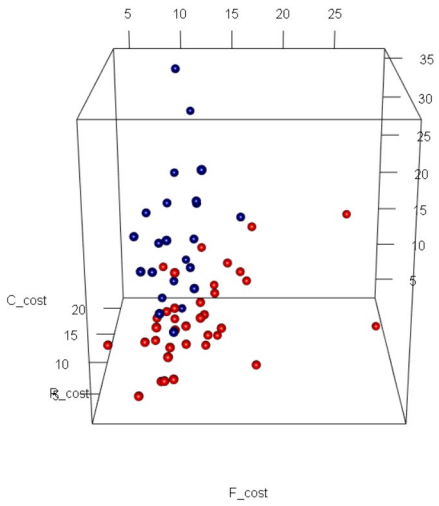
MULTIVARIATE ASSIGNMENT

Milk Transportation Data

Ayoushman Bhattacharya
Himadri Sekhar Manna
Ayan Paul



June 15, 2020

Contents

1. Introduction

We are given data on transporting milk from farms to dairy plants.Data is provided on Fuel Used(Gasoline or Diesel) and various costs- Fuel Cost(X_1), Repair Cost(X_2), and Capital Cost(X_3) , measured on cents per mile basis.
Following are few rows from our dataset.

Fuel Type	Fuel Cost	Repair Cost	Capital Cost
1	16.44	12.43	11.23
1	7.19	2.7	3.92
1	9.92	1.35	9.75
...	...	...	...
2	11.88	12.18	21.2
2	12.03	9.22	23.09

For transporters using Gasoline ,we have $n_1 = 36$ observations on $\underset{\sim}{X}^T = (X_1, X_2, X_3)$ and for transported using Diesel ,we have $n_2 = 23$ observations on the same observed variables.

Our goal is to seek answers of the following questions,

- *Can we consider 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 answer the above questions, we started with checking

1. if the the class-specific data on the costs are multivariate normal and if not, then if we can consider some transformation,
2. if the covariance matrices can be considered to equal,
3. if yes, then if we can consider the mean vectors of different classes(based on Fuel used) to be same or not.

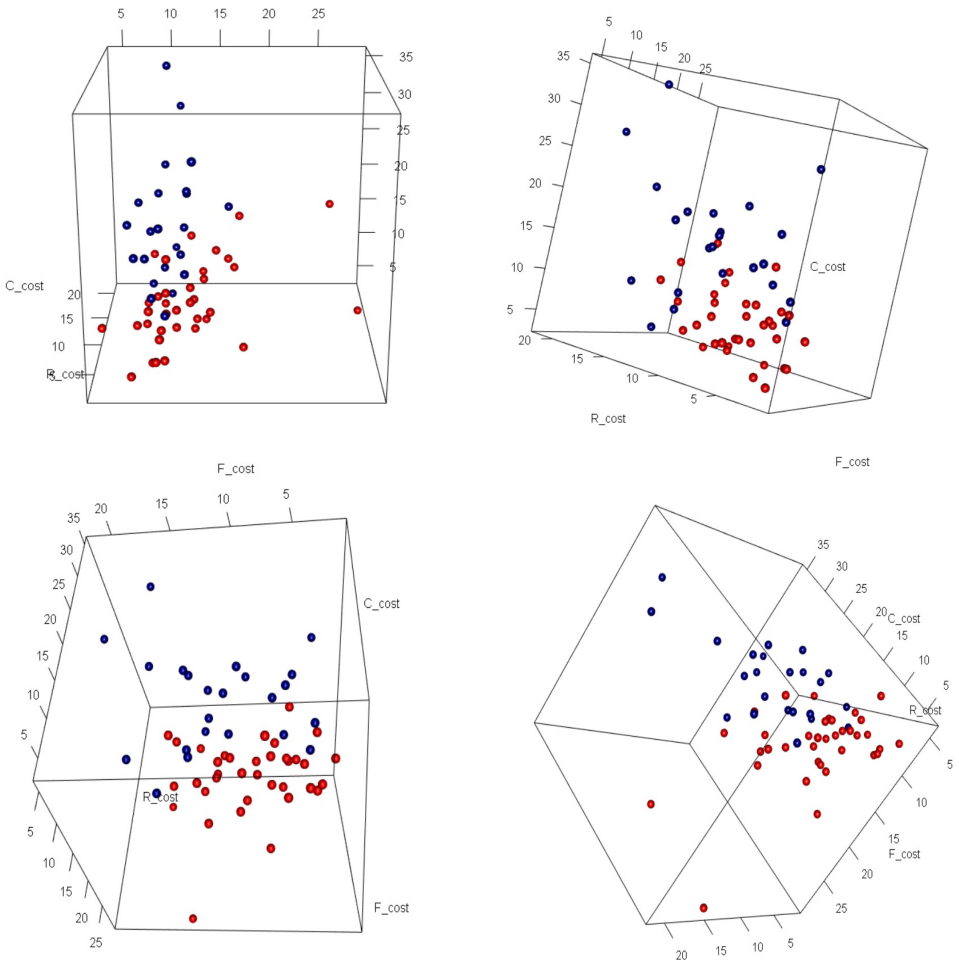
Please note that, this report contains 3D plots. To have better view to those plots please check the html file attached with the report. Download the html file and open the file with any web browser to access it. File link here

2. Data Visualisation

First we look at the scatterplot matrix of the data for two populations viz. Gasoline and Diesel.

Clearly, the plot shows that the marginal normality two populations may be doubted, especially the Gasoline one. Also, the plots show that there are possibility of presence of outliers.

Now, let us plot the three dimensional data and see if we can say anything about the data. While considering the plot we will treat the 'Fuel Type' as grouping variable. So 'Fuel cost' , 'Repair cost' and 'Capital cost' will be plotted as two groups Gasoline and Diesel.



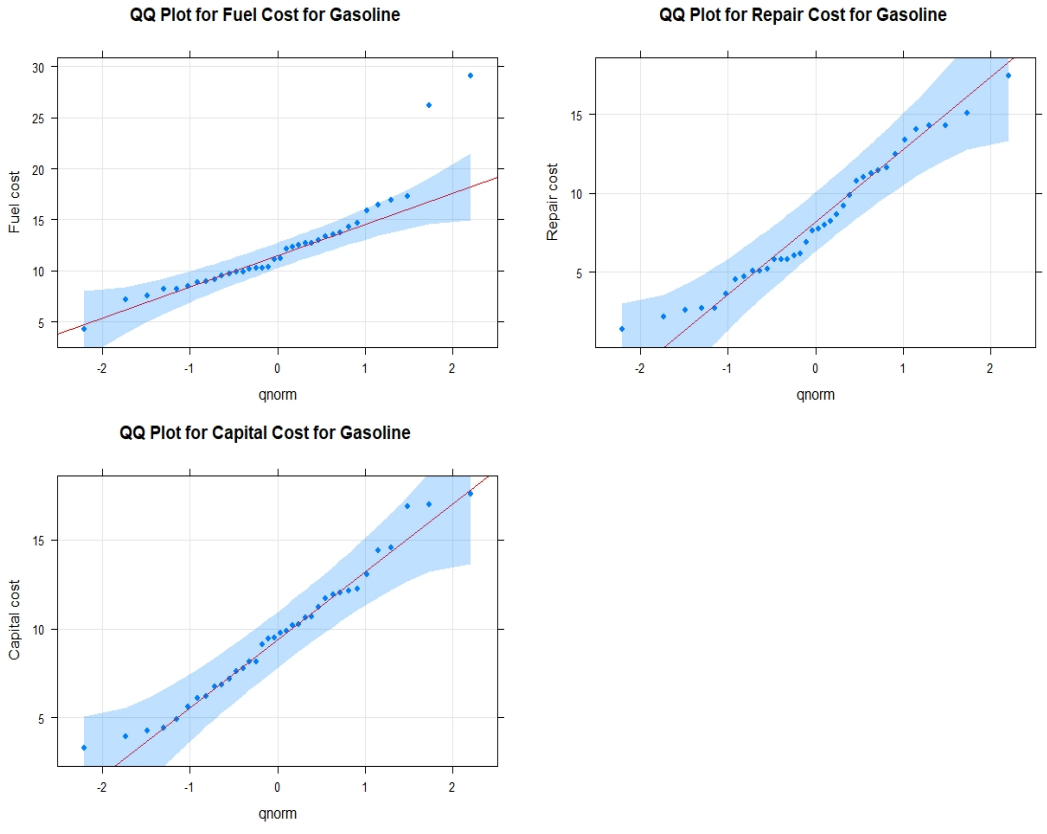
From the 3D scatter plots we see that the points are forming two different groups. At first look one can say that difference in Fuel type can affect other costs, an extensive study on discriminant analysis will reveal the properties of those population. Another thing note there are some points for of the groups which are place at distant from the bulk (centre) of the data, we can suspect these points as outliers.

3. Checking for Normality

In the later part of the analysis we will be doing Principal Component Analysis, Confidence interval for mean vector, Parallel Analysis, Discriminant Analysis. Although Principal Component Analysis does not require the data to be Normal but other procedures requires the normality of the data. So, at first we will check the normality of the data. For checking this we will perform Shapiro-Wilk test, Mardia test and QQ plot.

3.1 Normality of Gasoline

Let us draw the QQ plot of the individual variables for the car which uses Gasoline as fuel.



The results of Shapiro-Wilk test is given by,

Data	Shapiro -Wilk Test Statistic	p-value
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

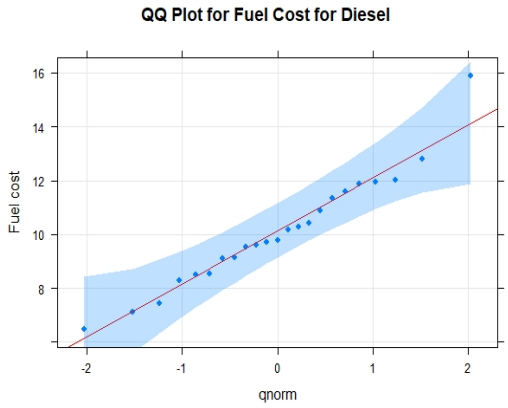
And the results of Mardia test is as follows:

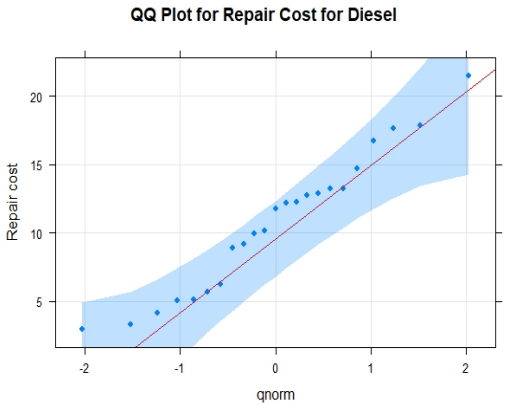
Test	Mardia Test Statistic	p-value
Skewness	37.9072	3.93898e-05
Kurtosis	2.77972	0.00544058

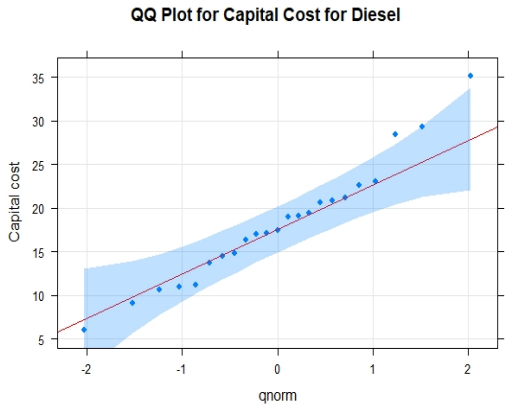
QQ plots and p-values of Shapiro-Wilk tests and Mardia test reject the assumption of normality of Gasoline data.

3.2 Normality of Diesel

Let us draw the QQ plot of the individual variables for the car which uses Diesel as fuel.







The results of Shapiro-Wilk test is given by,

Data	Shapiro -Wilk Test Statistic	p-value
Fuel Cost	0.96232	0.5117
Repair Cost	0.96177	0.5
Capital Cost	0.96872	0.6583
Multivariate Data	0.96557	0.7312

And the results of Mardia test is as follows:

Test	Mardia Test Statistic	p-value
Skewness	7.292359	0.6975862
Kurtosis	-0.430022	0.6671797

QQ plots and p-values of Shapiro-Wilk tests and Mardia test accept the assumption of normality of Diesel data.

4. Making Data Normal

As discussed in the previous section Gasoline data is normally distributed but Diesel data is normally distributed, 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 $\underset{\sim}{\lambda}$ that maximises the multivariate normal likelihood for gasoline data is

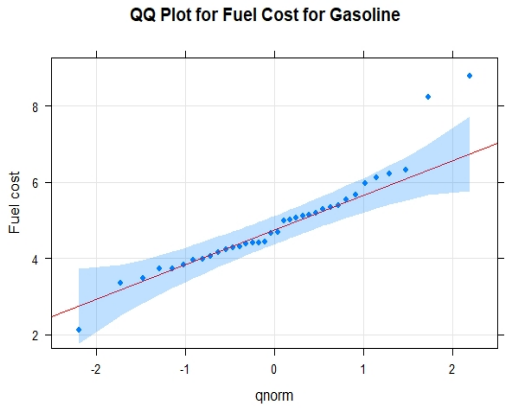
$$\underset{\sim}{\lambda} = (0.0644923, 0.6983734, 0.6457150)$$

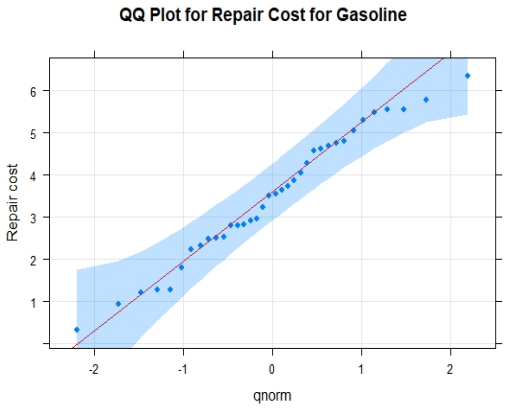
Since each column is cost column and transforming these columns with different λ doesn't make any sense as the resulting variable have no longer the same unit, it will create difficulty in comparison. So, we will consider the transformation as,

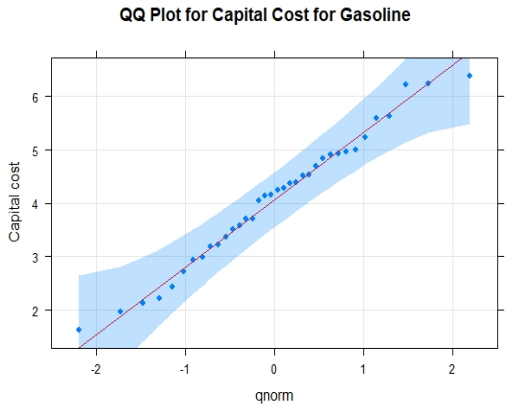
$$\underset{\sim}{\lambda} = (0.5, 0.5, 0.5)$$

We will again perform Shapiro-Wilk test, Mardia test and QQ plot to check Multivariate normality.

Let us draw the QQ plot of the individual variables for the car which uses Gasoline as fuel.







The results of Shapiro-Wilk test is given by,

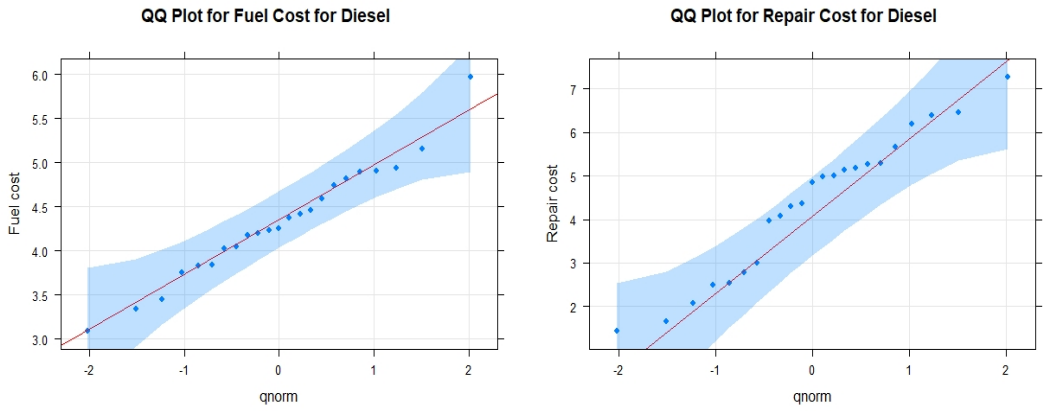
Data	Shapiro -Wilk Test Statistic	p-value
Fuel Cost	0.91708	0.01035
Repair Cost	0.97726	0.6525
Capital Cost	0.98092	0.7761
Multivariate Data	0.96421	0.2505

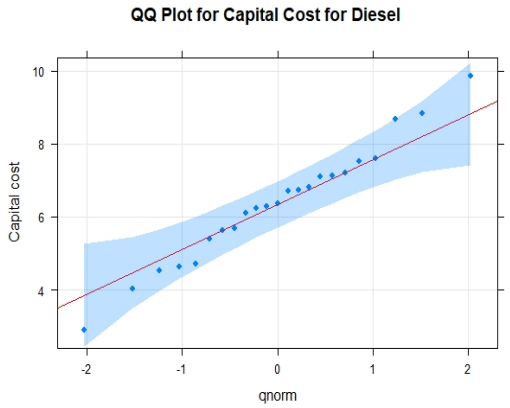
And the results of Mardia test is as follows:

Test	Mardia Test Statistic	p-value
Skewness	24.91345	0.00551199
Kurtosis	1.4879776	0.1367568

QQ plots and p-values of Shapiro-Wilk tests and Mardia test still **reject** the assumption of normality of Gasoline data.

Let us draw the QQ plot of the individual variables for the car which uses Diesel as fuel.





The results of Shapiro-Wilk test is given by,

Data	Shapiro -Wilk Test Statistic	p-value
Fuel Cost	0.97936	0.895
Repair Cost	0.95683	0.4024
Capital Cost	0.98687	0.9851
Multivariate Data	0.97142	0.8936

And the results of Mardia test is as follows:

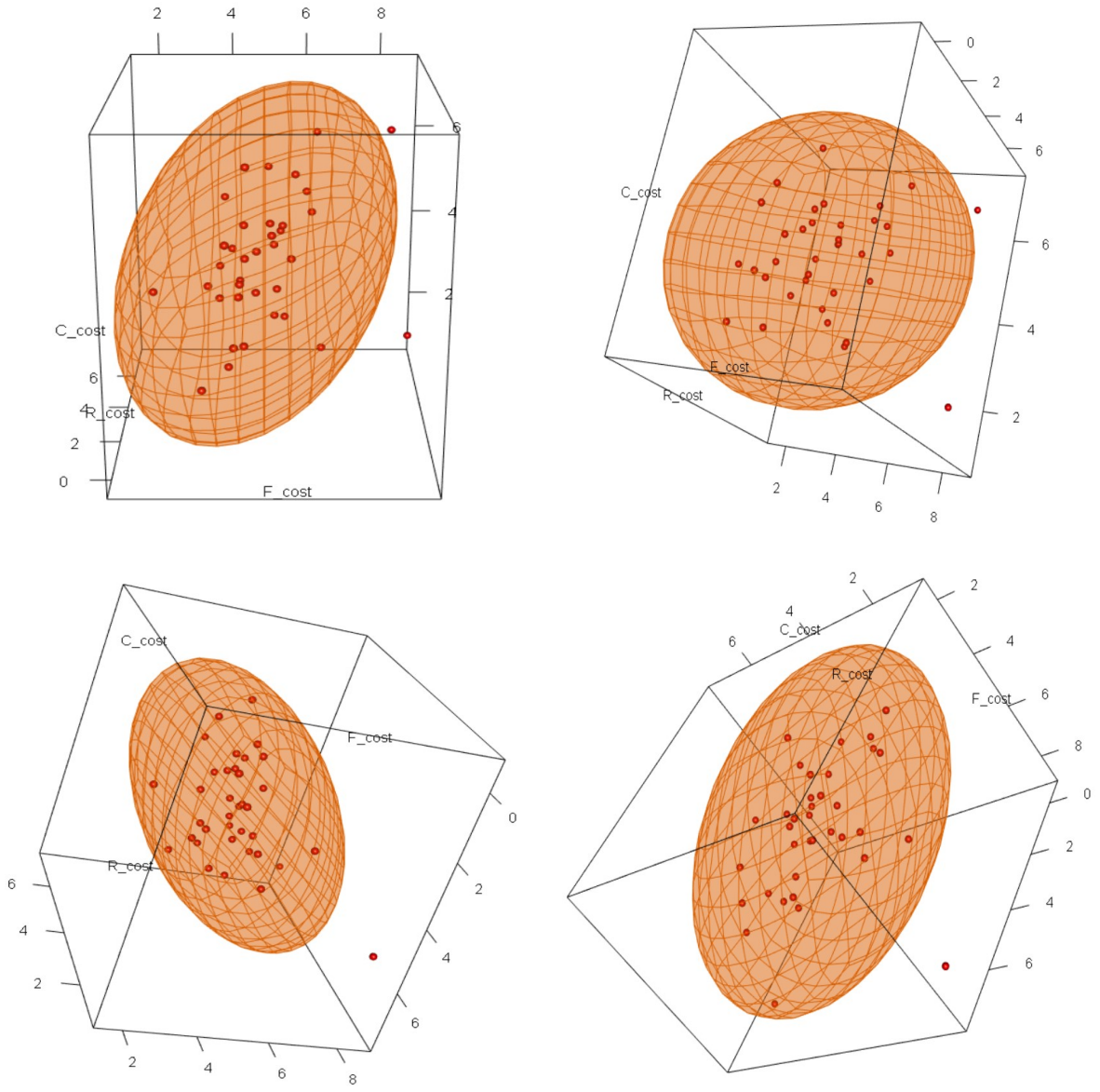
Test	Mardia Test Statistic	p-value
Skewness	5.18073	0.878782
Kurtosis	-0.92976	0.3524952

QQ plots and p-values of Shapiro-Wilk tests and Mardia test **accept** the assumption of normality of Diesel data.

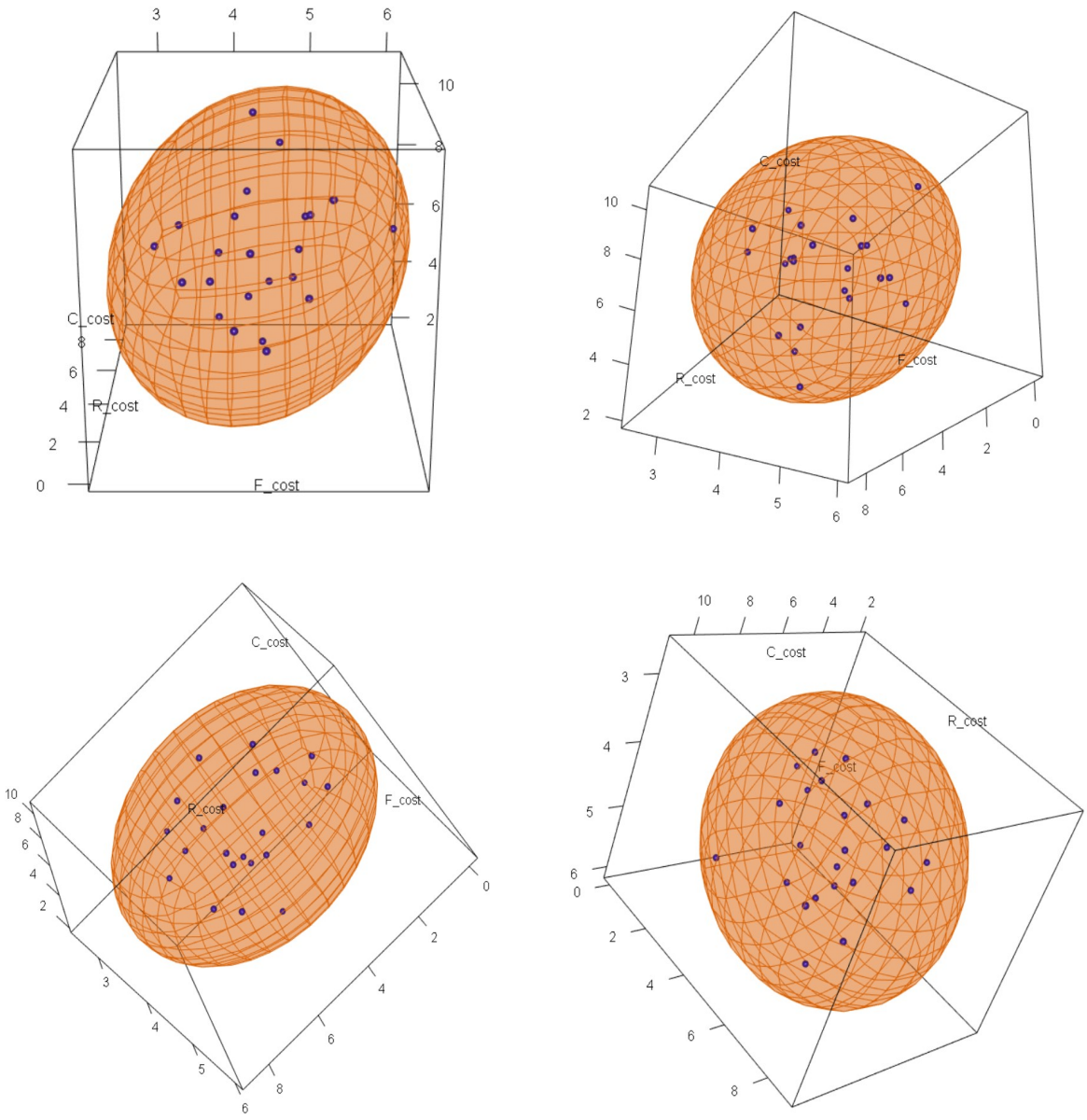
Since the same scenario prevails, now we highly doubt this non-normality is happening only due to presence of leverage (outliers) in the data set. To validate this point let us plot confidence ellipsoid for data for the Box Cox transformed data set $\underset{3\times 1}{\underset{\sim}{Y_{ij}}}$. 95% confidence ellipsoid for i^{th} population is given by,

$$(\underset{\sim}{Y_{ij}} - \underset{\sim}{\bar{Y}_i})^T {S_i}^{-1} (\underset{\sim}{Y_{ij}} - \underset{\sim}{\bar{Y}_i}) = {\chi^2}_p(0.95) \quad ; \; i = 1, 2$$

Data ellipse for Gasoline is the following,



Data ellipse for Diesel is the following,



So, our intuition was right these data ellipse clearly indicate the presence of outliers.

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.

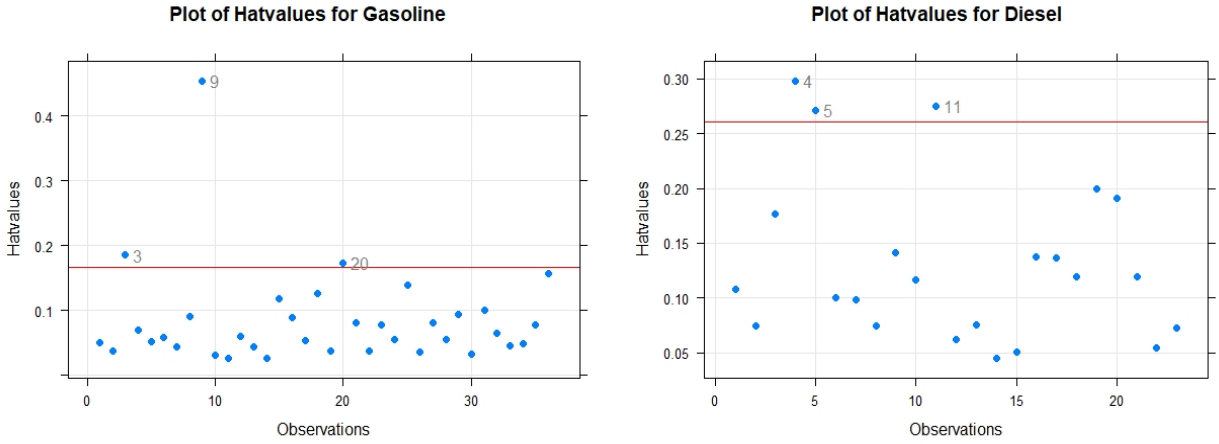
To detect the leverage points we will use Hat Matrix. Let us denote the coefficient matrix of i^{th} population by

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

and Hat Matrix is defined as follows,

$$H_i = Z_i(Z_i{}^T Z_i)^{-1} Z_i{}^T = ((h^{(i)}{}_{jk})) \quad ; \ i = 1, 2$$

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;



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.

Let us perform Shapiro-Wilk test and Mardia test on variables for the car which uses Gasoline as fuel.

The results of Shapiro-Wilk test is given by,

Data	Shapiro -Wilk Test Statistic	p-value
Fuel Cost	0.9544	0.1435
Repair Cost	0.9821	0.8126
Capital Cost	0.9793	0.7234
Multivariate Data	0.96421	0.2505

And the results of Mardia test is as follows:

Test	Mardia Test Statistic	p-value
Skewness	9.18957	0.51421
Kurtosis	0.14516	0.88458

The p-values of Shapiro-Wilk tests and Mardia test **accept** the assumption of normality of Gasoline data.

Let us perform Shapiro-Wilk test and Mardia test on variables for the car which uses Diesel as fuel.

The results of Shapiro-Wilk test is given by,

Data	Shapiro -Wilk Test Statistic	p-value
Fuel Cost	0.9641	0.5506
Repair Cost	0.9606	0.4755
Capital Cost	0.9690	0.6654
Multivariate Data	0.97142	0.8936

And the results of Mardia test is as follows:

Test	Mardia Test Statistic	p-value
Skewness	8.674548	0.563243
Kurtosis	-0.370027	0.71136

The p-values of Shapiro-Wilk tests and Mardia test **accept** the assumption of normality of 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 Principal Components for Gasoline data set

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

$$S_1 = \begin{pmatrix} 1.1409654 & 0.8393898 & 0.4938155 \\ 0.8393898 & 2.1955563 & 0.6527992 \\ 0.4938155 & 0.6527992 & 1.3383613 \end{pmatrix} \quad R_1 = \begin{pmatrix} 1 & 0.5303410 & 0.3996150 \\ 0.530341 & 1 & 0.3808208 \\ 0.399615 & 0.3808208 & 1 \end{pmatrix}$$

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 EV-EV pairs $(\hat{\lambda_1}, \underset{\sim}{\hat{e_1}})$ of R_1 is $\left(1.8774604, \begin{pmatrix} 0.6020909 \\ 0.5949829 \\ 0.53243020 \end{pmatrix}\right)$, $\left(0.6536039, \begin{pmatrix} -0.3328173 \\ -0.4191258 \\ 0.84472848 \end{pmatrix}\right)$, $\left(0.4689357, \begin{pmatrix} -0.7257542 \\ 0.6858053 \\ 0.05433116 \end{pmatrix}\right)$

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

Table 5.1: Contribution of the Variables to 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

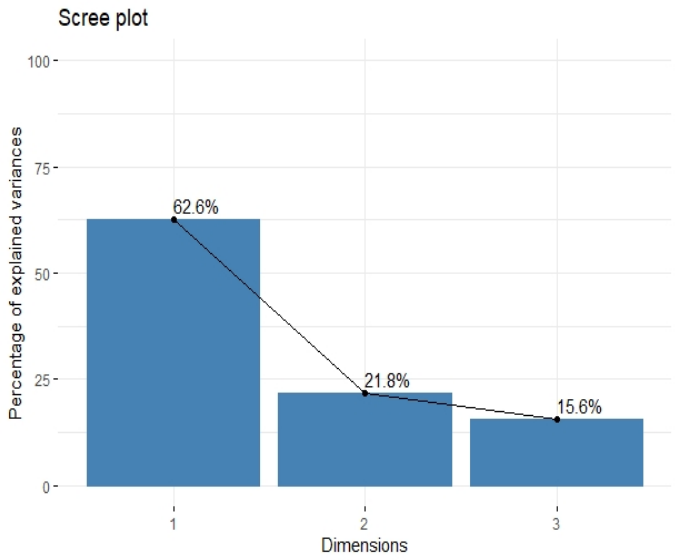
The principal components are given by

$$\begin{aligned} Z_{11} &= 0.6020909\ Y_{11} + 0.5949829\ Y_{12} + 0.53243020\ Y_{13} \\ Z_{12} &= -0.3328173\ Y_{11} - 0.4191258\ Y_{12} + 0.84472848\ Y_{13} \\ Z_{13} &= -0.7257542\ Y_{11} + 0.6858053\ Y_{12} + 0.05433116\ Y_{13} \end{aligned}$$

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

Table 5.2: Percentage of variance of Principal Components

	Eigenvalue	Percentage of variance	Cumulative percentage of variance
PC_1	1.8774604	62.58201	62.58201
PC_2	0.6536039	21.78680	84.36881
PC_3	0.4689357	15.63119	100.00000



- 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 84.4% variability.

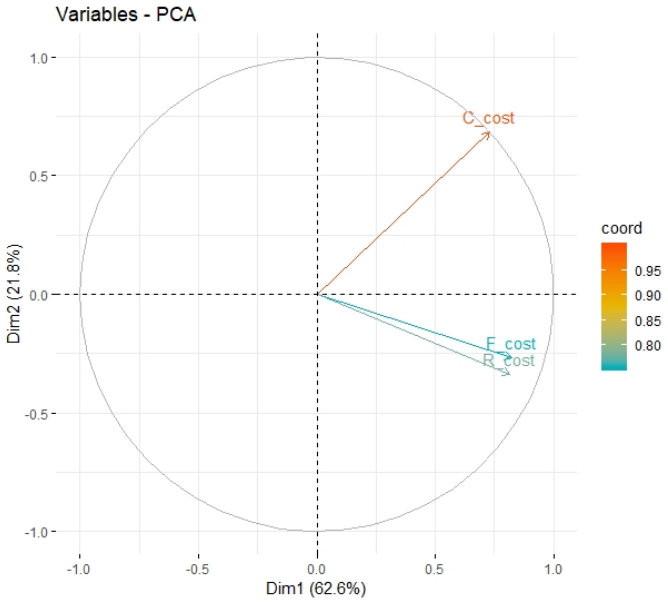
The following table represents the correlation between the variables and principal components.

Table 5.3: Correlation coefficients between 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

Correlation Circle Plot uses coordinates as the correlation between variables and the first two PC's having highest variance. The color is determined the square of distance of the point from the origin. Variables having higher color (length) define well representativity in the PC's. Arrows in the same direction shows high correlation among themselves. Arrows at 90° and 180° at each other shows zero correlation and negative correlation respectively.

The following diagram represents the correlation circle plot for Gasoline data set,



- Each of the variables have length (color) more than 0.75 , so all of them are well represented in the first two PC's. As a result one can drop the PC_3.
- Fuel cost and Repair cost are highly correlated and these variables have more or less zero correlation with Capital cost.

5.2 Principal Components for Diesel data set

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

$$S_2 = \begin{pmatrix} 0.38752239 & 0.1641008 & 0.08090999 \\ 0.16410077 & 2.1551809 & 0.50386795 \\ 0.08090999 & 0.5038679 & 1.22220283 \end{pmatrix} \quad R_2 = \begin{pmatrix} 1 & 0.1795645 & 0.1175661 \\ 0.1795645 & 1 & 0.3104583 \\ 0.1175661 & 0.3104583 & 1 \end{pmatrix}$$

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 $(\hat{\lambda_2}, \underset{\sim}{\hat{e_2}})$ of R_2 is $\left(1.4169940, \begin{pmatrix} 0.4520790 \\ 0.6496940 \\ 0.6111648 \end{pmatrix}\right)$, $\left(0.9012142, \begin{pmatrix} 0.8748847 \\ -0.1894780 \\ -0.4457296 \end{pmatrix}\right)$, $\left(0.6817919, \begin{pmatrix} 0.1737856 \\ -0.7362037 \\ 0.6540663 \end{pmatrix}\right)$

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

Table 5.4: Contribution of the Variables to 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

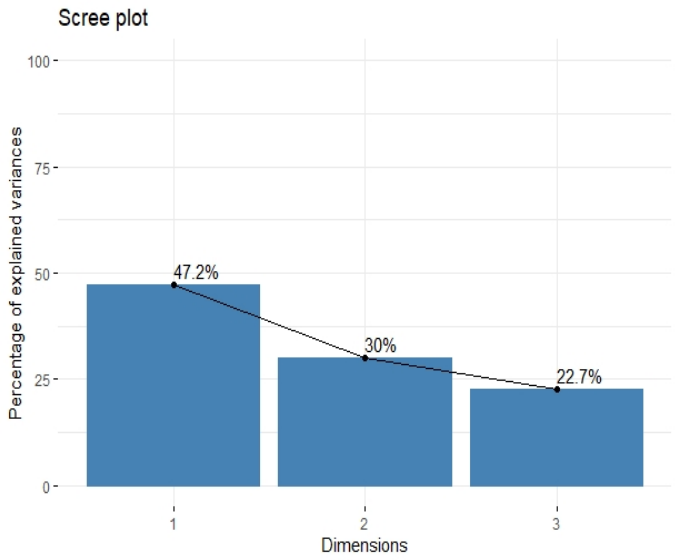
The principal components are given by

$$Z_{21} = 0.4520790\, Y_{21} + 0.6496940\, Y_{22} + 0.6111648\, Y_{23}$$
$$Z_{22} = 0.8748847\, Y_{21} - 0.1894780\, Y_{22} - 0.4457296\, Y_{23}$$
$$Z_{23} = 0.1737856\, Y_{21} - 0.7362037\, Y_{22} + 0.6540663\, Y_{23}$$

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

Table 5.5: Percentage of variance of Principal Components

	Eigenvalue	Percentage of variance	Cumulative percentage of variance
PC_1	1.4169940	47.23313	47.23313
PC_2	0.9012142	30.04047	77.27360
PC_3	0.6817919	22.72640	100.00000



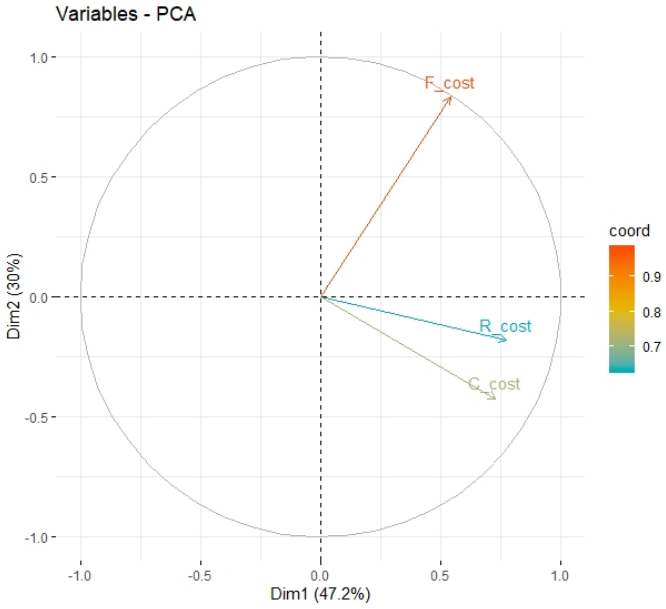
- Seeing the Scree plot and the table of percentage of variance of PC , it is difficult drop the PC_3 since the first two PC has about only 77% variability and there is no formation of elbow shape in scree plot.

The following table represents the correlation between the variables and principal components.

Table 5.6: Correlation coefficients between Variables and Principal Components

	PC_1	PC_2	PC_3
Fuel cost	0.5381437	0.8305481	0.1434959
Repair cost	0.7733798	-0.1798758	-0.6078884
Capital cost	0.7275156	-0.4231414	0.5400671

The following diagram represents the correlation circle plot for Diesel data set,



- Each of the variables have length (color) more than 0.6 . The least one is Repair cost that has distance 0.62 from origin. So, it is difficult to drop PC_3.
- Repair cost and Capital cost are moderately correlated whereas Fuel cost has more or less zero correlation with Capital cost.

5.3 Findings

- It is clear from the above analysis that PC_3 can be dropped from the Gasoline data but that is not possible for the Diesel data. 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.
- The Principal Components for these populations is not similar , so we have to analysis for each population separately.

6. Confidence Region for Mean Vector

Since we have achieved both univariate normality and multivariate normality of our dataset, we are now interested in doing confidence region for mean vector for Gasoline data and Diesel data.

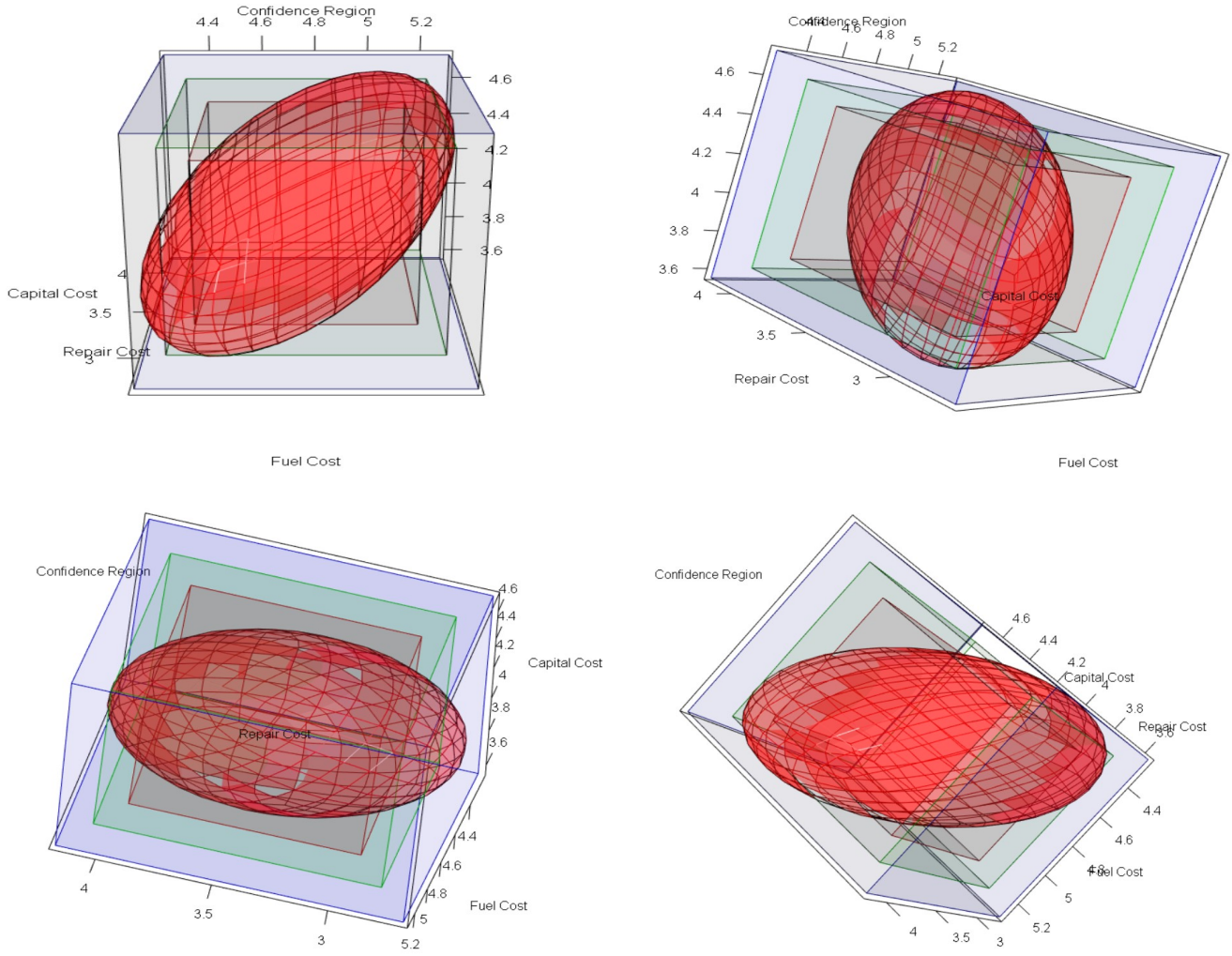
Let $\underset{\sim}{\mu_i} = \begin{pmatrix} \mu_{i_1} \\ \mu_{i_2} \\ \mu_{i_3} \end{pmatrix}$ be the mean response of the 3 variables for the i^{th} group (1 represents Gasoline , 2 represents Diesel).

6.1 Confidence Region for mean of Gasoline data

Different types of confidence region for mean vector $\underset{\sim}{\mu_1}$ of transformed variables are given below:

- $[4.406381, 5.129207] \times [2.926083, 3.928781] \times [3.752178, 4.535039]$ is the Individual Confidence Interval for $\underset{\sim}{\mu_1}$ with confidence coefficient 95% each.
- $[4.320138, 5.215449] \times [2.806448, 4.048415] \times [3.658773, 4.628444]$ is the Bonferroni Simultaneous Confidence Interval for $\underset{\sim}{\mu_1}$ with confidence coefficient at least 95%.
- $[4.227800, 5.307788] \times [2.678357, 4.176506] \times [3.558765, 4.728452]$ is the Simultaneous Confidence Interval for $\underset{\sim}{\mu_1}$ with confidence coefficient 95%.

We have plotted all these confidence regions along with confidence ellipsoid in the same graph.
The inner cuboid (color : pale red) is the Individual Confidence Interval, the middle one (color : pale green) is the Bonferroni Simultaneous Confidence Interval , the outer one (color : pale blue) is the Simultaneous Confidence Interval and the red ellipsoid is the 95% confidence ellipsoid for $\underset{\sim}{\mu_1}$.



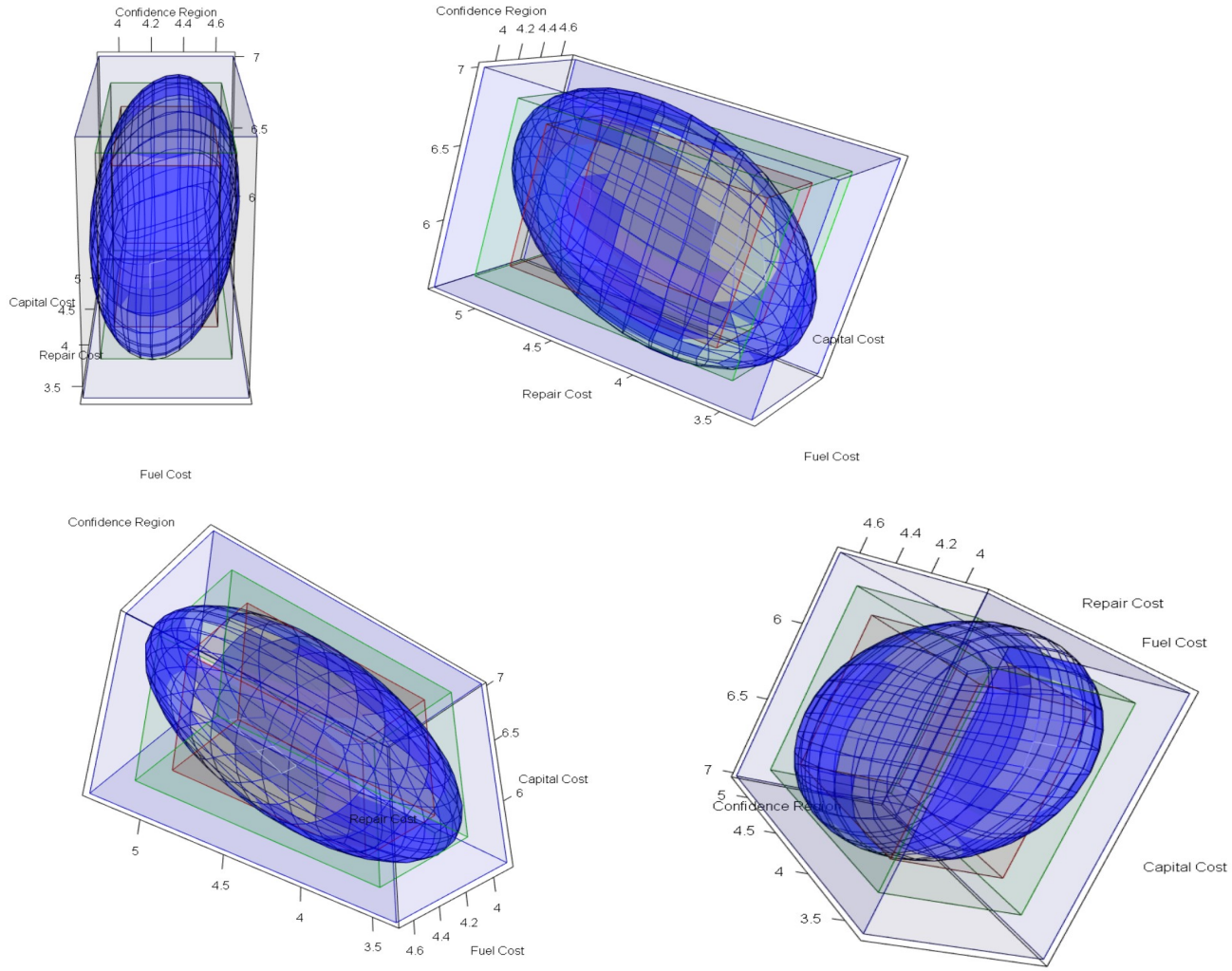
From the plot , we see that Simultaneous Confidence Interval is the largest Confidence Interval among these and individual Confidence Intervals are the smallest one. Simultaneous Confidence Interval is the projection of 95% Confidence Ellipsoid of $\underset{\sim}{\mu_1}$ on the respective axes.

6.2 Confidence Region for mean of Diesel data

Different types of confidence region for mean vector $\underset{\sim}{\mu_2}$ of transformed variables are given below:

- $[4.022195, 4.560585] \times [3.680056, 4.949723] \times [5.792567, 6.748703]$ is the Individual Confidence Interval for $\underset{\sim}{\mu_2}$ with confidence coefficient 95% each.
- $[3.955043, 4.627737] \times [3.521693, 5.108086] \times [5.673310, 6.867960]$ is the Bonferroni Simultaneous Confidence Interval for $\underset{\sim}{\mu_2}$ with confidence coefficient at least 95%.
- $[3.876331, 4.706449] \times [3.336069, 5.293710] \times [5.533524, 7.007746]$ is the Simultaneous Confidence Interval for $\underset{\sim}{\mu_2}$ with confidence coefficient 95%.

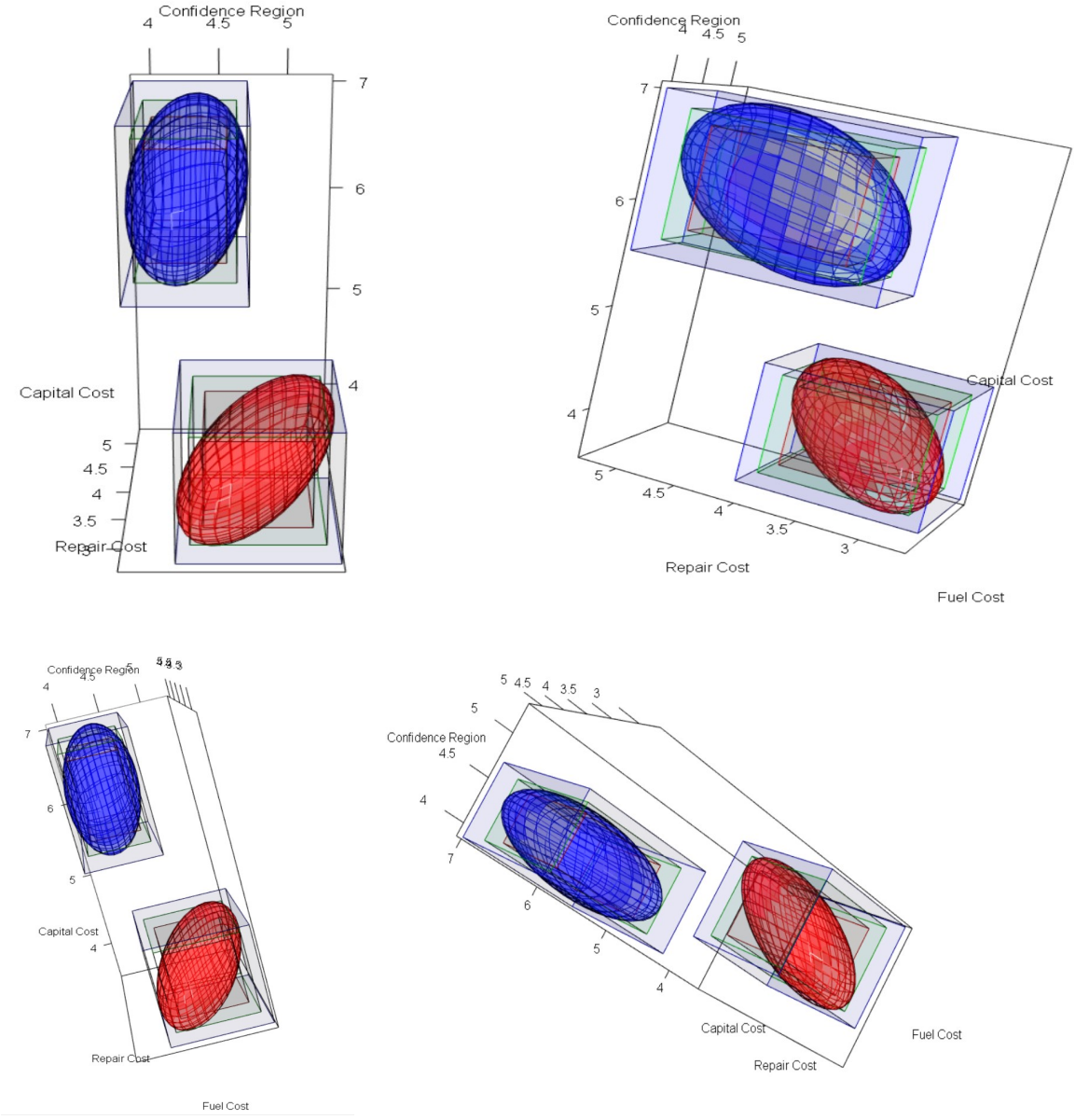
We have plotted all these confidence regions along with confidence ellipsoid in the same graph.
The inner cuboid (color : pale red) is the Individual Confidence Interval, the middle one (color : pale green) is the Bonferroni Simultaneous Confidence Interval , the outer one (color : pale blue) is the Simultaneous Confidence Interval and the blue ellipsoid is the 95% confidence ellipsoid for $\underset{\sim}{\mu_2}$.



From the plot , we see that Simultaneous Confidence Interval is the largest Confidence Interval among these and individual Confidence Intervals are the smallest one. Simultaneous Confidence Interval is the projection of 95% Confidence Ellipsoid of $\underset{\sim}{\mu_2}$ on the respective axes.

6.3 Comparative Study

To have the comparative study between two groups, let us plot the above two confidence region plots in same plot. Note the following color scheme: the populations will be identified by the color of the confidence ellipsoid i.e. red confidence ellipsoid will represent the Gasoline data and blue confidence ellipsoid will represent the Diesel data and the remaining colors will be same as previous.



The plot is very clear difference in mean of the variables for these two populations. But the Capital cost has most significant difference among populations. Capital cost for Diesel is much higher than that of Gasoline , which encourages us for performing different discriminant rules for the dataset.

7. Profile Analysis

Initially we will test, $H_0 : \Sigma_1 = \Sigma_2$ vs. $H_1 : \Sigma_1 \neq \Sigma_2$, where Σ_1 and Σ_2 are the dispersion matrices for Gasoline used group and Diesel used Group. We will use **Bartlett's Test** Statistic for testing this.
The p-value of the test is 0.2673>0.05.Hence H_0 is accepted at 5% level of significance.
Now, we will do profile analysis .Let $\underset{\sim}{\mu_i} = \begin{pmatrix} \mu_{i_1} \\ \mu_{i_2} \\ \mu_{i_3} \end{pmatrix}$ be the mean response of the 3 variables for the i^{th} group (1 represents Gasoline ,2 represents Diesel). We are interested in the answer of the following questions,

- Are the profiles parallel?
- Assuming the profiles are parallel ,are they coincident?
- If answer of the above questions is 'yes' ,then are the profiles level?

For answering the first question we will test,

$$H_0 : \mu_{1_i} - \mu_{1_{i-1}} = \mu_{2_i} - \mu_{2_{i-1}} \;\; ; \;\; i = 2, 3 \quad vs. \quad H_1 : H_0 \text{ is false}$$

The value of the test statistic is given by,

$$T^2 = (\underset{\sim}{\bar{Y_1}} - \underset{\sim}{\bar{Y_2}})^T \; C^T \left[\left(\frac{1}{n_1} + \frac{1}{n_2} \right) C \; S_{pooled} \; C^T \right]^{-1} \; C \; (\underset{\sim}{\bar{Y_1}} - \underset{\sim}{\bar{Y_2}})$$

where, C is of the following form

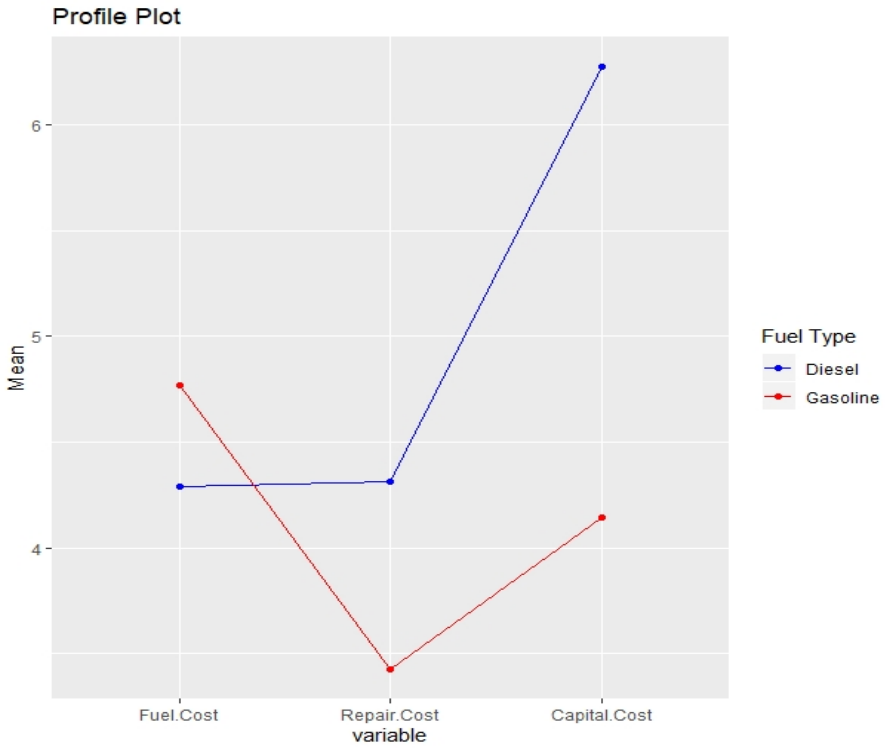
$$C = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}$$

Reject H_0 at α% significance

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

Now for the dataset,

$$T^2{}_{Observed} = 65.99988 \geq 6.436646 = \frac{(59-2)(3-1)}{59-3} \times F_{3-1, 59-3}(0.05)$$



Hence, the profiles aren't parallel. As the profiles aren't parallel ,the profiles can't be level also.
As a result, the mean vectors are class-specific . Thus, the two populations are different. Constructing Discriminant Rules for these two classes is appropriate here, as the two populations are different.

8. Discriminant Analysis

Since we are working on the transformed data set, which are jointly multivariate normal, we can use Linear Discriminant Analysis (as the variance-covariance matrices are equal at 5% level of significance using Box's M Test) and Quadratic Discriminant Analysis here.
Let $\boldsymbol{\pi_1}$ and $\boldsymbol{\pi_2}$ denote the population of transporters using gasoline and diesel fuels respectively.

To apply discriminant analysis initially the whole transformed dataset is considered as training set as well as validation set and then is partitioned in training set and validation(test) set . We have randomly done an approximate 75%-25% split of the dataset, the 75% part has been taken as the training set and the remaining 25% has been taken as the validation set. Doing this, the training set size has been obtained as 44 & the validation set size has been obtained as 15. Then, we will predict the fuel type used for each observation of the validation set using the discriminant rule obtained using the training set.

8.1 Linear Discriminant Analysis

According to this rule, allocate $\underset{\sim}{x_0}$ to $\boldsymbol{\pi_i}$ if,

$$\widehat{d_i(\underset{\sim}{x_0})} = max\{\widehat{d_1(\underset{\sim}{x_0})}, \widehat{d_2(\underset{\sim}{x_0})}\}$$

where,

$$\widehat{d_i(\underset{\sim}{x})} = \underset{\sim}{\bar{x}_i}^T S_{pooled}^{-1} \underset{\sim}{x} - \frac{1}{2} \underset{\sim}{\bar{x}_i}^T S_{pooled}^{-1} \underset{\sim}{\bar{x}_i} + ln(p_i) \;\; ; \quad i = 1, 2$$

and

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

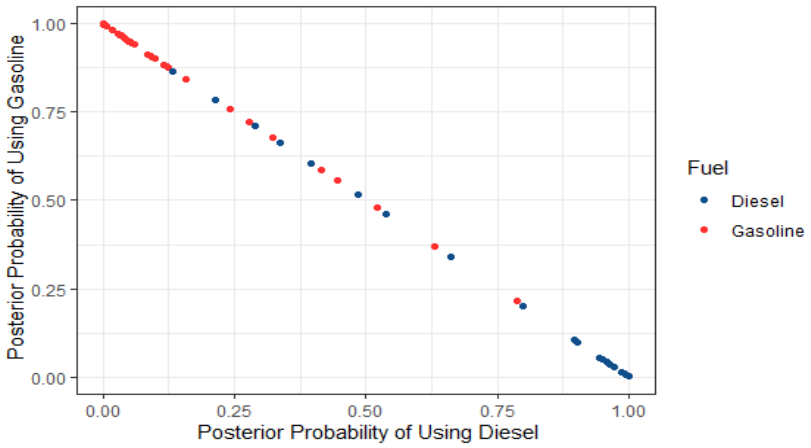
.

8.1.1 Discrimination Using Whole Data

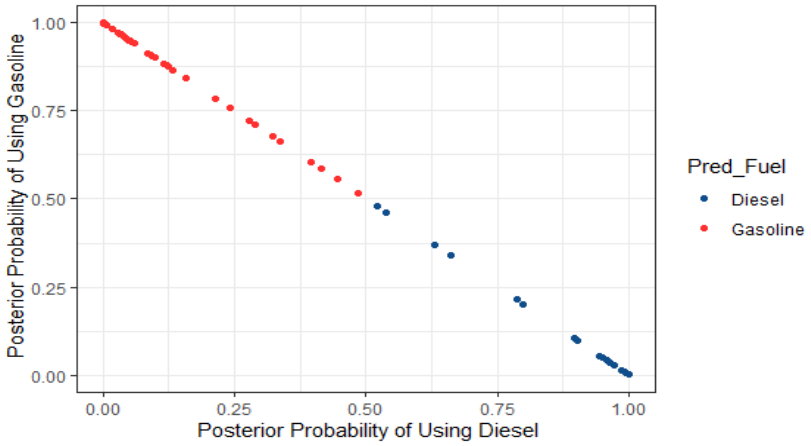
After constructing the discriminant rule based on the whole transformed data , now we wish to predict the fuel type used for each point of the dataset using this rule.

The linear discriminant rule maximizes posterior probability $P(\pi_k | \underset{\sim}{X} = \underset{\sim}{x})$, while assuming $\underset{\sim}{X}$ is drawn from a multivariate normal distribution , with a class-specific mean vector and a common covariance matrix.

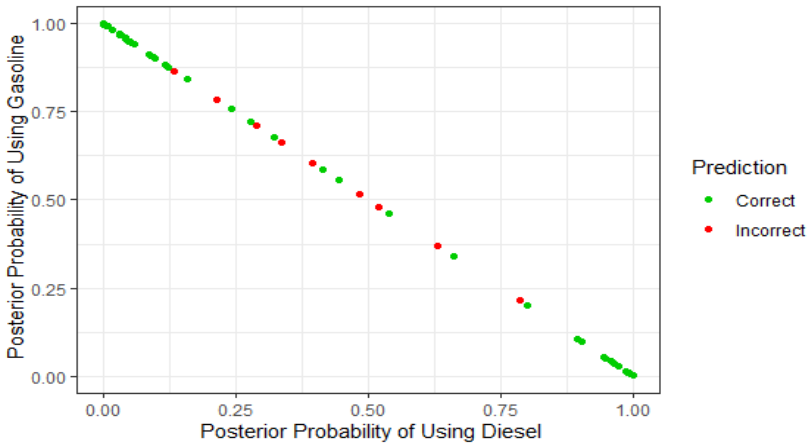
In the following diagram, the data points are plotted w.r.t. their corresponding posterior probabilities and coloured according as the Fuel used.



The following diagram is the same as above but coloured according as the Fuel predicted.



The following diagram is the same as above but coloured according as correctly classified or not.



The **confusion matrix** is as follows,

Predicted Fuel Type	*Actual Fuel Type*	
	Gasoline	Diesel
Gasoline	**33**	**6**
Diesel	**3**	**17**

In the table, the green cells indicates no. of data points classified correctly whereas the red cells indicates no. of data points classified incorrectly.
Thus, the Apparent Error Rate(APER)is= $\frac{6+3}{59} \times 100\% = 15.25424\%$

8.1.2 Discrimination Using the Training-Validation Split Up

To get better estimate of the actual error of misclassification, we use the training set to develop the discriminant rule and use it to predict the fuel used of each data point of the validation set.

The **confusion matrix** is given as follows,

Predicted Fuel Type	*Actual Fuel Type*	
	Gasoline	Diesel
Gasoline	**6**	**2**
Diesel	**1**	**6**

Thus,proportion of misclassification in the validation set = $\frac{3}{15} = 0.2$.

8.2 Quadratic Discriminant Analysis

According to this rule, allocate $\underset{\sim}{x_0}$ to $\boldsymbol{\pi_i}$ if,

$$\widehat{d_i^Q(\underset{\sim}{x_0})} = max\{\widehat{d_1^Q(\underset{\sim}{x_0})}, \widehat{d_2^Q(\underset{\sim}{x_0})}\}$$

where,

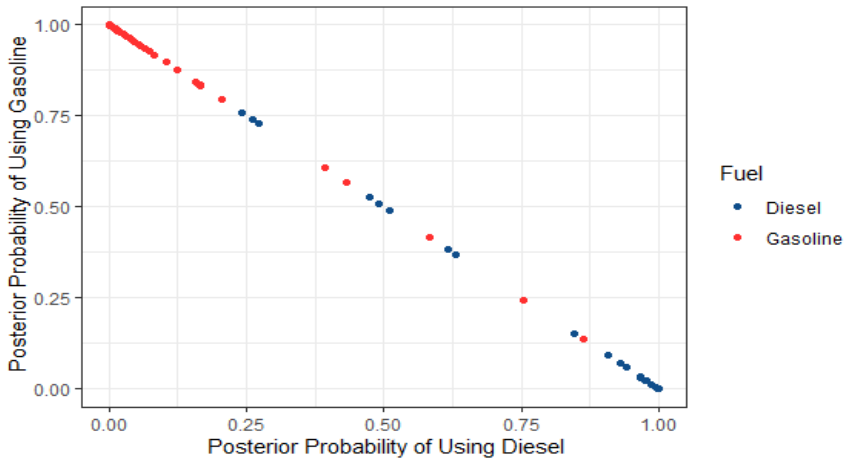
$$\widehat{d_i^Q(\underset{\sim}{x})} = -\frac{1}{2}ln|S_i| - \frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\bar{x}_i})^T S_i^{-1}(\underset{\sim}{x} - \underset{\sim}{\bar{x}_i}) + ln(p_i) \;\; ; \quad i = 1, 2$$

8.2.1 Discrimination Using Whole Data

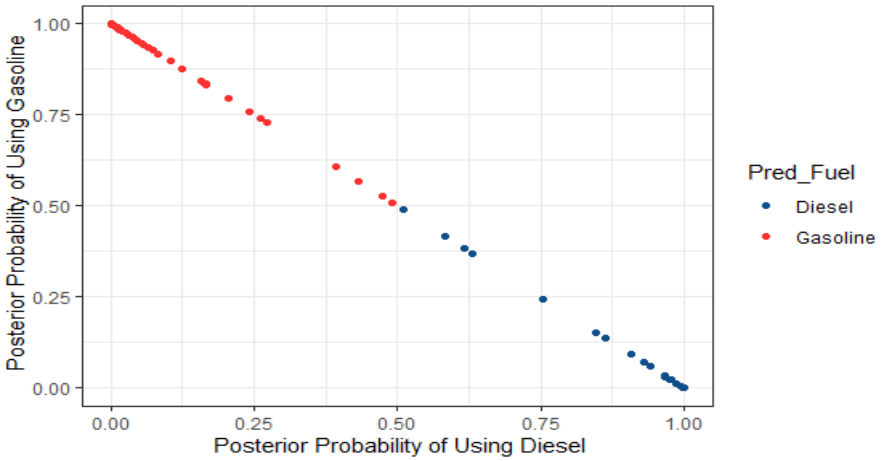
After constructing the discriminant rule based on the whole transformed data, now we predict the fuel type for the each observations of the dataset using this rule.

The quadratic discriminant rule maximizes posterior probability $P(\pi_k|\underset{\sim}{x})$,while assuming $\underset{\sim}{X}$ is drawn from a multivariate normal distribution ,with a class-specific mean vector and covariance matrix.

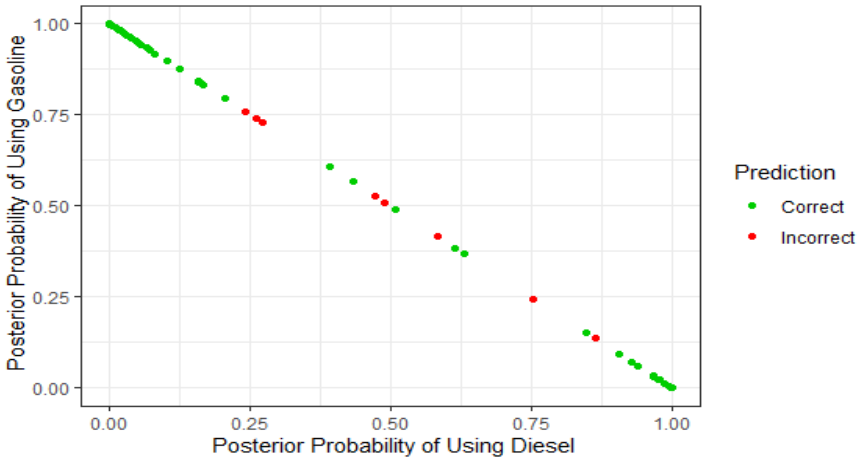
In the following diagram, the data points are plotted w.r.t. their corresponding posterior probabilities and coloured according to the Fuel used.



The following diagram is the same as above but coloured according to the Fuel predicted.



The following diagram is the same as above but coloured according to correct or incorrect classification.



The **confusion matrix** is as follows,

Predicted Fuel Type	*Actual Fuel Type*	
	Gasoline	Diesel
Gasoline	**33**	**5**
Diesel	**3**	**18**

Thus, the Apparent Error Rate(APER)is= $\frac{5+3}{59} \times 100\% = 13.55932\%$

8.2.2 Discrimination Using the Training-Validation Split Up

After constructing the discriminant rule based on the training set,now we predict the fuel used for the elements of the validation set using this rule. The **confusion matrix** is given as follows,

Predicted Fuel Type	*Actual Fuel Type*	
	Gasoline	Diesel
Gasoline	**7**	**2**
Diesel	**0**	**6**

Thus,proportion of misclassification in the validation set= $\frac{2}{15} = 0.13333333$.

8.3 Comparison

Clearly, based on the performance of Linear Discriminant rule and Quadratic Discriminant rule on the validation set, Quadratic Discriminant rule is better. Now, we will try some other methods and compare their performances on classifying data points with Linear Discriminant Analysis and Quadratic Discriminant Analysis.

8.3.1 Classification Using Logistic Regression

Here, we have tried to fit a logistic regression on the training dataset with the fuel type being the response variable and the three costs are the explanatory variables and then predict the fuel type of each data point of the validation set. The categorical response fuel type($\mathbf{Y}$) observes two values 1 & 2. We fit the following logistic regression model on the training set:

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

for $i = 1, 2, ..., 44$, $k = 0, 1$ and $\beta_0, \beta_1, \beta_2, \beta_3$ are unknown parameters. After fitting the estimates of the parameters are obtained 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 observations, Gasoline has been used for 7 observations & Diesel has been used for 8 observations. So, the prior probabilities of both the populations are almost same in the validation set.

Now we use the following prediction rule for the validation set:

- Compute the value of $\frac{exp(-8.9614-1.7093X_{1j}+1.0697X_{2j}+2.2689X_{3j})}{1+exp(-8.9614-1.7093X_{1j}+1.0697X_{2j}+2.2689X_{3j})}$ for every observation j in the validation set for $j = 1, 2, .., 15$.
- If the value for the j^{th} observation is ≤ 0.5, we predict $\hat{Y}_j - 1 = 0$ i.e. $\hat{Y}_j = 1$.
- Otherwise, we predict $\hat{Y}_j - 1 = 1$ i.e. $\hat{Y}_j = 2$.
- Now, we compare every $\hat{Y}_j$ with Y_j in the validation set to get the amount of wrong prediction(or, classification).

The **Confusion Matrix** obtained here is as follows:

Predicted	*Actual Fuel Type*	
Fuel Type	Gasoline	Diesel
Gasoline	**6**	**2**
Diesel	**1**	**6**

Thus,proportion of misclassification in the validation set $= \frac{3}{15} = 0.2$.

8.3.2 Classification Using k-Nearest Neighbours

In this method, each data point of the validation set is taken and the k data points from the training dataset is chosen which are nearest to this new data point (nearest in the sense 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.

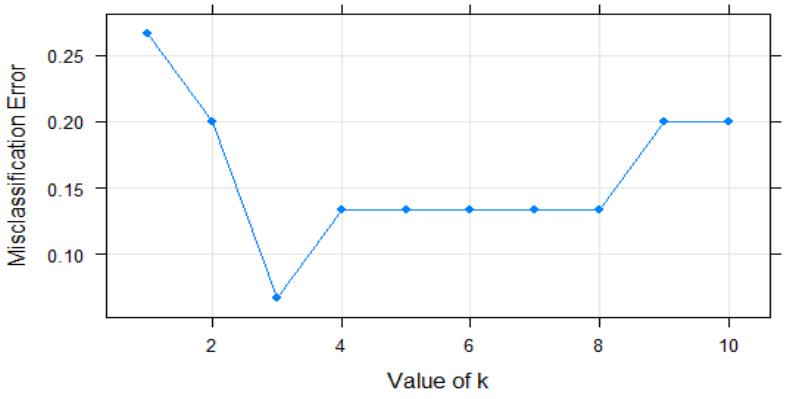


Figure 8.1: Misclassification Error for Various Choices of k

From the graph, it is evident that for k = 3, we have the lowest error. So, now we will predict the fuel type of each data points in the validation set using 3-nearest neighbours.

The **Confusion Matrix** is obtained as follows:

Predicted Fuel Type	*Actual Fuel Type*	
	Gasoline	Diesel
Gasoline	**7**	**1**
Diesel	**0**	**7**

Thus,proportion of misclassification in the validation set= $\frac{1}{15} = 0.0666667$.

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: So, for the validation dataset we obtained, 3-nearest neighbours performs the best in the predicting the

Method of Classification	Misclassification Error
Linear Discriminant Analysis	20%
Quadratic Discriminant Analysis	13.33%
Logistic Regression	20%
3-Nearest Neighbours	6.67%

fuel types of the data points in the validation set. Quadratic Discriminant Analysis comes next. And linear discriminant analysis and logisitc regression performs identically in predicting the fuel types and having the highest misclassification error among all the methods.

8.3.4 Lachenbruch's 'Holdout' Procedure

Though using a training-validation split up of the whole dataset is helpful to provide a more reliable estimate of **AER**(*Actual Error Rate*), it suffers from not having a large enough dataset and using a smaller no. of data points for constructing the discriminant function. So, now we use **Lachenbruch's Holdout** procedure. The whole method can be described in the following steps:

- Holdout the first observation of the population using Gasoline.
- Use the remaining 58 observations 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 no of hold-out observations in the Gasoline population which are misclassified and denote it by $n_{1M}^{(H)}$.
- Repeat the same 6 steps for the Diesel population and calculate $n_{2M}^{(H)}$ which is the no of hold-out observations in the Diesel population which are misclassified.
- Estimates of $\mathbf{P(2|1)}$ & $\mathbf{P(1|2)}$ is obtained as $\hat{P}(2|1) = \frac{n_{1M}^{(H)}}{36}$ and $\hat{P}(1|2) = \frac{n_{2M}^{(H)}}{23}$
- The estimate of expected actual error rate, $\mathbf{E(AER)}$ is obtained as $\hat{E}(AER) = \frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{59}$.

8.3.4.1 Result Using LDA

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

Predicted Fuel Type	*Actual Fuel Type*	
	Gasoline	Diesel
Gasoline	**32**	**6**
Diesel	**4**	**17**

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

$$n_{1M}^{(H)} = 4, n_{2M}^{(H)} = 6, \hat{P}(2|1) = \frac{4}{36} = 0.1111111, \hat{P}(1|2) = \frac{6}{23} = 0.2608696$$

And, $\hat{E}(AER) = \frac{4+6}{59} = 0.1694915$ i.e. the estimate of expected actual error rate is approximately 16.95%.

8.3.4.2 Result Using QDA

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

Predicted Fuel Type	*Actual Fuel Type*	
	Gasoline	Diesel
Gasoline	**32**	**7**
Diesel	**4**	**16**

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

$$n_{1M}^{(H)} = 4, n_{2M}^{(H)} = 7, \hat{P}(2|1) = \frac{4}{36} = 0.111111, \hat{P}(1|2) = \frac{7}{23} = 0.3043478$$

And, $\hat{E}(AER) = \frac{7+4}{59} = 0.1864407$ i.e. the estimate of expected actual error rate is approximately 18.64%. Clearly, Linear Discriminant rule is performing slightly better.

9. Summary

To sum up what we have done with the given data on three variables Fuel cost, Repair cost, Capital cost of transporting milk from farms to dairy plants which uses either Gasoline or Diesel as fuel in their cars;

- We have checked the normality of those two populations. Gasoline population was turned out to be non-normal. We performed Box-Cox transformation on the Gasoline data and applied the same transformation on the Diesel data so that we can have some comparative study.
- We checked for ouliers in the dataset with the help of hat matrix and removed those outliers with respective column means.
- We have done principal component analysis. It was found that the last PC for gasoline data can be dropped but that was not possible for Diesel data. Also the principal components of these populations were not similar. As a conclusion, to have further study we need to handle these two populations separately.
- While doing Confidence regions for mean vector of the transformed population, we found that there were difference in mean of variables for those population. But the major and significant difference was due to Capital cost. Capital cost for Diesel data very much higher than that of Gasoline data.
- Although from the confidence region it was clear that the mean of the populations are different, we perform the profile analysis to check whether they are parallel or not. The hypothesis of parallelism was also rejected . That motivated us for performing Discriminant Analysis.
- The transformed Gasoline & Diesel populations turned out to be homogeneous only through covariace matrix by the acceptance of **Box's M** test, so it was expected that both linear and quadratic discriminant analysis would perform equally good on separating the populations. But as we can see that QDA has a smaller APER as well as a smaller classification error than LDA while predicting the fuel type of the data points in the whole dataset and in the validation set respectively. From this we may conclude there may present some univariate abnormal value in some the variables which have been tackled by QDA more efficiently while classifying than LDA.
- While comparing different classification methods based on their respective classification error we see that 3-nearest neighbours is performing the best among the all while both LDA and logistic regression have the most classification error. As nearest neighbour is a very crude way of classification, so to classify the observations in the validation set we will be preferring QDA here.
- **Lachenbruch's** Holdout procedure however revealed that estimate of *expected actual error rate* approximately same for both LDA & QDA.