Indian Statistical Institute, New Delhi

Masters of Statistics

Milk Transportation Data

Multivariate Data Analysis Project Work

Name: Abhirup Sengupta | Subhasis Sur | Shantanu Nayek
Roll Number: MD2201 | MD2223 | MD2218
Supervisor : Prof. Swagata Nandi

Abstract

The Milk Transportation Data Multivariate Data Analysis project aims to analyze the transportation logistics of milk being transported by vehicles which uses either diesel or gasoline as fuel. The project focuses on utilizing multivariate statistical techniques to identify the factors that contribute to the variability of milk delivery mode of transportation, and to develop predictive models for optimizing milk transportation mode based on capital cost , fuel cost of the vehicles and repair cost of the vehicles. The data analysis will involve exploring the relationships between different variables such as cost of diesel and gasoline , repair cost of the two types of vehicles , capital cost and transportation modes. The project will be based on analysing the discrimination rule and classification rule for the two vehicle types based on the three variables mentioned. Moreover , we also observed how the two profiles vary and try to conclude the insights obtained from the data.

Contents

Chapter 1

Introduction

Milk is one of the most significant and complete drink which contains almost all kind of nutrients. It is mostly produced in rural areas . So, transportation of milk in the urban areas is a necessity for the well-being of the society. We are interested in analysing a particular aspect of milk transportation. We are here with a data which contains a categorical variable on the fuel type of a vehicle by which the milk is transported. The two typed of fuel that are present are namely Diesel and Gasoline. We intend to observe how the capital cost for transportation , fuel cost of the two fuels , repair cost of the vehicles are significantly affected by the fuel type. Moreover , we worked on analysing how the capital cost, fuel cost and the repair cost of the vehiclecle may play a significant role in discrimination and obtain a classification rule for the two kinds of the fuel type used in the vehicles.

Chapter 2

Description of the Dataset

• **Name :** Milk Transportation

• **Source Link:** https://drive.google.com/file/d/1hpvidd1Yh-t5a39GrpS-c8XanITTO SRE/view?usp=sharing

• **Description :** Dataset contains information about the Fuel type(wheather it is gasoline or diesel) , Fuel cost, Repair cost, Capital cost of milk transportation mode . There are 59 datapoints and 4 variables in the dataset.

• **Variable :** 1) Fuel type : Fuel type is denoted as 1 if fuel type is gasoline and it is denoted as 2 if fuel type is diesel .

2) Fuel cost :- Fuel cost means the cost of fuel consumed for the mode used for transportation of milk . It is measured in the unit cents/mi .

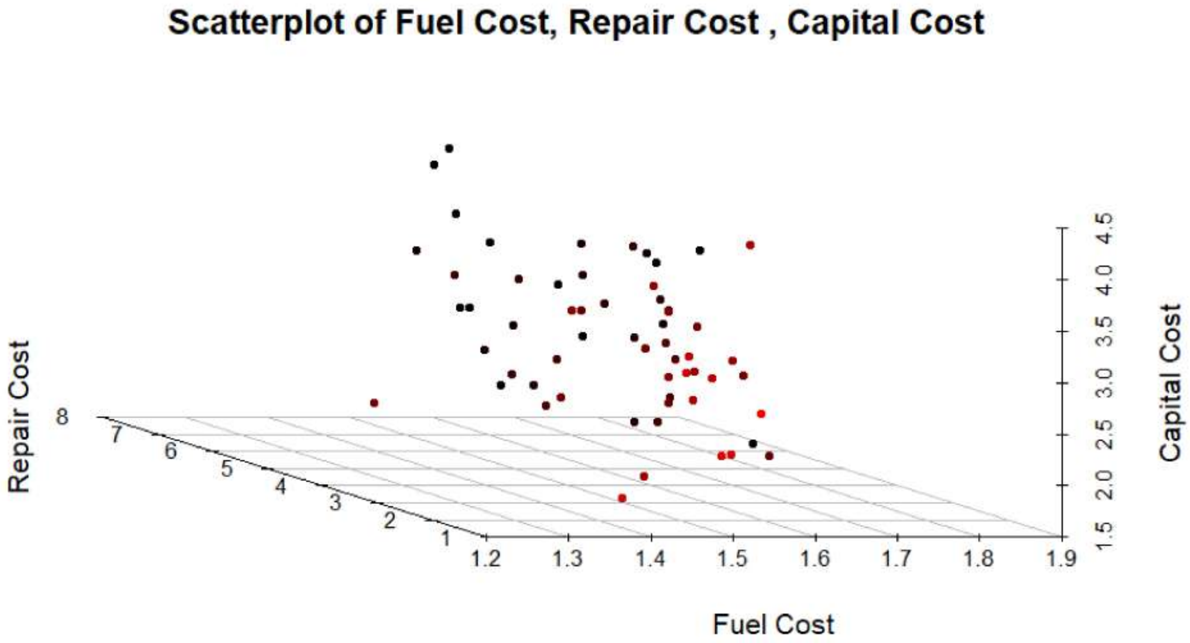
3) Repair cost :- Repair costs for transportation refer to the expenses incurred in repairing or maintaining vehicles used for transportation purposes. It is also measured in unit cents/mi .

4) Capital cost :- Costs applying to the physical assets of transportation, mainly infrastructures, terminals, and vehicles. It is also measured in unit cents/mi.

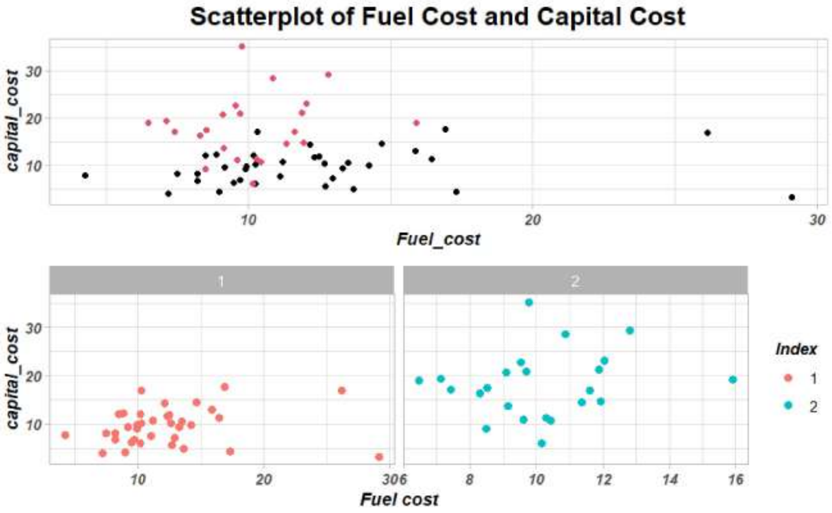
Chapter 3

Exploratory Data Analysis

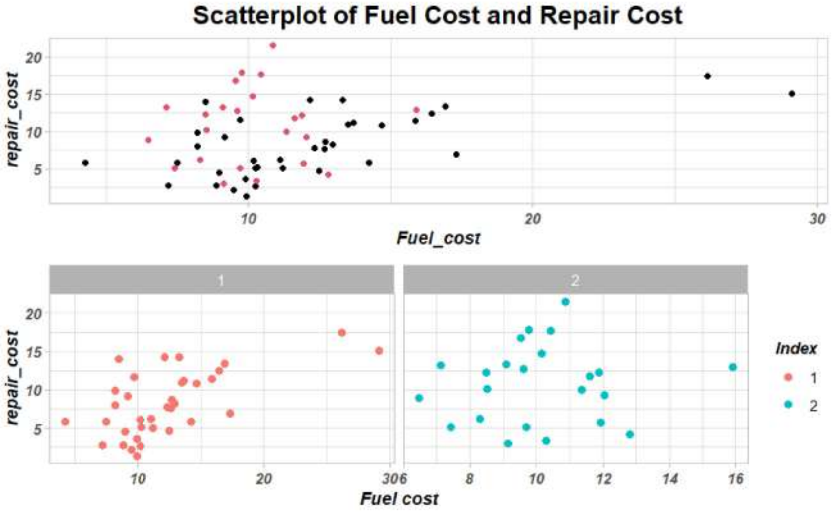
3.1 Scatter plot



From the below plot we get the presence of 2 outliers in the total dataset of Fuel

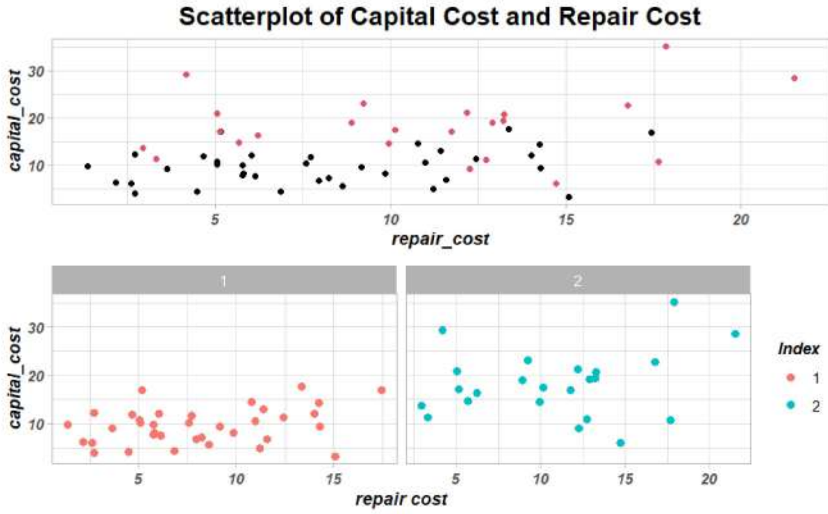


cost and capital cost. Also we get some outliers in the grouped dataset.

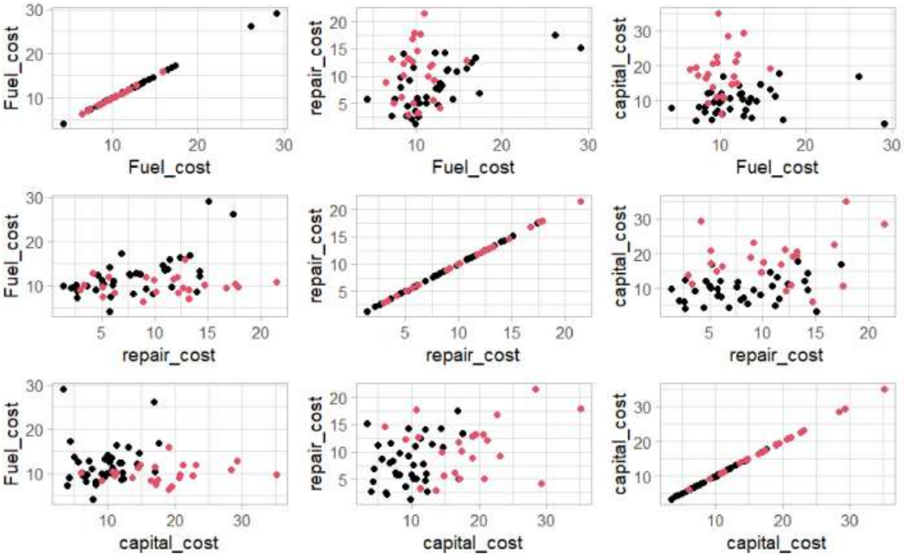


From the above plot we get the presence of 2 outliers in the total dataset of Fuel cost and repair cost. Also we get some outliers in the grouped dataset.

Similarly ,from the below plot we get the presence of 2 outliers in the total dataset

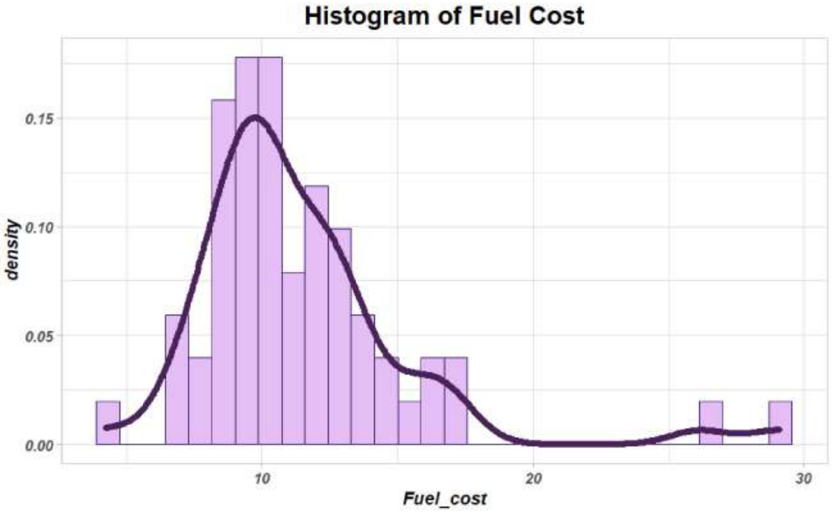


of Repair cost and Capital cost. Also we get some outliers in the grouped dataset.

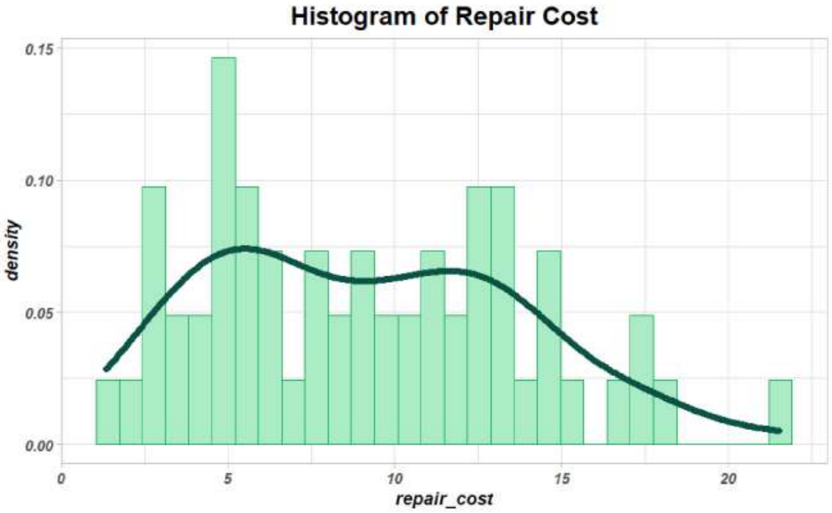


From these plot we get that, 3 covariates Fuel cost, Repair cost and capital cost are uncorrelated.

3.2 Histogram

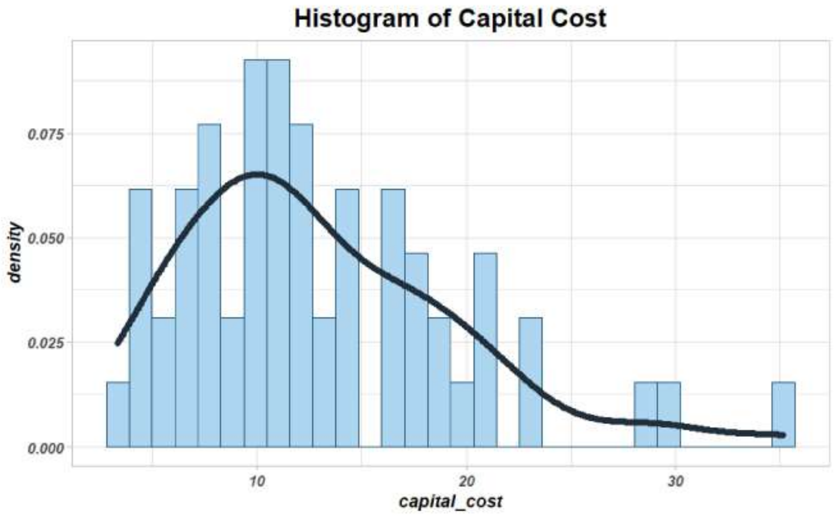


From the above plot we get that the data of fuel cost is positively skewed . It may due to presence of outliers.

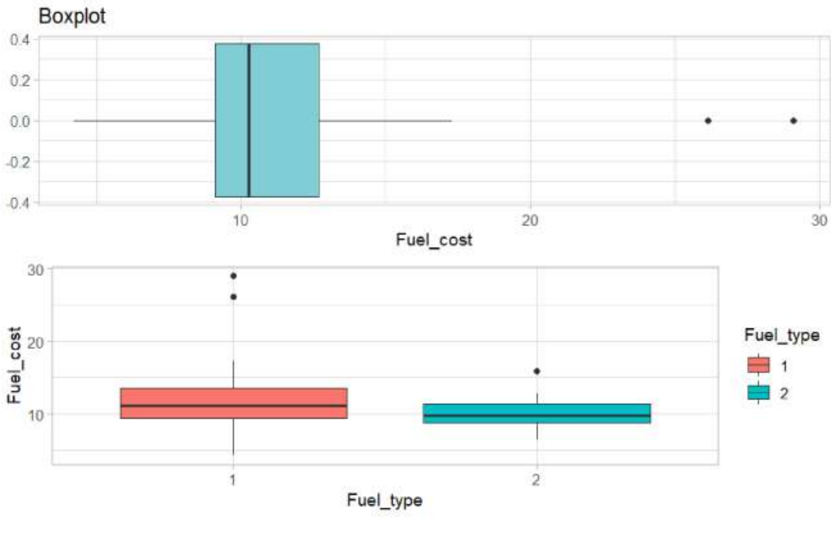


From the above plot we get that the data of repair cost is bimodal.

Similar with the fuel cost we get that, from the below plot, the data of capital cost is positively skewed . It may due to presence of outliers.



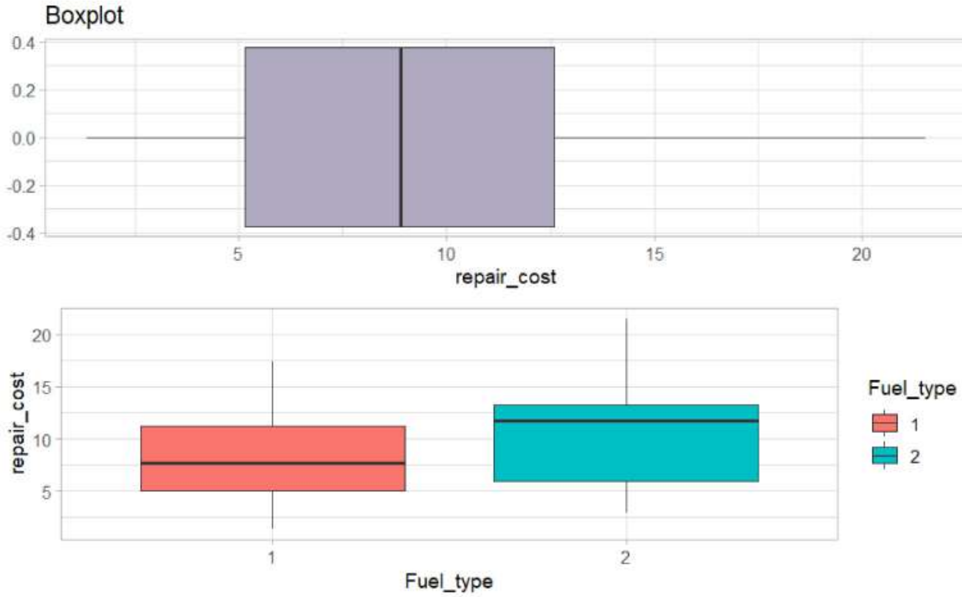
3.3 Box Plot and Outlier Detection



Also from the box plot we get a clear indication of the presence of 2 outliers in the Fuel cost data of the total dataset.There are also 2 outlier when fuel type 1 and

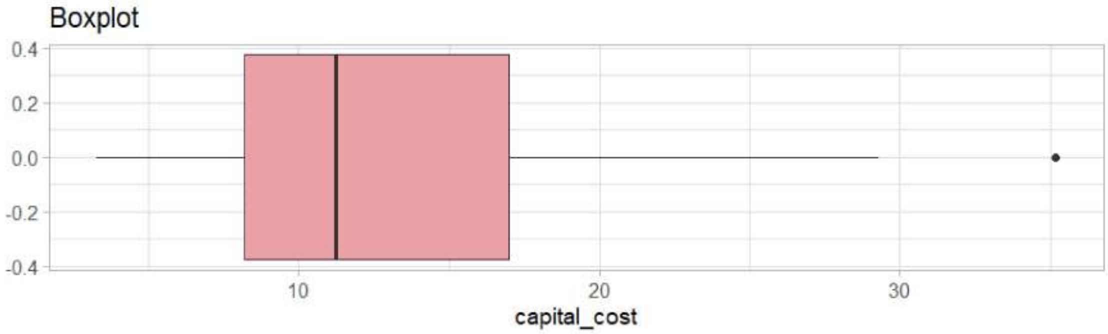
1 outlier when fuel type is 2.
• In the total dataset the fuel data is slightly positively skewed.
• Median of the fuel cost is high in the fuel type is 1 than the fuel type 2.

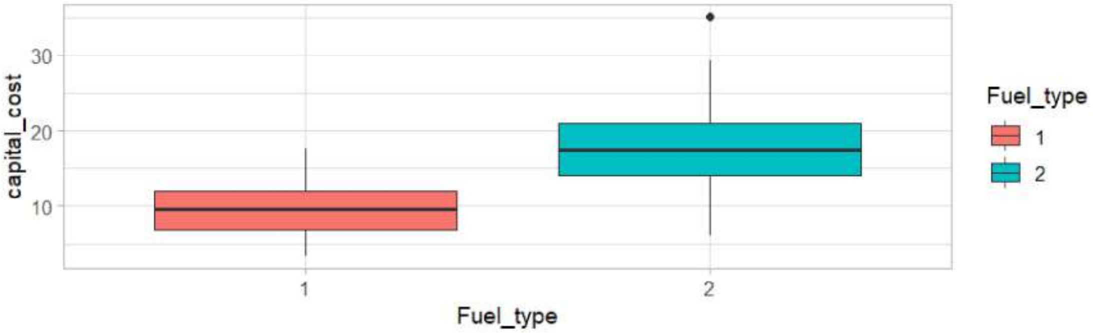
From the plot we get that repair cost has no outlier in both the total dataset and grouped dataset.



• In the total dataset repair cost is symmetric.
• In the case of Fuel type 1, data is symmetric but in the fuel type 2, data is negatively skewed. Average of the repair cost of the fuel type 2 is higher than the fuel type 1.

From the plot we get that capital cost has 1 outlier in both the total dataset and grouped dataset when fuel type is 2.
• In the total dataset capital cost is positively skewed.
• Average of the repair cost of the fuel type 2 is much higher than the fuel type 1.


Boxplot
0.4
0.2
0.0
-0.2
-0.4
10
20
30
capital_cost


capital_cost
30
20
10
1
2
Fuel_type
Fuel_type
1
2

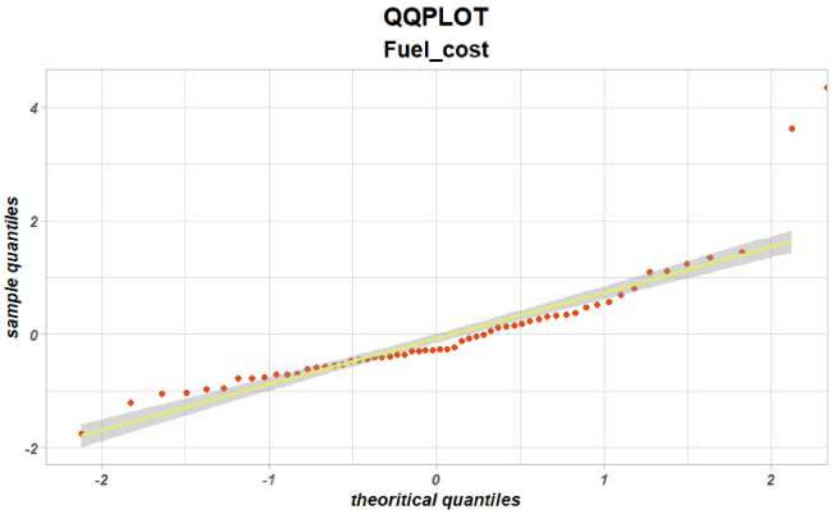
Chapter 4

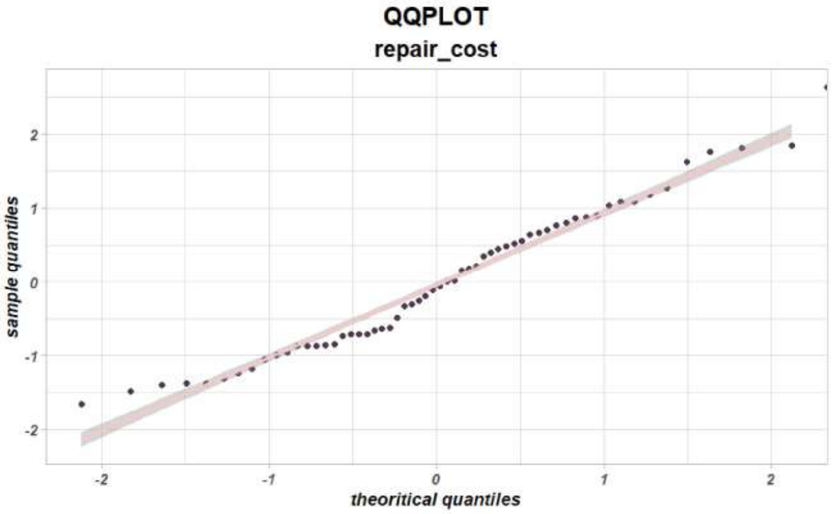
Checking Normality and Transformation of Variables

4.1 QQPlot

To check normality , we want to see Quantile-quantile plot first .

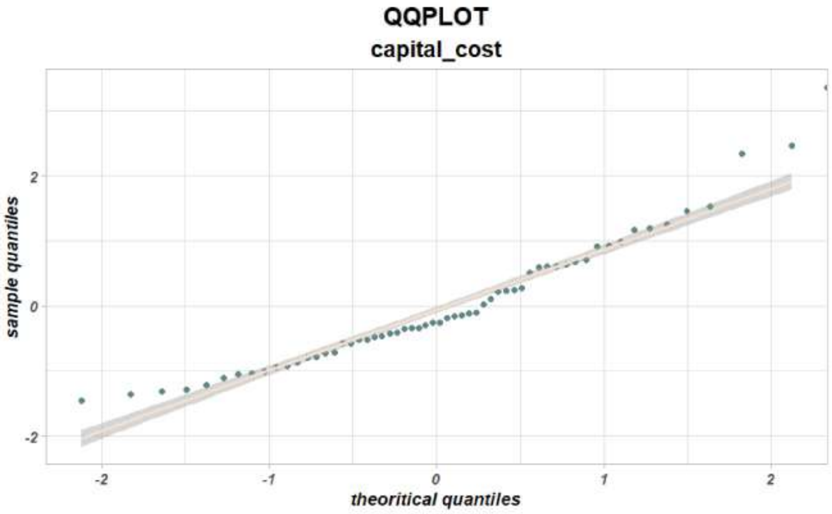
From the QQplot below, for the variable Fuel cost , it is observed that the distribution seems to be less normal .





From the QQplot above for the variable Repair cost , it is observed that the distribution seems to be also less normal .

From the QQplot below for the variable Capital cost , it is observed that the distribution seems to be also less normal .



4.2 Shapiro Wilk Test

Now, we want to check wheather there is a univariate normality in each of the three variables fuel cost, repair cost, capital cost . So, we are doing shapiro wilk normality test .

Shapiro Wilk Test Description :-

Assumption :- Here , $X_1, X_2, ..., X_n$ are random sample from a population .

Null Hypothesis :- $X_1, X_2, ..., X_n$ are i.i.d $N(\mu, \sigma^2)$ for some μ, σ^2 .

Test statistic :- W $= \frac{(\sum_{i=1}^{n} a_i X_{(i)})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

where , a $= \frac{m^T V}{\sqrt{m^T V^- 1 V^- 1 m}}$

with m and V the meaan vector and variance -covariance matrix of $(Z_{(1)}, ..., Z_{(n)})^T$

Rejection criteria :- Reject the null hypothesis for large values of the test statistic

.

Table 4.1: Shapiro Wilk

Variable	value of test statistic	p-value
Fuel Cost	0.81061	$2.99e-07$
Repair Cost	0.96621	0.1002
Capital Cost	0.92945	0.002082

From the above table of the Shapiro Wilk test , it is observed that p value for Fuel cost is very small and p value for repair cost and capital cost are 0.1002 and 0.002082 respectively . So, we reject the null hypothesis that they follow Normal distribution for Fuel Cost and capital cost at 5 percent level of significance but accept the null hypothesis for Repair cost .

We want to check wheather Multivariate Normality is present in the dataset or not . So ,we use mvn function in R.It gives Multivariate normality test as well as univariate normality.Here test statistics are Henze Zirklers and Anderson Darling respectively.

4.3 Anderson Darling Test

The Anderson darling (AD) test is a measure of how well a data fits a specified distributions

Assumption :- Here , $X_1, X_2, ..., X_n$ are random sample from a population .

Null Hypothesis :- The data comes from a specified distrubutions .

Alternate Hypothesis :- The data does not come from the specified distributions .

Test statistic :- AD = $-n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[lnF(X_i) + ln(1 - F(X_{n-i+1}))$

where , n = the sample size ; F(X) = the cdf of the specified distributions ; i = the i th sample , calculated when the data is sorted in ascending order .

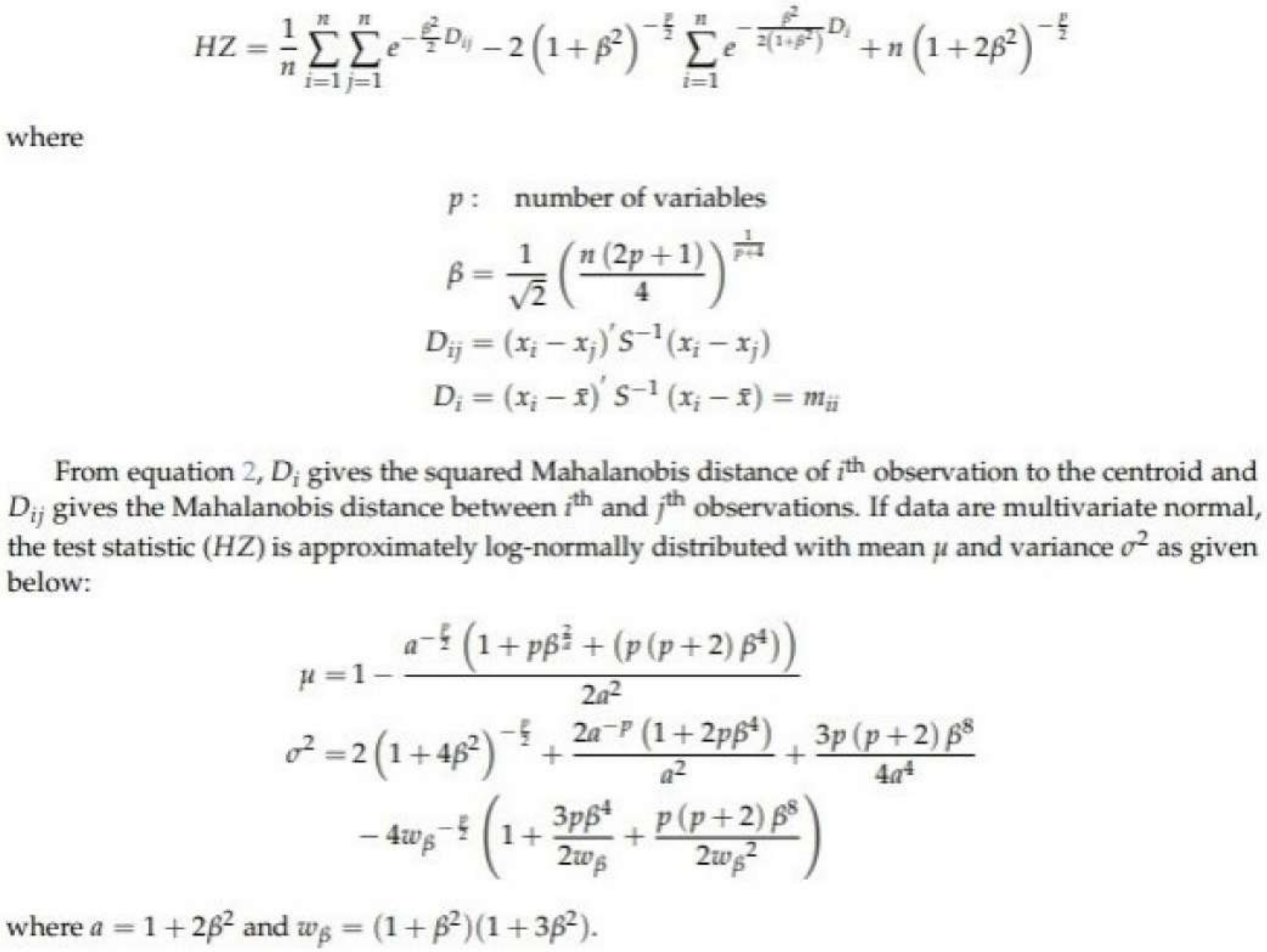
Decision :- Then , find the p value of the test statistic and if p value are less , we reject the null hypothesis .

Table 4.2: MVN Anderson Darling

Variable	value of test statistic	p-value
Fuel Cost	2.5898	< 0.001
Repair Cost	0.6070	0.1096
Capital Cost	1.0195	0.0102

From the above table of the Anderson Darling test , it is observed that p value for Fuel cost is less than 0.001 and p value for repair cost and capital cost are 0.1096 and 0.0102 respectively . So, we reject the null hypothesis that they follow Normal distribution for Fuel Cost and capital cost at 5 percent level of significance but accept the null hypothesis for Repair cost .

4.4 Henze Zirkler Test



$$HZ = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} e^{-\frac{\beta^2}{2}D_{ij}} - 2\left(1+\beta^2\right)^{-\frac{p}{2}} \sum_{i=1}^{n} e^{-\frac{\beta^2}{2(1+\beta^2)}D_i} + n\left(1+2\beta^2\right)^{-\frac{p}{2}}$$

where

$$\begin{aligned} p:&\quad \text{number of variables} \\ \beta &= \frac{1}{\sqrt{2}}\left(\frac{n\,(2p+1)}{4}\right)^{\frac{1}{p+4}} \\ D_{ij} &= (x_i - x_j)' S^{-1} (x_i - x_j) \\ D_i &= (x_i - \bar{x})' \, S^{-1} \, (x_i - \bar{x}) = m_{ii} \end{aligned}$$

From equation 2, D_i gives the squared Mahalanobis distance of i^{th} observation to the centroid and D_{ij} gives the Mahalanobis distance between i^{th} and j^{th} observations. If data are multivariate normal, the test statistic (HZ) is approximately log-normally distributed with mean μ and variance σ^2 as given below:

$$\begin{aligned} \mu &= 1 - \frac{a^{-\frac{p}{2}}\left(1 + p\beta^{\frac{2}{a}} + \left(p\,(p+2)\,\beta^4\right)\right)}{2a^2} \\ \sigma^2 &= 2\left(1+4\beta^2\right)^{-\frac{p}{2}} + \frac{2a^{-p}\left(1+2p\beta^4\right)}{a^2} + \frac{3p\,(p+2)\,\beta^8}{4a^4} \\ &\quad - 4{w_\beta}^{-\frac{p}{2}}\left(1 + \frac{3p\beta^4}{2w_\beta} + \frac{p\,(p+2)\,\beta^8}{2{w_\beta}^2}\right) \end{aligned}$$

where $a = 1 + 2\beta^2$ and $w_\beta = (1+\beta^2)(1+3\beta^2)$.

Table 4.3: MVN Henze Zirkler

value of test statistic	p-value
1.250848	0.00147892

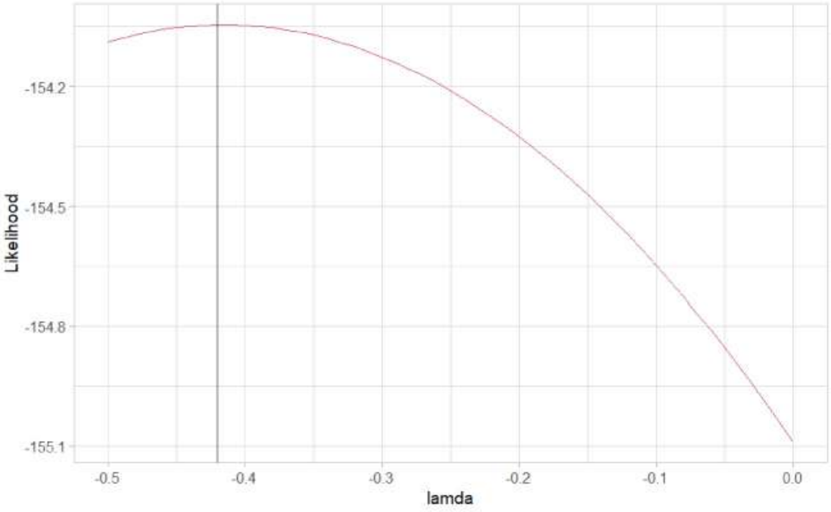
From the above table , it is observed that , the p value of the test is 0.00147892 , which is less than 0.05 . So , we reject the null hypothesis of the presence of multivariate normality in the dataset at 5 percent level of significance.

4.5 Transformation of the Variables

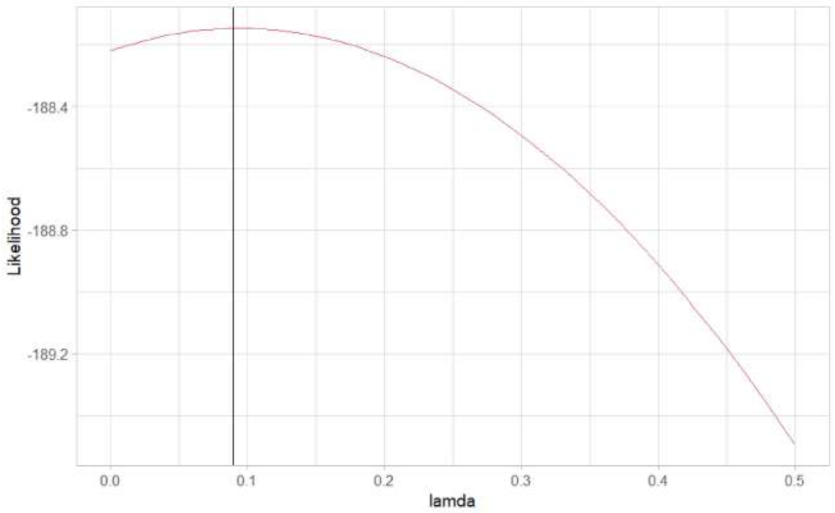
Now, our first work is to transform the variables to make each variable univariate normal and to make the data multivariate normal .

4.5.1 Power Transformation

Here , to make normality for the two variable Fuel cost and capital cost , we do power transformation to each variable . Now , first work is to choose λ and λ is chosen such that the log-likelihood is maximum.

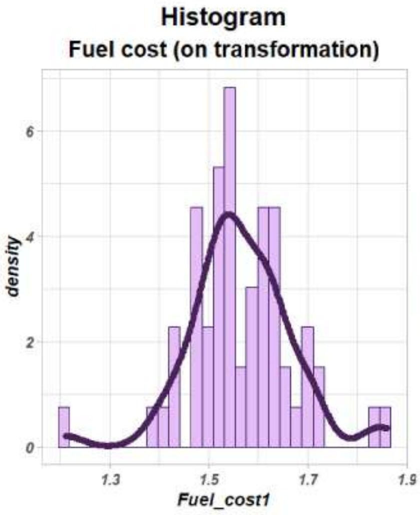


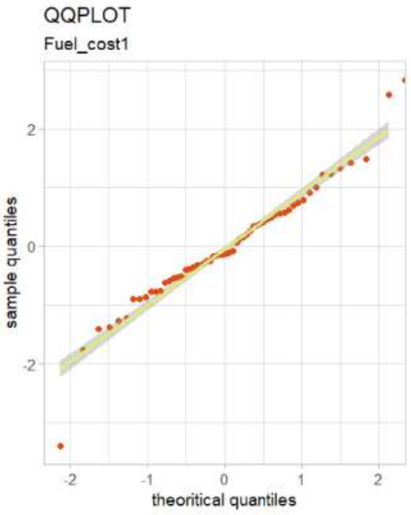
From the above plot , for the variable fuel cost, the value of λ is -0.4.

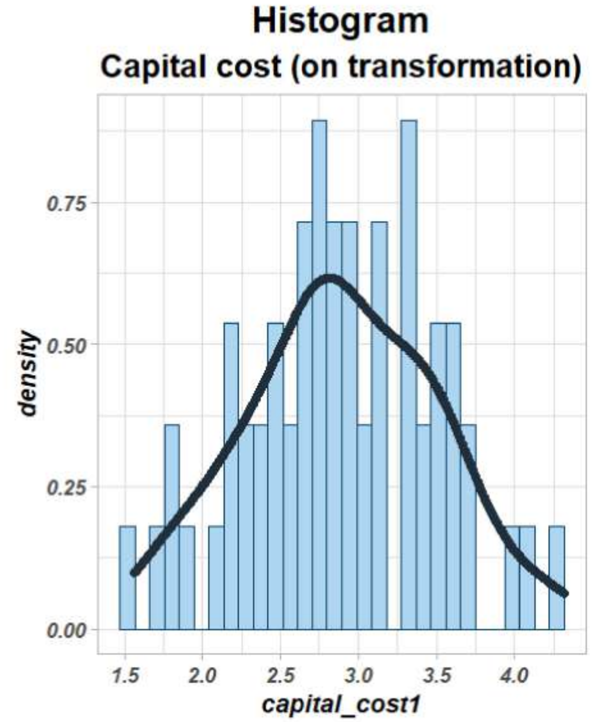


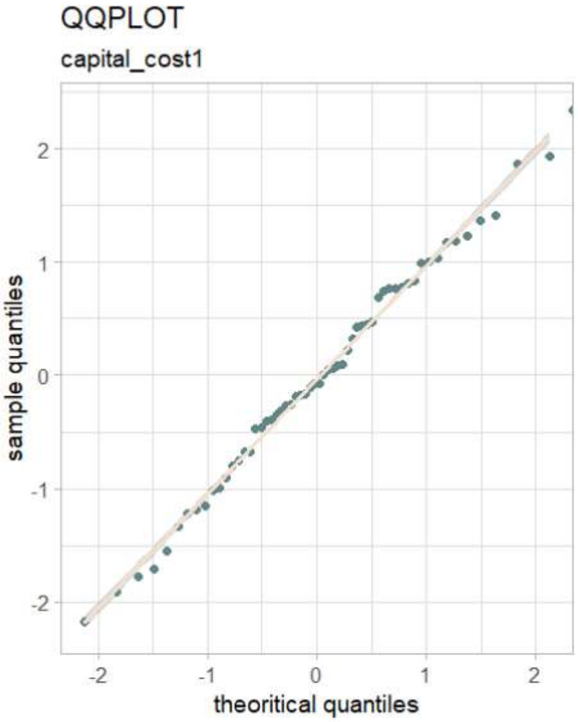
From the above plot , for the variable capital cost, the value of λ is 0.1.
So , we got the value of λ and change the variables by using Power Transformation .

From the below histogram and QQPlot ,it can be said that the transformed variable for fuel cost is more or less normal.









From the above histogram and QQPlot , it can be said that the transformed variable for capital cost is more or less normal .

Now , we want to check univariate normality and multivariate normality in the dataset by using mvn function in R.

Table 4.4: MVN Anderson Darling

Variable	value of test statistic	p-value
Fuel Cost	0.5610	0.1407
Repair Cost	0.6070	0.1096
Capital Cost	0.1464	0.9651

From the above table of the Anderson Darling test , it is observed that p value for Fuel cost, repair cost and capital cost are 0.1407, 0.1096 and 0.9651 respectively . So, we accept the null hypothesis that they follow Normal distribution for Fuel Cost and capital cost at 5 percent level of significance but accept the null hypothesis for Repair cost.

Table 4.5: MVN Henze Zirkler

value of test statistic	p-value
0.8714187	0.099428947

Again , From the above table for Henze Zirkler Test , it is observed that p value is 0.099428947 which is greater than 0.05 . So , we accept the null hypothesis of the presence of multivariate normality at 5 percent level of significance .

So, Now for the transformed variables in the data , univariate normality is present for each of the variables and multivariate normality is also present in the dataset

Chapter 5

Checking Homogeneity

5.1 BoxM test

Box's M test is a multivariate statistical test used to check the equality of multiple variance-covariance matrices.

Assumption :- $X_i(n_i * p)$ are $NDM_p(\mu_i, \Sigma_i)$ i = 1(1)k

Null Hypothesis :- all variance-covariance matrices are equal .

Alternate Hypothesis :- variance-covariance matrices are not equal .

Test statistic :- M = $\gamma[\sum_{i=1}^{k}(n_i - 1)log|S_{u_i}^{-1}S_u|]$

where , $S_u = \frac{nS}{n-k}$

$S_{u_i} = \frac{n_i S_i}{n_i - 1}$

$\gamma = 1$ - $\frac{2p^2+3p-1}{6(p+1)(k-1)}[\sum_{i=1}^{k}(\frac{1}{n_i-1} - \frac{1}{n-k})]$
Asymptotic distribution(under H_0 : Chi-square with df = $\frac{1}{2}p(p+1)(k-1)$

Decision :- Then , find the p value of the test statistic and if p value are less , we reject the null hypothesis.

Table 5.1: BoxM Test

value of chi.sq test statistic	df	p-value
10.633	6	0.1004

From the test we get the p-value 0.1004. So, we accept $H_0 : \Sigma_1 = \Sigma_2$ at 0.05 level of significance.

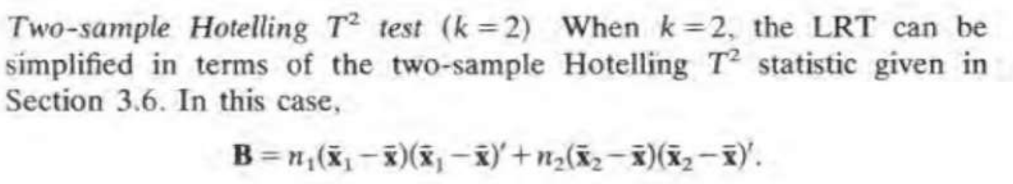
Chapter 6

MANOVA

To test $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ vs $H1 : \mu_i \neq \mu_j$ for some i,j when the covariance matrix of the k populations are equal.

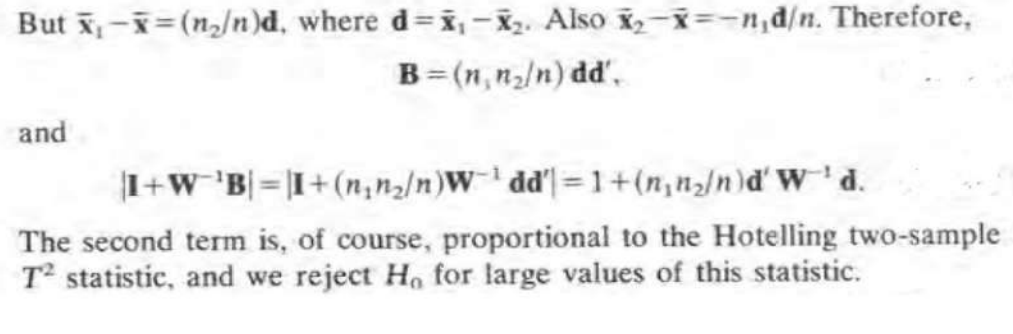
In our case, we are interested in analysis of two population having Multivariate Normal Distribution with equal variance- covariance matrix Σ.

It can be shown that the LRT statistic for the above testing when we have two population is equivalent to two sample Hotelling T^2.



Two-sample Hotelling T^2 test $(k=2)$ When $k=2$, the LRT can be simplified in terms of the two-sample Hotelling T^2 statistic given in Section 3.6. In this case,

$$\mathbf{B} = n_1(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}})(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}})' + n_2(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}})(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}})'.$$



But $\bar{\mathbf{x}}_1 - \bar{\mathbf{x}} = (n_2/n)\mathbf{d}$, where $\mathbf{d} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$. Also $\bar{\mathbf{x}}_2 - \bar{\mathbf{x}} = -n_1\mathbf{d}/n$. Therefore,

$$\mathbf{B} = (n_1 n_2/n)\,\mathbf{dd}',$$

and

$$|\mathbf{I} + \mathbf{W}^{-1}\mathbf{B}| = |\mathbf{I} + (n_1 n_2/n)\mathbf{W}^{-1}\,\mathbf{dd}'| = 1 + (n_1 n_2/n)\mathbf{d}'\,\mathbf{W}^{-1}\,\mathbf{d}.$$

The second term is, of course, proportional to the Hotelling two-sample T^2 statistic, and we reject H_0 for large values of this statistic.

6.1 Hotelling's T square

Given $\Sigma_1 = \Sigma_2$
To test H_0: μ_1=μ_2 vs H_1: $\mu_1 \neq \mu_2$

Test Statistic : $D^2 = (n_1 n_2/n)(\bar{x_1} - \bar{x_2})' S_u^{-1} (\bar{x_1} - \bar{x_2})$
where $S_u = (n_1 S_1 + n_2 S_2)/(n-2)$, $n = n_1 + n_2$
Under $H_0 : D^2 follows T^2(p, n-2)$

Table 6.1: Hotelling T square

value of test statistic	numerator df	denominator df	p-value
47.032	3	55	$2.685e-07$

From the testing ,we get the p-value $2.685e-07$. So, we reject H_0: μ_1=μ_2 at 0.05 level of significance.

Chapter 7

ANOVA

The main hypothesis of interest here is that of 'no difference in treatment effects',

$$\mathcal{H}_0 : \tau_1 = \tau_2 = \cdots = \tau_t.$$

This hypothesis can be rephrased as

$$\mathcal{H}_0 : \begin{pmatrix} \tau_1 - \tau_2 \\ \tau_1 - \tau_3 \\ \vdots \\ \tau_1 - \tau_t \end{pmatrix} = \mathbf{0},$$

or as $\boldsymbol{A\beta} = \mathbf{0}$ where

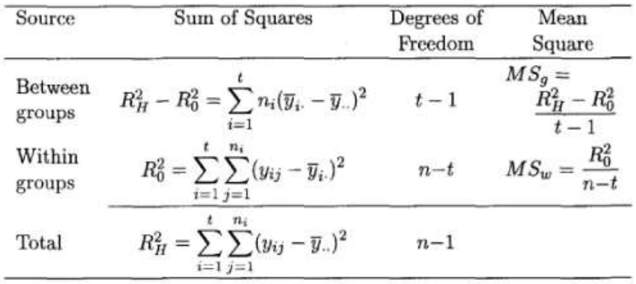
$$\boldsymbol{A}_{(t-1)\times(t+1)} = \begin{pmatrix} 0 & 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 1 & 0 & \cdots & 0 & -1 \end{pmatrix}.$$

The error sum of squares under the hypothesis (R_H^2) is very easy to calculate, since the restriction $\boldsymbol{A\beta} = \mathbf{0}$ reduces (6.2.1) to a model with common mean of all the observations. Therefore,

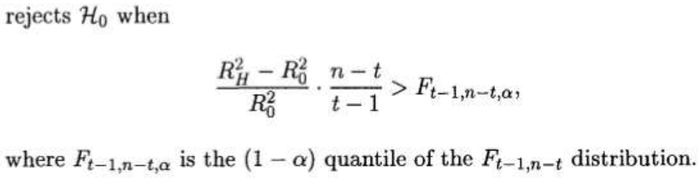
$$\begin{aligned} R_H^2 &= \min_{\boldsymbol{\beta}\,:\,\boldsymbol{A\beta} = \mathbf{0}} \|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 = \min_{\mu,\,\tau_1} \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \mu - \tau_1)^2 \\ &= \min_{\theta} \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \theta)^2 = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{..})^2. \end{aligned}$$

Thus, R_H^2 is the sum of squared deviations from the grand mean of all the observations. We can also find an interpretable expression for $R_H^2 - R_0^2$, as follows.

$$\begin{aligned} R_H^2 &= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{..})^2 = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i\cdot} + \overline{y}_{i\cdot} - \overline{y}_{..})^2 \\ &= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i\cdot})^2 + \sum_{i=1}^{t}\sum_{j=1}^{n_i}(\overline{y}_{i\cdot} - \overline{y}_{..})^2, \end{aligned}$$



Source	Sum of Squares	Degrees of Freedom	Mean Square
Between groups	$R_H^2 - R_0^2 = \sum_{i=1}^{t} n_i(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2$	$t-1$	$MS_g = \frac{R_H^2 - R_0^2}{t-1}$
Within groups	$R_0^2 = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i\cdot})^2$	$n-t$	$MS_w = \frac{R_0^2}{n-t}$
Total	$R_H^2 = \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{\cdot\cdot})^2$	$n-1$	

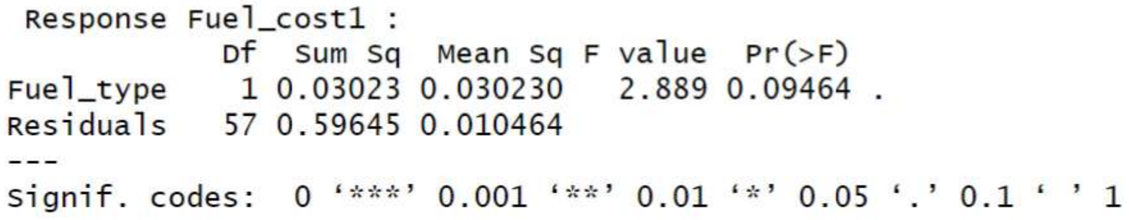


rejects $\mathcal{H}_0$ when

$$\frac{R_H^2 - R_0^2}{R_0^2} \cdot \frac{n-t}{t-1} > F_{t-1,n-t,\alpha},$$

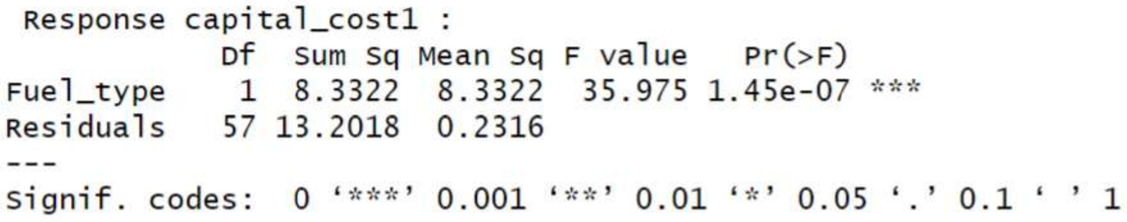
where $F_{t-1,n-t,\alpha}$ is the $(1-\alpha)$ quantile of the $F_{t-1,n-t}$ distribution.

Here we do anova individually for each variable taking as response with fuel type as treatment.



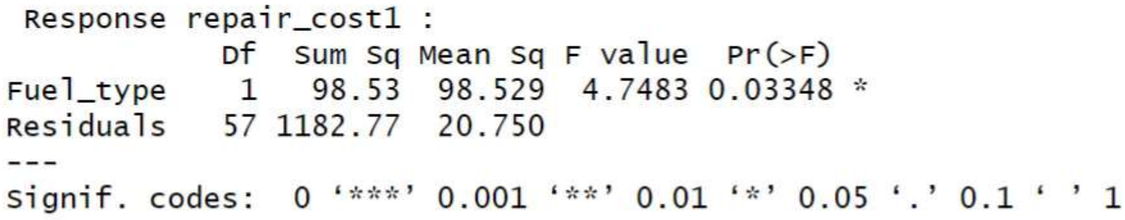
```
 Response Fuel_cost1 :
           Df  Sum Sq  Mean Sq F value  Pr(>F)  
Fuel_type   1 0.03023 0.030230   2.889 0.09464 .
Residuals  57 0.59645 0.010464                  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here , P-value is 0.09464.So we accept H_0 at 0.05 level of significance. So, there is no effect of fuel type in fuel cost.



```
 Response capital_cost1 :
            Df  Sum Sq Mean Sq F value   Pr(>F)    
Fuel_type    1  8.3322  8.3322  35.975 1.45e-07 ***
Residuals   57 13.2018  0.2316                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Here , P-value is $1.45e-07$.So we reject H_0 at 0.05 level of significance. So, there is a effect of fuel type in capital cost.



```
 Response repair_cost1 :
            Df  Sum Sq Mean Sq F value  Pr(>F)  
Fuel_type    1   98.53  98.529  4.7483 0.03348 *
Residuals   57 1182.77  20.750                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Here , P-value is 0.03348 .So we reject H_0 at 0.05 level of significance. So, there is a effect of fuel type in repair cost.

Chapter 8

PCA

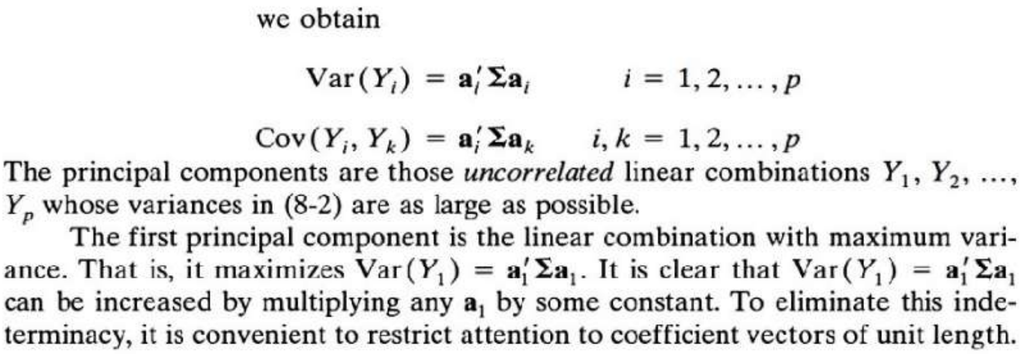
Algebraically, principal components are particular linear combinations of the p random variables $X_1, X_2, \ldots, X_p$. Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system with $X_1, X_2, \ldots, X_p$ as the coordinate axes. The new axes represent the directions with maximum variability and provide a simpler and more parsimonious description of the covariance structure.

As we shall see, principal components depend solely on the covariance matrix $\boldsymbol{\Sigma}$ (or the correlation matrix $\boldsymbol{\rho}$) of $X_1, X_2, \ldots, X_p$. Their development does not require a multivariate normal assumption. On the other hand, principal components derived for multivariate normal populations have useful interpretations in terms of the constant-density ellipsoids. Further, inferences can be made from the sample components when the population is multivariate normal. (See Section 8.5.)

Let the random vector $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ have the covariance matrix $\boldsymbol{\Sigma}$ with eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p \geqslant 0$.

Consider the linear combinations

$$\begin{aligned}
Y_1 &= \mathbf{a}_1'\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\
Y_2 &= \mathbf{a}_2'\mathbf{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\
&\vdots \qquad\qquad\qquad\qquad\qquad \vdots \\
Y_p &= \mathbf{a}_p'\mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p
\end{aligned}$$



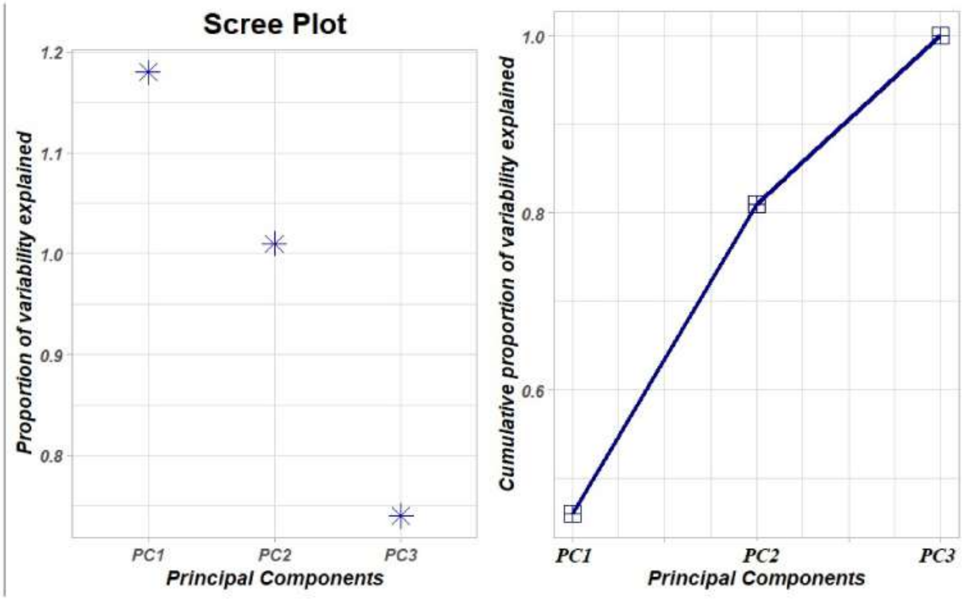
we obtain

$$\mathrm{Var}(Y_i) = \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_i \qquad i = 1, 2, \ldots, p$$

$$\mathrm{Cov}(Y_i, Y_k) = \mathbf{a}_i'\mathbf{\Sigma}\mathbf{a}_k \qquad i, k = 1, 2, \ldots, p$$

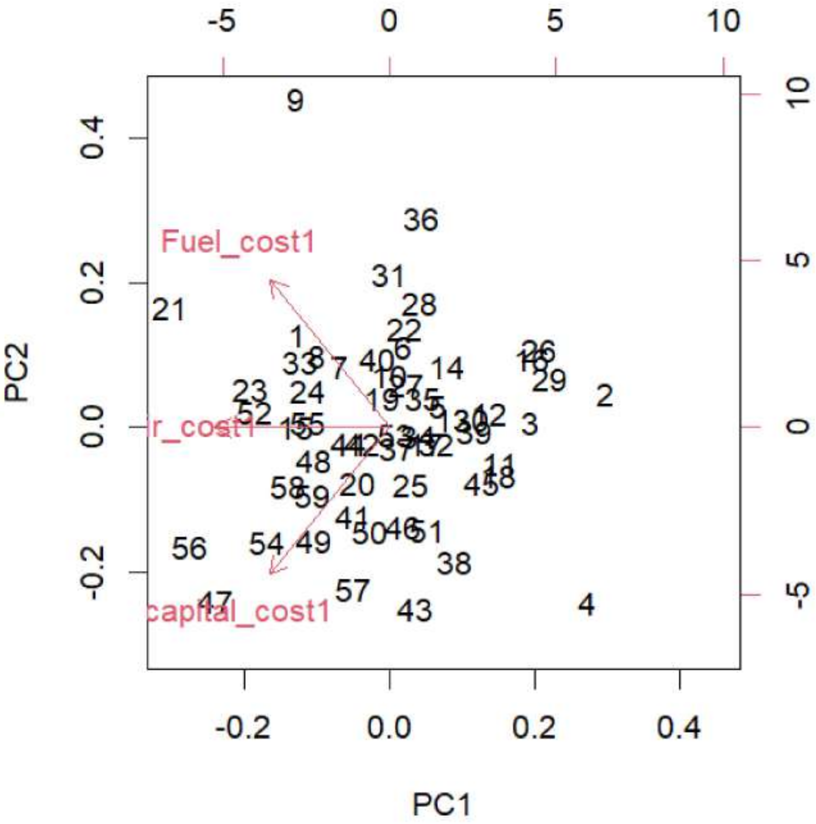
The principal components are those *uncorrelated* linear combinations Y_1, Y_2, ..., Y_p whose variances in (8-2) are as large as possible.

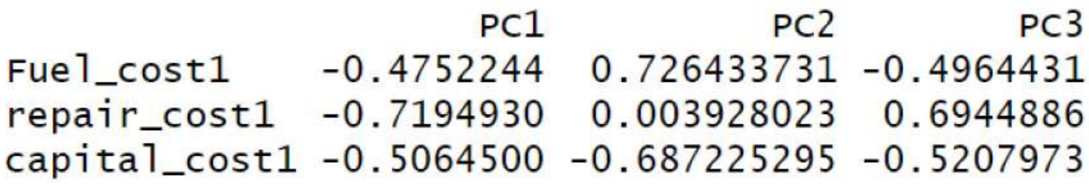
The first principal component is the linear combination with maximum variance. That is, it maximizes $\mathrm{Var}(Y_1) = \mathbf{a}_1'\mathbf{\Sigma}\mathbf{a}_1$. It is clear that $\mathrm{Var}(Y_1) = \mathbf{a}_1'\mathbf{\Sigma}\mathbf{a}_1$ can be increased by multiplying any $\mathbf{a}_1$ by some constant. To eliminate this indeterminacy, it is convenient to restrict attention to coefficient vectors of unit length.



From the scree plot we get that there is no so much variability between eigenvalues of the variance covariance matrix i.e. the variances of the principal component.

From the plot of cumulative proportions of eigen values ,we can explain only 80% data So we can't drop any one of the principal component.
Here we get that reapair cost has low contribution in PC2.





```
                     PC1          PC2        PC3
Fuel_cost1    -0.4752244  0.726433731 -0.4964431
repair_cost1  -0.7194930  0.003928023  0.6944886
capital_cost1 -0.5064500 -0.687225295 -0.5207973
```

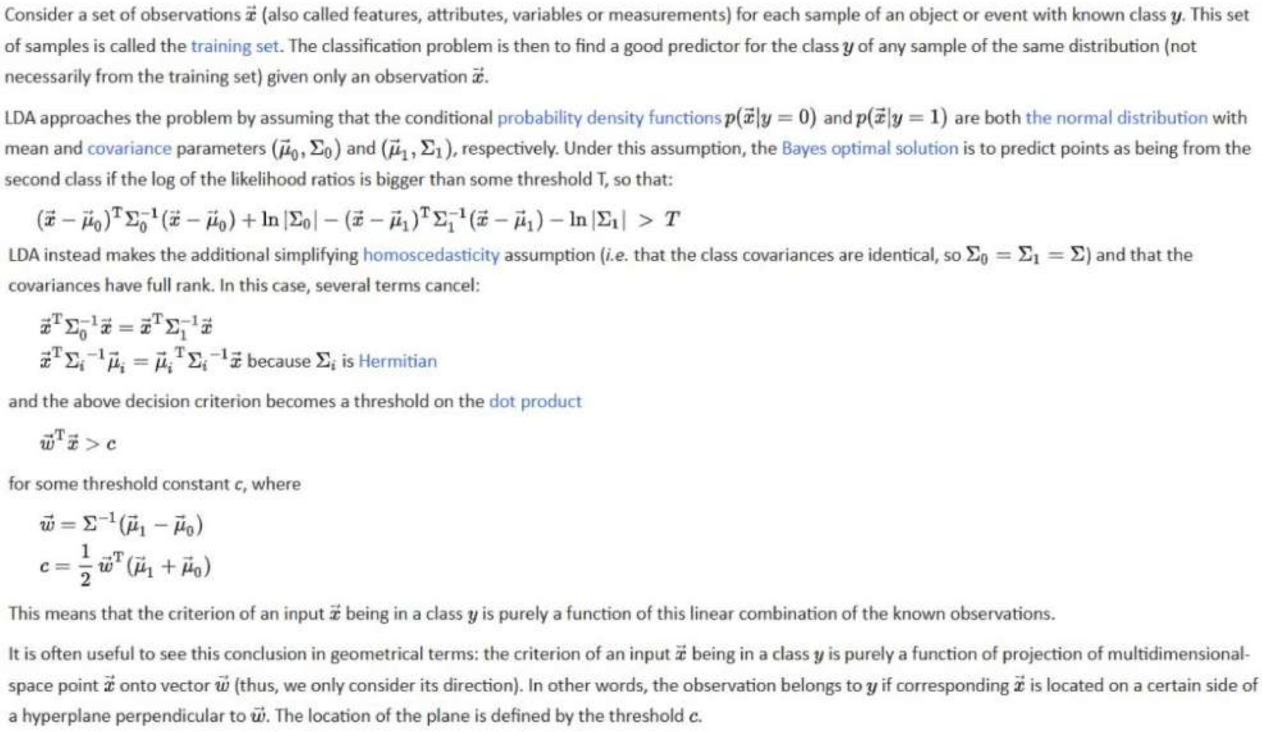
Chapter 9

Discriminant Analysis

Discriminant analysis is a technique that is used by the researcher to analyze the research data when the criterion or the dependent variable is categorical and the predictor or the independent variable is interval in nature. The term categorical variable means that the dependent variable is divided into a number of categories. The objective of discriminant analysis is to develop discriminant functions that are nothing but the linear combination of independent variables that will discriminate between the categories of the dependent variable in a perfect manner. It enables the researcher to examine whether significant differences exist among the groups, in terms of the predictor variables. It also evaluates the accuracy of the classification.

From the Box M test done previously , we observed the test for equality of covariance test gets accepted for the two population. In the light of the given data , we infer that the two population have same variance covariance matrix at 5

9.1 LDA



Consider a set of observations $\vec{x}$ (also called features, attributes, variables or measurements) for each sample of an object or event with known class y. This set of samples is called the training set. The classification problem is then to find a good predictor for the class y of any sample of the same distribution (not necessarily from the training set) given only an observation $\vec{x}$.

LDA approaches the problem by assuming that the conditional probability density functions $p(\vec{x}|y=0)$ and $p(\vec{x}|y=1)$ are both the normal distribution with mean and covariance parameters $(\vec{\mu}_0, \Sigma_0)$ and $(\vec{\mu}_1, \Sigma_1)$, respectively. Under this assumption, the Bayes optimal solution is to predict points as being from the second class if the log of the likelihood ratios is bigger than some threshold T, so that:

$$(\vec{x}-\vec{\mu}_0)^{\mathrm{T}}\Sigma_0^{-1}(\vec{x}-\vec{\mu}_0)+\ln|\Sigma_0|-(\vec{x}-\vec{\mu}_1)^{\mathrm{T}}\Sigma_1^{-1}(\vec{x}-\vec{\mu}_1)-\ln|\Sigma_1| \; > \; T$$

LDA instead makes the additional simplifying homoscedasticity assumption (*i.e.* that the class covariances are identical, so $\Sigma_0=\Sigma_1=\Sigma$) and that the covariances have full rank. In this case, several terms cancel:

$$\vec{x}^{\mathrm{T}}\Sigma_0^{-1}\vec{x}=\vec{x}^{\mathrm{T}}\Sigma_1^{-1}\vec{x}$$

$\vec{x}^{\mathrm{T}}{\Sigma_i}^{-1}\vec{\mu}_i={\vec{\mu}_i}^{\mathrm{T}}{\Sigma_i}^{-1}\vec{x}$ because Σ_i is Hermitian

and the above decision criterion becomes a threshold on the dot product

$$\vec{w}^{\mathrm{T}}\vec{x}>c$$

for some threshold constant *c*, where

$$\vec{w}=\Sigma^{-1}(\vec{\mu}_1-\vec{\mu}_0)$$
$$c=\frac{1}{2}\vec{w}^{\mathrm{T}}(\vec{\mu}_1+\vec{\mu}_0)$$

This means that the criterion of an input $\vec{x}$ being in a class y is purely a function of this linear combination of the known observations.

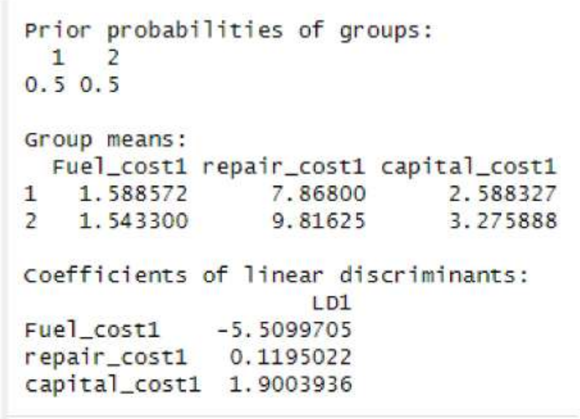
It is often useful to see this conclusion in geometrical terms: the criterion of an input $\vec{x}$ being in a class y is purely a function of projection of multidimensional-space point $\vec{x}$ onto vector $\vec{w}$ (thus, we only consider its direction). In other words, the observation belongs to y if corresponding $\vec{x}$ is located on a certain side of a hyperplane perpendicular to $\vec{w}$. The location of the plane is defined by the threshold *c*.

Here, first , we make a train dataset taking samples from each group (for Gasoline and diesel). And then fit a linear discriminant analysis to the train dataset , taking equal prior probabilities (0.5,0.5).

• Prior Probabilities of Groups :- These represents the proportion of each fuel type in the training set which is (0.5,0.5) here .

• Group means :- These display the mean values for each variables Fuel cost , Repair cost and capital cost for each Fuel type - gasoline and diesel .

• Coefficients of Linear Discriminants :- These display the linear combination of the variables Fuel cost , Repair cost and Capital Cost which is used to form the decision rule . Here for our data , the linear combination we got is :- -5.5099705*Fuel Cost +0.1195022*Repair Cost + 1.90036*Capital Cost.



```
Prior probabilities of groups:
  1   2 
0.5 0.5 

Group means:
  Fuel_cost1 repair_cost1 capital_cost1
1   1.588572      7.86800      2.588327
2   1.543300      9.81625      3.275888

Coefficients of linear discriminants:
                     LD1
Fuel_cost1    -5.5099705
repair_cost1   0.1195022
capital_cost1  1.9003936
```

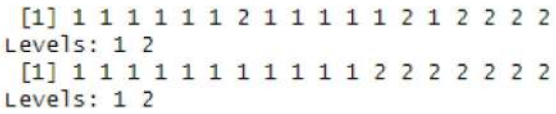
Figure 9.1:

9.1.1 APER

APER, Obtained by Hold one out Cross Validation Method –

Apparent Error Rate(APER) can be defined as the fraction of observations in the training set that are misclassified by the sample classification function. In out context , we obtain the Apparent Error Rate by -We split the total sample into training sample and validation sample. The training sample is used to find the classification rule and the validation sample is used to evaluate it. Since, here we have small sample size, it is not a good idea to split into train and validation set. A reasonable approach for this problem is the leave-one out cross validation.

Here , we used this lda model to validation dataset to evaluate the model .



```
 [1] 1 1 1 1 1 1 2 1 1 1 1 1 2 1 2 2 2 2
Levels: 1 2
 [1] 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2
Levels: 1 2
```

Figure 9.2:

And also, we want to find the misclassification error rate i.e. Apparent Error Rate(APER). We got the APER i.e. misclassification error rate is **0.1864407**.

Chapter 10

Logistic Regression

Here we use multinomial logistic regression. We want to predict the y(categorical response) the based on the several variable $X1, X2, X3$. If $P(Y = 1|X_1 = x_1, X_2 = x_2, X_3 = x_3) = P(\mathbf{X}) >$ some certain value then we assign the observation into the group $Y = 1$. We model the Probability as $logit(P(\mathbf{X}) = \beta_0 + \beta X$

We estimate the coefficients by maximum likelihood method.

10.0.1 Obtaining the estimates of the parameters

Table 10.1: Estimates of the parameters

Particulars	Estimate	Standard Error	z value	p value	
Intercept	3.1773	5.8686	0.541	0.588232	
Fuel cost	-11.6281	4.4999	-2.584	0.009764	**
repair cost	0.4682	0.3132	1.495	0.134972	
capital cost	4.2173	1.1861	3.556	0.000377	*

From the above table , we conclude that in the light of the given data Fuel Cost and Capital Cost are significant predictors for predicting Fuel Type. Using the estimate of the parameters , we next observe the fitted probabilities.

10.0.2 Fitted value of the probabilities

The above graph shows the fitted probabilities for the given values of the predictors and use these probabilities to obtain the confusion matrix taking threshold as the

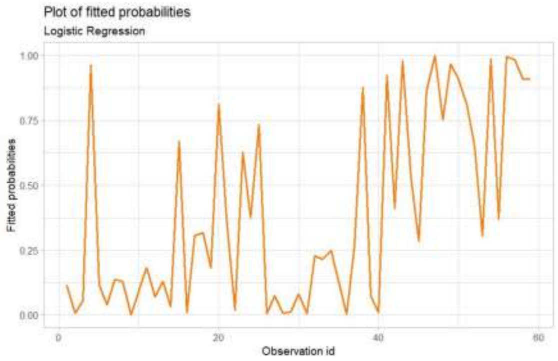


Figure 10.1:

median of the fitted probabilities.

10.0.3 Confusion Matrix

Type	Diesel(fitted)	Gasoline(Fitted)	Total
Diesel(actual)	27	9	36
Gasoline(actual)	2	21	23
Total	29	30	59

The confusion matrix shows that 11 out of 59 values are wrongly predicted. This leads to an error of 18 percent.But it is not an good idea to predict using the data used for training the model. So, we wish to study the misclassification error by method of hold out cross validation.

10.0.4 Misclassification Error : Cross validation

In the above graph , we observe change is misclassification error with change of threshold. Here we observe that for the value of threshold 0.267 and 0.76 , the misclassification error is lowest. So, if we are given a threshold , then we may work with that and observe the misclassification error, otherwise work with median of the fitted probabilities as threshold.

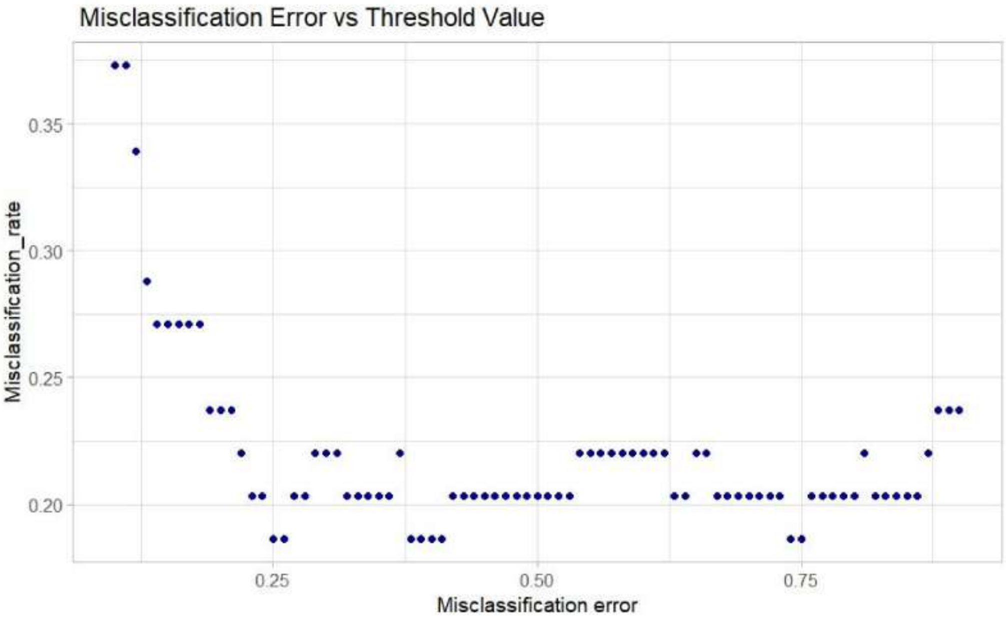
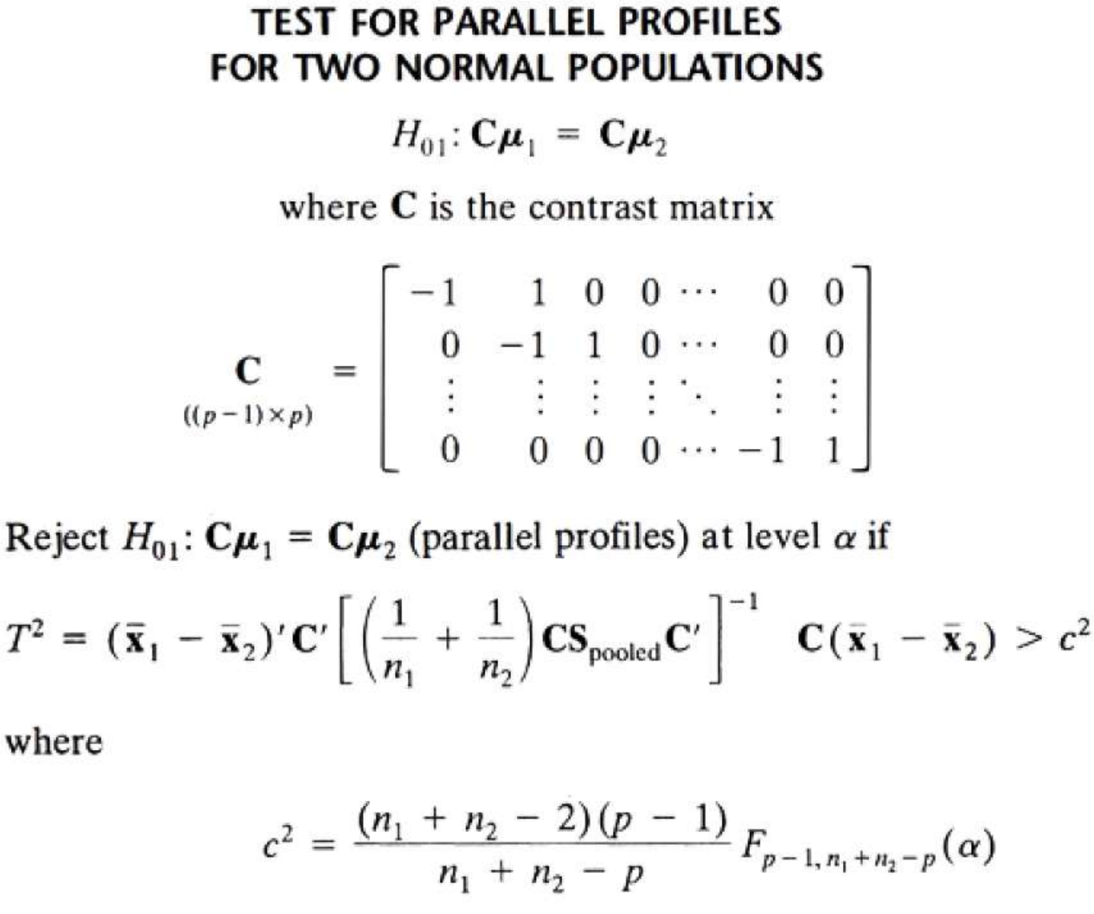

Misclassification Error vs Threshold Value
Misclassification_rate
0.35
0.30
0.25
0.20
0.25
0.50
0.75
Misclassification error

Figure 10.2:

Chapter 11

Profile Analysis



TEST FOR PARALLEL PROFILES FOR TWO NORMAL POPULATIONS

$$H_{01}: \mathbf{C}\boldsymbol{\mu}_1 = \mathbf{C}\boldsymbol{\mu}_2$$

where $\mathbf{C}$ is the contrast matrix

$$\underset{((p-1)\times p)}{\mathbf{C}} = \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}$$

Reject $H_{01}: \mathbf{C}\boldsymbol{\mu}_1 = \mathbf{C}\boldsymbol{\mu}_2$ (parallel profiles) at level α if

$$T^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{C}'\left[\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\mathbf{C}\mathbf{S}_{\text{pooled}}\mathbf{C}'\right]^{-1} \mathbf{C}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) > c^2$$

where

$$c^2 = \frac{(n_1 + n_2 - 2)(p - 1)}{n_1 + n_2 - p} F_{p-1, n_1+n_2-p}(\alpha)$$

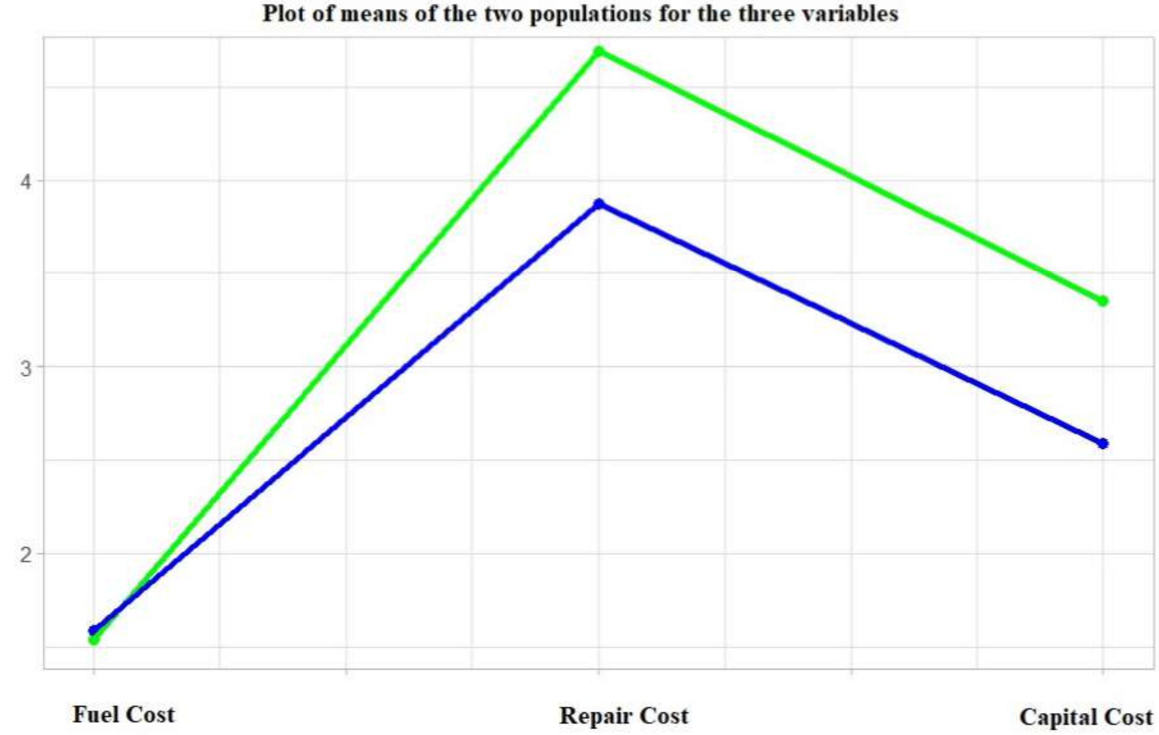


Table 11.1: Profile Analysis

value of test statistic	numerator df	denominator df	c^2	p-value
41.77441	2	56	6.436646	$3.669506e-12$

Here we reject H_0 at 0.05 level of significance. Profiles are not parallel which we also see that from the plot.

Chapter 12

Findings

• There are few outliers but we had not omitted it due to lesser number of data points .

• The variables fuel cost and repair cost deviate from Normality whereas Capital cost didn't . So, we used Box-Cox transformation to achieve normality for the first two variables .

• After Transformation, each of the variables achieved univariate normality and multivariate normality jointly .

• Here, we had two populations. The Box-M test shows the two populations are homogeneous .

• From MANOVA we can conclude in the light of the given data that the fuel type- gasoline and diesel have significant effect .

• From ANOVA we conclude in the light of the given data that the fuel type have significant effect on repair cost and capital cost . There is no as such effect of fuel type on fuel cost .

• Implementing Principal Component Analysis is not a good idea as 81% of the total variability is explained by the first two principal components where there are only three variables .

• On analysing linear discriminants and thereby classifying we obtain the estimate of the total misclassification probability approximately 0.186. .This is done by Hold-one-out cross validation since sample sizes are very small .

- From profile analysis, we conclude in the light of the given data that the two population profiles are not parallel . Hence, needless to check coincidence .

- Considering it to be a classification problem , on implementing logistic regression the estimate of the total probability of misclassification is approximately 0.18. This is done by Hold-one-out cross validation since sample sizes are very small .

Chapter 13

References

- Multivariate Analysis , Kantilal Varichand Mardia , J. M. Bibby , J. T. Kent

- Applied Multivariate Statistical Analysis , Richard A. Arnold , Dean W. Wichren

- An Introduction to Statistical Learning ,G. James , D. Witten, T. Hastie , R. Tibshirani

- Linear Models : An Integrated Approach , S. Rao Jammalamadak, Debasis Sengupta

- A new family of power transformations to improve normality or symmetry, IN KWON YEO, Department of Control and Instrumentation of Engineering, Kangwon National University,Chunchon, 200701, Korea and RICHARD A JOHNSON, Department of Statistics, University of Wisconsin Madison, Wisconsin 53706, U.S.A.

- MVN: An R Package for Assessing Multivariate Normality by Selcuk Korkmaz, Dincer Goksuluk and Gokmen Zararsiz