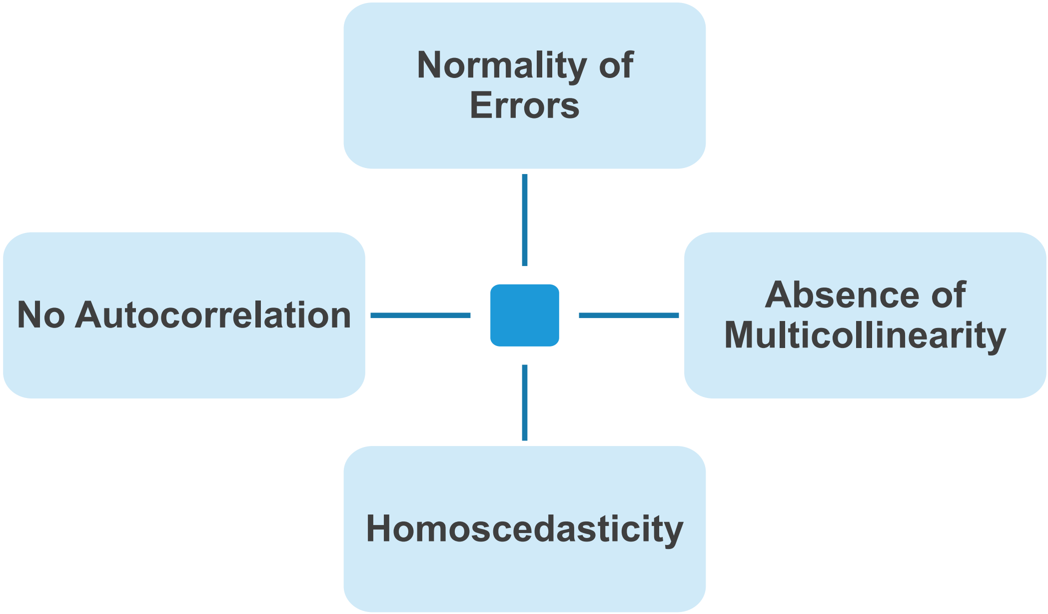
Multiple Linear Regression

Multicollinearity Problem

Contents

Key Assumptions of Multiple Linear Regression

Multiple Linear Regression **makes four key assumptions**



Violations of these assumptions may result in biased variable relationships, over or under-estimation of parameters (i.e. biased standard errors), and unreliable confidence intervals and significance tests

Problem of Multicollinearity

Multicollinearity exists if there is strong linear relationship among the independent variables

Multicollinearity has two serious consequences:

1. Highly Unstable Model Parameters

   As standard errors of their estimates are inflated

2. Model Fails to Accurately Predict for Out of Sample Data

Therefore, it is important to check for Multicollinearity in regression analysis

Indicators

- High Pairwise Correlation
  Among independent variables
- Significant F Value but Very Few Significant t Values for global testing and individual testing respectively

Detecting Multicollinearity Through VIF

VIF (Variance Inflation Factor) Method:

Dependent Variable : Y

Independent variables : X1, X2, X3, X4

Dependent Variable	Independent Variables	R^2	$1 - R^2$ = Tolerance	VIF = 1/(Tolerance)
X1	X2, X3, X4			
X2	X1, X3, X4			
X3	X1, X2, X4			
X4	X1, X2, X3			

Any VIF > 5, indicates presence of Multicollinearity

Detecting Multicollinearity in R

#Importing the Data, Fitting Linear Model

```
perindex<-read.csv("Performance Index.csv", header=TRUE)
jpimodel<-lm(jpi~aptitude+tol+technical+general, data=perindex)
```

#Variance Inflation Factor
#Install and load package “car”.

```
install.packages("car")
library(car)
```

car stands for Companion to Applied Regression and consists of several useful functions for advance regression analysis.

```
vif(jpimodel)
```

vif() in package car calculates VIFs.

* Continuing with same dataset “Performance Index”

Detecting Multicollinearity in R

```
# Output
```

```
aptitude       tol technical   general
1.179906  1.328205  2.073907  2.024968
```

Interpretation :
All VIFs are less than 5, Multicollinearity is not present.

Multicollinearity – Remedial Measures

The problem of Multicollinearity can be solved by different approaches:

- Drop one of the independent variables, which is explained by others
- Use Principal Component Regression in case of severe Multicollinearity
- Use Ridge Regression

*
Dropping a variable may not be a good idea if many VIFs are large.
Principal Component Method will be discussed in detail under Data Reduction and Segmentation

Case Study - Modelling Resale Price of Cars

Background

- A car garage has old cars for resale. They keep records for different models of cars and their specifications.

Objective

- To predict the resale price based on the information available about the engine size, horse power, weight and years of use of the cars

Available Information

- Records -26
- Independent Variables: ENGINE SIZE, HORSE POWER, WEIGHT AND YEARS
- Dependent Variable: RESALE PRICE

Data Snapshot

ridge regression data

Dependent variable

Independent variables

MODEL	RESALE PRICE	ENGINE SIZE	HORSE POWER	WEIGHT	YEARS
Daihatsu Cuore	3870	846	32	650	2.9
Suzuki Swift 1.0 GL	4163	993	39	790	2.9
Fiat Panda Mambo L	3490	899	29	730	3.1

Observations

Columns	Description	Type	Measurement	Possible values
MODEL	Model of the car	character	-	-
RESALE PRICE	Resale price	numeric	Euro	positive values
ENGINE SIZE	Size of the engine	numeric	cc	positive values
HORSE POWER	Power of the engine	numeric	kW	positive values
WEIGHT	Weight of the car	numeric	kg	positive values
YEARS	Number of years in use	numeric	-	positive values

Correlation Matrix

```
# Importing the Data
```

```
ridgedata<-read.csv("ridge regression data.csv", header=TRUE)
```

```
# Graphical representation of data
# Install and load package “GGally”
```

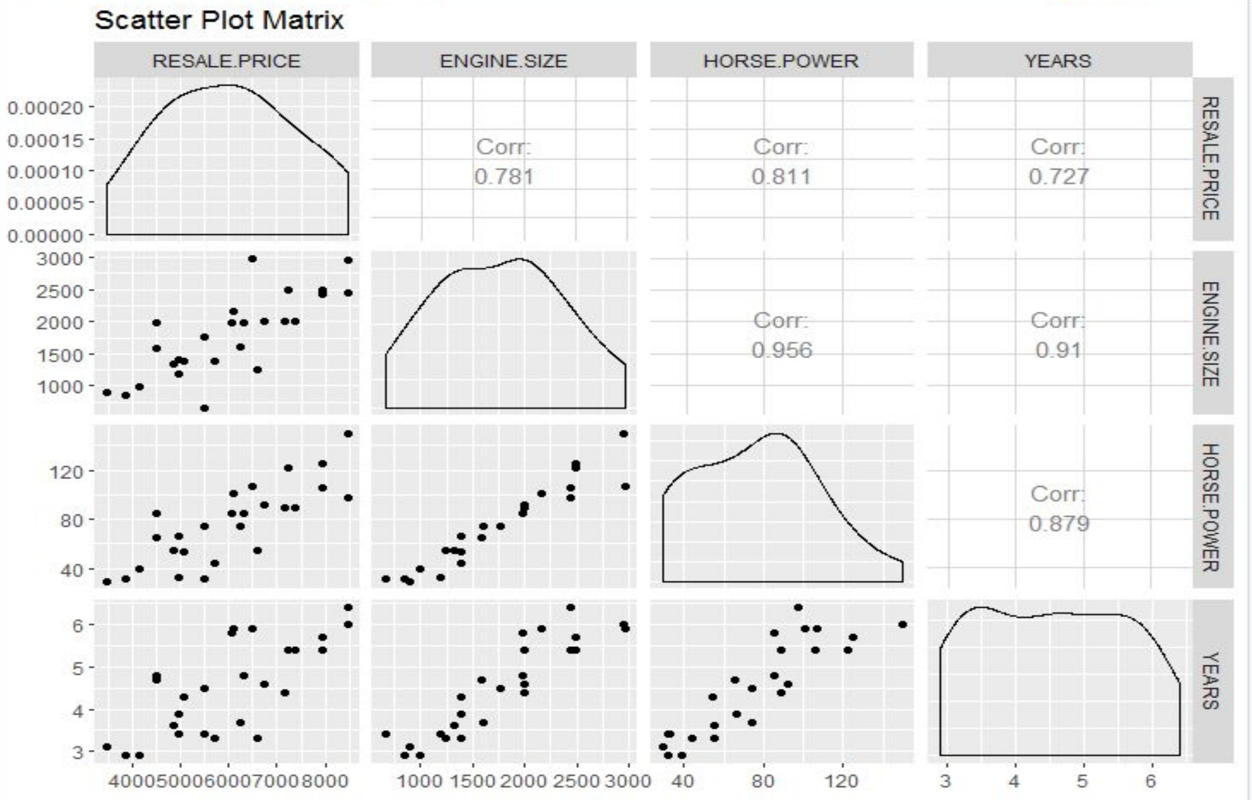
```
install.packages("GGally")
library(GGally)

ggpairs(ridgedata[,c("RESALE.PRICE","ENGINE.SIZE","HORSE.POWER","YEARS
")],title="Scatter Plot Matrix", columnLabels = c("RESALE.PRICE",
"ENGINE.SIZE","HORSE.POWER","YEARS"))
```

- *ggpairs() in the package GGally is used to plot the scatter plot matrix.*

Correlation Matrix

```
# Output
```



Interpretation :

- *The independent variables have high positive correlation among themselves .*

Detecting Multicollinearity in R

```
#Fitting Linear Model
```

```
model<-lm(RESALE.PRICE ~ ENGINE.SIZE +HORSE.POWER +WEIGHT+ YEARS,
data=ridgedata)
```

```
#Variance Inflation Factor
```

```
library(car)
vif(model)
```

```
# Output
```

```
ENGINE.SIZE HORSE.POWER      WEIGHT       YEARS
  15.759113   12.046734    9.113045   13.978640
```

Interpretation:

- *VIF values for all the variables are greater than 5, hence we can conclude that there exist Multicollinearity between the independent variables.*

Quick Recap

This session explained the **problem of Multicollinearity, along with its consequences and remedial measures:**

Multicollinearity Exists	• When independent variables have strong linear relationship
Results in	• Unstable model parameters • Inaccurate predictions for out of sample data
Indicators	• High pairwise correlation • Significant F value but very few significant t values
Checking in R	• Variance Inflation Factor **`vif()`** function in package **car**
Remedial Measures	• Drop variables • Use Principal Component Regression • Ridge regression