A BEGINNER'S GUIDE TO

FIRST DATA ANALYSIS PROJECT

USING LINEAR REGRESSION IN R

KANIKA TAYAL
ANUBHAV DUBEY

To

Our families without whom we are nothing
Our teachers who have given us the light to walk our paths
Our friends whose questions became the motivation for this book

Preface

Ours is not a better way, it is merely another way.
Neale Donald Walsch

When we started working on our first project, we were lost and lacked direction. We reached out to our seniors and searched on web about ways to do a project, both of which helped us immensely in getting started. However, we realised that the transition from knowledge based mostly in theory to research and application can be a daunting one if there hasn't been any prior attempt.

Since epistemology is best left to the branch of philosophy, we, for now, deal only with the application part.

This book is an attempt to provide a basic framework towards approaching the first project in Data Analysis. The pages that follow are neither theoretically deep nor exhaustive in methods; however, they are intended to act as a bridge between the theory and applications especially with respect to statistical techniques.

Data Science is an ever growing and ever evolving field. Hence, any journey into this wonderful realm can never fully cover all its aspects. However, it is important to begin because like all journeys, this will be exhilarating to say the least and we believe that this book will be a good starting point.

To great beginnings
Kanika Tayal
Anubhav Dubey

TABLE OF CONTENTS

Overview

Linear Regression: The technique of modelling a linear relationship between a response (dependent) variable and one or more explanatory (independent) variables is called linear regression.

The equation of multiple linear regression is given by:

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ \ldots \ + \beta_p X_p + \varepsilon}$$

where, Y is the response variable and X_i's (i = 1 to p) are explanatory variables and ε is the error term.

We estimate the (p+1) parameters $\beta_0, \beta_1, \ldots, \beta_p$ using the method of least squares, that is, by minimising the error sum of squares.

Assumptions:

1. **Normality of residuals:** ε are independent and identically distributed $N(0,\sigma^2)$ variates
2. **Linearity:** The relationship between X and Y is linear
3. **Homoscedasticity:** The variance of error is constant for any value of X
4. **Independence:** Observations are independent of each other (no multi-collinearity)

t-test for significance of parameters:

Let $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ be the estimated parameters. We now test the following hypothesis:

Null Hypothesis, H_0: $\beta_i = 0$ vs.
Alternate Hypothesis, H_1: $\beta_i \neq 0$

Test Statistic:

$$t = \frac{\widehat{\boldsymbol{\beta}_i}}{se(\widehat{\boldsymbol{\beta}_i})} \sim t_{(n-(p+1))}$$

where, se $(\hat{\beta}_i)$ is the standard error of estimated parameter

Test Criteria: Reject H_0 at α% level of significance if $|t| > t_{(\alpha/2)}$. Otherwise, do not reject H_0.

Confidence Interval:

100(1-α)% C.I. for β_i is given by $\left(\widehat{\beta}_i \pm t_{\frac{\alpha}{2}}\sqrt{se(\widehat{\beta}_i)}\right)$

During the course of this book, we will perform data analysis using linear regression on the Carseats dataset which is contained in ISLR package. The **steps** followed in this book are mentioned below:

1. Load the dataset

2. Univariate analysis – exploring the descriptive statistics of variables using methods of summarization and visualization

3. Bivariate analysis – to find associations or differences between two variables and measure the significance of such associations or differences
    3.1 Simple Linear regression along with regression diagnostics

4. Multivariate Analysis – to find association between one response and more than one explanatory variable
    4.1 Multiple Linear regression along with regression diagnostics
    4.2 Comparing models using adjusted R^2, AIC and ANOVA
    4.3 Stepwise regression
    4.4 Parallel slopes model
    4.5 Using Interaction terms for modelling

1. Dataset

We use the Carseats dataset which is a part of the ISLR package.

The dataset is simulated and it is described in the lines that follow:

It is in the format of a data frame with 400 observations on the following 11 variables:

1. Sales - Unit sales (in thousands) at each location
2. CompPrice - Price charged by competitor at each location
3. Income - Community income level (in thousands of dollars)
4. Advertising - Local advertising budget for company at each location (in thousands of dollars)
5. Population - Population size in region (in thousands)
6. Price - Price company charges for car seats at each site
7. ShelveLoc - A factor with levels Bad, Good and Medium indicating the quality of the shelving location for the car seats at each site
8. Age - Average age of the local population
9. Education - Education level at each location
10. Urban - A factor with levels No and Yes to indicate whether the store is in an urban or rural location
11. US – A factor with levels No and Yes to indicate whether the store is in US or not

1.1. Loading the dataset

Since the dataset is contained in the ISLR package, we first load the package and then the dataset.

Note: If your dataset is stored in a csv file then, you can use read.csv() function. You can also use read.table() command to read data from an excel file.

```
>library(ISLR)
```

Task 1: Store the dataset in another variable and examine its structure

```
>prodata <- Carseats
>attach(prodata)  #columns of the data frame can be accessed directly
```

str() function displays the internal structure of an object in R. It also displays first few observations of all the variables along with its datatype.

```
>str(prodata)

'data.frame':     400 obs. of  11 variables:
 $ Sales      : num  9.5 11.22 10.06 7.4 4.15 ...
 $ CompPrice  : num  138 111 113 117 141 124 115 136 132 132 ...
 $ Income     : num  73 48 35 100 64 113 105 81 110 113 ...
 $ Advertising: num  11 16 10 4 3 13 0 15 0 0 ...
 $ Population : num  276 260 269 466 340 501 45 425 108 131 ...
 $ Price      : num  120 83 80 97 128 72 108 120 124 124 ...
 $ ShelveLoc  : Factor w/ 3 levels "Bad","Good","Medium": 1 2 3 3 1 1 3 2 3
3 ...
 $ Age        : num  42 65 59 55 38 78 71 67 76 76 ...
 $ Education  : num  17 10 12 14 13 16 15 10 10 17 ...
 $ Urban      : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 1 1 ...
 $ US         : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 1 2 1 2 ...
```

Q1. Explain the structure of the dataset.

Ans. Carseats dataset contains 400 observations on 8 numeric and 3 factor variables. Each column represents a variable and each row corresponds to an observation on all these variables.

Q2. What are the levels of the three factor variables?

Ans. The variable ShelveLoc has 3 levels: Good, Medium and Bad. On the other hand, US and Urban have 2 levels each: Yes and No.

2. Univariate Analysis

a. What – As the name suggests, Univariate (uni – one) analysis is concerned with the analysis of a single variable.
b. Why – It is useful to describe the variable in question by summarizing it and trying to find some patterns indicative about its nature.
c. How – We now describe the ways in which we can perform this analysis.

Task 2: Summarise the data

Summary command has a wide-ranging application in R. When applied on a data frame, it returns mean along with quartiles of all the variables.

```
>summary(prodata)
```

```
   Sales          CompPrice       Income        Advertising
Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000
1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000
Median : 7.490   Median :125   Median : 69.00   Median : 5.000
Mean   : 7.496   Mean   :125   Mean   : 68.66   Mean   : 6.635
3rd Qu.: 9.320   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000
Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000

 Population        Price        ShelveLoc        Age          Education
Min.   : 10.0   Min.   : 24.0   Bad   : 96   Min.   :25.00   Min.   :10.0
1st Qu.:139.0   1st Qu.:100.0   Good  : 85   1st Qu.:39.75   1st Qu.:12.0
Median :272.0   Median :117.0   Medium:219   Median :54.50   Median :14.0
Mean   :264.8   Mean   :115.8                Mean   :53.32   Mean   :13.9
3rd Qu.:398.5   3rd Qu.:131.0                3rd Qu.:66.00   3rd Qu.:16.0
Max.   :509.0   Max.   :191.0                Max.   :80.00   Max.   :18.0

 Urban       US
No :118   No :142
Yes:282   Yes:258
```

Q3. Is there any evidence that the distributions of Sales and Price are somewhat symmetric?

Ans. As seen above, the mean and median values for these two variables are almost equivalent, indicating that their distributions are somewhat symmetric.

2.1. Distribution of variables

Task 3: Visualise the distributions of the variables Sales, Price, Advertising and Income with a stem-and-leaf plot and histogram.

Stem-and-leaf plot is a device that helps in analysing the distribution of a quantitative data in a graphical format. The idea is to make a special table where each data value is split into a "**stem**" (the first digit or digits) and a "**leaf**" (usually the last digit).

The **stem()** function in R allows us to obtain the stem-and-leaf plot.

```
>stem(Sales)
```

```
The decimal point is at the |

   0 | 02459
   1 | 48
   2 | 112357799
   3 | 0011224555666779999
   4 | 011111222223444445556667777788889999
   5 | 000000011111222233333333344444555666666677777889999999
   6 | 00000011122222244444444555555556666677777788999999999
   7 | 000002223344444444555555556666666677777778888888899
   8 | 00000011122222333344445556666677777777788888889999
   9 | 0000001111222333333444445555555566677
  10 | 000011111233334445555666677888
  11 | 0012222233355567778999
  12 | 00001133455566679
  13 | 01344469
  14 | 49
  15 | 6
  16 | 3
```

```
>stem(Price)
```

```
  The decimal point is 1 digit(s) to the right of the |

   2 | 4
   3 |
   4 | 9
   5 | 345
   6 | 3489
   7 | 024789
   8 | 012346789
   9 | 0123456789
  10 | 0123456789
  11 | 0123456789
  12 | 0123456789
  13 | 0123456789
  14 | 013456789
```

```
  15 | 012456789
  16 | 02346
  17 | 13
  18 | 5
  19 | 1
```

```
>stem(Advertising)
```

```
  The decimal point is at the |

   0 | 00000000000000000000000000000000000000000000000000000000000000000000000000+76
   2 | 000000000000000000000
   4 | 00000000000000000000000000000
   6 | 000000000000000000000
   8 | 00000000000000000000
  10 | 0000000000000000000000000000000000000000000000000
  12 | 000000000000000000000000000000000
  14 | 00000000000000000000
  16 | 000000000000000000
  18 | 0000000000000
  20 | 00000
  22 | 0000
  24 | 00
  26 | 0
  28 | 0
```

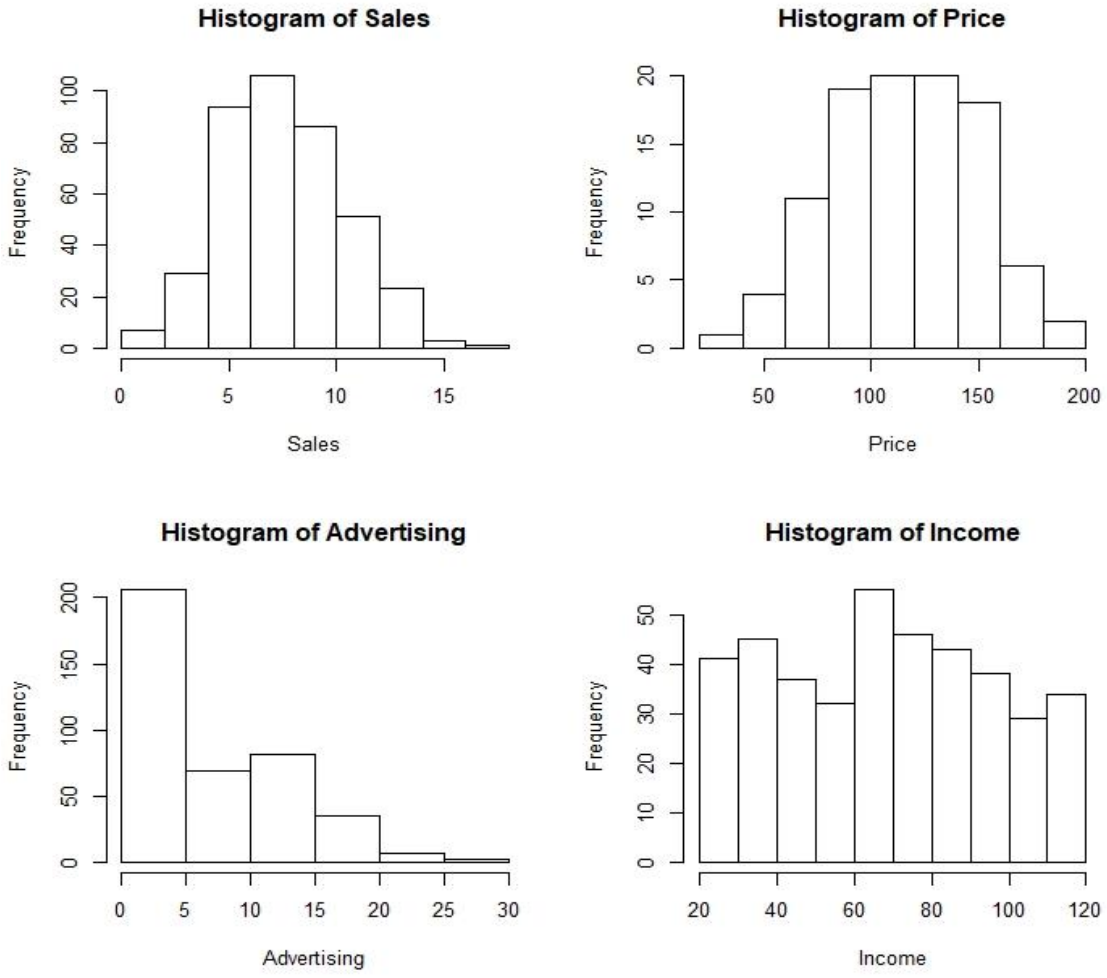
```
>stem(Income)
```

```
  The decimal point is 1 digit(s) to the right of the |

   2 | 1111222234444
   2 | 55555566666678888899999
   3 | 000001112222222333333344
   3 | 555555666677778888999
   4 | 00001111222222222344444
   4 | 555566666778888
   5 | 00111222334444
   5 | 5666778888899
   6 | 000000011112222233333444444
   6 | 55556667777778888899999999999
   7 | 0000111122223333334444
   7 | 555566667777888889999
   8 | 00000011111122222333334444444
   8 | 66777888889999
   9 | 000001112223333334444
   9 | 5666778888899
  10 | 00000000122223333344
  10 | 55555566677789
  11 | 000111112333334
  11 | 555677777788889999
  12 | 0000
```

Histograms provide a visual interpretation of numerical data by indicating the number of data points that lie within a range of values. These ranges of values are called classes or bins. The frequency of the data that falls in each class is depicted by the use of a bar. The higher the bar, greater the frequency of data values in that bin.

The **hist()** function in R allows us to plot the histogram for the required variable.
We also use **par()** to generate a matrix of plots.

```
>par(mfrow=c(2,2))
>hist(Sales)
>hist(Price)
>hist(Advertising)
>hist(Income)
>par(mfrow=c(1,1))
```



Q4. Describe the skew and kurtosis of the distribution of Sales, Price, Advertising and Income.

Ans. For each of the given variables, we have the following observations:

- Sales: it appears that the distribution is symmetric and leptokurtic.
- Price: the distribution seems to be again symmetric and leptokurtic. Thus, we have a graphical confirmation of our observations regarding symmetry earlier.
- Advertising: we see a leptokurtic right-skew in the distribution.
- Income seems to be uniformly distributed.

Since the distributions of Sales and Price seem to be normally distributed, we check if the assumption of normality holds true or not.

Task 4: Create Q-Q Plots to observe normality/log-normality of the Sales and Price variates.

A **quantile-quantile plot** (Q-Q plot) is a scatter plot where we plot the dataset values vs normal distribution values for quantiles determined from the dataset. The y-coordinate of a Q-Q plot are the dataset values, the x coordinates are values from the normal distribution. It is a method to check for normality in a dataset through visualization.

We use **qqnorm()** command to obtain the required scatter plot and use **qqline()** command to fit the a line to the points.

```
>par(mfrow=c(2,1))
```

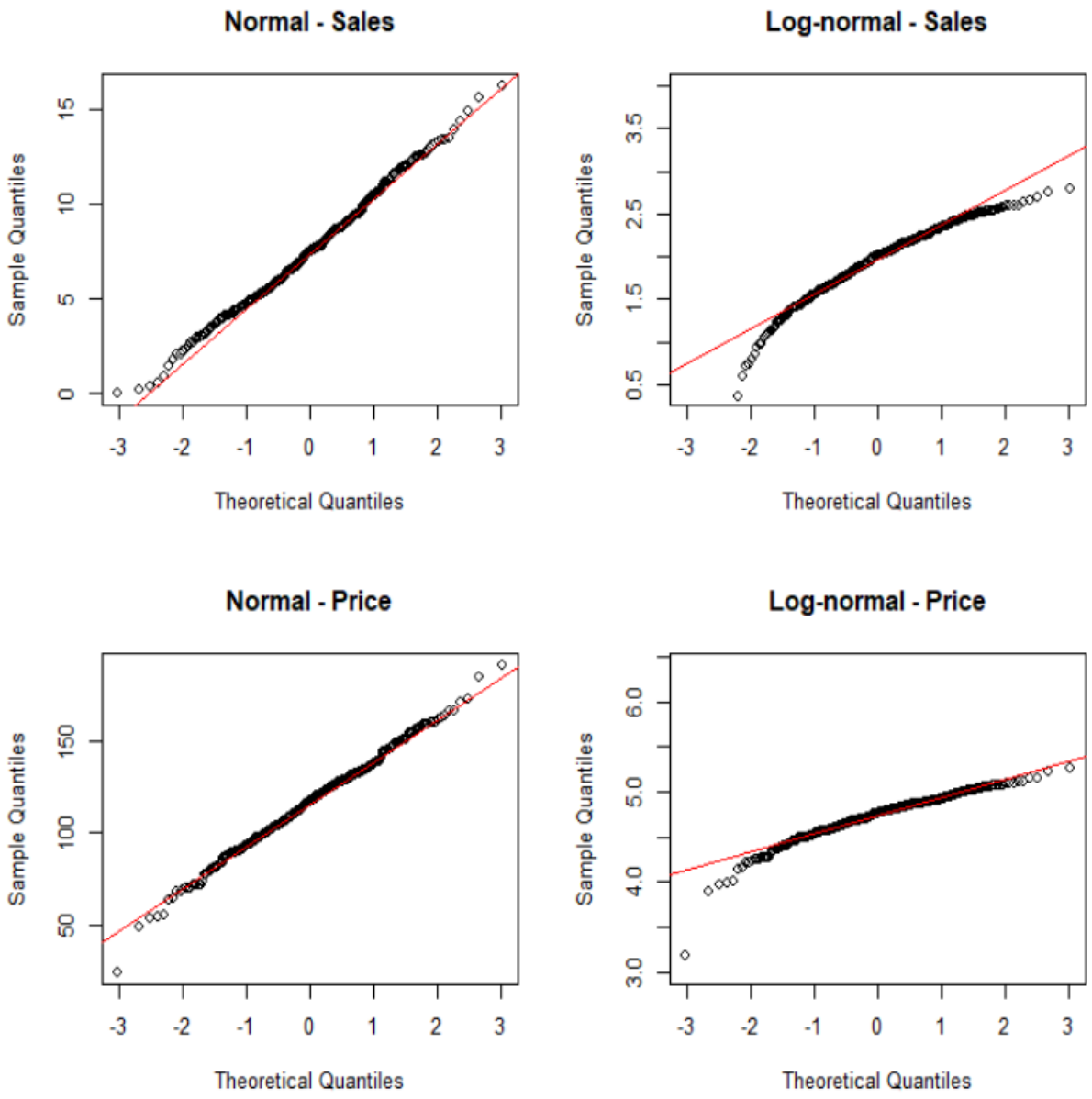
i) Sales

```
>qqnorm(Sales, main = "Normal - Sales")
>qqline(Sales,col="red")

>qqnorm(log(Sales),ylim=c(0.4,4), main = "Log-normal - Sales")
>qqline(log(Sales),col="red")
```

ii)Price

```
>qqnorm(Price, main = "Normal - Price")
>qqline(Price,col="red")
```

```
>qqnorm(log(Price),ylim=c(3,6.4), main = "Log-normal - Price")
>qqline(log(Price),col="red")
```



Q5. Does the distribution of Sales, Price appear to be normal or log-normal?

Ans. As evidenced by the plots above, the scatter plot of theoretical and sample quantiles fit the red line better for normal distribution.
Hence, both Sales and Price seem to have a normal distribution.

2.2. Inference about the mean

We will now try to make inference about the population mean of Sales and Price. To do so, we perform one sample t-test.

Note: Price and Sales are normally distributed hence the assumption of t-test is satisfied

One sample t-test is a statistical procedure used to determine whether a sample of observations could have been generated by a process with a specific population mean.

Hypothesis:

Null hypothesis, H_0: $\mu = \mu_0$ vs.
Alternate hypothesis, H_1: $\mu \neq \mu_0$

Test Statistic:

$$t = \frac{(\bar{X} - \mu_0)}{s/\sqrt{n}} \sim t_{(n-1)}$$

where, n is the number of observations, $\bar{X}$ is the sample mean and s is the sample standard deviation.

Test Criteria:
If $|t| > t_{(\alpha/2)}$ then we reject the null hypothesis of $\mu = \mu_0$. Otherwise, do not reject H_0.

We use **t.test()** function in R to apply the t-test.

Task 5: **(i)** Apply t-test for the null hypothesis that true mean of **Sales** is 7.

```
>mean(Sales)
```

```
[1] 7.496325
```

t-test to test μ = 7 with a conf.level = 0.95

```
>t.test(Sales, mu = 7, conf.level = 0.95)
```

```
        One Sample t-test

data:  Sales
t = 3.5149, df = 399, p-value = 0.0004904
```

```
alternative hypothesis: true mean is not equal to 7
95 percent confidence interval:
 7.218725 7.773925
sample estimates:
mean of x
 7.496325
```

Q6. What is the estimated population mean and its 95% confidence interval?

Ans. The estimated population mean is 7.496 and the 95% CI is (7.219, 7.774).

Q7. What is the probability that we would commit a Type I error?

Ans. With only 5% chance of being wrong we assert that the true mean of Sales lies between 7.22 and 7.77.

Q8. Explain the result obtained by t-test.

Ans. Since the p-value is less than 0.05, we reject the null hypothesis and conclude that the true Sales mean is not equal to 7.

(ii) Apply t-test for the null hypothesis that true mean of **Price** is 115.

```
>mean(Price)

[1] 115.4851
```

t-test to test μ= 115 with a conf.level = 0.95

```
>t.test(Price, mu = 115, conf.level = 0.95)

        One Sample t-test

data:  Price
t = 0.14505, df = 100, p-value = 0.885
alternative hypothesis: true mean is not equal to 115
95 percent confidence interval:
 108.8492 122.1211
sample estimates:
mean of x
 115.4851
```

We will now attempt to separate the data on the basis of one of the factor variables and then analyse each of the subsets obtained.

Task 6: Subset the dataset on the basis of whether the store is located in US or not.

filter() function enables us to filter the rows of the data table that meet certain criteria creating a new data subset. Unlike base subsetting with [,], rows where the condition evaluates to NA are dropped automatically. It is a function of the **dplyr** package so we load the package using the library command.

```
>library(dplyr)
```

```
>proUSno<- prodata %>% filter(US=="No") %>% droplevels()    # non-US data
>proUSyes<- prodata %>% filter(US=="Yes") %>% droplevels()  # US data
```

```
>summary(proUSno)
```

```
    Sales          CompPrice        Income        Advertising
 Min.   : 0.000   Min.   : 77.0   Min.   : 22.00   Min.   : 0.000
 1st Qu.: 5.080   1st Qu.:115.0   1st Qu.: 39.00   1st Qu.: 0.000
 Median : 6.660   Median :124.0   Median : 66.50   Median : 0.000
 Mean   : 6.823   Mean   :124.6   Mean   : 65.28   Mean   : 0.507
 3rd Qu.: 8.523   3rd Qu.:134.0   3rd Qu.: 84.00   3rd Qu.: 0.000
 Max.   :14.900   Max.   :159.0   Max.   :120.00   Max.   :11.000

   Population        Price        ShelveLoc       Age          Education
 Min.   : 10.0   Min.   : 24.0   Bad   :34   Min.   :25.00   Min.   :10.00
 1st Qu.:113.8   1st Qu.: 98.0   Good  :24   1st Qu.:38.00   1st Qu.:12.00
 Median :244.0   Median :116.5   Medium:84   Median :54.50   Median :14.00
 Mean   :252.8   Mean   :114.0               Mean   :53.13   Mean   :14.18
 3rd Qu.:398.2   3rd Qu.:129.8               3rd Qu.:65.75   3rd Qu.:16.00
 Max.   :508.0   Max.   :185.0               Max.   :80.00   Max.   :18.00

  Urban      US
 No :46   No:142
 Yes:96
```

```
>summary(proUSyes)
```

```
    Sales          CompPrice        Income        Advertising
 Min.   : 0.370   Min.   : 85.0   Min.   : 21.00   Min.   : 0.00
 1st Qu.: 5.763   1st Qu.:115.2   1st Qu.: 45.00   1st Qu.: 5.00
 Median : 7.790   Median :125.0   Median : 70.00   Median :10.00
 Mean   : 7.867   Mean   :125.2   Mean   : 70.52   Mean   :10.01
 3rd Qu.: 9.988   3rd Qu.:135.0   3rd Qu.: 93.00   3rd Qu.:14.00
 Max.   :16.270   Max.   :175.0   Max.   :120.00   Max.   :29.00

   Population        Price        ShelveLoc       Age          Education
 Min.   : 12.0   Min.   : 55.0   Bad   : 62   Min.   :25.00   Min.  :10.00
 1st Qu.:148.2   1st Qu.:101.0   Good  : 61   1st Qu.:41.25   1stQu.:11.00
```

```
Median :281.5   Median :118.0   Medium:135   Median :54.50   Median:14.00
Mean   :271.5   Mean   :116.8                Mean   :53.43   Mean  :13.75
3rd Qu.:397.5   3rd Qu.:131.0                3rd Qu.:66.00   3rdQu.:16.00
Max.   :509.0   Max.   :191.0                Max.   :80.00   Max.  :18.00

 Urban       US
No : 72   Yes:258
Yes:186
```

Q9. Is there any difference between the advertising budgets inside and outside the US?

Ans. According to the output obtained above, there seems to be a lot of difference in the advertising budgets inside and outside the US.

Q10. Is there any difference between Price, Sales inside and outside the US?

Ans. We will have to perform two sample t-test for equality of means to answer this question.

Task 7: Perform two sample t-test for equality of means to determine the significance of difference in the Price, Sales inside and outside the US.

Two sample t-test is used to determine if two population means are equal. Here, we conduct t-test for unpaired data.

The hypothesis is

Null hypothesis, H_0: $\mu_1 = \mu_2$ vs.
Alternate hypothesis, H_1: $\mu_1 \neq \mu_2$

Test Statistic:

1) Unequal variances,

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim \text{t distribution with } d.f = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1}\right) + \left(\frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}\right)}$$

This test is also known as Welch's t-test.

2) Equal variances,

$$t = \frac{\overline{X_1} - \overline{X_2}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(n_1+n_2-1)}$$

where, $s_p^2 = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 1}}$

n_1 and n_2 are the sample sizes, $\overline{X_1}$ and $\overline{X_2}$ are sample means and s_1^2 and s_2^2 are sample variances.

Test Criteria:

If $|t| > t_{(\alpha/2)}$ then we reject the null hypothesis of no difference (in other words, equality of means of two independent data). Otherwise, do not reject H_0.

```
>t.test(proUSno$Sales, proUSyes$Sales)

        Welch Two Sample t-test

data:  proUSno$Sales and proUSyes$Sales
t = -3.6956, df = 316.03, p-value = 0.0002585
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.5996161 -0.4881261
sample estimates:
mean of x mean of y
6.823028  7.866899
```

```
>t.test(proUSno$Price, proUSyes$Price)

        Welch Two Sample t-test

data:  proUSno$Price and proUSyes$Price
t = -1.1164, df = 262.24, p-value = 0.2653
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.902698  2.183952
sample estimates:
mean of x mean of y
113.9507  116.8101
```

For Sales, since the p-value is less than 0.05, we reject the hypothesis of no difference and conclude that sales differ significantly inside and outside the US.

For Price, since the p-value is greater than 0.05, we do not reject the hypothesis of no difference and cannot conclude that prices differ significantly inside and outside the US in the given sample.

Task 8: Subset the dataset on the basis of the variable ShelveLoc.

```
>proSLbad <- prodata %>% filter(ShelveLoc=="Bad") %>% droplevels()
>proSLgood <- prodata %>% filter(ShelveLoc=="Good") %>% droplevels()
>proSLmed <- prodata %>% filter(ShelveLoc=="Medium") %>% droplevels()
```

```
>summary(proSLbad)
```

```
   Sales          CompPrice        Income        Advertising
 Min.   : 0.370   Min.   : 86.0   Min.   : 21.00   Min.   : 0.000
 1st Qu.: 4.053   1st Qu.:116.0   1st Qu.: 52.75   1st Qu.: 0.000
Median : 5.210   Median :123.5   Median : 73.00   Median : 4.500
 Mean   : 5.523   Mean   :124.0   Mean   : 72.24   Mean   : 6.219
 3rd Qu.: 7.463   3rd Qu.:134.0   3rd Qu.: 93.00   3rd Qu.:11.000
 Max.   :11.670   Max.   :161.0   Max.   :120.00   Max.   :23.000

   Population        Price        ShelveLoc      Age          Education
 Min.   : 10.0   Min.   : 64.00   Bad:96    Min.   :25.00   Min.   :10.00
 1st Qu.:145.5   1st Qu.: 96.75             1st Qu.:38.00   1st Qu.:12.00
 Median :296.0   Median :113.50             Median :52.00   Median :14.00
 Mean   :275.3   Mean   :114.27             Mean   :52.05   Mean   :13.96
 3rd Qu.:400.5   3rd Qu.:130.00             3rd Qu.:68.00   3rd Qu.:16.00
 Max.   :501.0   Max.   :191.00             Max.   :80.00   Max.   :18.00

  Urban       US
 No :22   No :34
 Yes:74   Yes:62
```

```
>summary(proSLgood)
```

```
     Sales          CompPrice        Income        Advertising
 Min.   : 3.58   Min.   : 89.0   Min.   : 21.00   Min.   : 0.000
 1st Qu.: 8.33   1st Qu.:115.0   1st Qu.: 41.00   1st Qu.: 0.000
 Median :10.50   Median :123.0   Median : 70.00   Median : 7.000
 Mean   :10.21   Mean   :125.8   Mean   : 67.98   Mean   : 7.353
 3rd Qu.:11.96   3rd Qu.:137.0   3rd Qu.: 93.00   3rd Qu.:12.000
 Max.   :16.27   Max.   :157.0   Max.   :117.00   Max.   :24.000

   Population      Price       ShelveLoc      Age          Education
 Min.   : 14   Min.   : 53.0   Good:85   Min.   :25.00   Min.   :10.00
 1st Qu.:176   1st Qu.:103.0             1st Qu.:40.00   1st Qu.:11.00
 Median :272   Median :122.0             Median :54.00   Median :14.00
 Mean   :267   Mean   :117.9             Mean   :52.61   Mean   :13.75
 3rd Qu.:353   3rd Qu.:132.0             3rd Qu.:63.00   3rd Qu.:16.00
 Max.   :503   Max.   :173.0             Max.   :80.00   Max.   :18.00

  Urban       US
 No :28   No :24
 Yes:57   Yes:61
```

```
>summary(proSLmed)
```

```
     Sales          CompPrice        Income        Advertising
 Min.   : 0.000   Min.   : 77.0   Min.   : 21.00   Min.   : 0.000
 1st Qu.: 5.625   1st Qu.:115.0   1st Qu.: 42.00   1st Qu.: 0.000
```

```
Median : 7.380   Median :125.0   Median : 69.00   Median : 5.000
Mean   : 7.307   Mean   :125.1   Mean   : 67.35   Mean   : 6.539
3rd Qu.: 8.775   3rd Qu.:135.0   3rd Qu.: 88.50   3rd Qu.:12.000
Max.   :13.360   Max.   :175.0   Max.   :120.00   Max.   :29.000

   Population        Price        ShelveLoc        Age          Education
Min.   : 12.0   Min.   : 24.0   Medium:219   Min.   :25.00   Min.  :10.00
1st Qu.:124.0   1st Qu.:101.0                1st Qu.:42.00   1stQu.:12.00
Median :261.0   Median :117.0                Median :55.00   Median:14.00
Mean   :259.4   Mean   :115.7                Mean   :54.16   Mean  :13.93
3rd Qu.:405.0   3rd Qu.:131.0                3rd Qu.:66.00   3rdQu.:16.00
Max.   :509.0   Max.   :185.0                Max.   :80.00   Max.  :18.00

 Urban       US
No : 68   No : 84
Yes:151   Yes:135
```

Q11. Which variable seems to differ the most across the three levels of ShelveLoc?

Ans. Sales seems to be clearly impacted by ShelveLoc.

The questions that we have asked ourselves have given us a little insight into what might be useful while modelling say the Sales variable. We will hence use these variables for further analysis in the upcoming sections of the book.

3. Bivariate Analysis

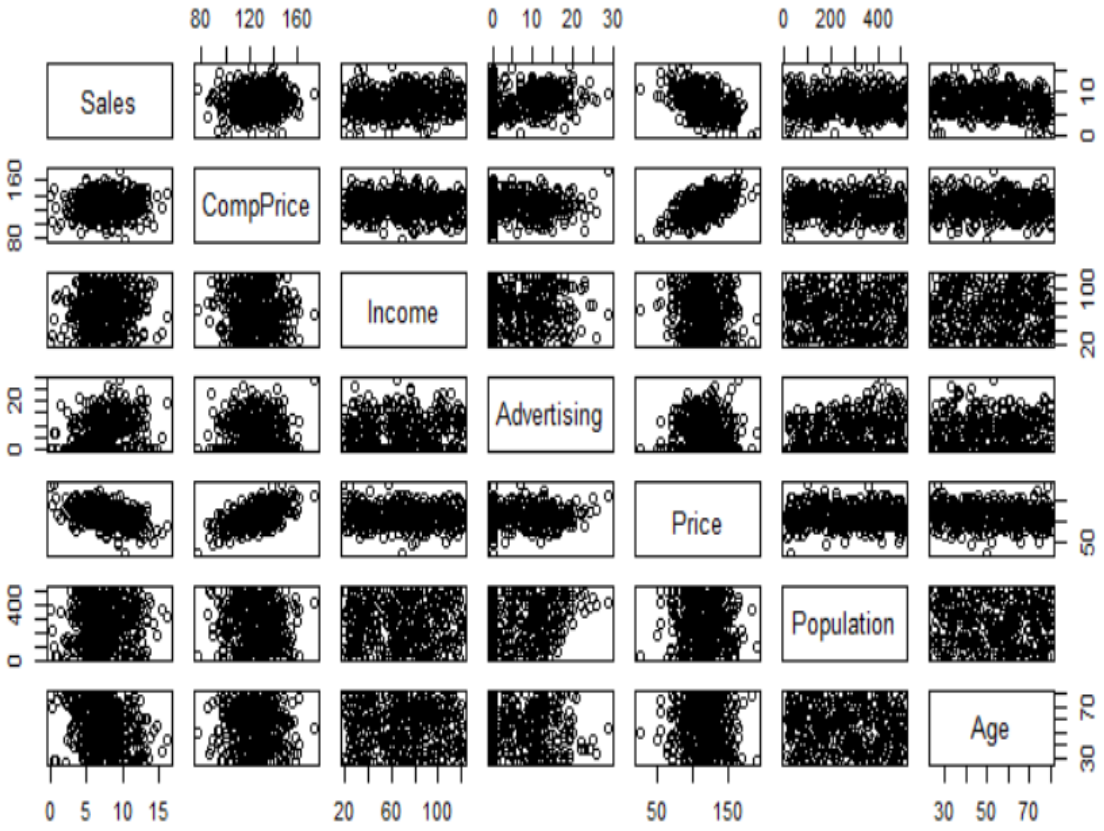
a. What – As the name suggests, Bivariate (bi – two) analysis is concerned with the analysis of relationship of two variable.
b. Why – It is useful in finding associations or differences between two variables and measure the significance of such associations or differences.
c. How – We now describe the ways in which we can perform this analysis.

3.1. Correlation Analysis

Task 9: Visualise the correlation among few pairs of the variables in the dataset.

We will plot all the quantitative variables in Carseats dataset using the plot command.

```
>plot(prodata[,c("Sales","CompPrice","Income","Advertising","Price",
"Population","Age")])
```



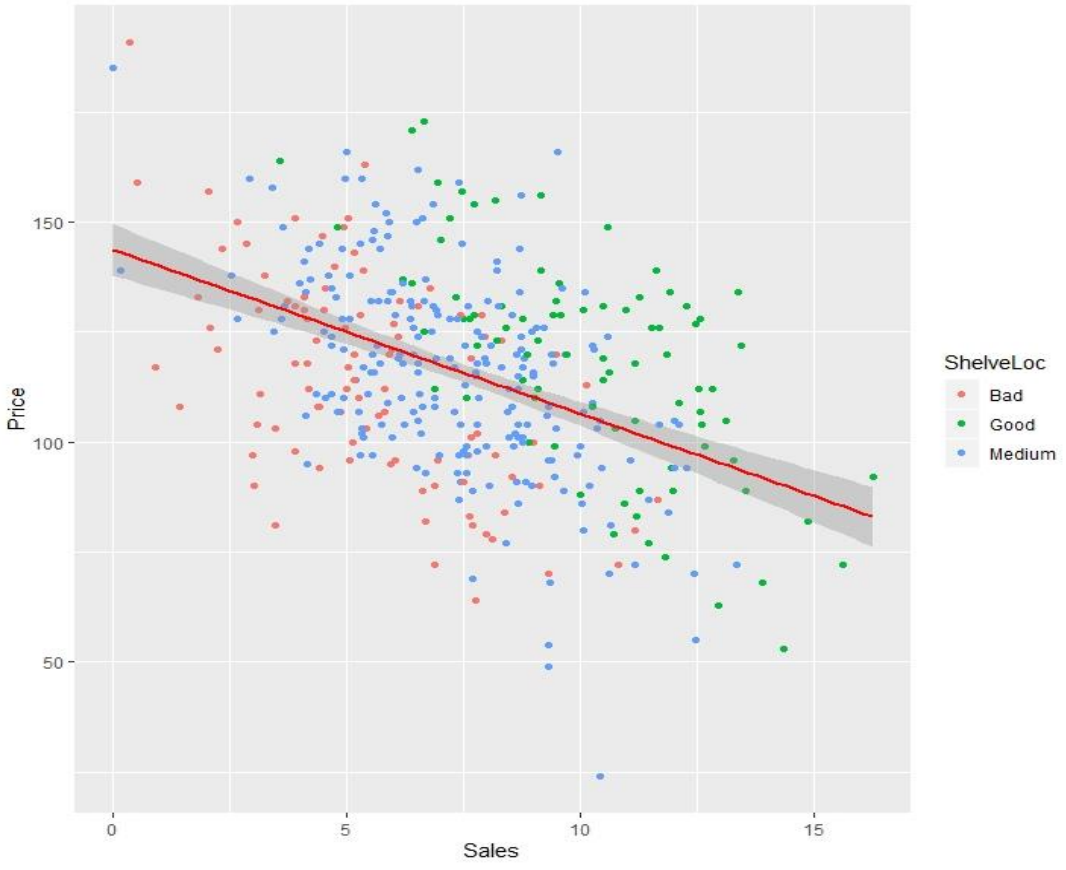
We see a negative correlation between Sales and Price whereas a positive correlation between Price and CompPrice. We will explore these associations using ggplot2 package.

ggplot2 is a R package dedicated to data visualization. It can greatly improve the quality and aesthetics of your graphics. We highly recommend this package as it allows one to create complex plots with only few lines of code.

i) Sales and Price

We have previously noted that Sales is impacted by ShelveLoc. Thus, we create a plot to check if there exists a difference in degree of association between Price and Sales for different levels of ShelveLoc.

```
>library(ggplot2)

>ggplot(prodata, aes(x = Sales, y = Price)) +
 geom_point(aes(color = ShelveLoc)) +
 geom_smooth(method = "lm", col = "Red")
```



Q12. Describe the relation in words.

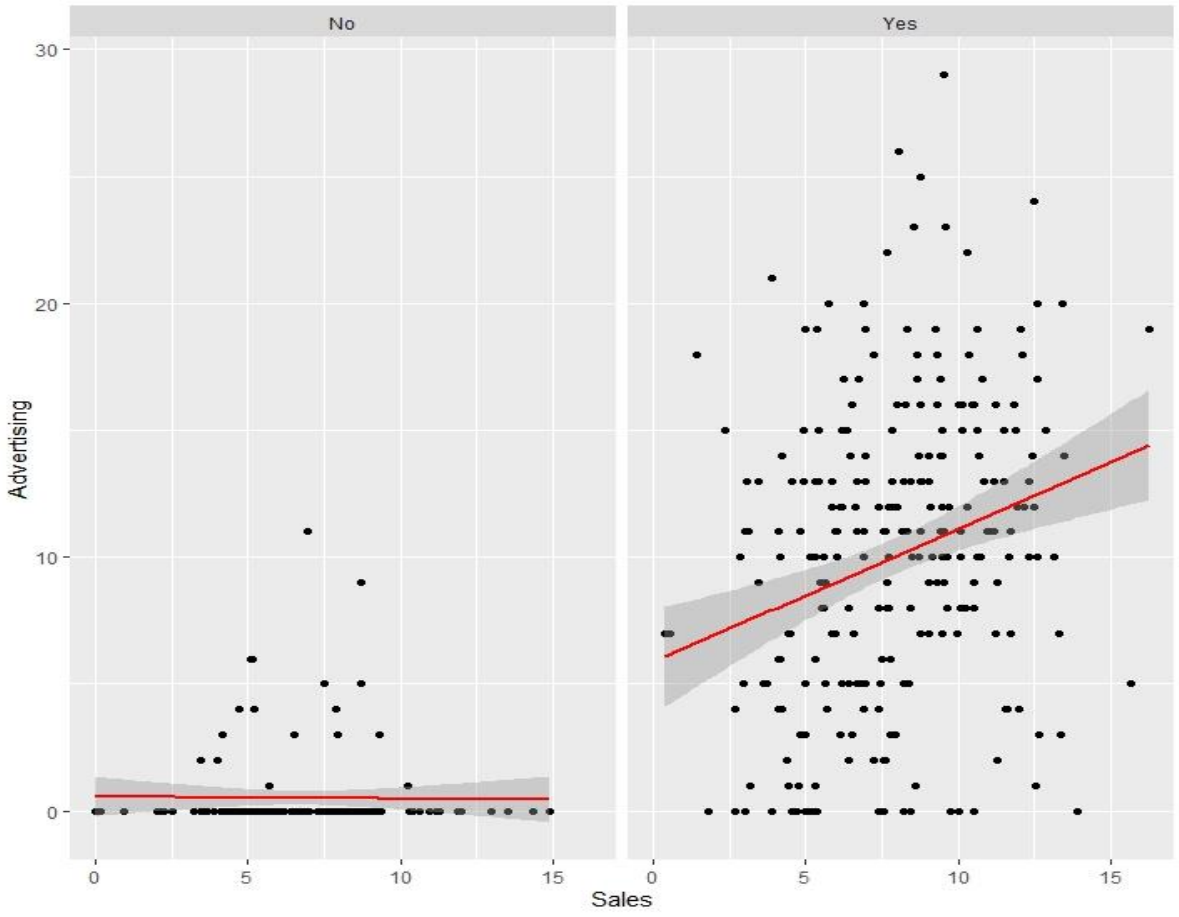
Ans. The relation between Price and Sales is negative. The plot does not suggest a difference in degree of association between price and Sales for different levels of Shelveloc.

ii) Sales and Advertising

Advertising budget is affected by US variable. We now create a plot to check if the correlation between Sales and Advertising is impacted by US.

```
>ggplot(prodata, aes(x = Sales, y = Advertising)) +
 geom_point() + geom_smooth(method = "lm", col = "Red") + facet_wrap(US)
```

Note: facet_wrap() creates separate plots on the basis of a factor variable

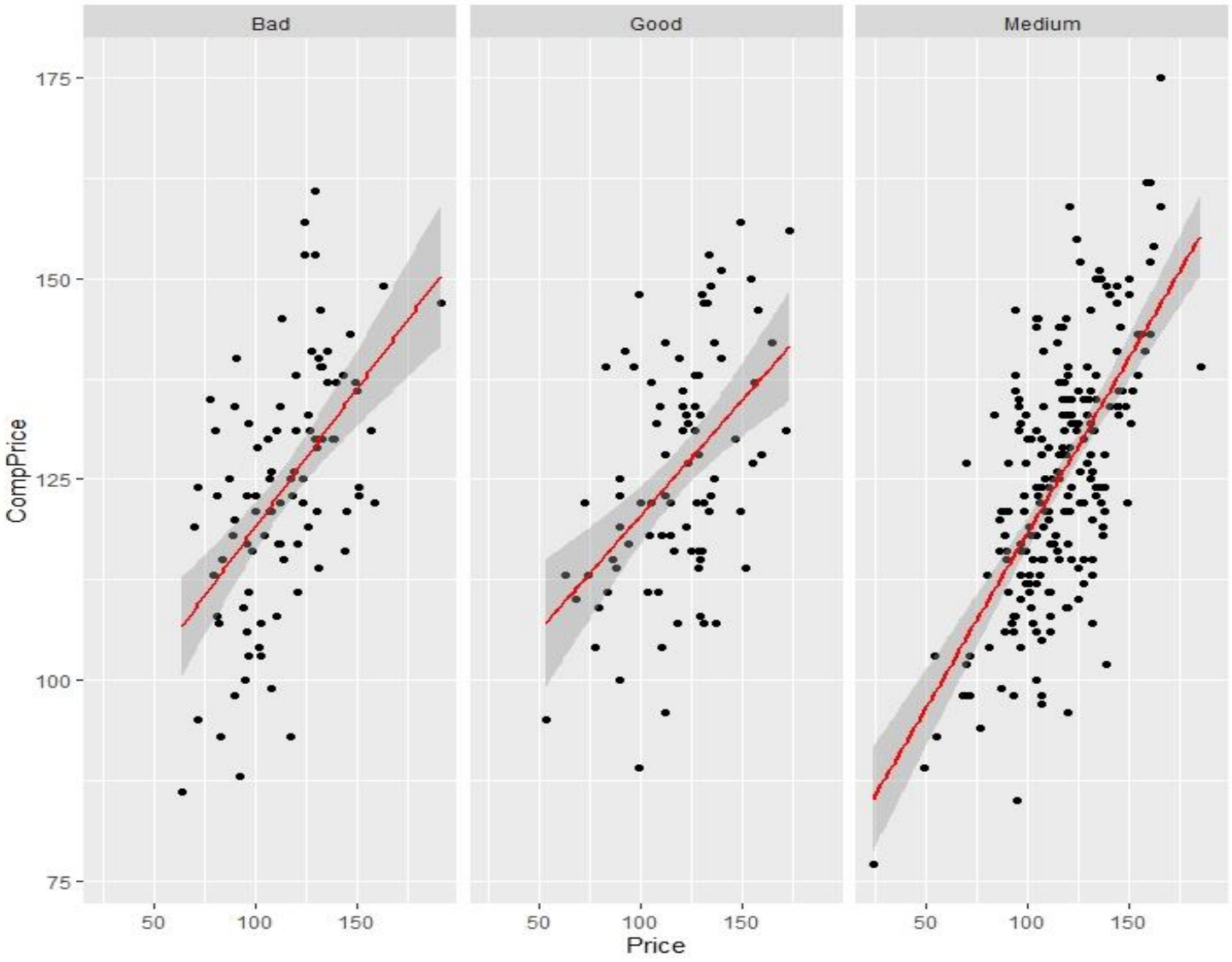


Q13. Describe the relation in words.

Ans. There is a mildly positive association between Advertising and Sales in US.

iii) Price and CompPrice

```
>ggplot(prodata, aes(x = Price, y = CompPrice)) +
 geom_point() + geom_smooth(method = "lm", col = "Red") +
 facet_wrap(ShelveLoc)
```

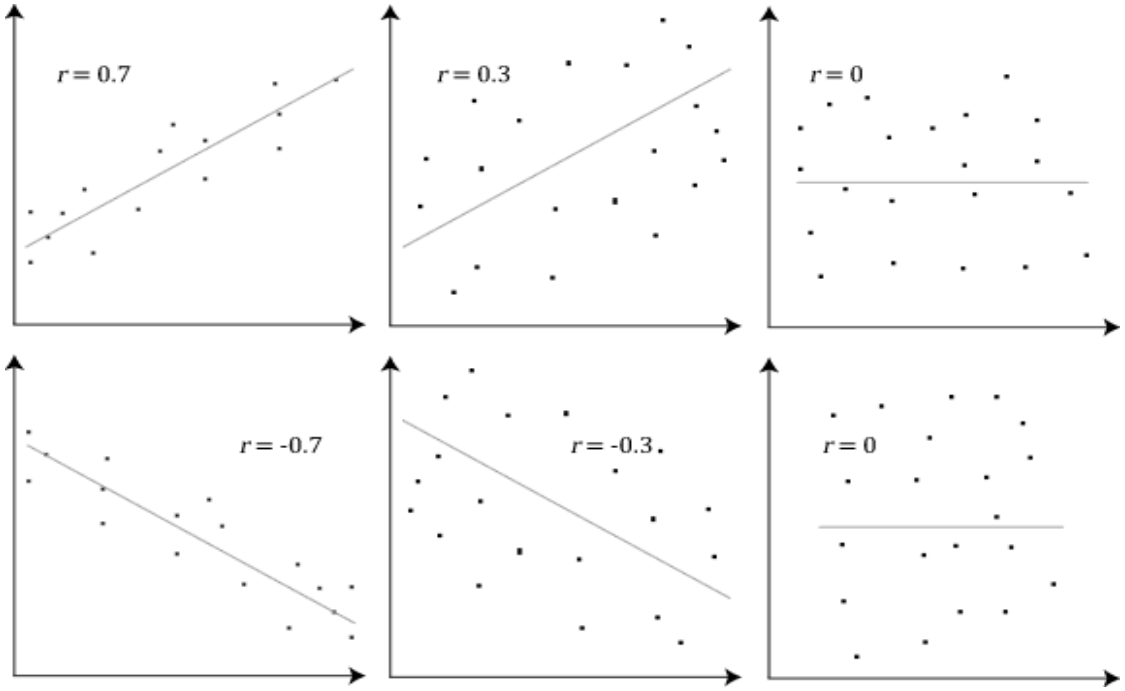


Q14. Describe the relation in words.

Ans. Price and CompPrice are positively correlated for all the levels of ShelveLoc. However, the line of best fit shows a difference in slopes for the 3 plots obtained.

Task 10: Perform bivariate correlation analysis for the pairs for which the visualisation was done.

Pearson's Correlation Coefficient (or Pearson Product moment correlation) is a measure of the strength of a linear association between two variables. Basically, correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit. It ranges from -1 to 1.



t-test for testing the significance of population correlation coefficient

Let ρ denote the population correlation coefficient and r be the sample correlation coefficient.

Test Hypothesis:

Null Hypothesis, H_0: $\rho = 0$ vs.
Alternate Hypothesis, H_1: $\rho \neq 0$

Test Statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{(n-2)}$$

where, n is the number of observations in each variable.

Test Criteria:
If p-value < α then reject the null hypothesis at α% level of significance for the given data. Otherwise, we fail to reject the null hypothesis.

In R, we can easily calculate correlation coefficient using the **cor()** function. You can compute different types of correlation coefficients by specifying the "method" argument of cor() function.

To conduct the t-test for significance of correlation, we use **cor.test()** function and pass the observation vectors as arguments.

i) Sales and Price

```
>cor(prodata$Sales,prodata$Price)
[1] -0.4449507
```

As shown above by the scatter plot between Sales and Price, the correlation coefficient also suggests a mildly negative correlation.

```
>cor.test(prodata$Sales,prodata$Price)

        Pearson's product-moment correlation

data:  Sales and prodata$Price
t = -9.912, df = 398, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.5203026 -0.3627240
sample estimates:
cor
-0.4449507
```

Q15. Analyse the result of the correlation test.

Ans. The correlation is almost certainly different from zero, since the p-value is almost zero. Thus there is certainly a relation between Sales and Price.

ii) Sales and Advertising

```
>cor(Sales,Advertising)
[1] 0.2695068

>cor(proUSyes$Sales,proUSyes$Advertising)
[1] 0.2557332

>cor(proUSno$Sales,proUSno$Advertising)
[1] -0.01260417
```

The correlation coefficient suggests a mildly positive correlation between Sales and Advertising in US. However, it appears to be insignificant outside US.

```
>cor.test(proUSyes$Sales,proUSyes$Advertising)

        Pearson's product-moment correlation

data:  proUSyes$Sales and proUSyes$Advertising
t = 4.2325, df = 256, p-value = 3.224e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1379154 0.3664146
sample estimates:
cor
0.2557332
```

In case of US, the p-value below 0.05 supports our observation about some positive correlation between Advertising and Sales.

```
>cor.test(proUSno$Sales,proUSno$Advertising)

        Pearson's product-moment correlation

data:  proUSno$Sales and proUSno$Advertising
t = -0.14915, df = 140, p-value = 0.8817
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1769640  0.1524396
sample estimates:
cor
-0.0126041
```

Outside US, the p-value > 0.8817, suggests that we fail to reject the null hypothesis. Hence, we conclude that the correlation coefficient is not significantly different from zero based on the sample data.

iii) Price and CompPrice

```
>cor(prodata$Price,prodata$CompPrice)
[1] 0.5848478

>cor(proSLmed$Price,proSLmed$CompPrice)
[1] 0.6463412

>cor(proSLbad$Price,proSLbad$CompPrice)
[1] 0.5386937

>cor(proSLgood$Price,proSLgood$CompPrice)
[1] 0.4825042
```

Q16. What is the order of correlation between Price and CompPrice for different levels of ShelveLoc?

Ans. The correlation between Price and CompPrice is highest for the Medium level of ShelveLoc followed by Bad and Good levels.

3.2. Bivariate Linear Regression Analysis

Task 11: Fitting a bivariate linear regression model

Regression Analysis is the mathematical measure of the underlying relationship between two or more variables.

When we study a variable (**dependent**) in terms of another variable (**independent**) through a **linear** relationship between them, it is called Bivariate Linear Regression Analysis.
If two or more independent variables are used then it is called Multiple Linear Regression Analysis.

In non-deterministic models, regression is always an approximation. Hence, the presence of errors is unavoidable. However, by **minimising the sum of squares of errors**, we find the best equation that can explain the dependent variable in terms of the independent variables by obtaining the closest estimates to intercept and coefficients of the explanatory variables.

Coefficient of determination (R^2) measures the proportion of variation in the dependent variable that can be predicted from the set of independent variables in a regression equation. It varies from 0 to 1. A value closer to 1 indicates that the variability in the dependent variable can be explained well by independent variables whereas a value closer to 0 implies that most of the variance results from chance causes or the absence of some other explanatory variable in the model.

$$R^2 = 1 - \frac{SSE}{TSS}$$

where, SSE (Sum of Squares due to Error) is the variability left unexplained by the model
TSS (Total Sum of Squares) is the total variability in the independent variable.

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \quad and \quad SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

However, addition of an explanatory variable almost always leads to an increase in the value of R^2, that is, R^2 is a non-decreasing function of number of regressors. Therefore, we use **adjusted R^2** which takes into account the number of predictors in the model. We divide both SSE and TSS with their degrees of freedom.

$$adjusted\ R^2 = \bar{R}^2 = 1 - \frac{SSE/(n-(p+1))}{TSS/(n-1)}$$

where, p is the number of explanatory variables in the model

Note: SSE has n-(p+1) degrees of freedom since, we estimate 1 intercept and p slope coefficients in the model.
Now, we will describe the steps of Bivariate Linear Regression Analysis by regressing Price on CompPrice and Sales on Price.

In R, the **lm()** (linear model) function can be used to create a linear regression model. This function accepts two arguments formula and data. The formula argument specifies the response and explanatory variables in the model separated by a "~" (tilde).

(i) Price vs. CompPrice

```
>lmPr_CP <- lm(Price ~ CompPrice, data = prodata)

>summary(lmPr_CP)


Call:
lm(formula = Price ~ CompPrice, data = prodata)

Residuals:
    Min      1Q  Median      3Q     Max
-48.473 -12.183   0.197  12.925  56.540

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.94110    7.90436   0.372     0.71
CompPrice    0.90301    0.06278  14.384<2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.23 on 398 degrees of freedom
Multiple R-squared:  0.342,    Adjusted R-squared:  0.3404
F-statistic: 206.9 on 1 and 398 DF,  p-value: < 2.2e-16
```

Q17. What is the best predictive equation for Price given CompPrice? Explain the relation.

Ans. The best predictive equation is: Price = 2.9411 + 0.903*CompPrice
Using this equation, we can say that for every unit increase in CompPrice, the expected Price increases by 0.903 units.

Q18. How much variability in Price is explained by CompPrice?

Ans. Adjusted R^2 of the above model is 0.3404. Hence,34.04% of the variability in Price is explained by CompPrice.

(ii) Sales vs. Price

```
>lmS_P <- lm(Sales~Price, data = prodata)

>summary(lmS_P)

Call:
lm(formula = Sales ~ Price, data = prodata)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5224 -1.8442 -0.1459  1.6503  7.5108

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.641915   0.632812  21.558<2e-16 ***
Price       -0.053073   0.005354  -9.912   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.532 on 398 degrees of freedom
Multiple R-squared:  0.198,    Adjusted R-squared:  0.196
F-statistic: 98.25 on 1 and 398 DF,  p-value: < 2.2e-16
```

Q19. What is the best predictive equation for Sales given Price? What is its interpretation?

Ans. Sales = 13.642 - 0.0531*Price
For a unit increase in Price, Sales is expected to decrease by 53.1 units.

Note: Sales in Carseats dataset is given in thousands.

3.3. Regression Diagnostics

In the Introduction, we mentioned about the assumptions of a linear model. Thus, the model obtained is the best fit to the data subject to the condition that these assumptions hold. Therefore, we use certain techniques, elucidated below, to check for these assumptions.

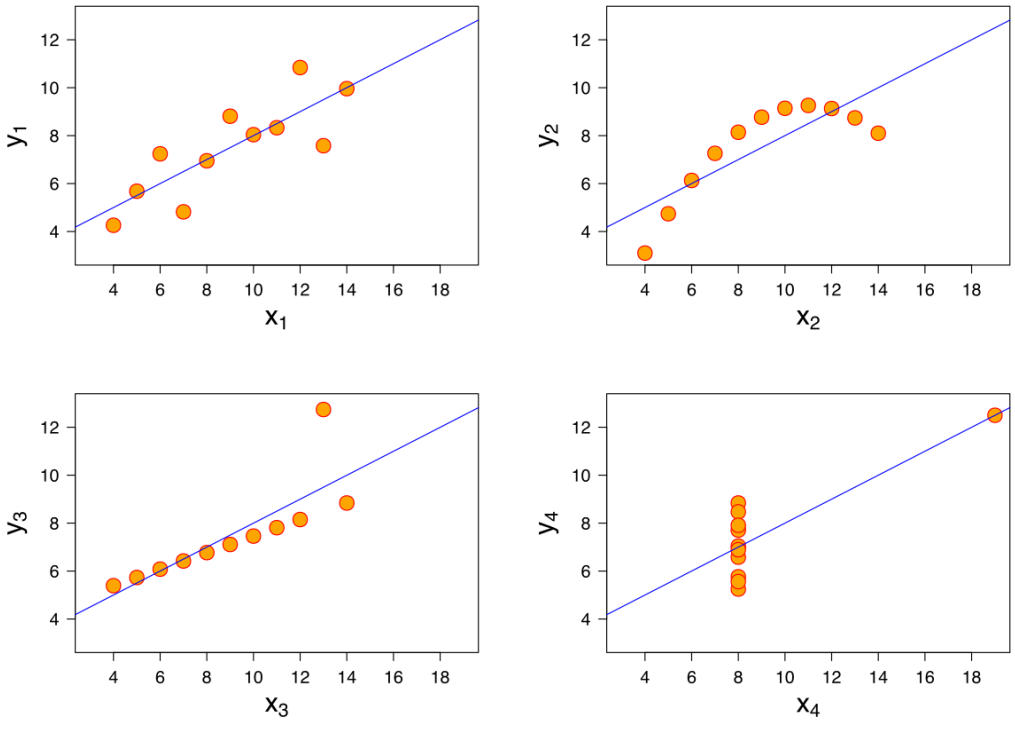
Task 12: Plot the actual values along with the fitted regression line

A perfect model should predict the values of the dependent variable without any error. However, we have the best model which minimises the errors instead of a perfect model.

Also, according to our assumption, the expected value of the error terms should be zero. In such a scenario, we look at the visualisation of the original values with the fitted regression line and expect to find an almost equal scatter of values on both sides of the line throughout its range.

We show, with the help of a plot, that simply having a linear fit is not sufficient. We must verify this assumption. Otherwise, the results can be misleading.

For example, consider Anscombe's quartet. It contains four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when visualised. However, we obtain the same linear fit in all the four cases.



Clearly, only the first graph satisfies the assumption we have made.

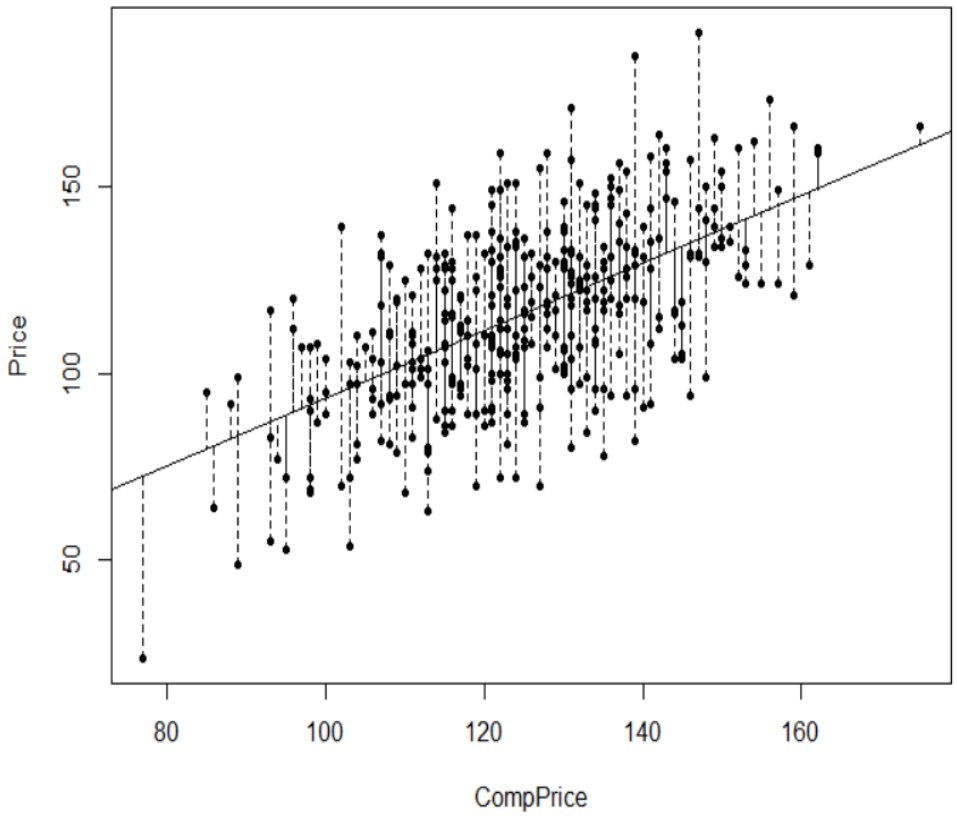
For the second plot, a quadratic fit seems to be more appropriate than a linear fit whereas in the third and fourth plot, due to influence of a single point, we have been deprived of a much better fit to the datasets.

Hence, verification of the model assumptions is very important.

(i) Price vs. CompPrice

```
>plot(prodata$CompPrice,prodata$Price, pch = 20)  #pch = 20 for black dots
>abline(lmPr_CP)
>segments(CompPrice,Price,CompPrice,fitted(lmPr_CP),lty = 2)
```

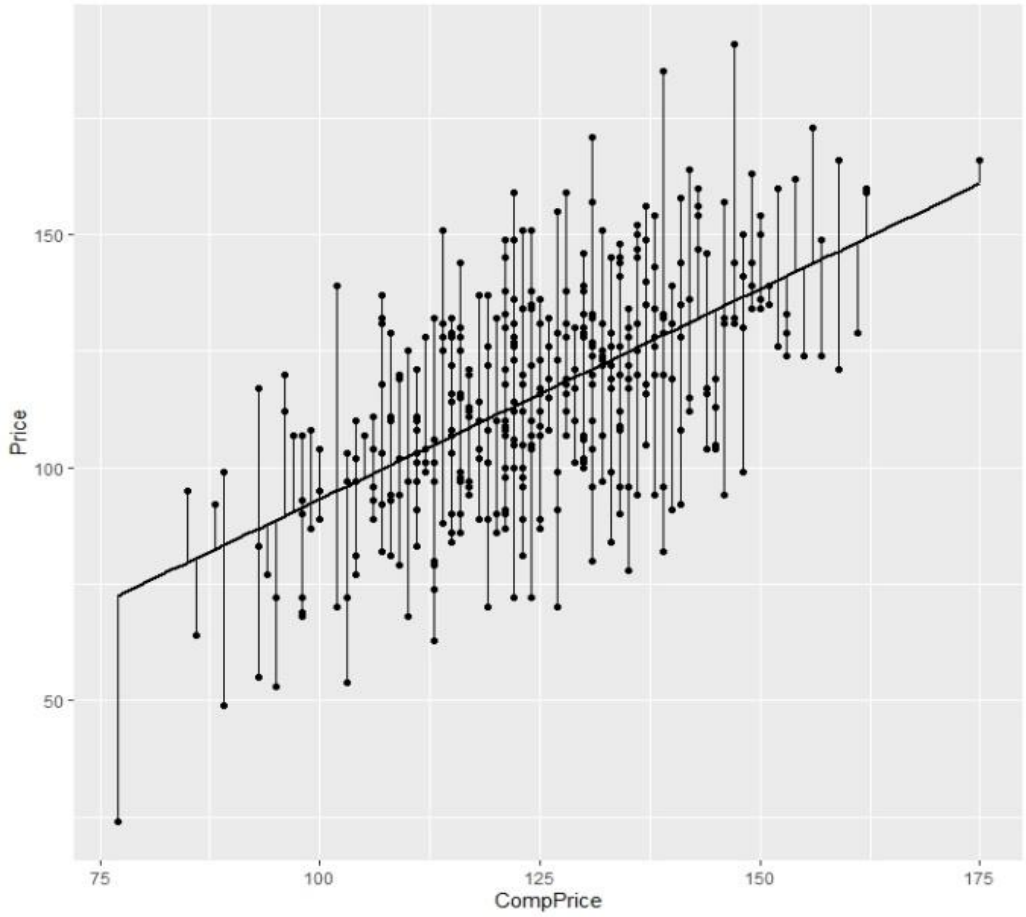
segments (x_0,y_0,x_1,y_1) function draws line segments between pairs of points whose coordinates are given by (x_0,y_0) and (x_1,y_1). Hence, the above segments command plots line segments between pairs of points (CompPrice,Price) and (CompPrice,$\widehat{\text{Price}}$).



augment() function in the **broom** package allows us to convert a model object into a data frame containing the residuals, fitted values along with the response and explanatory variables. This is useful for a neat representation of data as well as in functions which take data frame as input such as the ggplot.

The following code is for the execution of this task using ggplot:

```
> library(broom)

>ggplot(augment(lmPr_CP),aes(x = CompPrice, y = Price)) + geom_point() +
 geom_smooth(method = "lm", se = FALSE, col = "black") +
 geom_segment(aes(x = CompPrice, y = Price, xend = CompPrice, yend =
 .fitted))
```

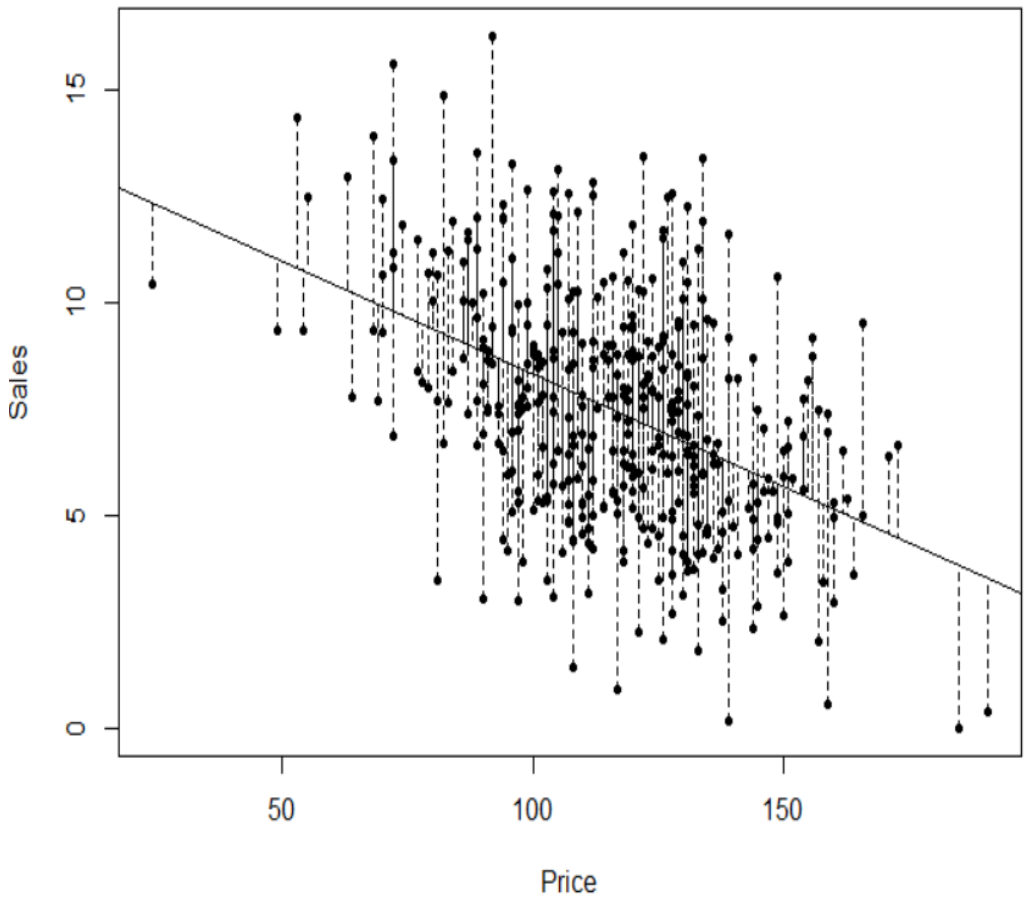


Q20. What should be the ideal relationship between the predicted and the observed values?

Ans. They should be identical, that is, they should fall on the 1:1 line. Of course they are not equal because of the presence of errors in model fitting. In any case they should be symmetric about a 1:1 line (i.e. the length of the residual segments should be approximately equal above and below the line) throughout the range.

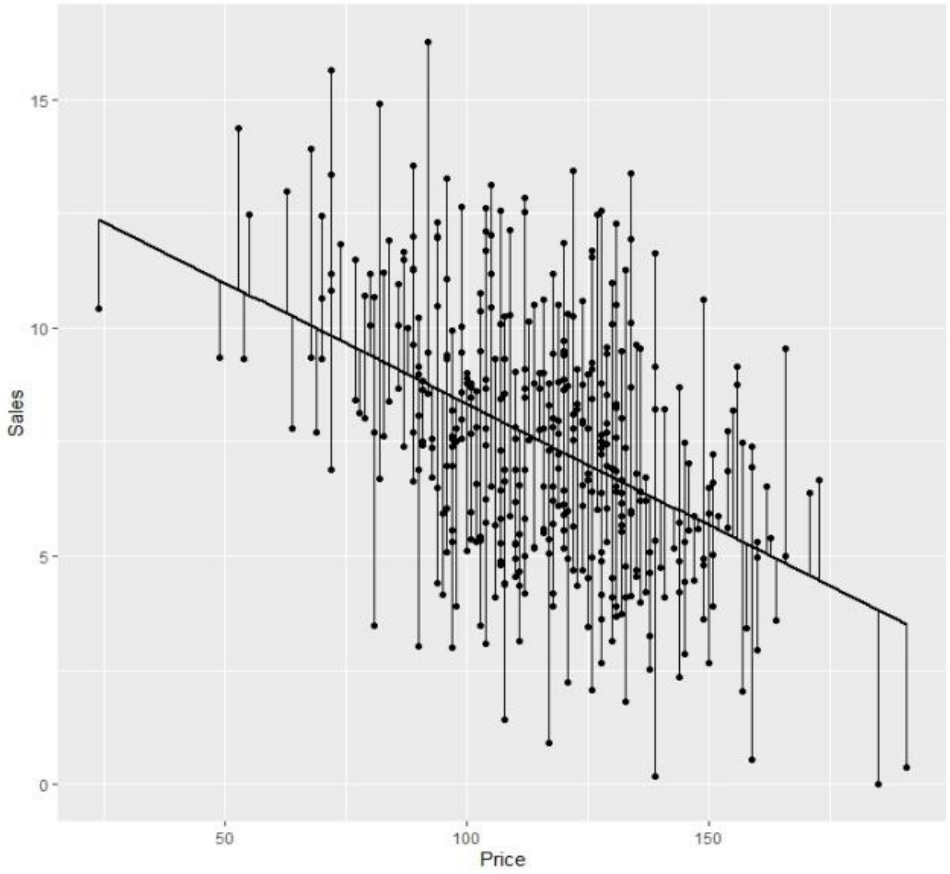
(ii) Sales vs. Price

```
>plot(prodata$Price,prodata$Sales,pch = 20)

>abline(lmS_P)

>segments(Price,Sales,Price,fitted(lmS_P),lty = 2)
```



Execution using ggplot:

```
>ggplot(augment(lmS_P),aes(x = Price, y = Sales)) + geom_point() +
 geom_smooth(method = "lm", se = FALSE, col = "black") +
 geom_segment(aes(x = Price, y = Sales, xend = Price, yend = .fitted))
```

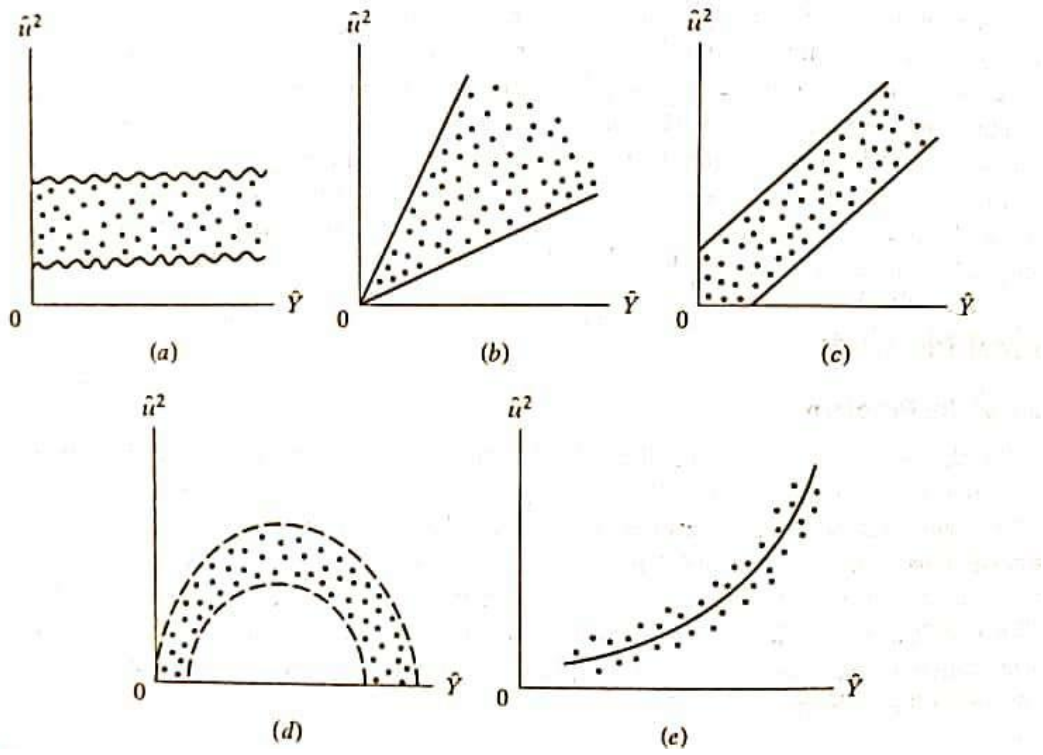


Task 13: Plot Residuals vs. Predicted with 3,2,1 sigma limits (to test the presence of heteroscedasticity).

Our assumption regarding constant variance of error terms (homoscedasticity) may not hold in the original data.

To check for this, we plot residuals against the predicted values.

If the scatter is sufficiently random, that is, there is no pattern that can be identified in the plot then we can say that heteroscedasticity is not significant.



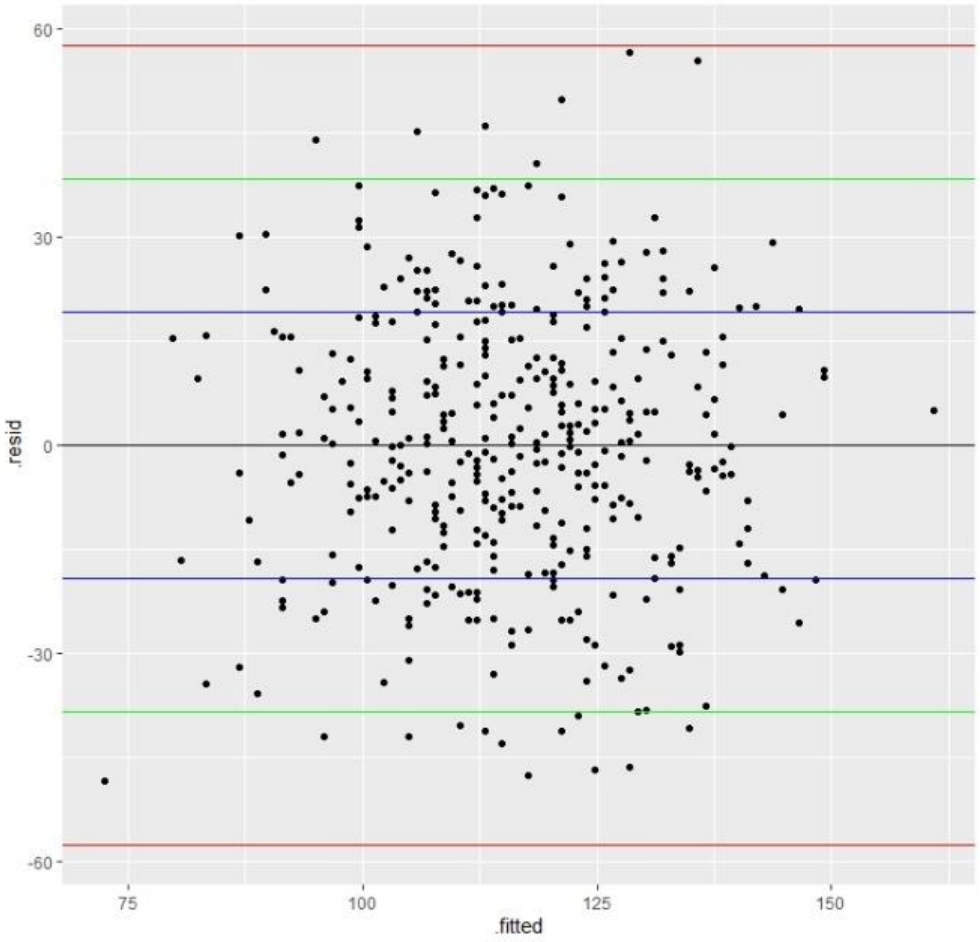
Apart from the first one, all the other plots show signs of presence of heteroscedasticity or non-constant variance of the error terms.
Therefore, a model showing these kinds of outputs on the Residuals vs. Fitted values plot violates the assumption of homoscedasticity.

(i) Price vs. CompPrice

Using ggplot:

Since we have to plot sigma limits of the residuals, we first create a vector containing ±3, ±2, ±1 sigma values and then use it as the y-intercept to plot the horizontal lines.

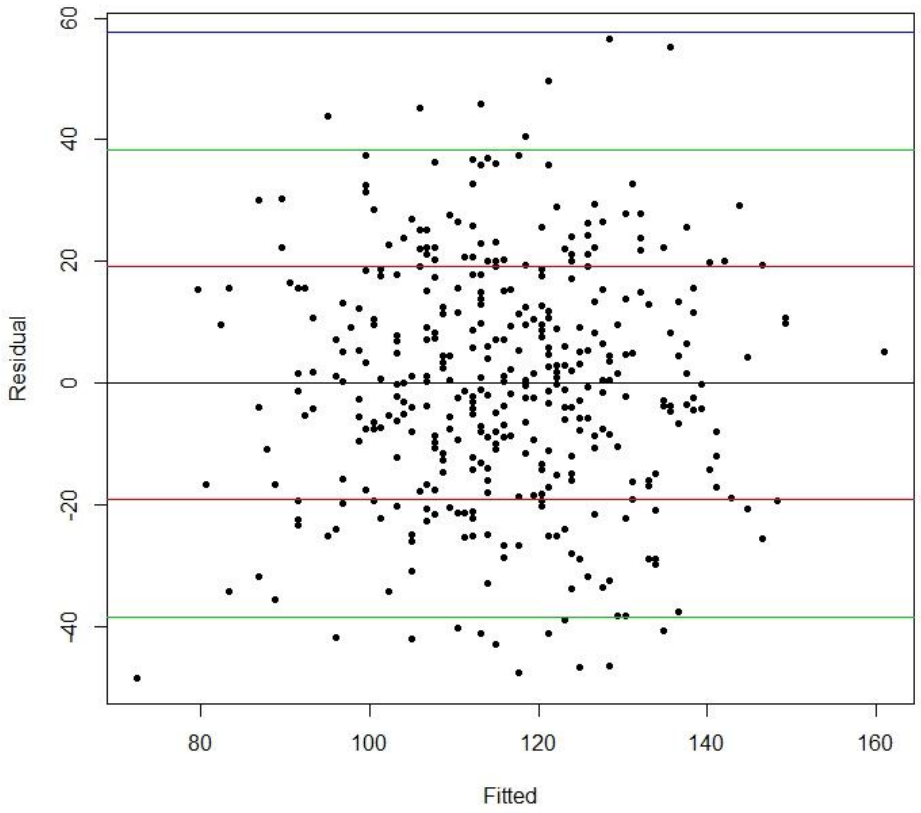
```
>std_devA <- sd(resid(lmPr_CP))

>std_vecA <- std_devA*seq(-3,3)

>ggplot(augment(lmPr_CP), aes(x = .fitted, y = .resid)) +
 geom_point() +
 geom_hline(yintercept = std_vecA,col=c("red","green","blue","black",
 "blue","green","red"))
```



Using plot and abline:

We use a for loop command to plot ±3, ±2, ±1 sigma lines on the residuals vs. fitted plot.

```
>plot(fitted(lmPr_CP), resid(lmPr_CP), pch=20, xlab="Fitted",
 ylab="Residual")

>for (j in -3:3) abline(h=j*std_devA, col=abs(j)+1)

>rm(std_devA, std_vecA)
```



Q21. What should this relation be? Do we see this expected relation?

Ans. All the residuals should ideally fall on the 0 horizontal line. However, they don't fall on that line due to the presence of error. But in any case, they should be symmetric about this line throughout the range, and have the same degree of spread.
There is no visible pattern in the plot and the spread seems to be random.

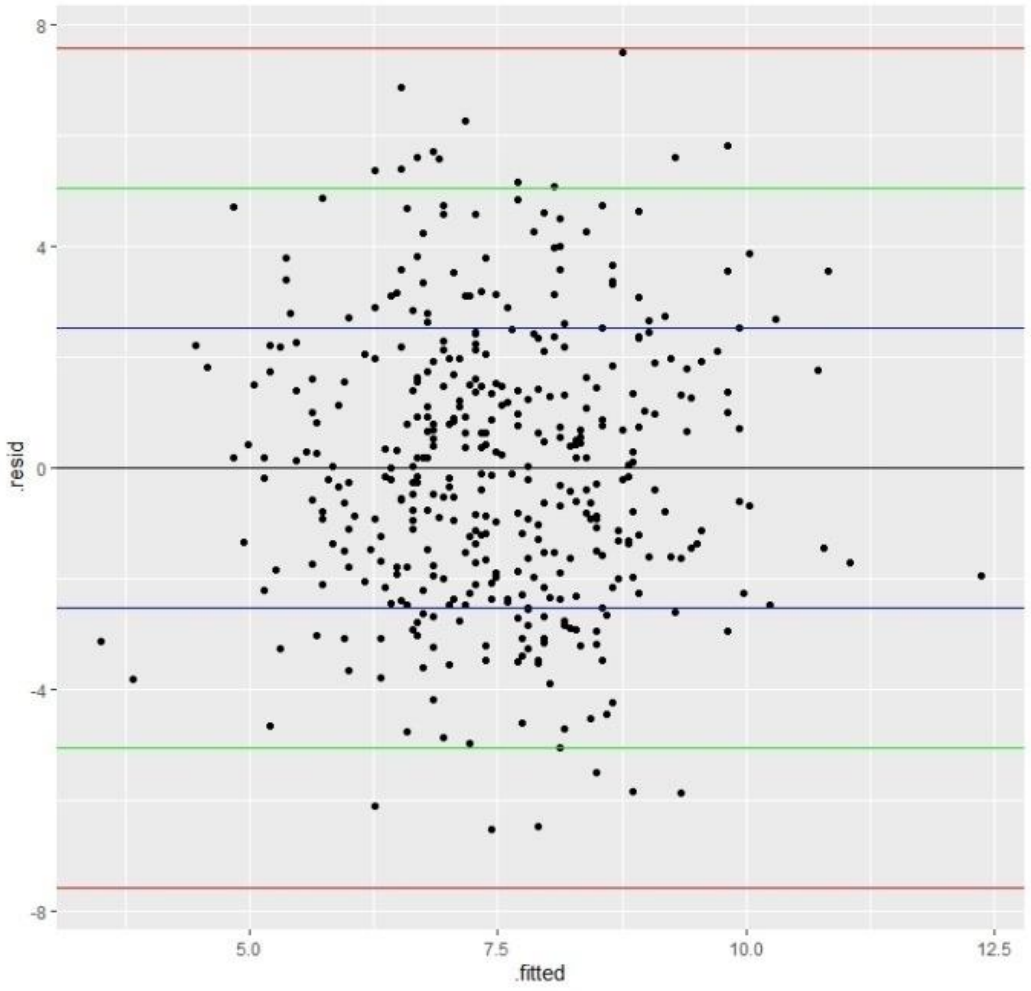
Note: The points on the upper and lower half of the plot are almost equal.
sum(resid(lmPr_CP) > 0) is the total number of points on the upper half of the plot
sum(resid(lmPr_CP) < 0) is the total number of points on the lower half of the plot

(ii) Sales vs. Price

Using ggplot:

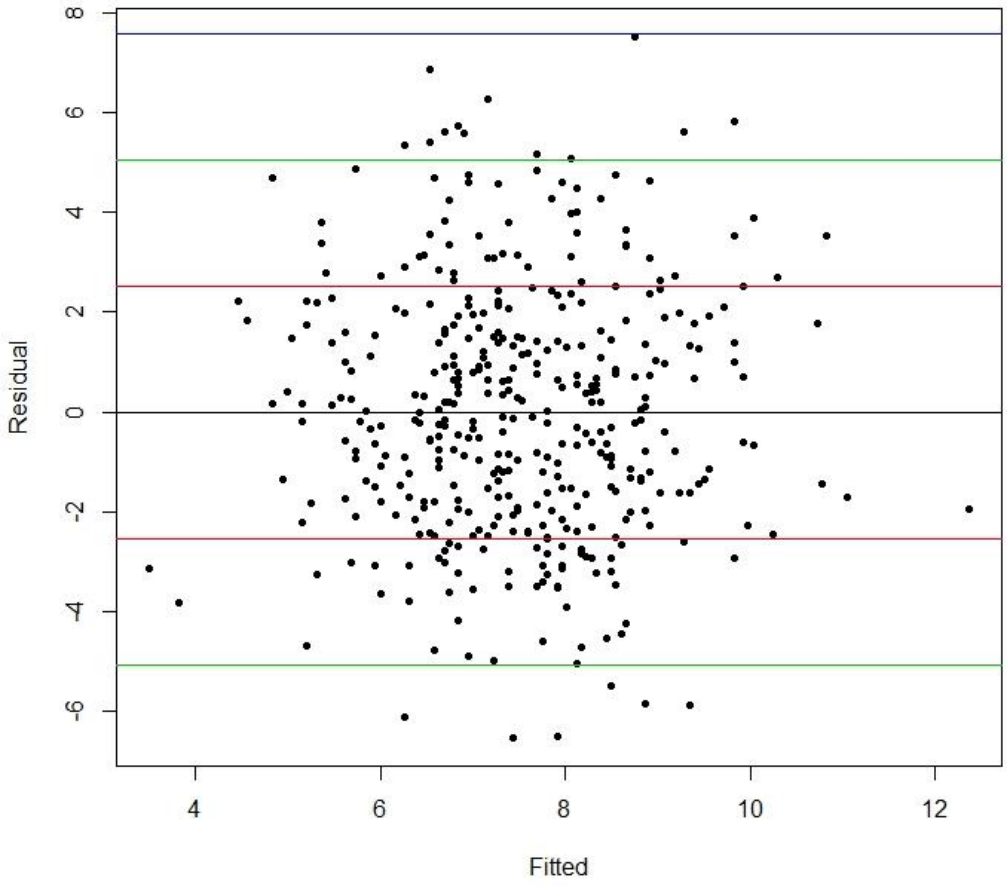
```
>std_devB<- sd(resid(lmS_P))
>std_vecB<- std_devB*seq(-3,3)
```

```
>ggplot(augment(lmS_P), aes(x = .fitted, y = .resid)) + geom_point() +
 geom_hline(yintercept = std_vecB,col=c("red","green","blue","black",
 "blue","green","red"))
```



Using plot and abline:

```
>plot(fitted(lmS_P), resid(lmS_P), pch=20, xlab="Fitted", ylab="Residual")
>for (j in -3:3) abline(h=j*std_devB, col=abs(j)+1)
>rm(std_devB,std_vecB)
```



The plots seem to be symmetric about the zero residual.

Task 14: Visualise the distribution of residuals (To test for the normality of error terms).

The assumption of the normality of the error terms should reflect in their visualisation. Therefore, a histogram or a stem-and-leaf plot of residuals should show bars that can be approximated to a bell shaped curve.

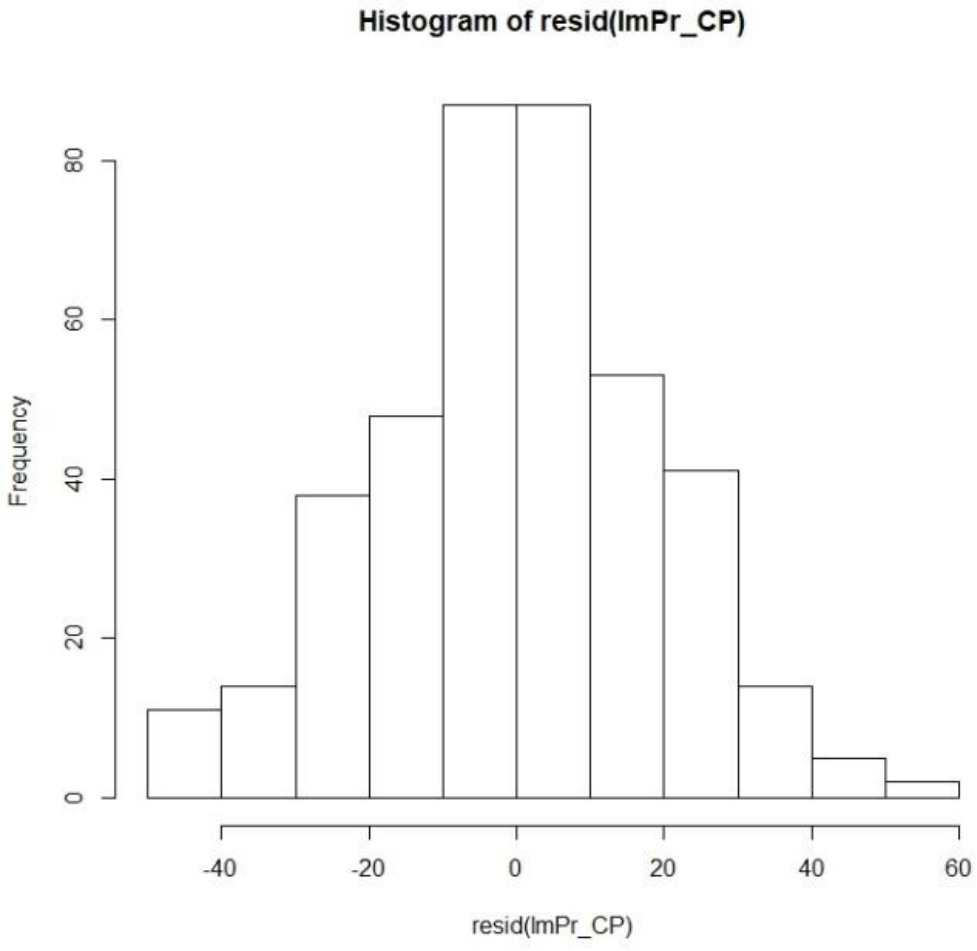
Along with this, QQ Plot of the residuals should show most of the points lying on or very close to the qqline throughout the range.

(i) Price vs. CompPrice

```
>stem(resid(lmPr_CP))
```

```
  The decimal point is 1 digit(s) to the right of the |

  -4 | 88763221110
  -3 | 988864444322210
  -2 | 99998776655555554433222222111111100000
  -1 | 9999999888888777777666665555444433322222211111000O
  -0 | 999999988888888877777776666665555555554444444444444333333322222+7
   0 | 000001111111112222222222223333333334444445555555555556666667777+4
   1 | 00000011111111222222333333445555555666666778888889999999
   2 | 000000001111122222222334444455666666778889999
   3 | 00123366667777
   4 | 0456
   5 | 057
```

```
>hist(resid(lmPr_CP))
```



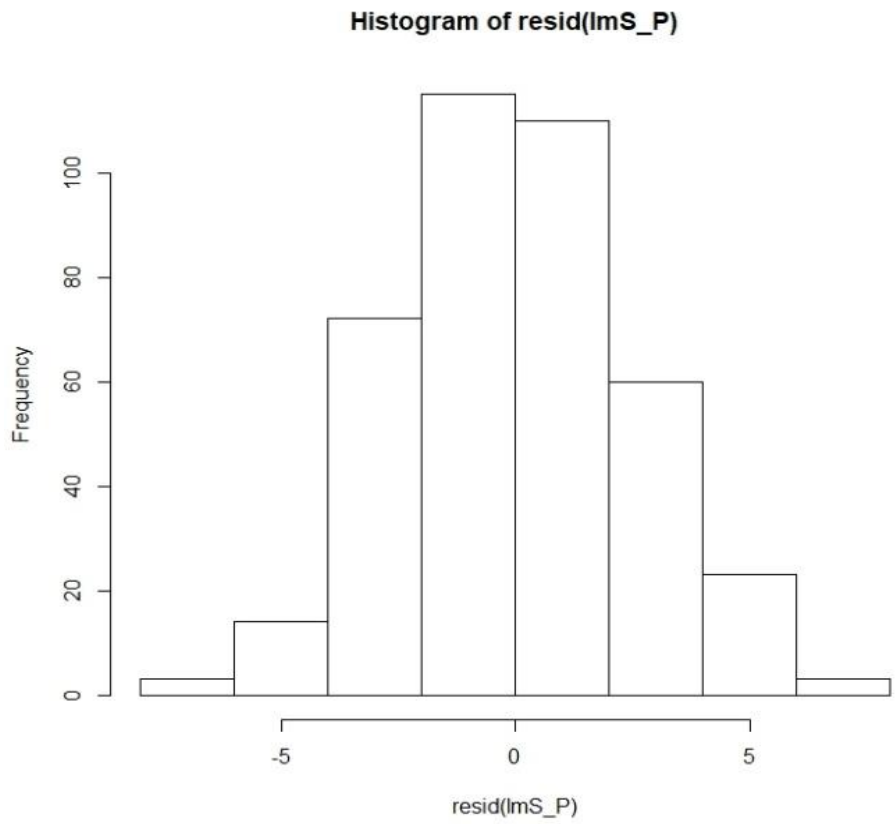
Q22. What can be said about the distribution of residuals?

Ans. On the basis of the above plotted histogram, the residuals are symmetric about zero and seem to be normally distributed.

(ii) Sales vs. Price

```
>stem(resid(lmS_P))

  The decimal point is at the |

  -6 | 551
  -5 | 98510
  -4 | 987765422
  -3 | 98876655555433222221111100
  -2 | 99999988887777666555555444444433333322222111100000
  -1 | 999999888887777666666555555444444433322222211111110000
  -0 | 99999999988888888776666666655554444333333322222222211110
   0 | 000011222222222333334444444455566666777777778888888999999
   1 | 00000111112223333334444445555555666667778888999
   2 | 0000111111222222233344444455556677788899
   3 | 1111111233445556668889
   4 | 0023356666777789
   5 | 124466678
   6 | 39
   7 | 5
```

```
>hist(resid(lmS_P))
```

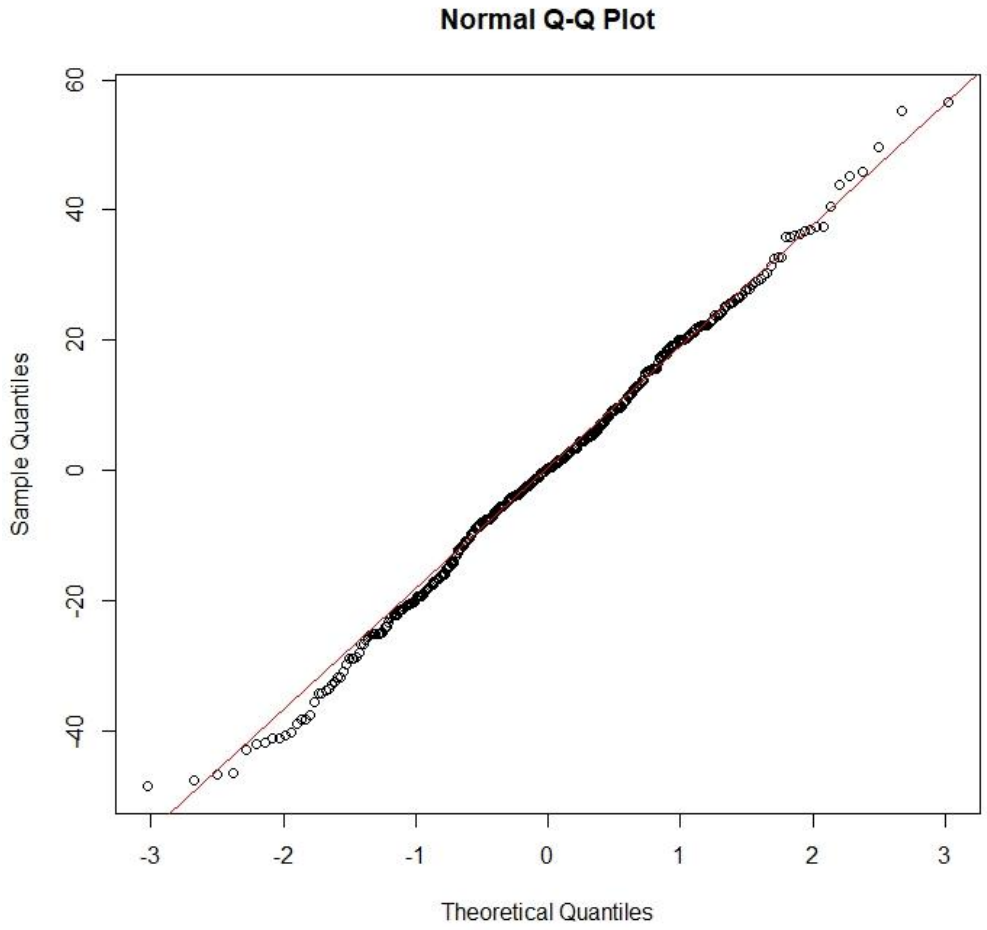


Again, the residuals are symmetric about zero and seem to be normally distributed.

Task 15: Create QQ-Plot of the residuals

(i) Price vs. CompPrice

```
>qqnorm(resid(lmPr_CP))
>qqline(resid(lmPr_CP), col = "red")
```

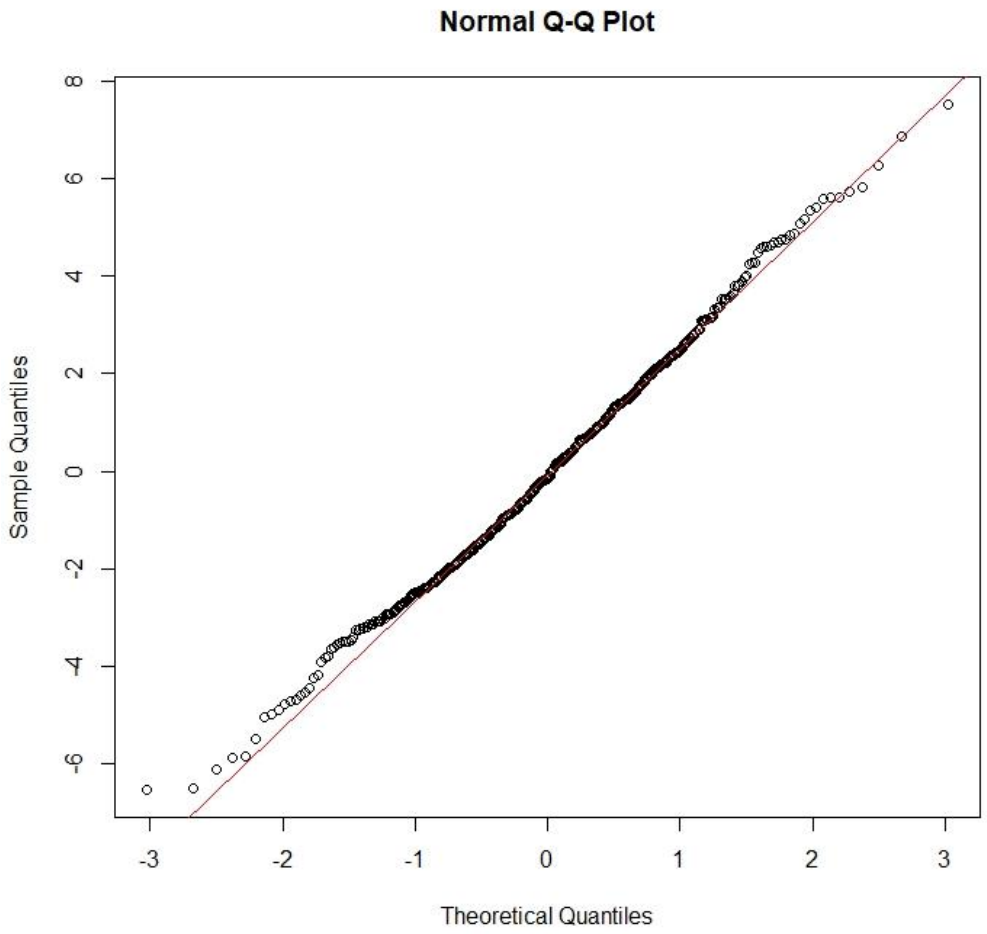


Q23. Do the residuals follow the normal distribution? Where are the discrepancies, if any?

Ans. For most of the part, the residuals follow the theoretical normal distribution well: they are very close to the normal line. However, things are not so smooth towards the tails.

(ii) Sales vs. Price

```
>qqnorm(resid(lmS_P))
>qqline(resid(lmS_P), col = "red")
```



Even in this case, we see the violation of the assumption of normal errors only towards the tails.

3.4. RMSE

Task 16: Find the values of RMSE and coefficient of determination for both the models.

Root Mean Squared Error (RMSE) is the standard deviation of the residuals. RMSE is a measure of how spread out the residuals are. In other words, it tells us how concentrated the data is around the line of best fit.

RMSE is always non-negative, and a value of 0 (almost never achieved in practice) would indicate a perfect fit to the data. In general, a lower RMSE is better than a higher one.

However, comparisons across different types of data would be invalid because the measure is dependent on the scale of the numbers used.

It is given by the formula $\boldsymbol{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} u_i^2}{n}}$

where $u_i = (\hat{y}_i - y_i)$ (i = 1(1)n) denotes the i^{th} residual and n is the number of observations.

(i) Price vs. CompPrice

```
>(RMSE_1 <- sqrt(mean(resid(lmPr_CP)^2)))
[1] 19.18114

>(R2_1 <- summary(lmPr_CP)$r.squared)
[1] 0.3420469
```

(ii) Sales vs. Price

```
>(RMSE_2 <- sqrt(mean(resid(lmS_P)^2)))
[1] 2.525987

>(R2_2 <- summary(lmS_P)$r.squared)
[1] 0.1979812
```

Q24. What can you say about both the models based on their RMSE and R^2values?

Ans. The two models cannot be compared using RMSE as their measuring units are different. Hence, a lower RMSE need not necessarily imply that Sales vs. Price model is better than Price vs. CompPrice model.

However, R^2 can overcome the drawback of RMSE and can be used to compare different models having same number of independent variables.

3.5. Prediction of the response variable

The whole idea behind modelling is to be able to predict the unknown from that which is known with a particular margin of error.

In this section, we use the models created previously to predict the dependent variables using two different methods in R.

Task 17: Predict the values for Sales and Price for the mean values of Price and CompPrice respectively.

a) Method of Regression Equation

We simply substitute the value of the explanatory variable in the regression equation obtained above to find the value of the response variable. Though this works well for a single value, we would need to apply a loop if we have many values to predict.

```
>mean(Price)
[1] 115.795

>mean(CompPrice)
[1] 124.7397

>mean(Sales)
[1] 7.496325
```

i) Price when CompPrice = 124.975

```
>lmPr_CP$coefficients
(Intercept)   CompPrice
  2.9411048   0.9030118

intercept = 2.9411048, slope = 0.9030118
equation1 :  Price = 2.9411048 +  0.9030118*CompPrice

>(PredPrice1 <- 2.9411048 +  0.9030118*124.975)
[1] 115.795
```

ii) Sales when Price = 115.795

```
>lmS_P$coefficients
(Intercept)       Price
13.64191518 -0.05307302

intercept = 13.64191518, slope = -0.05307302
equation2: Sales = 13.64191518 - 0.05307302*Price
```

```
>(PredSales1 <- 13.64191518 - 0.05307302*115.795)
[1] 7.496325
```

b) Using the predict function

predict() on taking an "lm" object as input and a specified data frame as the newdata argument returns the predicted values. It is also helpful in finding confidence and prediction intervals.

i) Price when CompPrice = 124.975

```
> (pred_S <- predict(lmPr_CP,data.frame(CompPrice = 124.975), se.fit =
   TRUE))

$fit
    1
115.795

$se.fit
[1] 0.9614637

$df
[1] 398

$residual.scale
[1] 19.22927
```

ii) Sales when Price = 115.795

```
> (pred_S <- predict(lmS_P,data.frame(Price = 115.795), se.fit = TRUE))

$fit
     1
7.496325

$se.fit
[1] 0.1266163

$df
[1] 398

$residual.scale
[1] 2.532326
```

The output returns fitted value of the response variable, standard error of the estimate and degrees of freedom.

3.6. Confidence and Prediction Interval

Confidence interval of the estimated value of Price

A **confidence interval** is a range of values associated with a population parameter for a given confidence level. For example, a confidence interval having 95%confidence level is a range of values that you can be 95% certain contains the true mean of the population.

Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu,\sigma^2)$ where $\sigma^2$is unknown. Let $\bar{X}$ denote the sample mean and s be the standard error.

Then,

$$P\left(-t_{(\alpha/2)} < \frac{\bar{X}-\mu}{s/\sqrt{n}} < t_{(\alpha/2)}\right) = 1-\alpha$$

Hence, 100(1-α)% CI for population mean (μ) is given by

$$\left(\bar{X} - t_{(\alpha/2)}\frac{s}{\sqrt{n}}, \bar{X} + t_{(\alpha/2)}\frac{s}{\sqrt{n}}\right)$$

```
> (PredP$fit + qt(c(.025,.975), PredP$df) * PredP$se.fit
[1] 113.9048 117.6852
```

Q25. What is the 95% C.I. for Price for the given CompPrice?

Ans. The 95% C.I. for the estimated Price is (113.95, 117.69)

ii) Sales when Price = 115.795

```
>(PredS<- predict(lmS_P,data.frame(Price = 115.795), se.fit = T))
$fit
       1
7.496325

$se.fit
[1] 0.1266163

$df
[1] 398

$residual.scale
[1] 2.532326
```

Confidence interval of the estimated value of Sales

```
>(PredS$fit + qt(c(.025,.975), PredS$df) * PredS$se.fit)
[1] 7.247405 7.745245
```

Q26. What is the 95% C.I. for Sales for the given Price?

Ans. The 95% C.I. for Price is (7.25, 7.75)

Task 18: Make a prediction data frame for the unique values of CompPrice and Price.

A **prediction interval** is a range of values that is likely to contain the value of a single new observation given specified settings of the predictors. For example, for a 95% prediction interval of (5,10), you can be 95% confident that the next new observation will fall within this range.

A prediction interval is always wider than the corresponding confidence interval.

(i) Price vs. CompPrice

We will use the **predict()** function and its interval argument to find the confidence and prediction intervals.

To begin with we create a data frame of the unique values of CompPrice variable.

```
>cpframe<- data.frame(CompPrice<- unique(prodata$CompPrice))

>str(cpframe)

'data.frame':  73 obs. of  1 variable:
 $ CompPrice....unique.prodata.CompPrice.: num  138 111 113 117 141 124 115
136 132 121 ...
```

Now, we find the Confidence Interval by specifying interval = “confidence”.
Pred.p1 contains the fitted values of Price along with lower and upper values of the confidence interval.

```
>pred.p1 <- predict(lmPr_CP, cpframe, interval = "confidence", level=.95)

>str(pred.p1)

 num [1:73, 1:3] 128 103 105 109 130 ...
 - attr(*, "dimnames")=List of 2
```

```
..$ : chr [1:73] "1" "2" "3" "4" ...
..$ : chr [1:3] "fit" "lwr" "upr"
```

```
>pred.p1[124.975 +1-min(CompPrice),]
```

```
 fit      lwr          upr
82.40614 77.46680 87.34548
```

To find the Prediction interval, we specify interval = "prediction". Pred.p2 contains the fitted values of Price along with lower and upper values of the prediction interval.

```
>pred.p2 <- predict(lmPr_CP, cpframe, interval = "prediction", level=.95)
```

```
>str(pred.p2)
```

```
 num [1:73, 1:3] 128 103 105 109 130 ...
 - attr(*, "dimnames")=List of 2
..$ : chr [1:73] "1" "2" "3" "4" ...
..$ : chr [1:3] "fit" "lwr" "upr"
```

```
>pred.p2[124.975 +1-min(CompPrice),]
```

```
fit       lwr   upr
82.40614  44.28118 120.53110
```

Q27. For the unique values of CompPrice in dataset, what are the 95% Confidence and Prediction Intervals for Price?

Ans. 95% confidence interval for Price is (77.46680, 87.34548) whereas 95% prediction interval for Price is(44.28118, 120.53110).
We see that prediction interval is wider than the confidence interval.

(ii) Sales vs. Price

```
>pframe <- data.frame(Price <- unique(prodata$Price))
```

```
>pred.s1 <- predict(lmS_P, pframe, interval = "confidence", level=.95)
```

```
>str(pred.s1)
 num [1:101, 1:3] 7.27 9.24 9.4 8.49 6.85 ...
 - attr(*, "dimnames")=List of 2
..$ : chr [1:101] "1" "2" "3" "4" ...
..$ : chr [1:3] "fit" "lwr" "upr"
```

```
>pred.s1[115.795 +1-min(Price),]
     fit      lwr      upr
5.256378 4.747126 5.765630
```

```
>pred.s2 <- predict(lmS_P, pframe, interval = "prediction", level=.95)
>str(pred.s2)
 num [1:101, 1:3] 7.27 9.24 9.4 8.49 6.85 ...
 - attr(*, "dimnames")=List of 2
..$ : chr [1:101] "1" "2" "3" "4" ...
..$ : chr [1:3] "fit" "lwr" "upr"


>pred.s2[115.795 +1-min(Price),]
fit        lwr upr
5.2563783  0.2519933 10.2607632
```

Q28. For the unique values of Price in dataset, what are the 95% Confidence and Prediction Intervals for Sales?

Ans. 95% confidence interval for Sales is (4.747126, 5.765630)whereas 95% prediction interval for Salesis (0.2519933, 10.2607632).
We note that even in this case prediction interval is wider than confidence interval.

Task 19: Graphical Visualisation of Confidence and Prediction Intervals

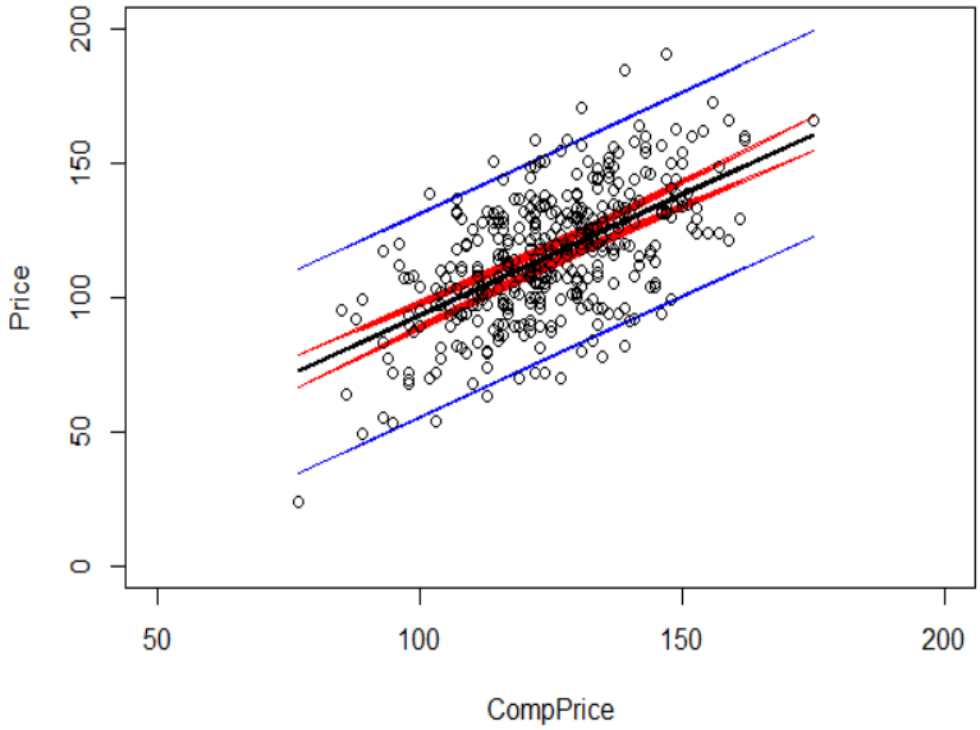
In this section, we will reaffirm that prediction interval is wider than confidence interval by the method of visualization. For doing so we will plot the lines of confidence interval and prediction interval along with the original data points to compare.

(i) Price vs. CompPrice

Using plot:

```
>plot(x=cpframe$CompPrice,type="n",y=pred.p1[,"fit"],ylim=c(0,200),
  xlim=c(50,200), xlab = "CompPrice", ylab = "Price")

>lines(cpframe$CompPrice,pred.p1[,"fit"],lwd=2)

>lines(cpframe$CompPrice,pred.p1[,"lwr"],col=2,lwd=1.5)

>lines(cpframe$CompPrice,pred.p1[,"upr"],col=2,lwd=1.5)

>lines(cpframe$CompPrice,pred.p2[,"lwr"],col=4,lwd=1.5)

>lines(cpframe$CompPrice,pred.p2[,"upr"],col=4,lwd=1.5)

>points(prodata$CompPrice, prodata$Price)
```

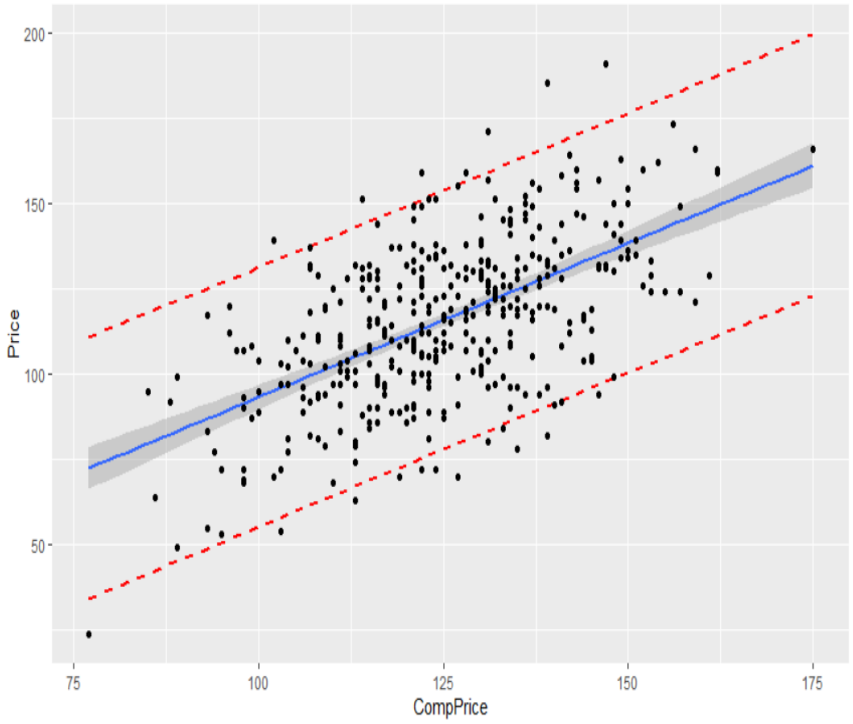
Blue lines represent the prediction interval whereas the red lines are the confidence interval of the predicted values of Price.



Using ggplot:

We can also use the ggplot command to prepare the above graph. You can use **geom_smooth()**to plot the line of best fit and use "se = TRUE" to plot the confidence interval.

```
>data1 <- cbind(prodata,predict(lmPr_CP, interval = "prediction"))

>ggplot(data1, aes(x = CompPrice, y = Price)) +
 geom_smooth(method = "lm", se = TRUE) +
 geom_line(aes(y = upr), col = "red", linetype = "dashed", lwd = 1) +
 geom_line(aes(y = upr), col = "red", linetype = "dashed", lwd = 1) +
 geom_point()
```

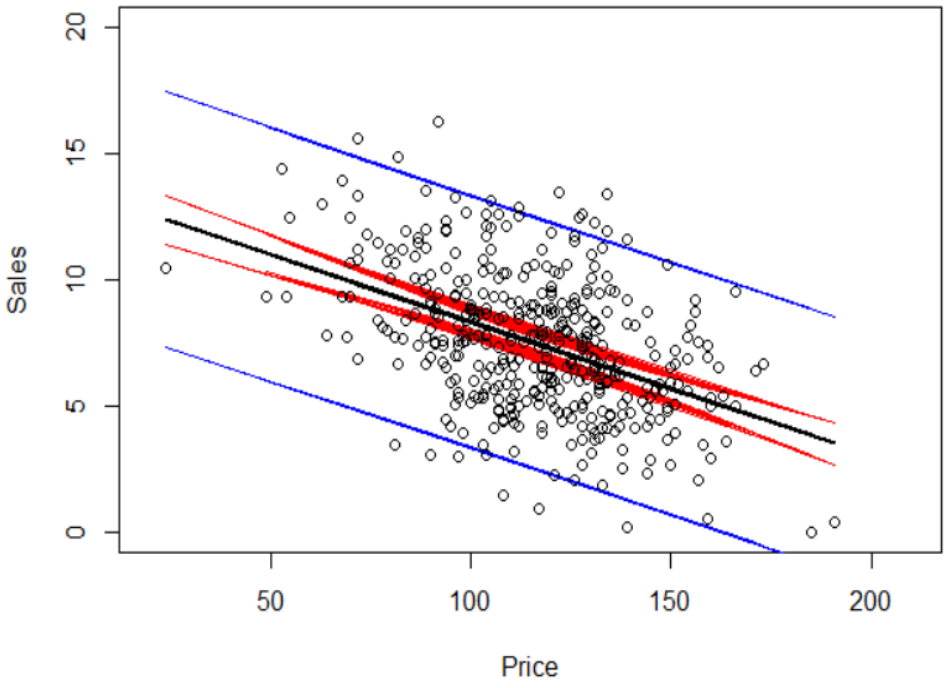


We clearly see that prediction interval is much wider than confidence interval.

(ii) Sales vs. Price

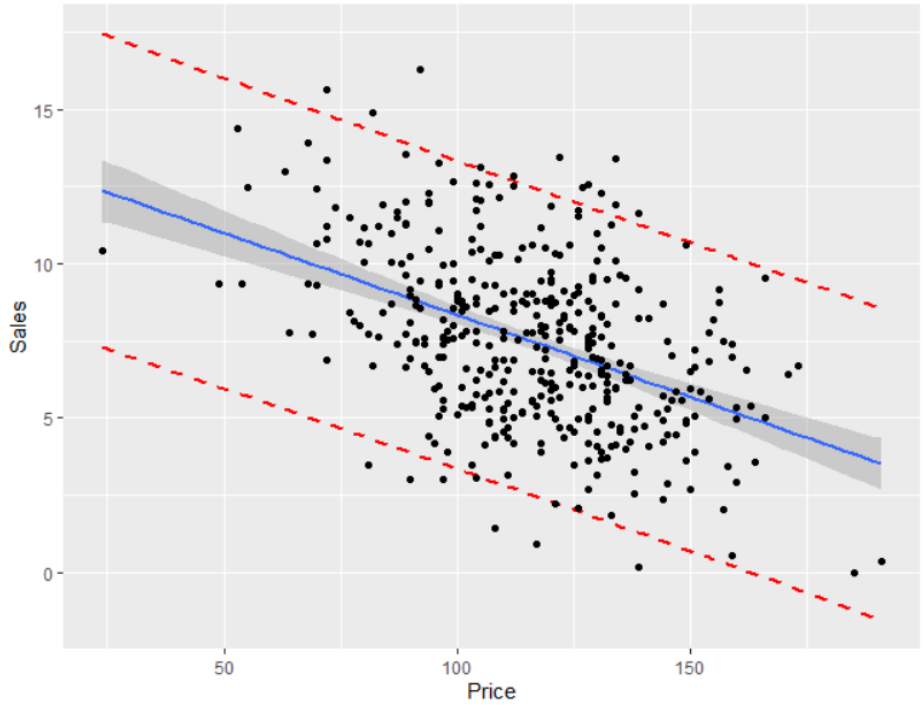
Using plot:

```
>plot(x=pframe$Price,type="n",y=pred.s1[,"fit"],ylim=c(0,20),
 xlim =c(20,210), xlab = “Price”, ylab = “Sales”)

>lines(pframe$Price,pred.s1[,"fit"],lwd=2)

>lines(pframe$Price,pred.s1[,"lwr"],col=2,lwd=1.5)

>lines(pframe$Price,pred.s1[,"upr"],col=2,lwd=1.5)

>lines(pframe$Price,pred.s2[,"lwr"],col=4,lwd=1.5)

>lines(pframe$Price,pred.s2[,"upr"],col=4,lwd=1.5)

>points(prodata$Price,prodata$Sales)
```



Using ggplot:

```
>data2 <- cbind(prodata,predict(lmS_P, interval = "prediction"))

>ggplot(data2, aes(x = Price, y = Sales)) +
 geom_smooth(method = "lm", se = TRUE) +
 geom_line(aes(y = upr), col = "red", linetype = "dashed", lwd = 1) +
 geom_line(aes(y = lwr), col = "red", linetype = "dashed", lwd = 1) +
 geom_point()
```



3.7. Robust Regression

Robust means resilient to change. Often, the assumptions made for the model may not hold. Thus, we might need a regression method that should work even if certain assumptions are violated. Such a regression is called robust regression.

Many of the problems with parametric regression can be avoided by fitting a so-called robust regression line. There are many variants of this. Here we just explore one method: lqs in the MASS package; this fits a regression to the "good" points in the dataset (as defined by some criterion), to produce a regression estimator with a high "breakdown" point, that is, even if certain assumptions don't hold, the performance of the estimator would not change much.

Task 20: Load the MASS package and compute a robust regression. Compare the fitted lines and the coefficient of determination (R^2) of this with those from the least-squares fit.

```
>library(MASS)
```

(i) Price vs. CompPrice

```
>lmPr_CP.r <- lqs(prodata$Price ~ prodata$CompPrice)

>summary(lmPr_CP.r)

 Length Class      Mode
crit            1    -none-     numeric
sing            1    -none-     character
coefficients    2    -none-     numeric
bestone         2    -none-     numeric
fitted.values 400    -none-     numeric
residuals     400    -none-     numeric
scale           2    -none-     numeric
terms           3    terms      call
call            2    -none-     call
xlevels         0    -none-     list
model           2    data.frame list

>lmPr_CP$coefficients
(Intercept)   CompPrice
  2.9411048   0.9030118

>lmPr_CP.r$coefficients
(Intercept) prodata$CompPrice
-2.648756          0.950000
```

Now we compute the residual sum of squares of both the models.

```
>sum(lmPr_CP$residuals^2)
[1] 147166.5

>sum(lmPr_CP.r$residuals^2)
[1] 147405.5
```

We see that residual sum of squares is slightly higher for robust regression.

(ii) Sales vs. Price

```
>lmS_P.r <- lqs(prodata$Sales ~ prodata$Price)

>summary(lmS_P.r)
             Length Class      Mode
crit           1    -none-     numeric
sing           1    -none-     character
coefficients   2    -none-     numeric
bestone        2    -none-     numeric
fitted.values 400   -none-     numeric
residuals     400   -none-     numeric
scale          2    -none-     numeric
terms          3    terms      call
call           2    -none-     call
xlevels        0    -none-     list
model          2    data.frame list


>lmS_P$coefficients
(Intercept)       Price
13.64191518 -0.05307302

>lmS_P.r$coefficients
  (Intercept) prodata$Price
  11.24377488   -0.03270833

>sum(lmS_P$residuals^2)
[1] 2552.244

>sum(lmS_P.r$residuals^2)
[1] 2645.646
```

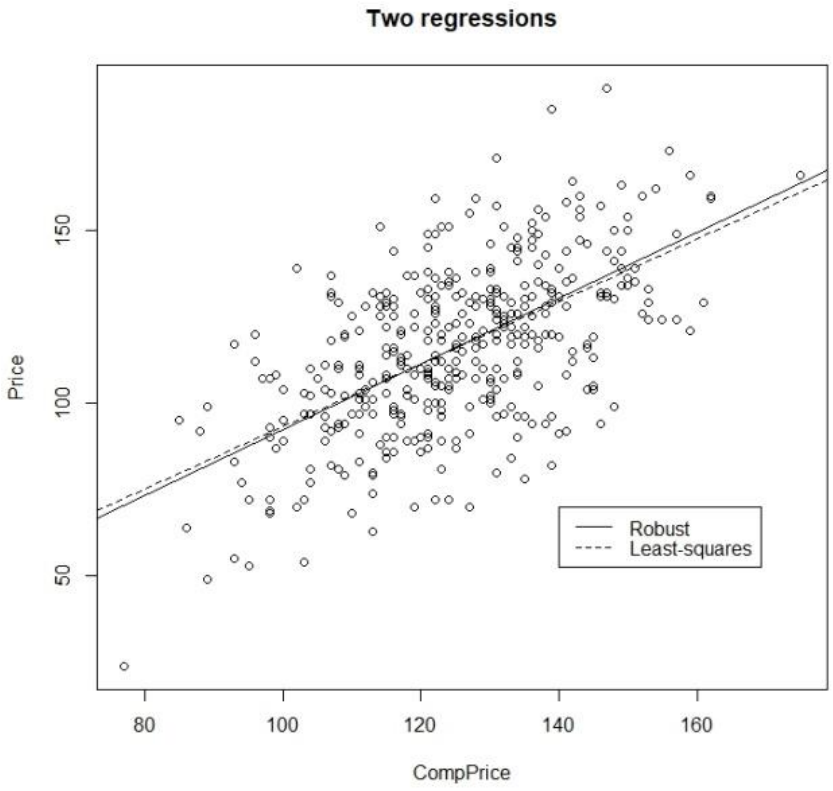
Q29. How does the sum of squares of residuals differ in case of Robust regression from that obtained by the Least Squares method?

Ans. The sum of squares of residuals is slightly higher in case of Robust regression.

Comparison of fitted lines:

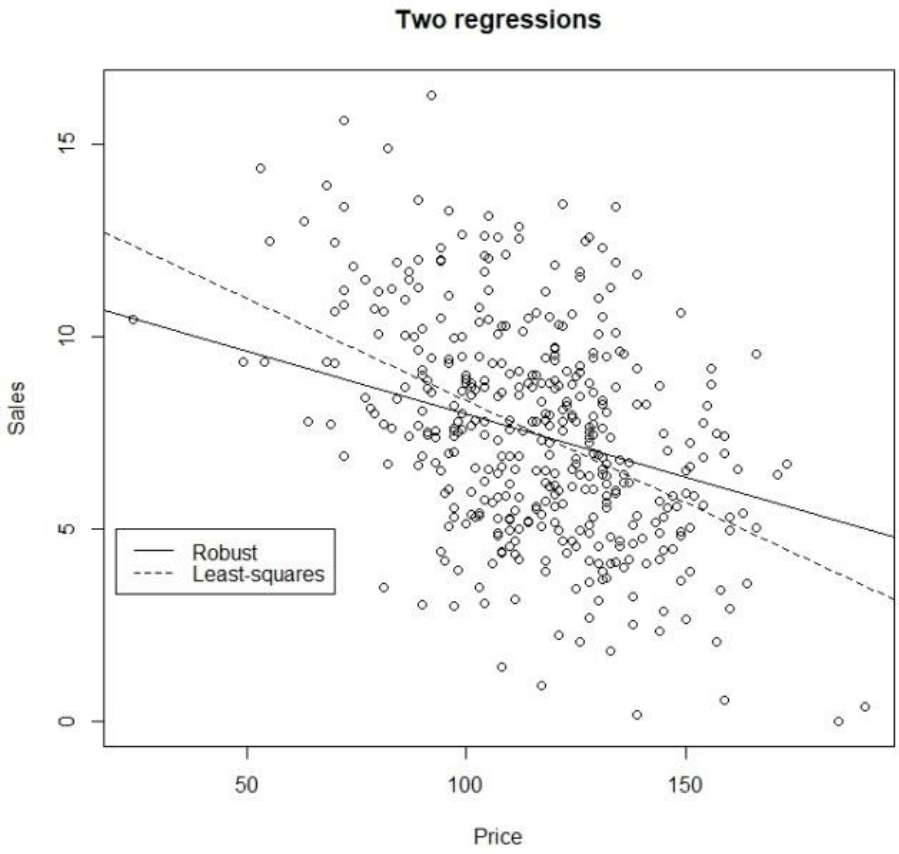
(i) Price vs. CompPrice

```
>plot(Price ~ CompPrice, data=prodata, main="Two regressions")
>abline(lmPr_CP, lty=2)
>abline(lmPr_CP.r, lty=1)
>legend(140,70, legend=c("Robust","Least-squares"), lty=1:2)
```



(ii) Sales vs. Price

```
>plot(Sales ~ Price, data=prodata, main="Two regressions")
>abline(lmS_P, lty=2)
>abline(lmS_P.r, lty=1)
>legend(20,5, legend=c("Robust","Least-squares"), lty=1:2)
```



Q30. What seems to be the advantage of Robust Regression?

Ans. It seems to give a better internal fit by compromising a little on the sum of squares of residuals. Specifically, it provides much better regression coefficient estimates when outliers are present in the data.
Outliers violate the assumption of normally distributed residuals in least squares regression. They tend to distort the least squares coefficients by having more influence than they deserve.

4. Multivariate Correlation and Regression

a. What – As the name suggests, Multivariate (multi – more than two) analysis is concerned with the analysis of relationship of more than two variables.

b. Why – It is useful in finding association between one response and more than one explanatory variable

c. How – We now describe the ways in which we can perform this analysis.

To fit a multiple linear regression model, we have to figure out the possible explanatory variables that can impact our response variable. To do so, we again start by analysing the scatter plot between variables of the Carseats dataset.

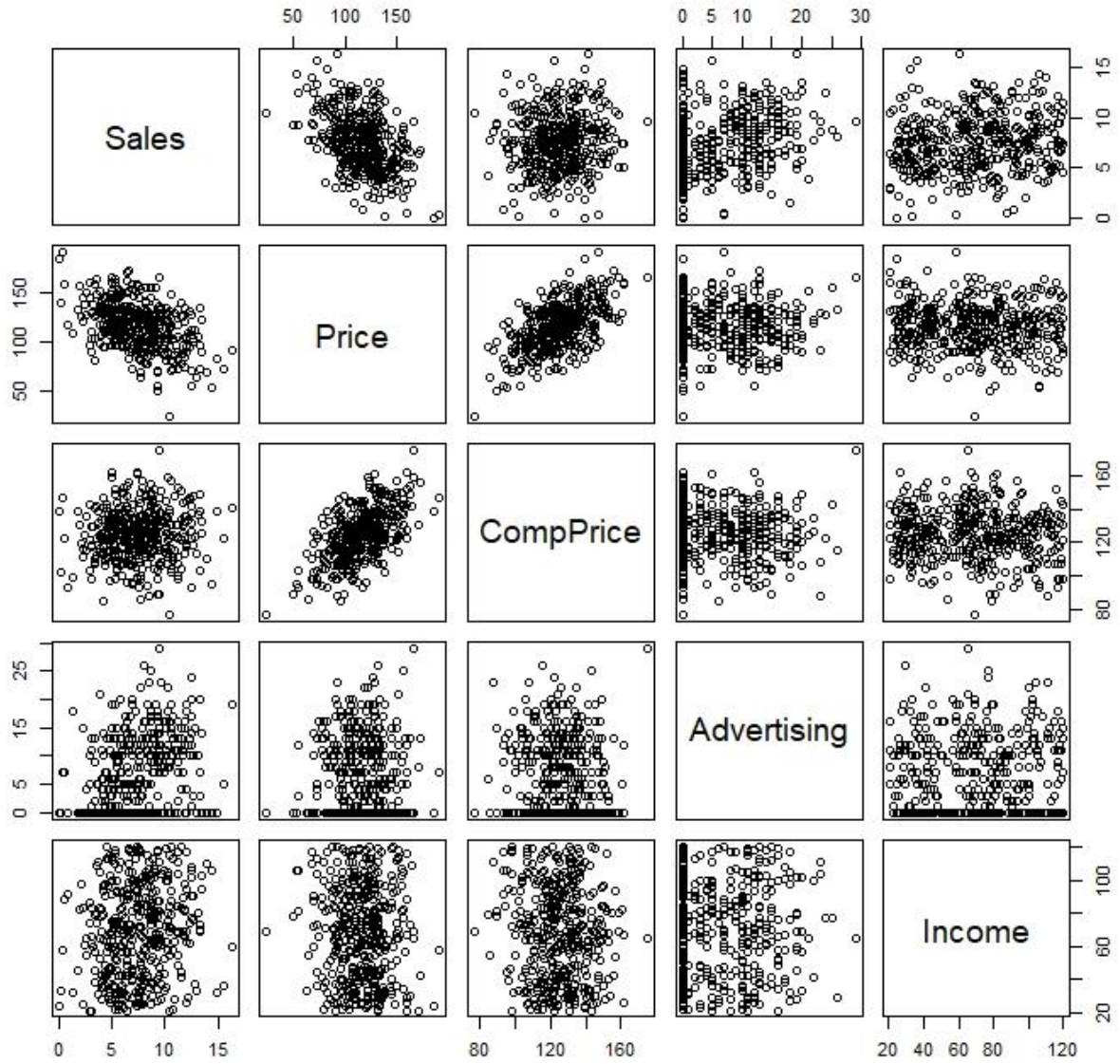
4.1. Multivariate Correlation Analysis

Task 21: Plot pair wise scatter plots for all numeric variables in the data. Also calculate these values.

The **pairs() function** returns a plot matrix, consisting of scatterplots for each variable-combination of a data frame**.**

Note: It can also be done using plot() function (as depicted in task 9)

```
>pairs(~ Sales + Price + CompPrice + Advertising + Income, data = prodata)
```



Q31. Describe the relationship between these variables

Ans. Apart from the positive association between Price and CompPrice, and negative association between Sales and Price, the other pairs seem to be uncorrelated.

We now find pairwise simple correlations using the **cor()** function.

```
>cor(prodata[c("Sales","Price","Advertising", "Income", "CompPrice")])
```

```
                Sales       Price Advertising      Income   CompPrice
Sales       1.00000000 -0.44495073  0.26950678  0.15195098  0.06407873
Price      -0.44495073  1.00000000  0.04453687 -0.05669820  0.58484777
Advertising 0.26950678  0.04453687  1.00000000  0.05899471 -0.02419879
Income      0.15195098 -0.05669820  0.05899471  1.00000000 -0.08065342
CompPrice   0.06407873  0.58484777 -0.02419879 -0.08065342  1.00000000
```

Q32. Describe the output in words.

Ans.

- Sales has a moderate negative correlation with Price, a low positive correlation with Advertising and Income and a very low positive correlation with CompPrice.
- Price has a very low positive and negative correlation with Advertising and Income while showing moderately strong correlation with CompPrice.
- Advertising has a very low positive and negative correlation with Income and CompPrice.
- Income has very low negative correlation with CompPrice.

4.2 Multiple Regression Analysis

Task 22: Null, simple and multiple regression models.

A **null model** contains no predictors and only an intercept term.
A **simple regression model** contains one predictor while a **multiple regression model** contains more than one predictor.

In this section, we will compare these models so as to observe if there is any improvement in terms of their capacity to explain the dependent variable.

In order to do so, we use three methods: comparison of adjusted R^2, comparison of AIC values and ANOVA.

We begin by fitting the required regression models and then proceed to find more information about each of them using the **summary()** function.

(i) Sales as the response variable

a) Null Model

A null model has no predictors. It just contains one intercept where the intercept is the mean of Y.We fit a null model by passing 1 as the explanatory variable.

```
>lmS_null <- lm(Sales ~ 1, data = prodata)

>summary(lmS_null)


Call:
lm(formula = Sales ~ 1, data = prodata)

Residuals:
    Min      1Q  Median      3Q     Max
-7.4963 -2.1063 -0.0063  1.8237  8.7737

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.4963     0.1412   53.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.824 on 399 degrees of freedom
```

Note: Null model does not return any R^2 or adjusted R^2 value as it has no independent variable in the model. Hence, total sum of squares is equal to the error sum of squares. Thus, $R^2 = 1 - \frac{SSE}{TSS} = 0$

b) Simple Regression Model (Sales ~ Price)

We have already fitted a simple linear regression model for Sales on price in the previous section of the book. Hence, we use summary function on the model object lmS_P

```
>summary(lmS_P)

Call:
lm(formula = Sales ~ Price, data = prodata)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5224 -1.8442 -0.1459  1.6503  7.5108

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.641915   0.632812  21.558<2e-16 ***
Price       -0.053073   0.005354  -9.912   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.532 on 398 degrees of freedom
Multiple R-squared:  0.198,    Adjusted R-squared:  0.196
F-statistic: 98.25 on 1 and 398 DF,  p-value: < 2.2e-16
```

Simple Regression Model (Sales ~ Advertising)

We fit another simple linear regression model with Sales as the response variable and Advertising as the explanatory variable.

```
>lmS_Ad <- lm(Sales ~ Advertising, data = prodata)

>summary(lmS_Ad)

Call:
lm(formula = Sales ~ Advertising, data = prodata)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3770 -1.9634 -0.1037  1.7222  8.3208

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.7370     0.1925  35.007  < 2e-16 ***
Advertising   0.1144     0.0205   5.583 4.38e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 2.723 on 398 degrees of freedom
Multiple R-squared:  0.07263,  Adjusted R-squared:  0.0703
F-statistic: 31.17 on 1 and 398 DF,  p-value: 4.378e-08
```

c) Multiple Regression Model

We now fit a MLRM using both Price and Advertising as the independent variables. It seems logical that advertising budget can influence the sales.

```
>lmS_P.Ad <- lm(Sales ~ Price + Advertising, data = prodata)

>summary(lmS_P.Ad)

Call:
lm(formula = Sales ~ Price + Advertising, data = prodata)

Residuals:
    Min      1Q  Median      3Q     Max
-7.9011 -1.5470 -0.0223  1.5361  6.3748

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.003427   0.606850  21.428< 2e-16 ***
Price       -0.054613   0.005078 -10.755  < 2e-16 ***
Advertising  0.123107   0.018079   6.809 3.64e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.399 on 397 degrees of freedom
Multiple R-squared:  0.2819,   Adjusted R-squared:  0.2782
F-statistic: 77.91 on 2 and 397 DF,  p-value: < 2.2e-16
```

Q33. How much of the total variability of the predictand is explained by each of the models? Give the three predictive equations, rounded to two decimals.

Ans. a) On the basis of the R^2 values of the models, we conclude the following

- 19.8% variability of Sales has been explained by Price
- 7.26% variability of Sales has been explained by the Advertising budget
- 28.19% variability of Sales is explained by a multiple regression model ofPrice and Advertising

b) Null: Sales = 7.496
Note that a null model always predicts sample mean of the response variable.
Simple: Sales = 13.642 - 0.059*Price
Simple: Sales = 6.737 + 0.114*Advertising
Multiple: Sales = 13.003 - 0.054*Price + 0.123*Advertising

(ii) Price as the response variable

a) Null Model

```
>lmPr_null <- lm(Price ~ 1, data = prodata)

>summary(lmPr_null)

Call:
lm(formula = Price ~ 1, data = prodata)

Residuals:
    Min      1Q  Median      3Q     Max
-91.795 -15.795   1.205  15.205  75.205

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  115.795      1.184   97.81   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.68 on 399 degrees of freedom
```

b) Simple Regression Model (Price ~ CompPrice)

```
>summary(lmPr_CP)

Call:
lm(formula = Price ~ CompPrice, data = prodata)

Residuals:
    Min      1Q  Median      3Q     Max
-48.473 -12.183   0.197  12.925  56.540


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.94110    7.90436   0.372     0.71
CompPrice    0.90301    0.06278  14.384<2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.23 on 398 degrees of freedom
Multiple R-squared:  0.342,    Adjusted R-squared:  0.3404
F-statistic: 206.9 on 1 and 398 DF,  p-value: < 2.2e-16
```

Simple Regression Model (Price ~ Sales)

```
>lmPr_S <- lm(Price ~ Sales, data = prodata)

>summary(lmPr_S)

Call:
lm(formula = Price ~ Sales, data = prodata)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-80.851 -15.332   0.528  13.315  57.791

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 143.7589     3.0143  47.692<2e-16 ***
Sales        -3.7304     0.3763  -9.912   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.23 on 398 degrees of freedom
Multiple R-squared:  0.198,    Adjusted R-squared:  0.196
F-statistic: 98.25 on 1 and 398 DF,  p-value: < 2.2e-16
```

c) Multiple Regression Model

```
>lmPr_CP.S <- lm(Price ~ CompPrice + Sales, data = prodata)

>summary(lmPr_CP.S)
Call:
lm(formula = Price ~ CompPrice + Sales, data = prodata)

Residuals:
    Min      1Q  Median      3Q     Max
-44.714 -11.112  -0.402   9.729  44.983

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.39563    6.56684   4.172 3.71e-05 ***
CompPrice    0.95094    0.05058  18.801< 2e-16 ***
Sales       -4.06122    0.27463 -14.788  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.46 on 397 degrees of freedom
Multiple R-squared:  0.5757,   Adjusted R-squared:  0.5736
F-statistic: 269.4 on 2 and 397 DF,  p-value: < 2.2e-16
```

Q34. How much of the total variability of the predictand is explained by each of the models? Give the three predictive equations, rounded to two decimals.

Ans.a) On the basis of the R^2 values of the models, we conclude the following

- 34.2% variability of Price has been explained by CompPrice
- 19.8% variability of Price has been explained by the Sales
- 57.57% variability of Price is explained by a multiple regression model of Sales and CompPrice

b) Null: Price = 115.795
Simple: Price = 2.941 + 0.903*CompPrice
Simple: Price = 143.759 - 3.7304* Sales
Multiple: Sales = Price = 27.396 + 0.951*CompPrice - 4.061*Sales

4.3 Comparing Models using adjusted R^2, AIC and ANOVA

1) Comparing regression models with the adjusted R^2

We have already discussed the concept of adjusted R2 in the previous sections of this book. We will now use it to compare models of both Sales and Price.

(i) Sales

```
>summary(lmS_null)$adj.r.squared
[1] 0

>summary(lmS_P)$adj.r.squared
[1] 0.195966

>summary(lmS_Ad)$adj.r.squared
[1] 0.07030384

>summary(lmS_P.Ad)$adj.r.squared
[1] 0.2782377
```

Q35. What is the effect of adding a variable to the simple model?

Ans. Adding Advertising variable to linear regression of Sales~Price increased adjusted R^2 from 0.196 to 0.278.

(ii) Price

```
>summary(lmPr_null)$adj.r.squared
[1] 0

>summary(lmPr_CP)$adj.r.squared
[1] 0.3403938

>summary(lmPr_S)$adj.r.squared
[1] 0.195966

>summary(lmPr_CP.S)$adj.r.squared
[1] 0.573605
```

Q36. What is the effect of adding a variable to the simple model?

Ans. Adding Sales variable to linear regression of Price~CompPrice increased adjusted R^2 from 0.3404 to 0.5736.

2) Comparing Regression Models with AIC

Akaike information criterion (AIC) is a technique based on in-sample fit to estimate the likelihood of a model to predict/estimate the future values. A good model is the one that has minimum AIC among all the other models. A **lower AIC** value indicates a **better** fit.

Let k be the number of estimated parameters in the model. Let be the maximum value of the likelihood function for the model. Then the AIC value of the model is given by
$AIC = 2k - 2\ln(\hat{L})$

To apply AIC in practice, we start with a set of candidate models, and then find the models' corresponding AIC values.We wish to select, from among the candidate models, the model that minimizes the information loss.

AIC() function from the stats package can be used to calculate AIC of the models. You have to pass model object as the argument of the AIC function.

(i) Sales

```
>AIC(lmS_null)
[1] 1968.706

>AIC(lmS_P)
[1] 1882.456

>AIC(lmS_Ad)
[1] 1940.543

>AIC(lmS_P.Ad)
[1] 1840.272
```

Q37. Which model is the best according to AIC?

Ans. We observe that multiple linear regression model has lowest AIC and hence it is the best.

(ii) Price

```
>AIC(lmPr_null)
[1] 3669.742

>AIC(lmPr_CP)
[1] 3504.293

>AIC(lmPr_S)
[1] 3583.492
>AIC(lmPr_CP.S)
```

```
[1] 3330.776
```

We again see that multiplelinear regression has lowest AIC and hence it is the best among the models under study.

3) Comparing regression models with ANOVA

We can use ANOVA to compare two models by using F-test. To perform it in R, you have to pass the objects of the models to be compared as two separate arguments.

ANOVA to compare two regression models

Let us say that we have to compare two models, a full model and a reduced model. The full model might have more variables than the reduced model.

For example, in the multiple regression case we have:
Reduced model $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$ and
Full model $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \beta_{k+1} x_{k+1} + \cdots + \beta_p x_p + \varepsilon$

We want to test the hypothesis that the full model adds explanatory value over the reduced model i.e. full model is better than the reduced model.

Mathematically, the **hypothesis** is

$$H_0: \beta_{k+1} = \cdots = \beta_p = 0 \text{ vs.}$$
$$H_1: \beta_i \neq 0 \text{ for at least one i (i = (k+1) to p)}$$

Test Statistic:

$$F = \frac{(SSE_{reduced} - SSE_{full})/(p-k)}{(SSE_{full})/(n-p-1)} \sim F_{(p-k,\ n-p-1)}$$

Test Criteria:
Reject H_0 at α% level of significance if p-value < α. Otherwise, do not reject H_0.

(i) Sales

```
>(a1 <- anova(lmS_P.Ad,lmS_P))
Analysis of Variance Table

Model 1: Sales ~ Price + Advertising
Model 2: Sales ~ Price
Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    397 2285.3
```

```
2    398 2552.2 -1   -266.91 46.367 3.64e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>diff(a1$RSS)/a1$RSS[2]
[1] 0.1045792
```

Q38. How does the RSS of multiple regression model compare with simple model of Sales ~ Price?

Ans. Simple model has one higher degrees of freedom as it has one less independent variable in the model.We also see a 10.5% reduction in Residual Sum of Squares as compared to simple model.

```
>(a2 <- anova(lmS_P.Ad,lmS_Ad))
Analysis of Variance Table

Model 1: Sales ~ Price + Advertising
Model 2: Sales ~ Advertising
Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    397 2285.3
2    398 2951.1 -1    -665.8 115.66 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>diff(a2$RSS)/a2$RSS[2]
[1] 0.2256085
```

22.5% reduction in Residual Sum of Squares as compared to simple model of Sales ~ Advertising.

We see that p-value < 0.05 for both the ANOVA tests. Hence, we reject the null hypothesis and conclude that multiple regression model of Sales is better than both the simple models.

(ii) Price

```
>(a3 <- anova(lmPr_CP.S, lmPr_CP))
Analysis of Variance Table

Model 1: Price ~ CompPrice + Sales
Model 2: Price ~ CompPrice
Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    397  94895
2    398 147166 -1    -52271 218.68 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>diff(a3$RSS)/a3$RSS[2]
[1] 0.3551855
```

Q39. How does the RSS compare with simple model of Price ~ CompPrice?

Ans. 35.5% reduction in Residual Sum of Squares as compared to simple model.

```
>(a4 <- anova(lmPr_CP.S, lmPr_S))
Analysis of Variance Table

Model 1: Price ~ CompPrice + Sales
Model 2: Price ~ Sales
Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    397  94895
2    398 179390 -1    -84495 353.49 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>diff(a4$RSS)/a4$RSS[2]
[1] 0.4710128
```

47.1% reduction in Residual Sum of Squares as compared to simple model of Price ~ Sales. On conducting ANOVA, we conclude that multiple regression model is better even in case of Price as the response variable.

4.4. Regression Diagnostics

Task 23: Analyse the regression diagnostics for the best model.

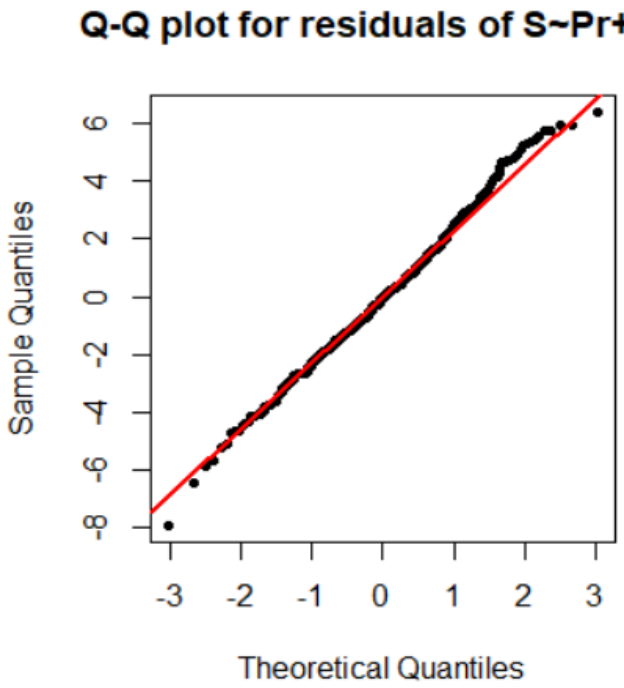
Since we have found multiple regression model to be the best one for the given dataset, we now try to check if the assumptions made are applicable or not.

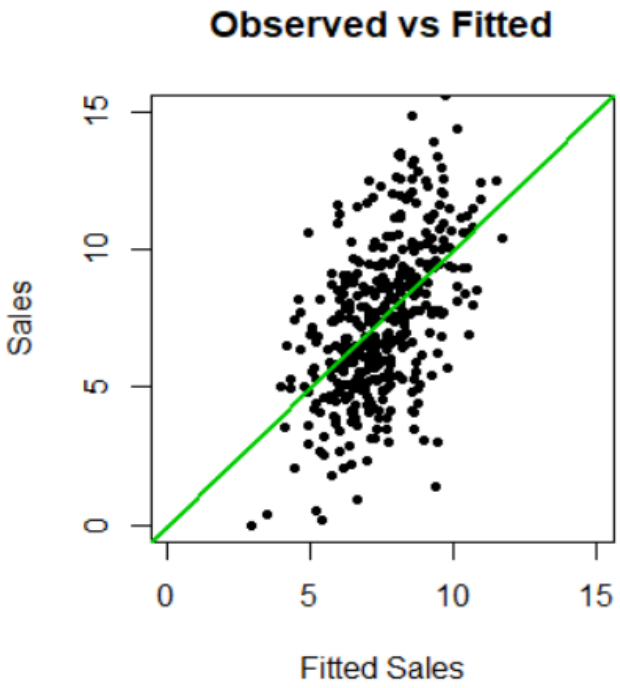
To plot the diagnostics of a multiple regression model, we will make 3 plots.

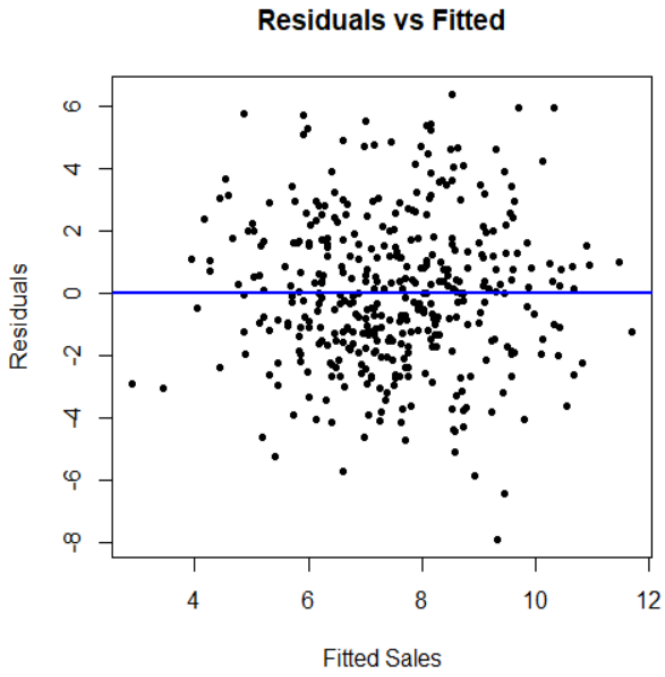
1) To for normality of residuals, we make a Q-Q plot of the residuals
2) To assess the fit of the model, we plot fitted vs. observed values of the response variable and compare it with a line having intercept 0 and slope 1.
3) To check for the assumption of homoscedasticity (constant variance), we plot residuals vs. fitted values and check if the scatter is random or not.

(i) Sales

```
> #1) Q-Q plot of residuals

>par(mfrow = c(1,2))

>qqnorm(residuals(lmS_P.Ad), pch=20, main = "Q-Q plot for residuals of
 S~Pr+Ad")

>qqline(residuals(lmS_P.Ad), col = 2, lwd = 2)


>#2) Fitted vs Actual Sales

>plot(Sales,fitted(lmS_P.Ad), pch = 20, xlim = c(0,15), ylim = c(0,15),
 main = "Observed vs Fitted", ylab = "Fitted Sales")

>abline(0,1, lwd = 2, col = 3)

>par(mfrow = c(1,1))


>#3) Residuals vs Fitted Sales

> plot(fitted(lmS_P.Ad), residuals(lmS_P.Ad), pch = 20, main = "Residuals
vs Fitted", xlab = "Fitted Sales", ylab = "Residuals")

>abline(h = 0, col = 4, lwd = 2)
```



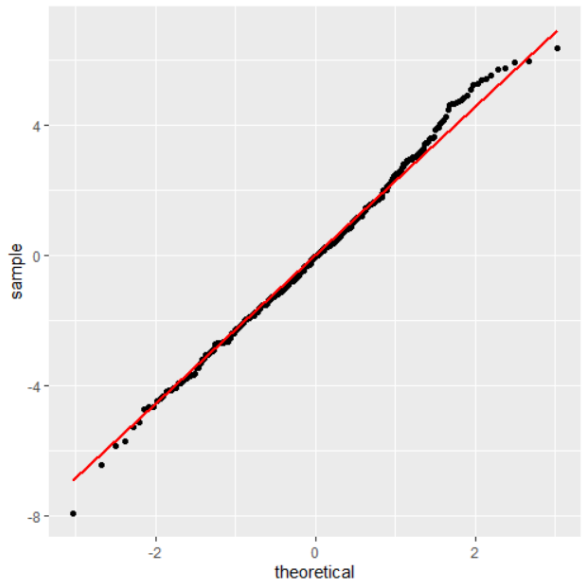


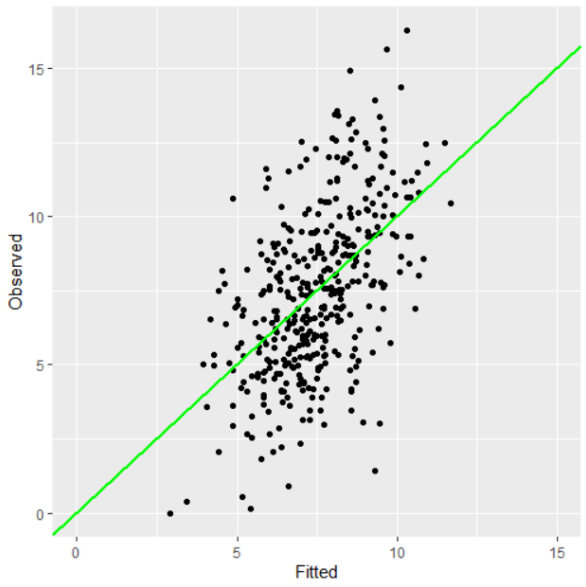


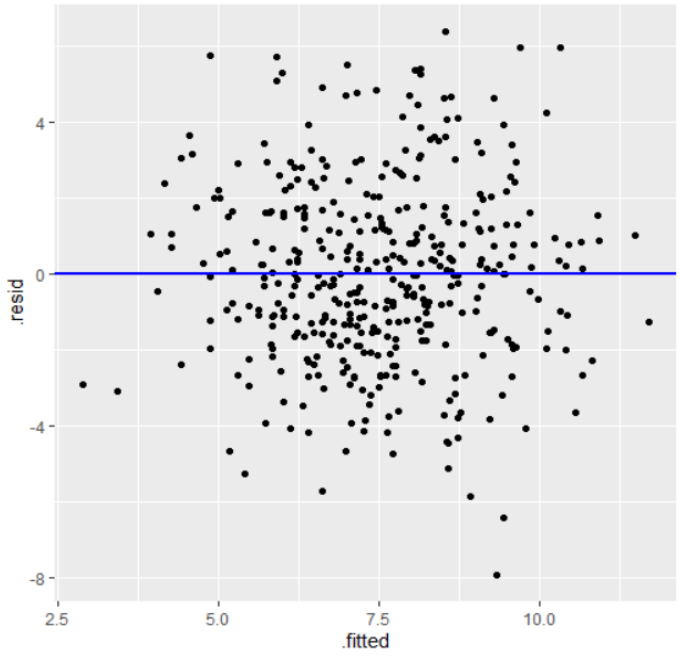
Alternatively using ggplot

```
>#1) Q-Q plot of residuals

>ggplot(augment(lmS_P.Ad)) +
geom_qq(aes(sample = .resid)) +
geom_qq_line(aes(sample = .resid), col = "red", lwd = 1)
```

```
> #2) Fitted vs Actual Sales

>ggplot(augment(lmS_P.Ad), aes(x = .fitted, y = Sales)) +
geom_point() + geom_abline(intercept= 0, slope = 1, col = "green",
lwd = 1) + scale_x_continuous(limits = c(0,15)) +xlab("Fitted") +
ylab("Observed")


> #3) Residuals vs Fitted Sales

>ggplot(augment(lmS_P.Ad), aes(x = .fitted, y = .resid)) +
geom_point() +
geom_abline(intercept = 0, slope = 0, col = "blue", lwd = 1)
```







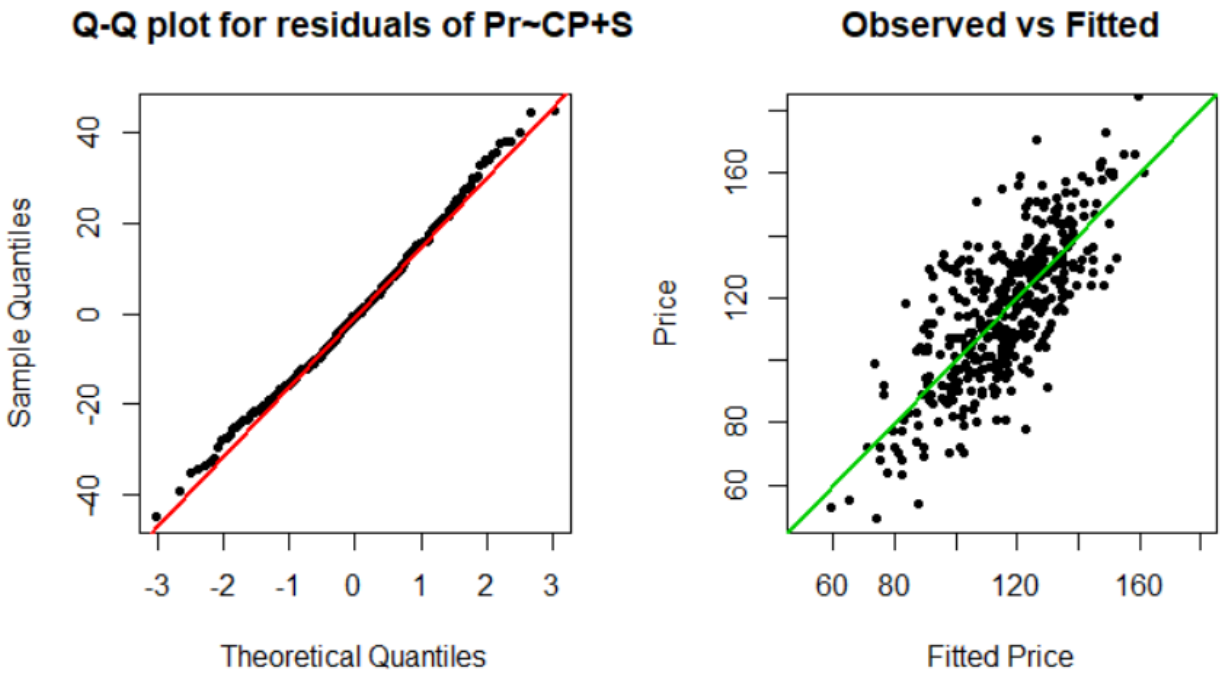
Q40. Are the residuals normally distributed? What can you conclude from these diagnostic plots?

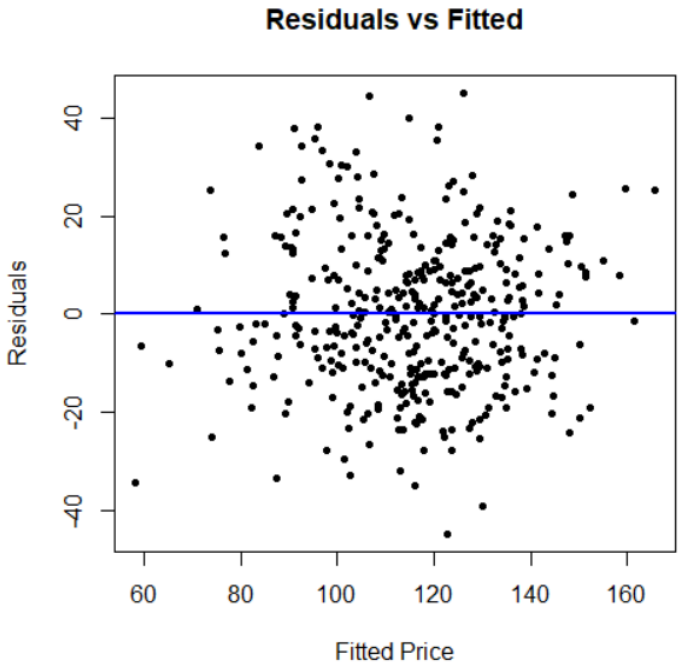
Ans. The distribution seems to be normal except towards the two ends.
On looking at the Observed vs. Fitted curve, the scatter is too much to give a good enough despite an upward sloping line being fit.

There is no pattern in residuals vs. fitted values plot. Hence, we can conclude that our model is homoscedastic.

(ii) Price

```
>#1) Q-Q plot of residuals

>par(mfrow = c(1,2))

>qqnorm(residuals(lmPr_CP.S), pch=20, main = "Q-Q plot for residuals of
Pr~CP+S")

>qqline(residuals(lmPr_CP.S), col = 2, lwd = 2)


> #2) Fitted vs Actual Price

> plot(fitted(lmPr_CP.S),prodata$Price, pch = 20,xlim = c(50,180),
ylim = c(50,180), main = "Observed vs Fitted", xlab = "Fitted Price",
ylab = "Price")

>abline(0,1, col = 3, lwd =2)

>par(mfrow = c(1,1))


>#  3) Residuals vs Fitted Sales

> plot(fitted(lmPr_CP.S), residuals(lmPr_CP.S), pch = 20, main =
   "Residuals vs Fitted", xlab = "Fitted Price", ylab = "Residuals")

>abline(h = 0, col = 4, lwd = 2)
```

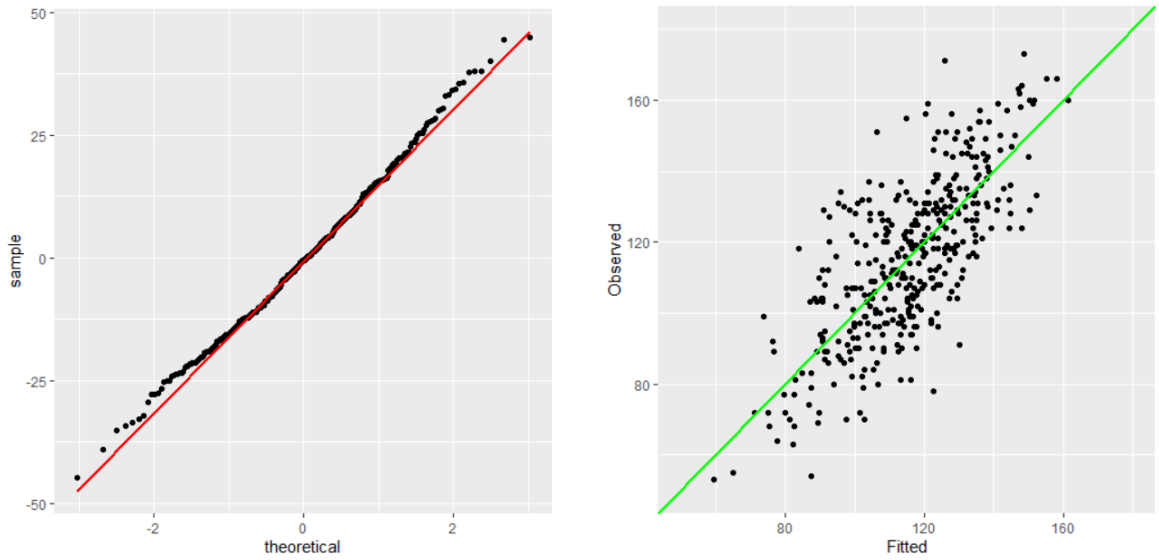


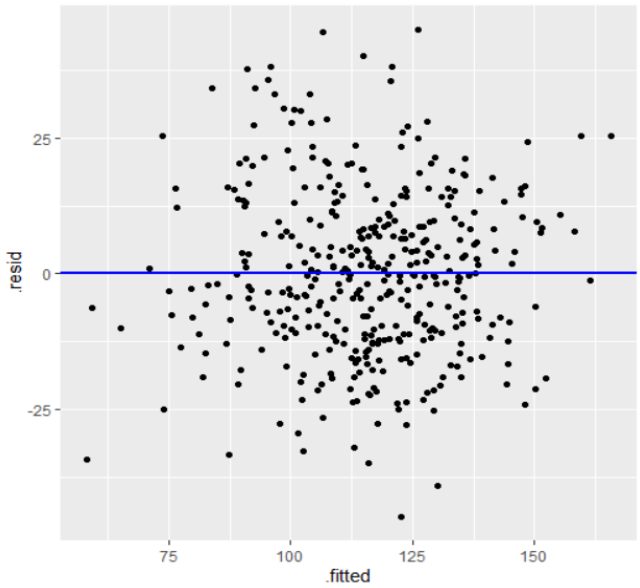


Alternatively using ggplot

```
>#1) Q-Q plot of residuals

>ggplot(augment(lmPr_CP.S)) +
geom_qq(aes(sample = .resid)) +
geom_qq_line(aes(sample = .resid), col = "red", lwd = 1)


> #2) Predicted vs Actual Price
```

```
>ggplot(augment(lmPr_CP.S), aes(x = .fitted, y = Price)) +
geom_point() + geom_abline(intercept= 0, slope = 1, col = "green",
lwd = 1) + scale_x_continuous(limits = c(50,180)) +
scale_y_continuous(limits = c(50,180)) +
xlab("Fitted") + ylab("Observed")


>#3) Residuals vs Fitted Price

>ggplot(augment(lmPr_CP.S), aes(x = .fitted, y = .resid)) +
geom_point() +
geom_abline(intercept = 0, slope = 0, col = "blue", lwd = 1)
```





The distribution seems to be normal except towards the two ends.
On looking at the Observed vs. Fitted curve, the upward sloping line seems to be a good fit.
There is no pattern in residuals vs. Fitted values plot.

4.5. Stepwise Regression

While dealing with datasets, it is important to find the variables that are most important in predicting the outcome variable. It is necessary due to the computational costs as well as the bias-variance trade-off.
There are many approaches to decide which variables should be retained in the model. We discuss one such method of variable selection.

Task 24: Apply backward stepwise regression on the full model.

In backward stepwise regression, we start with a full model, that is, a model that uses all explanatory variables, say k, to predict the dependent variable. We then proceed to remove a variable on the basis of some criteria (here we use AIC). We fit the model again with k-1 explanatory variables and repeat the procedure until the loss of information is prevented by removing none of the remaining variables.

In R, we use **step()** function to perform backward stepwise regression. We pass the model object having all the remaining variables as the independent variable. A shortcut is to use "." in order to specify all the variables in the data frame as the explanatory variables.

(i) Sales

```
>lms <- step(lm(Sales~.,data=prodata))
Start:  AIC=26.82
Sales ~ CompPrice + Income + Advertising + Population + Price +
    ShelveLoc + Age + Education + Urban + US

Df Sum of Sq     RSS    AIC
- Population   1      0.33  403.16  25.15
- Education    1      1.19  404.02  26.00
- Urban        1      1.23  404.06  26.04
- US           1      1.57  404.40  26.38
<none>402.83  26.82
- Income       1     76.16  478.99  94.09
- Advertising  1    127.14  529.97 134.54
- Age          1    217.44  620.27 197.48
- CompPrice    1    519.91  922.74 356.35
- ShelveLoc    2   1053.20 1456.03 536.80
- Price        1   1323.23 1726.06 606.85

Step:  AIC=25.15
Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
    Age + Education + Urban + US

Df Sum of Sq     RSS    AIC
- Urban        1      1.15  404.31  24.29
- Education    1      1.36  404.52  24.49
- US           1      1.89  405.05  25.02
<none>403.16  25.15
```

```
- Income       1     75.94  479.10  92.18
- Advertising  1    145.38  548.54 146.32
- Age          1    218.52  621.68 196.38
- CompPrice    1    521.69  924.85 355.27
- ShelveLoc    2   1053.18 1456.34 534.89
- Price        1   1323.51 1726.67 605.00

Step:  AIC=24.29
Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
    Age + Education + US

Df Sum of Sq     RSS    AIC
- Education    1      1.44  405.76  23.72
- US           1      1.85  406.16  24.12
<none>404.31  24.29
- Income       1     76.64  480.96  91.73
- Advertising  1    146.03  550.34 145.63
- Age          1    217.59  621.91 194.53
- CompPrice    1    526.17  930.48 355.69
- ShelveLoc    2   1053.93 1458.25 533.41
- Price        1   1322.80 1727.11 603.10

Step:  AIC=23.72
Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
    Age + US

Df Sum of Sq     RSS    AIC
- US           1      1.63  407.39  23.32
<none>405.76  23.72
- Income       1     77.87  483.62  91.94
- Advertising  1    145.30  551.06 144.15
- Age          1    217.97  623.73 193.70
- CompPrice    1    525.25  931.00 353.92
- ShelveLoc    2   1056.88 1462.64 532.61
- Price        1   1322.83 1728.58 601.44

Step:  AIC=23.32
Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc +
    Age

Df Sum of Sq     RSS    AIC
<none>407.39  23.32
- Income       1     76.68  484.07  90.30
- Age          1    219.12  626.51 193.48
- Advertising  1    234.03  641.42 202.89
- CompPrice    1    523.83  931.22 352.01
- ShelveLoc    2   1055.51 1462.90 530.68
- Price        1   1324.42 1731.81 600.18
```

Best model for Sales (given by backward stepwise regression) contains Income, Age, Advertising, CompPrice, ShelveLoc and Price.

```
>lmsbest <- lm(Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc
+ Age, data = prodata)

>summary(lmsbest)
```

```
Call:
lm(formula = Sales ~ CompPrice + Income + Advertising + Price +
    ShelveLoc + Age, data = prodata)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7728 -0.6954  0.0282  0.6732  3.3292

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.475226   0.505005   10.84   <2e-16 ***
CompPrice       0.092571   0.004123   22.45   <2e-16 ***
Income          0.015785   0.001838    8.59   <2e-16 ***
Advertising     0.115903   0.007724   15.01   <2e-16 ***
Price          -0.095319   0.002670  -35.70   <2e-16 ***
ShelveLocGood   4.835675   0.152499   31.71   <2e-16 ***
ShelveLocMedium 1.951993   0.125375   15.57   <2e-16 ***
Age            -0.046128   0.003177  -14.52   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.019 on 392 degrees of freedom
Multiple R-squared:  0.872,    Adjusted R-squared:  0.8697
F-statistic: 381.4 on 7 and 392 DF,  p-value: < 2.2e-16
```

Q41. What is the best model with Sales as the response? How much variability is explained by this model?

Ans. Sales ~ CompPrice + Income + Advertising + Price + ShelveLoc + Age
86.97% of the variability is explained by the model as given by the adjusted R^2 value

Q42. Are all the explanatory variables in the best model for Sales significant?

Ans. Since, p-value of the t-test of significance of all the regression coefficients is almost zero. We conclude that all the explanatory variables in the best model are significant.

(ii) Price

```
>lmp <- step(lm(Price~.,data=prodata))
Start:  AIC=1800.61
Price ~ Sales + CompPrice + Income + Advertising + Population +
    ShelveLoc + Age + Education + Urban + US


Df Sum of Sq    RSS    AIC
- Population   1        63  34024 1799.3
- US           1        68  34030 1799.4
- Education    1        69  34031 1799.4
- Urban        1       103  34064 1799.8
<none>                     33962 1800.6
- Income       1      4502  38464 1848.4
```

```
- Advertising  1      8182  42144 1885.0
- Age          1     13902  47864 1935.9
- ShelveLoc    2     44578  78539 2132.0
- CompPrice    1     82501 116462 2291.5
- Sales        1    111558 145519 2380.6

Step:  AIC=1799.35
Price ~ Sales + CompPrice + Income + Advertising + ShelveLoc +
    Age + Education + Urban + US

Df Sum of Sq    RSS    AIC
- Education    1        87  34111 1798.4
- Urban        1        93  34118 1798.4
- US           1        96  34120 1798.5
<none>                      34024 1799.3
- Income       1      4480  38504 1846.8
- Advertising  1      9541  43566 1896.2
- Age          1     14021  48046 1935.4
- ShelveLoc    2     44537  78562 2130.1
- CompPrice    1     82675 116700 2290.4
- Sales        1    111697 145722 2379.2

Step:  AIC=1798.37
Price ~ Sales + CompPrice + Income + Advertising + ShelveLoc +
    Age + Urban + US

Df Sum of Sq    RSS    AIC
- US           1        84  34195 1797.3
- Urban        1        99  34210 1797.5
<none>                      34111 1798.4
- Income       1      4538  38649 1846.3
- Advertising  1      9488  43599 1894.5
- Age          1     14027  48138 1934.1
- ShelveLoc    2     44565  78677 2128.7
- CompPrice    1     82589 116700 2288.4
- Sales        1    111610 145722 2377.2

Step:  AIC=1797.34
Price ~ Sales + CompPrice + Income + Advertising + ShelveLoc +
    Age + Urban

Df Sum of Sq    RSS    AIC
- Urban        1        96  34291 1796.5
<none>                      34195 1797.3
- Income       1      4470  38665 1844.5
- Age          1     14052  48247 1933.0
- Advertising  1     15193  49387 1942.4
- ShelveLoc    2     44482  78677 2126.7
- CompPrice    1     82548 116743 2286.5
- Sales        1    111556 145751 2375.3

Step:  AIC=1796.46
Price ~ Sales + CompPrice + Income + Advertising + ShelveLoc +
    Age

Df Sum of Sq    RSS    AIC
<none>                      34291 1796.5
- Income       1      4516  38807 1844.0
- Age          1     13979  48270 1931.2
- Advertising  1     15311  49602 1942.1
```

```
- ShelveLoc    2     44408  78699 2124.8
- CompPrice    1     83470 117761 2288.0
- Sales        1    111479 145770 2373.3
```

Best model for Price (given by backward stepwise regression)containsIncome, Age, Advertising, ShelveLoc, CompPrice and Sales.

```
>lmpbest<- lm(Price ~ CompPrice + Income + Advertising + Sales + ShelveLoc
+ Age, data = prodata)

>summary(lmpbest)

Call:
lm(formula = Price ~ CompPrice + Income + Advertising + Sales +
    ShelveLoc + Age, data = prodata)

Residuals:
    Min      1Q  Median      3Q     Max
-23.352  -6.626   0.285   6.311  27.765

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     45.35149    4.75990   9.528  < 2e-16 ***
CompPrice        0.95336    0.03086  30.890< 2e-16 ***
Income           0.12413    0.01728   7.185 3.41e-12 ***
Advertising      0.97808    0.07393  13.230< 2e-16 ***
Sales           -8.02319    0.22475 -35.699  < 2e-16 ***
ShelveLocGood   39.22310    1.74755  22.445< 2e-16 ***
ShelveLocMedium 15.76152    1.22781  12.837< 2e-16 ***
Age             -0.38509    0.03046 -12.641  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.353 on 392 degrees of freedom
Multiple R-squared:  0.8467,   Adjusted R-squared:  0.844
F-statistic: 309.3 on 7 and 392 DF,  p-value: < 2.2e-16
```

Q43. What is the best model with Price as the response? How much variability is explained by this model?

Ans. Price ~ CompPrice + Income + Advertising + Sales + ShelveLoc + Age
Adjusted R^2 = 0.844, Hence 84.4% variability in Price is explained by the best model.

4.6 Multicollinearity

Task 25: Diagnosing Multicollinearity

Multicollinearity refers to a situation in which two or more explanatory variables in a multiple regression model are highly linearly related.
The following are the consequences of Multicollinearity:

- It is difficult to obtain estimates of the parameters in a model.
- If obtained, the estimates of regression coefficients have large standard errors.
- Hence, we almost always get insignificant t-ratios of regression coefficients.

VIF approach to detect Multicollinearity:
VIF or Variance Inflation Factor is defined as $\boldsymbol{VIF = \frac{1}{1 - R_j^2}}$
where, R_j^2 is the multiple correlation coefficient of j^{th} explanatory variable on all the remaining explanatory variables.

In general, a VIF of 5 or 10 and above indicates a multicollinearity problem.

In R, we use VIF() function from the car package to check for multicollinearity.

```
>library(car)
```

(i) Sales

```
>vif(lmsbest)
                GVIF Df GVIF^(1/(2*Df))
CompPrice1   .534883  1        1.238904
Income      1.015448  1        1.007694
Advertising 1.012935  1        1.006447
Price       1.534425  1        1.238719
ShelveLoc1   .015139  2        1.003763
Age         1.016830  1        1.008380
```

Q44. According to the VIF>10 criterion, which variables are highly correlated with the others?

Ans. Using VIF>10, variables do not appear to be highly collinear.

(ii) Price

```
>vif(lmpbest)
                GVIF Df GVIF^(1/(2*Df))
CompPrice1   .021632  1        1.010758
Income      1.066169  1        1.032555
Advertising 1.102532  1        1.050015
Sales       1.837543  1        1.355560
ShelveLoc1   .588323  2        1.122625
Age         1.110891  1        1.053988
```

There appears to be no multicollinearity in the data.

4.7. Parallel Slopes Model

Task 26: Combine a continuous and a discrete explanatory variable to predict Sales.

A regression in which we have additive effects of a numerical and categorical predictor is termed as a **parallel slopes model**. In parallel regression the lines of best fit, corresponding to distinct classes of the categorical variable, have same slope but different intercepts i.e. there is only one regression line, which is displaced up or down for each class of the categorical predictor. Hence, the name parallel slopes.

In R, a simple linear regression formula looks like y ~ x, where y is the name of the response variable, and x is the name of the explanatory variable. Here, we will simply extend this formula to include multiple explanatory variables.

A parallel slopes model has the form y ~ x + z, where z is a categorical explanatory variable and x is a numerical explanatory variable.

While performing univariate analysis, we had noted that Sales is impacted by ShelveLoc. Thus, we will use ShelveLoc as the categorical explanatory variable in parallel slopes analysis.

a) Simple model (Sales ~ Price)

```
>summary(lmS_P)

Call:
lm(formula = Sales ~ Price, data = prodata)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5224 -1.8442 -0.1459  1.6503  7.5108

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.641915   0.632812  21.558<2e-16 ***
Price       -0.053073   0.005354  -9.912   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.532 on 398 degrees of freedom
Multiple R-squared:  0.198,    Adjusted R-squared:  0.196
F-statistic: 98.25 on 1 and 398 DF,  p-value: < 2.2e-16
```

b) Simple model (Sales ~ ShelveLoc)

```
>lmS_sh<- lm(Sales ~ ShelveLoc, data = prodata)

>summary(lmS_sh)
```

```
Call:
lm(formula = Sales ~ ShelveLoc, data = prodata)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3066 -1.6282 -0.0416  1.5666  6.1471

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.5229     0.2388  23.131  < 2e-16 ***
ShelveLocGood     4.6911     0.3484  13.464< 2e-16 ***
ShelveLocMedium   1.7837     0.2864   6.229  1.2e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.339 on 397 degrees of freedom
Multiple R-squared:  0.3172,   Adjusted R-squared:  0.3138
F-statistic: 92.23 on 2 and 397 DF,  p-value: < 2.2e-16
```

c) Combined model (Sales ~ Price + ShelveLoc)

```
>lmS_P.sh <- lm(Sales ~ Price + ShelveLoc, data = prodata)

>summary(lmS_P.sh)

Call:
lm(formula = Sales ~ Price + ShelveLoc, data = prodata)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8229 -1.3930 -0.0179  1.3868  5.0780

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     12.001802   0.503447  23.839  < 2e-16 ***
Price           -0.056698   0.004059 -13.967  < 2e-16 ***
ShelveLocGood    4.895848   0.285921  17.123< 2e-16 ***
ShelveLocMedium  1.862022   0.234748   7.932 2.23e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.917 on 396 degrees of freedom
Multiple R-squared:  0.5426,   Adjusted R-squared:  0.5391
F-statistic: 156.6 on 3 and 396 DF,  p-value: < 2.2e-16
```

Q45. Which model has the highest adjusted R^2? By how much?

Ans. Combined model has the highest adjusted R^2. On adding Shelveloc to the model adjusted R^2 increases from 0.196 to 0.5391.

Q46. Is ShelveLoc significant in the combined model?

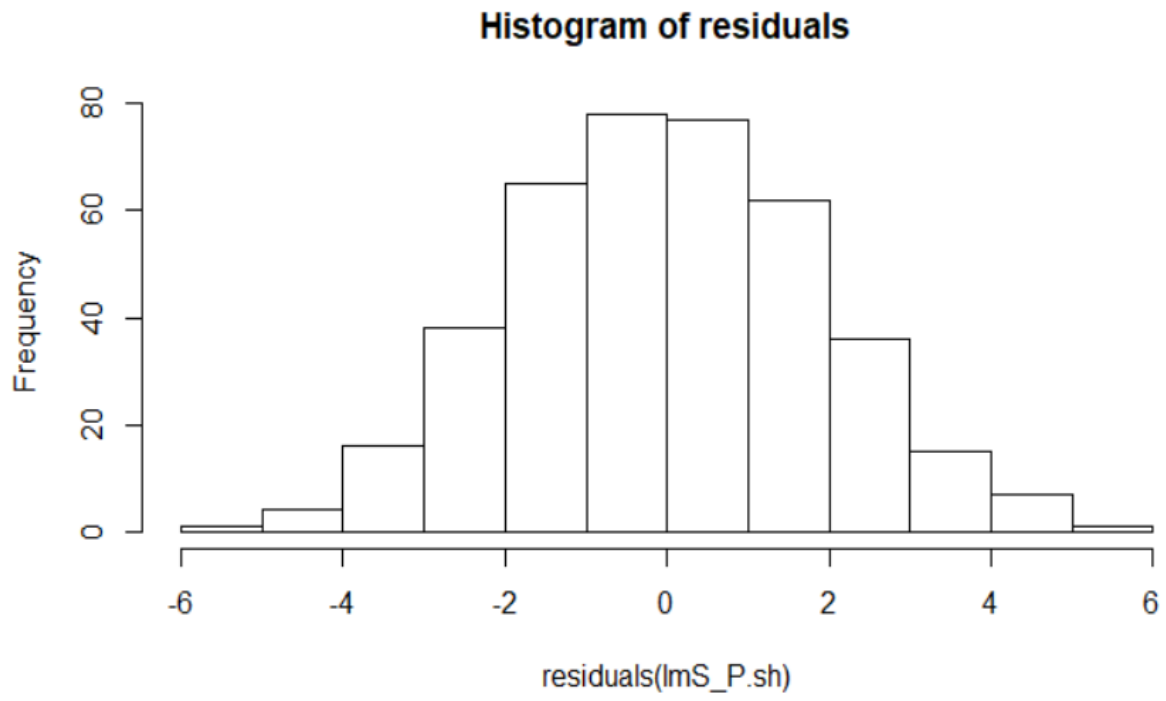
Ans. Yes, since the p-value of ShelveLoc is almost zero.

4.7.1. Diagnostics of Parallel Slopes Model

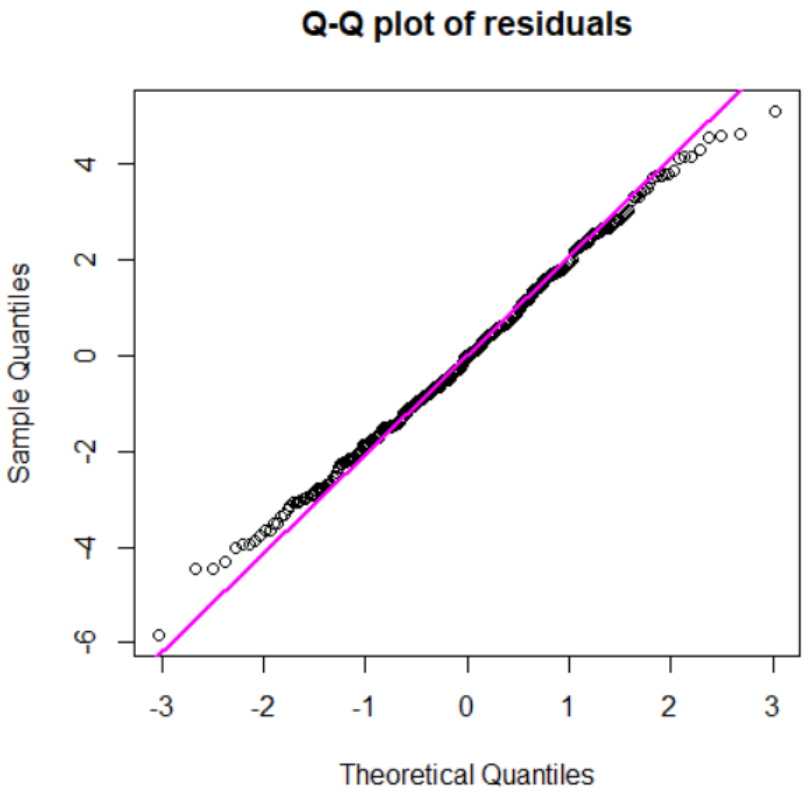
Task 27: Visualise the residuals of the combined model and check for the normality.

```
>stem(residuals(lmS_P.sh))

  The decimal point is at the |

  -5 | 8
  -4 | 5530
  -3 | 9997765543211110000
  -2 | 9998888877766555333222222111111000
  -1 | 99999988888777777766555555555555555444444433322221111111111100000+3
  -0 | 99999998888888888777777777777666665555555555444443333333333322221
   0 | 000011111111111222222333333344444444555555566666666666666777777788+3
   1 | 00000111111112222233334444444555566666667777777777888888899999
   2 | 000222223333344445555566666667788888899
   3 | 0023334556777888
   4 | 1113566
   5 | 1
```

```
>hist(residuals(lmS_P.sh), main = "Histogram of residuals")
```



```
>qqnorm(residuals(lmS_P.sh), main = "Q-Q plot of residuals")
>qqline(residuals(lmS_P.sh), col = 6, lwd = 2)
```



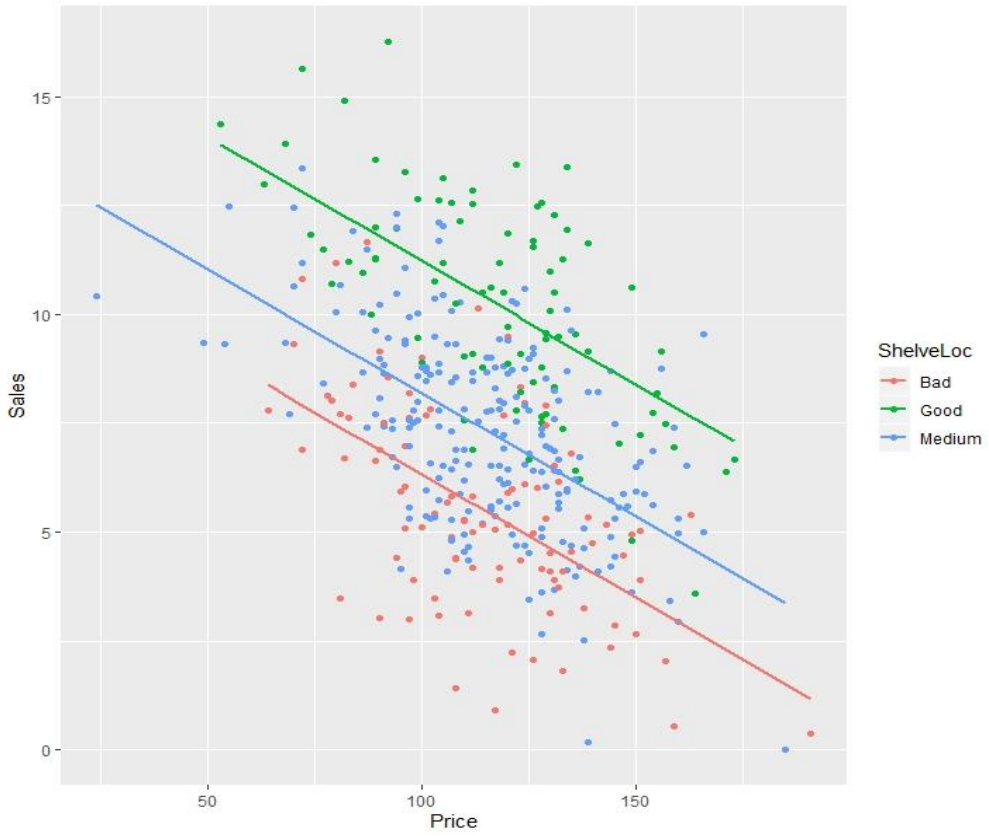
Q47. Do the residuals appear to be normally distributed?

Ans. On the basis of histogram and Q-Q plot, the residuals appear to have a normal distribution.

Task 28: Use discrete variable to plot parallel lines of regression

In this task we will plot the data points of Sales and Price along with its line of best fit as determined by the parallel slopes model.

```
>ggplot(augment(lmS_P.sh), aes(x = Price, y = Sales, col = ShelveLoc)) +
 geom_point() +
 geom_line(aes(y = .fitted, x = Price), size = 1)
```



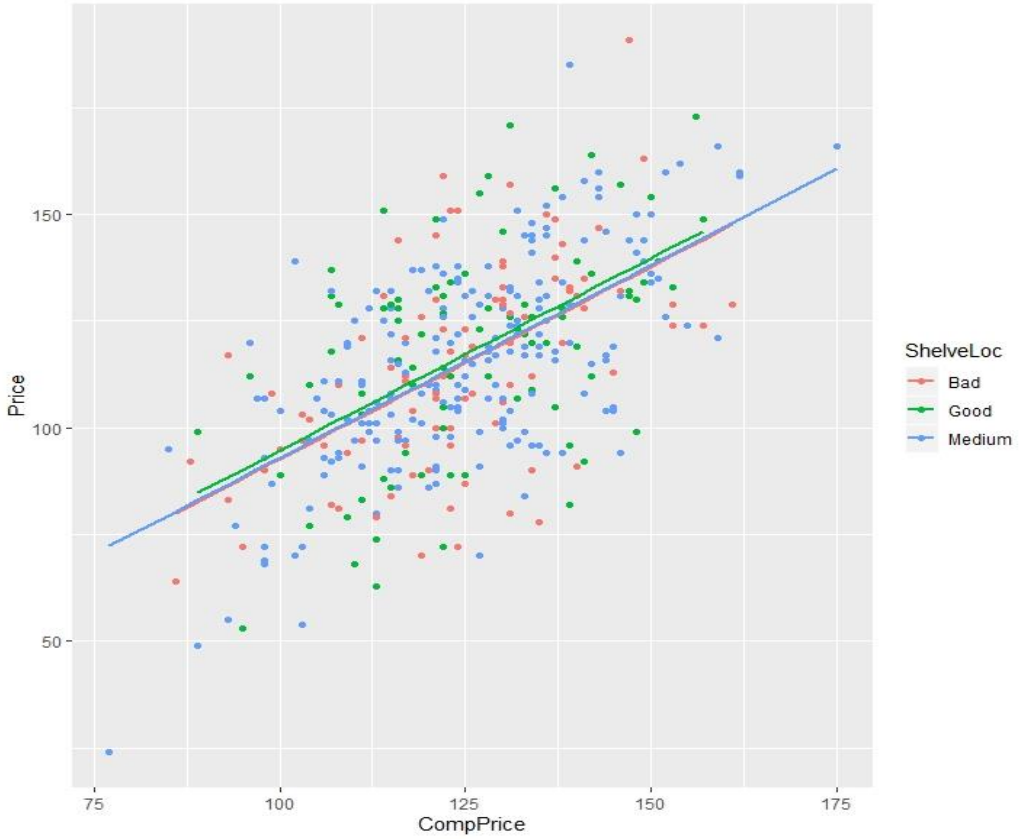
Q48. Do the lines of best fit appear to be different across ShelveLoc?

Ans. Yes, the intercept of the lines are different though slope is same.

Note:

If parallel slopes regression is performed with price as the response variable and ShelveLoc as the categorical variable, we get the following result:

```
>lmPr_CP.Sh <- lm(Price ~ CompPrice + ShelveLoc, data = prodata)

>ggplot(augment(lmPr_CP.Sh), aes(y = Price, x = CompPrice, col =
 ShelveLoc)) +
 geom_point() +
 geom_line(aes(y = .fitted, x = CompPrice), size = 1)
```



Q49. Do the lines of best fit appear to be different across ShelveLoc?

Ans. The lines are almost overlapping implying that the intercepts do not differ much. Hence, ShelveLoc does not play a significant role in explaining price.

You can also check that
summary(lmPr_CP)$adj.r.squared > summary(lmPr_CP.Sh)$adj.r.squared which also validates the results obtained by plotting parallel slopes model.

4.8. Interactions

In regression, an interaction effect exists when the effect of an independent variable on a dependent variable changes, depending on the value(s) of one or more other independent variables. In a regression equation, an interaction effect is represented as the product of two or more independent variables.

For example, here is a typical regression equation *without* an interaction:

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

where $\hat{y}$ is the predicted value of a dependent variable, X_1 and X_2 are independent variables, and β_0, β_1, and β_2 are regression coefficients.

And here is the same regression equation *with* an interaction:

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

Here, β_3 is a regression coefficient, and X_1X_2 is the interaction. The interaction between X_1 and X_2 is called a two-way interaction, because it is the interaction between two independent variables.

Task 29: Use interactions to model response

(i) Sales

```
>lms_i <- lm(Sales ~ Price*ShelveLoc, data = prodata)

>summary(lms_i)

Call:
lm(formula = Sales ~ Price * ShelveLoc, data = prodata)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9037 -1.3461 -0.0595  1.3679  4.9037

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           11.832984   0.965788  12.252  < 2e-16 ***
Price                 -0.055220   0.008276  -6.672 8.57e-11 ***
ShelveLocGood          6.135880   1.392844   4.405 1.36e-05 ***
ShelveLocMedium        1.630481   1.171616   1.392    0.165
Price:ShelveLocGood   -0.010564   0.011742  -0.900    0.369
Price:ShelveLocMedium  0.001984   0.010007   0.198    0.843
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.918 on 394 degrees of freedom
Multiple R-squared:  0.5444,   Adjusted R-squared:  0.5386
F-statistic: 94.17 on 5 and 394 DF,  p-value: < 2.2e-16
```

We observe that the p-value corresponding to the coefficient estimates of interaction terms is greater than 0.05. Hence, we can conclude that interaction of Price and ShelveLoc does not play a significant role in explaining Sales in model.

Q50. How much of the variation in Sales is explained by this model? Is it better than the additive model?

Ans. adjusted R^2 = 0.5386 which is almost same as the additive model.

(ii) Price

```
>lmp_i <- lm(Price~CompPrice*ShelveLoc,data = prodata)

>summary(lmp_i)


Call:
lm(formula = Price ~ CompPrice * ShelveLoc, data = prodata)

Residuals:
    Min      1Q  Median      3Q     Max
-47.477 -11.509   0.132  12.444  57.331

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 9.63233   16.28256   0.592    0.554
CompPrice                   0.84379    0.13034   6.474 2.84e-10 ***
ShelveLocGood               6.44992   24.11715   0.267    0.789
ShelveLocMedium           -13.82554   19.41044  -0.712    0.477
CompPrice:ShelveLocGood    -0.03426    0.19164  -0.179    0.858
CompPrice:ShelveLocMedium   0.11425    0.15496   0.737    0.461
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.29 on 394 degrees of freedom
Multiple R-squared:  0.3448,   Adjusted R-squared:  0.3365
F-statistic: 41.47 on 5 and 394 DF,  p-value: < 2.2e-16
```

Q51. How much of the variation in Price is explained by this model? Is it better than the additive model?

Ans. adjusted R^2 = 0.3365 which is almost same as the additive model

Annexure

1.Measures of Central Tendency

A *measure of central tendency* is a single value that attempts to describe a set of data by identifying the central position within that set of data.

a) Arithmetic Mean – The sum of all observations divided by the number of observations.

b) Median – The value that divides a sorted numerical data into two equal halves.

c) Mode – The value that appears most often in the data.

2. Partition Values

These are the values that divide the data into a number of equal parts.

a) Quartiles – Three points that divide the data into 4 equal parts.

b) Deciles – Nine points that divide the data into 10 equal parts.

c) Percentiles – Ninety nine points that divide the data into 100 equal parts.

3. Measures of Dispersion

A measure of dispersion is useful in gauging the extent of spread in the given data.

a) Range – The difference between the maximum and minimum values of a data.

b) Quartile Deviation – The difference between the first and third quartiles divided by 2.

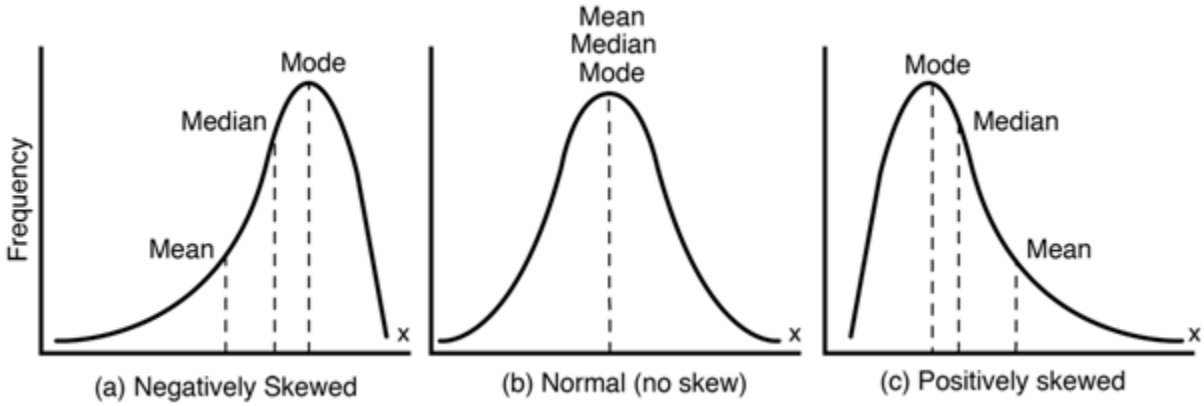
c) Standard Deviation – The positive square root of arithmetic mean of squares of deviations of given values from their arithmetic means.

d) Variance – Square of standard deviation.

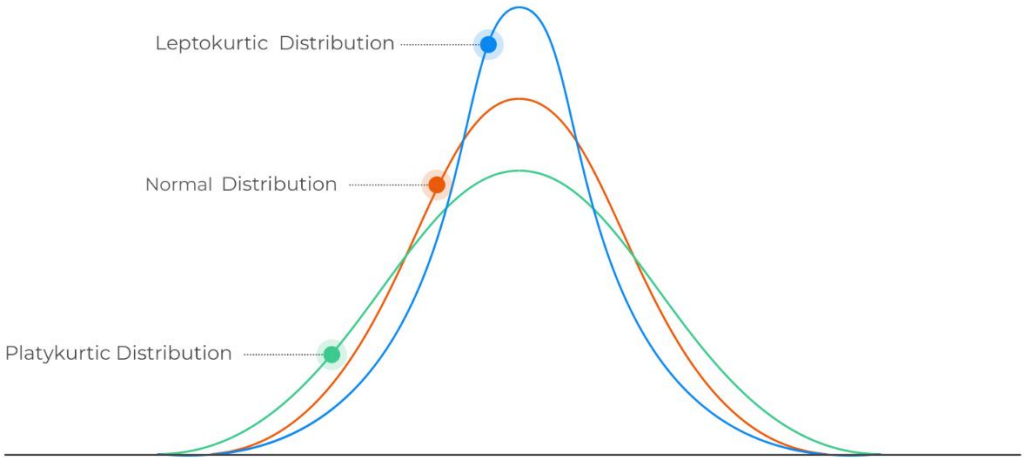
4. Skewness and Kurtosis

a) Skewness allows us to measure the symmetry or the lack of it for a given distribution.
Symmetric distribution: Mean = Median = Mode

Positively skewed distribution: Mean > Median > Mode
Negatively skewed distribution: Mean < Median < Mode



b) Kurtosis gives us an insight about the "peakedness" or "flatness" of the frequency curve of the given data.



5. Normal Distribution

The normal distribution is a continuous probability distribution. A random variable X is said to follow normal distribution if the equation of the distribution is given by

$$f(x) = \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}$$

where μ is the population mean and σ is the population standard deviation.

We use the following form to denote that X follows a normal distribution with mean μ and standard deviation σ: $X \sim N(\mu,\sigma^2)$

If $X \sim N(0,1)$ then X is said to follow a standard normal distribution.

A *log-normal* (or *lognormal*) *distribution* is a continuous probability *distribution* of a random variable whose *logarithm* is normally *distributed*. Thus, if the random variable X is *log*-normally *distributed*, then Y = ln(X) has a *normal distribution*.

6. Chi-square Distribution

Degrees of freedom (d.f./ d.o.f.) – These are the number of values that can independently vary in a statistical analysis, that is, the number of values left after accounting for every constraint involved in an analysis.

A random variable Y is said to follow a chi-square distribution with n d.o.f., denoted by $\chi^2_{(n)}$ if its distribution is the sum of squares of n independent standard normal variables.

7. t-distribution (Studentised t-distribution)

The ratio of independent standard normal variate and the square root of chi-square variate divided by its degrees of freedom follows t distribution.

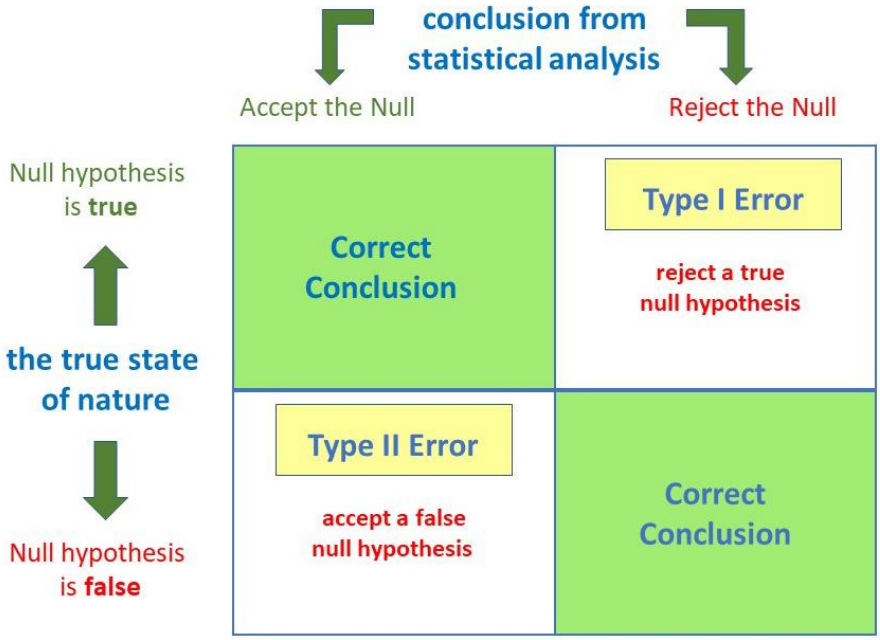
8. F-distribution

The ratio of two independent chi-square variates divided by its degrees of freedom follows F-distribution. Let U and V be independent chi-square random variables with r_1 and r_2 degrees of freedom, respectively. Then, $F = \frac{U/r_1}{V/r_2} \sim F(r_1, r_2)$

9. Errors in testing

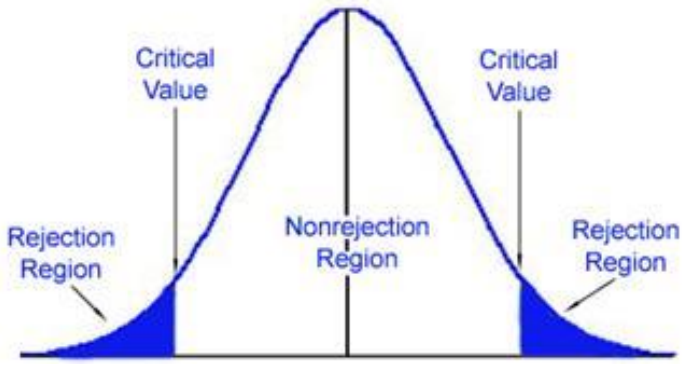
a) Type I error, also known as a false positive, occurs when we incorrectly reject a true null hypothesis. This means that we report that our findings are significant when in fact they have occurred by chance. It is denoted by α.

b) Type II error, also known as a false negative, occurs when we fail to reject a null hypothesis which is really false. Here we conclude there is not a significant effect, when actually there really is. It is denoted by β.



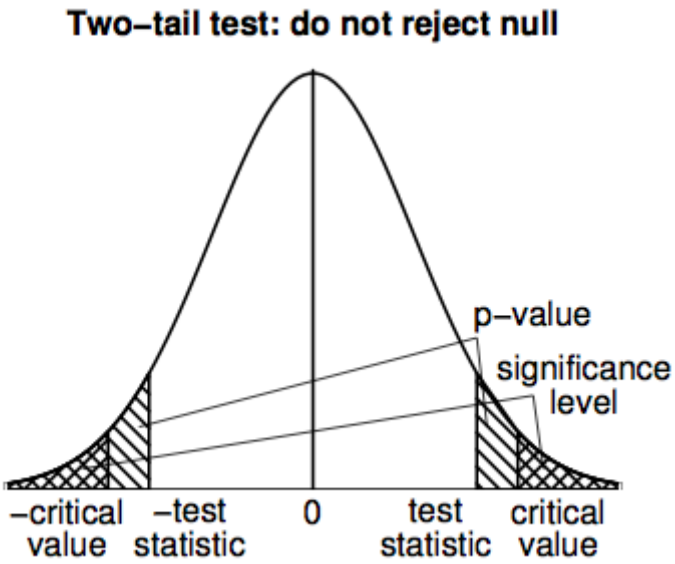
10. Critical Value (Tabulated Value)

In hypothesis testing, a critical value is a point on the test distribution that is compared to the test statistic to determine whether to reject the null hypothesis.



11. p-value

In statistics, the p-value is the probability of obtaining results as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. The p-value provides the smallest level of significance at which the null hypothesis would be rejected. A smaller p-value means that there is stronger evidence in favour of the alternative hypothesis.


Two–tail test: do not reject null
p–value
significance
level
–critical
value
–test
statistic
0
test
statistic
critical
value

References

1. Class Notes by Dr. Priyanka Aggarwal (HOD, Department of Statistics, Hindu College, Delhi University)
2. Gupta, S.C. and Kapoor, V.K. (2014). Fundamentals of Mathematical Statistics
3. Montgomery, D., Peck, E. and Vining, G. (2006). Introduction to Linear Regression Analysis.
4. James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017). An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics)