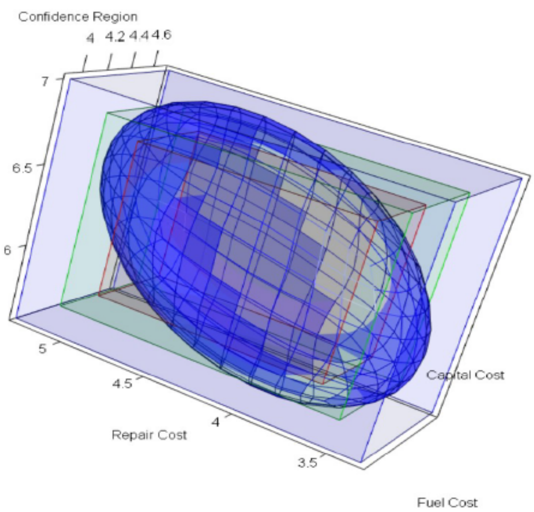

Figure 15: 95% Confidence Ellipsoid for Diesel

As in the previous case, we observe that the largest and smallest confidence regions are given by the Simultaneous and Individual Confidence Intervals respectively.

6.3 Comparisons

To compare between the two types of Fuel, we plot the two sets of confidence regions on the same graph together with the two confidence ellipsoids.

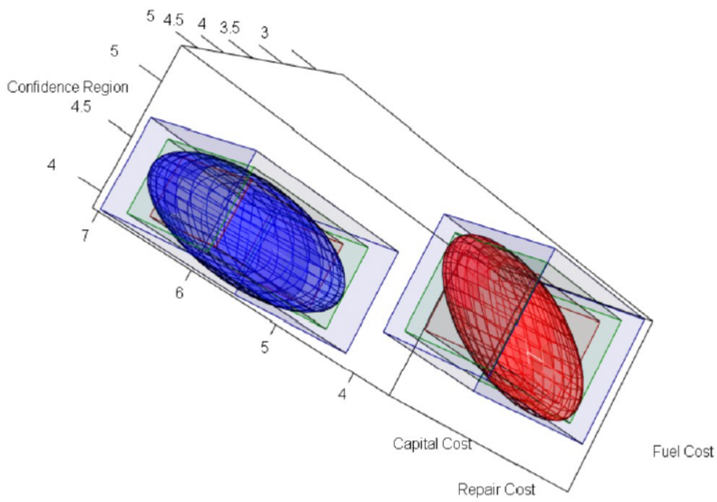

Figure 16: 95% Confidence Ellipsoids

From the figure, we can easily conclude that the mean vectors of the Gasoline and Diesel data are significantly different. We further note that among the three variables, the Capital Cost has the largest difference (much higher for Diesel than Gasoline). Based on these observations, we conclude that it is meaningful to perform discriminant analysis on this dataset.

7 Profile Analysis

Our primary task is to test for the equality of the covariance matrices for the two groups using Gasoline and Diesel, denoted by Σ_1 and Σ_2, respectively. The testing problem is given by

$$H_0 : \Sigma_1 = \Sigma_2 \text{ ag. } H_1 : \Sigma_1 \neq \Sigma_2$$

Hence, we use the Bartlett's Test, for which the p-value turns out to be $0.2673 (> 0.05)$. This leads to us accepting H_0 at 5% level of significance.
Henceforth, the following questions arise:

- Are the profiles parallel?
- If yes, then, are they coincient?
- If yes, then, are they level?

Thus, to test if the profiles are parallel, we write the problem as

$$H_0 : \mu_{1i} - \mu_{1(i-1)} = \mu_{2i} - \mu_{2(i-1)},\ i = 2, 3 \ \text{ ag. } H_1 : \text{not } H_0$$

The test statistic that we use is

$$T^2 = (\bar{Y_1} - \bar{Y_2})' C' \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) C S_{pooled} C' \right]^{-1} C(\bar{Y_1} - \bar{Y_2})$$

where

$$C = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}$$

We reject H_0 at $\alpha\%$ level of significance if

$$T^2 \geq \frac{(n_1 + n_2 - 1)(p - 1)}{n_1 + n_2 - p} F_{(p-1),(n_1+n_2-p);\alpha}$$

Based on the data, we find

$$T^2_{obs} = 65.998 > \frac{(59 - 1)(3 - 1)}{59 - 3} F_{(3-1),(59-3);0.05} (= 6.436)$$

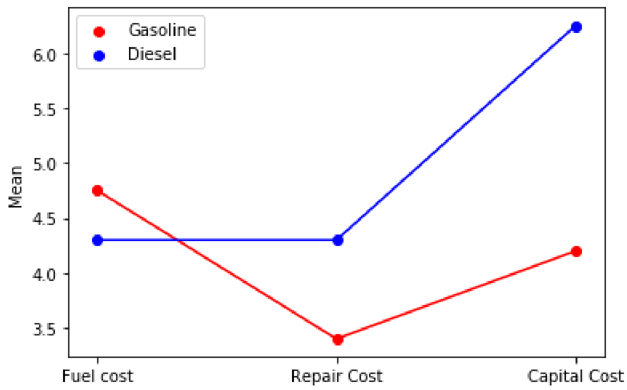

Figure 17: Profile Plot

Thus, we can conclude that the profiles are not parallel and hence, neither are they level. Therefore, the mean vectors and class-specific and the two populations are distinct and so, we can construct appropriate Discrimination Rules for the two groups.

8 Discriminant Analysis

Owing to normality of the data and equality of the covariance matrices for Gasoline and Diesel, we can use Linear Discriminant Analysis and also, Quadratic Discriminant Analysis for the purpose of classification. For this, we denote the populations of transporters using Gasoline and Diesel by π_1 and π_2.
While applying discriminant analysis to our dataset, the implementation is done in two different methods - firstly, the whole transformed dataset is considered as both the training and the validation(test) set, and secondly, we partition the dataset into training and test sets . For the latter, approximately 75% part of the data is taken as the training set and the remaining 25% as the validation set. Hence, the size of the training set has been obtained to be 44 and the validation set size is, thus, 15.

8.1 Linear Discriminant Analysis

Define

$$S_{pooled} := \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

$$d_i(x) := \bar{x}_i{}' S_{pooled}^{-1} x - \frac{1}{2} \bar{x}_i{}' S_{pooled}^{-1} \bar{x}_i + ln(p_i), \; i = 1, 2$$

Hence, we assign x_{test} to π_i if

$$d_i(x_{test}) = \max\{d_1(x_{test}), d_2(x_{test})\}$$

This method is designed to maximise the posterior probability $\mathbb{P}(\pi_i | X = x)$ under the assumptions of normality and equal covariance matrices.

8.1.1 Using Entire Data

Here, we construct the classification rule based on the entire dataset and proceed to use the rule on each point in the dataset.
In the figures below, we plot the ordered pairs of the two posterior probabilities for each data point. The points are coloured differently in the three graphs according to their original class, predicted class and correctness of classification.

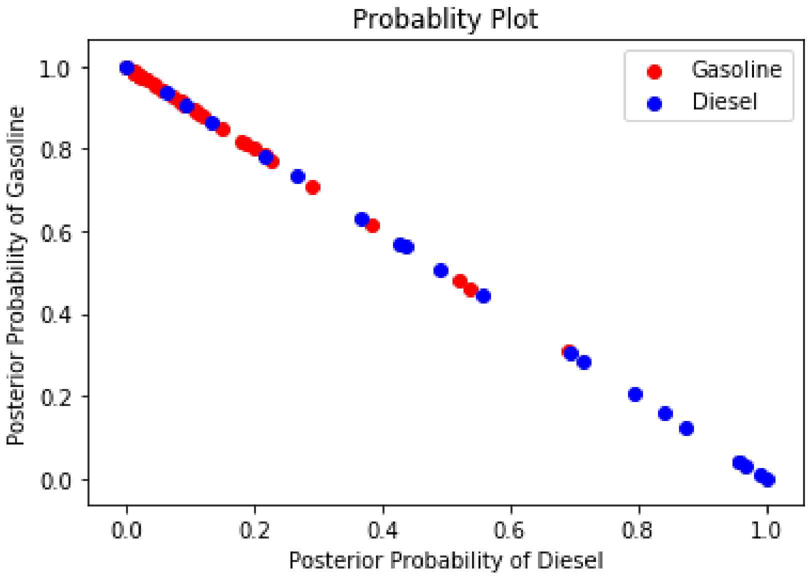

Figure 18: Posterior Prediction Probability

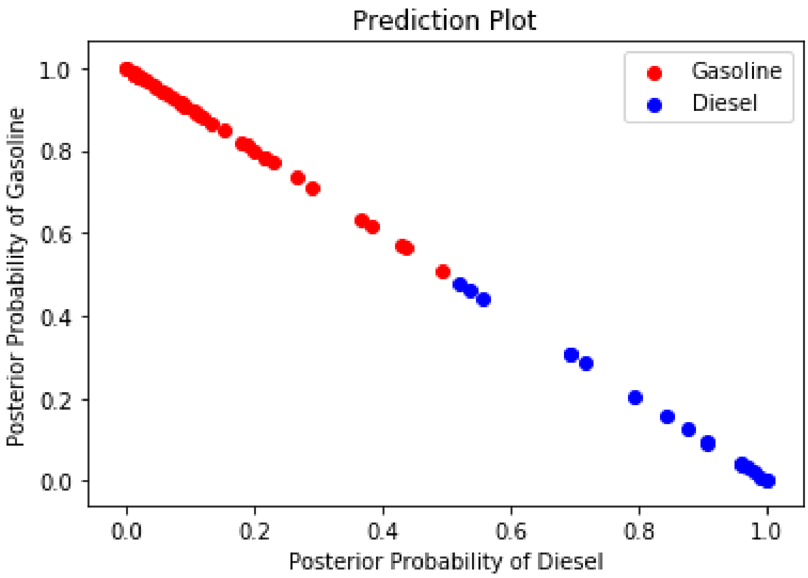

Figure 19: Prediction

The confusion matrix obtained from our observations is

Predicted Fuel Type	Original Fuel Type Gasoline	Diesel
Gasoline	33	6
Diesel	3	17

Hence, the observed error rate is

$$\left(\frac{6+3}{59}\times 100\right)\% \approx 15.25\%$$

8.1.2 Using Training-Validation Split

However, to obtain a more realistic estimate of the probability of misclassification, we split the data into training and test sets as discussed earlier. After constructing the rule using the training data, we apply it on the validation set to obtain the following confusion matrix.

Predicted Fuel Type	Original Fuel Type Gasoline	Diesel
Gasoline	6	2
Diesel	1	6

Hence, the observed rate, here, is

$$\left(\frac{3}{15}\times 100\right)\% = 20\%$$

8.2 Quadratic Discriminant Analysis

Define

$$d_i^Q(x) = -\frac{1}{2}ln|S_i| - \frac{1}{2}(x-\bar{x_i})^T S_i^{-1}(x-\bar{x_i}) + ln(p_i), i = 1,2$$

assign x_{test} to π_i if

$$d_i^Q(x_{test}) = \max\{d_1^Q(x_{test}), d_2^Q(x_{test})\}$$

This method is also designed to maximise the posterior probability $\mathbb{P}(\pi_i|X = x)$ under the assumptions of normality and equal covariance matrices.

8.2.1 Using Entire Data

Here, we construct the classification rule based on the entire dataset and proceed to use the rule on each point in the dataset.

In the given figures, we plot the ordered pairs of the two posterior probabilities for each data point. The points are coloured differently in the three graphs according to their original class, predicted class and correctness of classification.

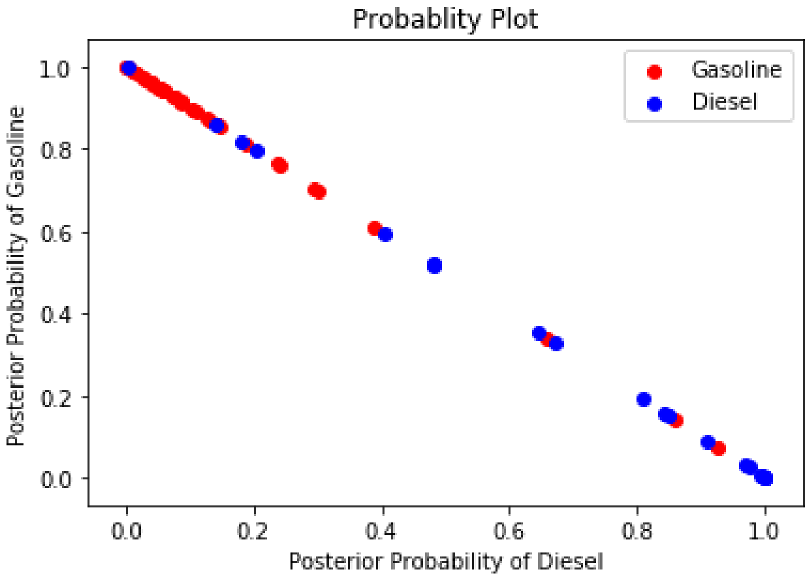

Figure 20: Posterior Prediction Probability

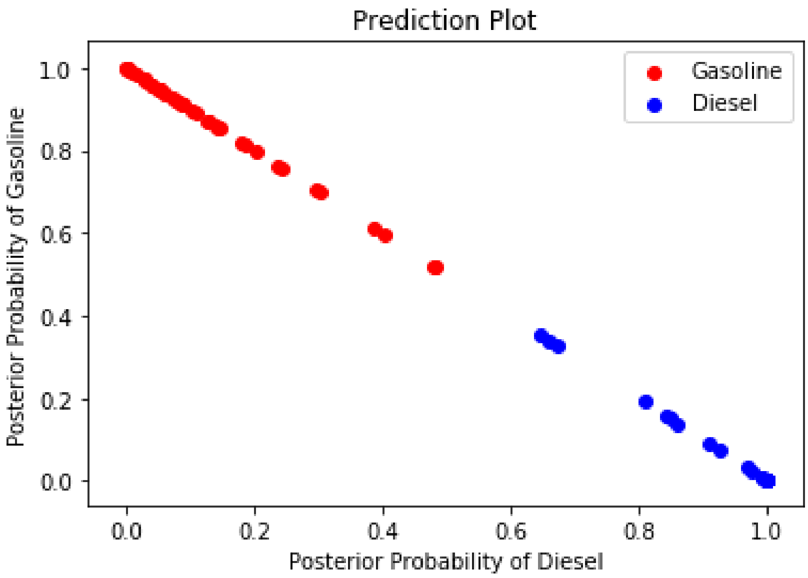

Figure 21: Prediction

The confusion matrix obtained from our observations is

Predicted Fuel Type	Original Fuel Type Gasoline	Diesel
Gasoline	33	5
Diesel	3	18

Hence,the observed error rate is

$$\left(\frac{5+3}{59}\times 100\right)\% \approx 13.56\%$$

8.2.2 Using Training-Validation Split

After constructing the discriminant rule using the training set,we apply it on validation set to predict the fuel used for the elements.The confusion matrix is obtained as follows,

Predicted Fuel Type	Original Fuel Type Gasoline	Diesel
Gasoline	7	2
Diesel	0	6

Hence, the observed error rate is

$$\left(\frac{2}{15}\times 100\right)\% \approx 13.33\%$$

8.3 Comparison

Comparing the performances of Linear Discriminant rule and Quadratic Discriminant rule on the validation set, we found Quadratic Discriminant rule better.
Now, we will look into other classification methods like logistic regression & k-nearest neighbors and compare their performance with Linear Discriminant Analysis and Quadratic Discriminant Analysis.

8.3.1 Logistic Regression

Here,we fit logistic regression on the training dataset with the fuel type (Y) as the response variable and the cost (X) as the explanatory variables and then predict the fuel type for each data point of the validation set.Here, Y takes two values 1 and 2.
Define,

$$p = P[Y_i - 1 = k] = \frac{exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i})}{1 + exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i})}$$

$i = 1, 2, ..., 44$ $k = 0, 1$.
where $\beta_0, \beta_1, \beta_2, \beta_3$ are the unknown parameters. After fitting the model we get the

estimates of the parameters as $\hat{\beta_0} = -8.9614, \hat{\beta_1} = -1.7093, \hat{\beta_2} = 1.0697, \hat{\beta_3} = 2.2689$ In the validation set,out of 15 obsevations, Gasoline has been used for 7 observations and Diesel has been used for 8 obsevations.So, the prior probabilities of both the populations are almost same in the validation set. The prediction rule used for the validation set is as follows:

- Compute the value of $\frac{exp(-8.9614-1.7093X_{1j}+1.0697X_{2j}+2.2689X_{3j})}{1+exp(-8.9614-1.7093X_{1j}+1.0697X_{2j}+2.2689X_{3j})}; j = 1, 2, .., 15$ for each j in the validation set.
- If $p \leq 0.5$ for some j,then $\hat{Y_j} - 1 = 0$ i.e $\hat{Y_j} = 1$
- Else take $\hat{Y_j} - 1 = 1$ i.e $\hat{Y_j} = 2$.
- Now, compare Y_j with $\hat{Y_j}$ in the validation set.

The confusion matrix obtained from our observation is as follows

Predicted Fuel Type	Original Gasoline	Fuel Type Diesel
Gasoline	6	2
Diesel	1	6

Hence,the observed error rate for the validation set is

$$\left(\frac{3}{15} \times 100\right)\% = 20\%$$

8.3.2 K-Nearest Neighbors

In this method, each data point of the validation set is taken and from the training dataset, k data points are chosen which are nearest to this new data point (nearest in terms of some distance measure like Euclidean Distance etc.) & for the new data point, its fuel type will be the one which is most common among the k chosen data points from the training set.Here k is a prefixed integer within 1 & 10.We firstly plotted the misclassification errors against different choices of k to see which k will be the most suitable for our data and we get the following plot.

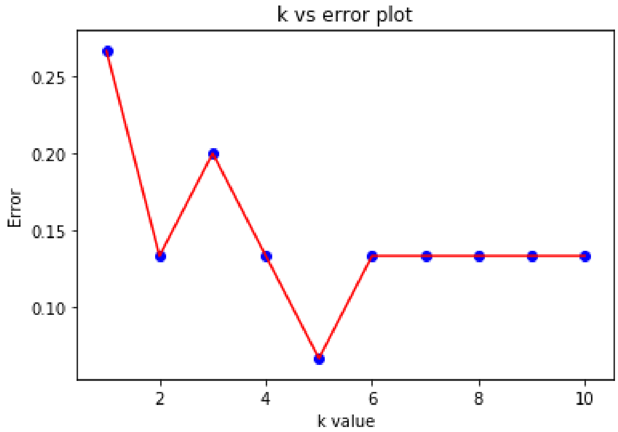

Figure 22: Misclassification Error for different choice of k

From the graph, it is observed that for k=5, we have the lowest error.So, now we will predict the fuel type of each data points in the validation set using 5-nearest neighbours. The confusion matrix obtained from our observation is as follows

Predicted Fuel Type	Original Gasoline	Fuel Type Diesel
Gasoline	7	1
Diesel	0	7

Hence,the observed error rate for the validation set is

$$\left(\frac{1}{15} \times 100\right)\% \approx 6.67\%$$

8.3.3 Comparison of all the Methods

Based on the training and validation dataset obtained from the transformed dataset we obtained the following table:

Method of classification	Error rate
Linear Discriminant Analysis	20%
Quadratic Discriminant Analysis	13.33%
Logistic Regression	20%
5-Nearest Neighbours	6.67%

So, for the validation dataset we obtained 5-nearest neighbour performs the best in predicting the fuel types of the data points in the validation set.And linear discriminant analysis and logisitc regression performs identically in predicting the fuel types and have the highest misclassification error among all the methods.

8.3.4 Lachenbruch's 'Holdout' Procedure

Though using a training-validation split of the dataset is useful in providing more reliable estimate of actual error rate, but have a drawback of not having large enough dataset consequently less number of data points are used for constructing the discriminant function. To overcome this we use Lachenbruch's 'Holdout' Procedure described in the following steps:-

- Holdout the first observation of the population using Gasoline.
- Use the remaining 58 observation of the whole transformed dataset as the training set.
- Obtain a classification rule based on the training set.
- Make prediction for the observation which was held out.
- Repeat the above steps for other 35 observations of the Gasoline population.
- Calculate the number of hold-out observations in the Gasoline population which are wrongly classified and denote it as $n_{1M}^{(H)}$.
- Repeat the above steps for the Diesel population and calculate the number of hold-out observations in the diesel population which are misclassified denoted as $n_{2M}^{(H)}$.
- Estimates of $P(2|1)$ & $P(1|2)$ is obtained as $P(\hat{2}|1) = \frac{n_{1M}^{(H)}}{36}$ & $P(\hat{1}|2) = \frac{n_{2M}^{(H)}}{59}$.
- The estimate of expected actual error rate E(AER) is obtained as $E(\hat{AER}) = \frac{n_{1M}^{(H)}+n_{2M}^{(H)}}{59}$

8.3.5 Result using LDA

The confusion matrix (for Linear Discriminant rule) generated using Holdout Procedure is as follows,

Predicted Fuel Type	Original Fuel Type Gasoline	Diesel
Gasoline	32	6
Diesel	4	17

For Linear Discriminant Analysis,using Lachenbruch's 'Holdout' procedure, we get the following estimates-

$$n_{1M}^{(H)} = 4, n_{2M}^{(H)} = 6, P(\hat{2}|1) = \frac{4}{36} \approx 0.111, P(\hat{1}|2) = \frac{6}{23} \approx 0.261,$$

and $E(\hat{AER}) = \frac{6+4}{59} \approx 0.169$ i.e.the estimate of expected actual error rate is approximately 16.95%.

8.3.6 Result using QDA

The confusion matrix (for Quadratic Discriminant rule) generated using Holdout Procedure is as follows,

Predicted Fuel Type	Original Gasoline	Fuel Type Diesel
Gasoline	32	7
Diesel	4	16

For Quadratic Discriminant Analysis,using Lachenbruch's 'Holdout' procedure, we get the following estimates-

$$n_{1M}^{(H)} = 4, n_{2M}^{(H)} = 7, P(\hat{2}|1) = \frac{4}{36} \approx 0.111, P(\hat{1}|2) = \frac{7}{23} \approx 0.304,$$

and $E(\hat{AER}) = \frac{7+4}{59} \approx 0.186$ i.e.the estimate of expected actual error rate is approximately 18.64%.Clearly, Linear Discriminant rule is performing slightly better.

9 Conclusion

The questions, we have raised in the beginning, are nicely answered through out the project.

- At least for Gasoline data we can construct 2 linear combinations of the variables to explain the variability
- We have constructed three types of confidence region for mean vector.
- Efficiently we have constructed rule for discriminating the fuel based on observed costs.

10 Acknowledgement

We would like to thank our professor Dr. Minerva Mukhopadhyay for giving us this opportunity. Her classnotes helped us a lot in doing the project and through this project we have learned the real life application of multivariate data.

11 References

Necessary codes and data file can be found in the following link :
`https://github.com/souvik2019/Multivariate-Project`