> library(boot)

> library(car)

> library(QuantPsyc)

> library(lmtest)

> library(sandwich)

> library(vars)

> library(nortest)

> library(MASS)

> head(data)

MarketID MarketSize LocationID AgeOfStore Promotion week SalesInThousands

1 1 Medium 1 4 3 1 33.73

2 1 Medium 1 4 3 2 35.67

3 1 Medium 1 4 3 3 29.03

4 1 Medium 1 4 3 4 39.25

5 1 Medium 2 5 2 1 27.81

6 1 Medium 2 5 2 2 34.67

> #checking the structure of data

> str(data)

'data.frame': 548 obs. of 7 variables:

$ MarketID : int 1 1 1 1 1 1 1 1 1 1 ...

$ MarketSize : chr "Medium" "Medium" "Medium" "Medium" ...

$ LocationID : int 1 1 1 1 2 2 2 2 3 3 ...

$ AgeOfStore : int 4 4 4 4 5 5 5 5 12 12 ...

$ Promotion : int 3 3 3 3 2 2 2 2 1 1 ...

$ week : int 1 2 3 4 1 2 3 4 1 2 ...

$ SalesInThousands: num 33.7 35.7 29 39.2 27.8 ...

> #checking the mean, median, quartile of data

> summary(data)

MarketID MarketSize LocationID AgeOfStore Promotion

Min. : 1.000 Length:548 Min. : 1.0 Min. : 1.000 Min. :1.000

1st Qu.: 3.000 Class :character 1st Qu.:216.0 1st Qu.: 4.000 1st Qu.:1.000

Median : 6.000 Mode :character Median :504.0 Median : 7.000 Median :2.000

Mean : 5.715 Mean :479.7 Mean : 8.504 Mean :2.029

3rd Qu.: 8.000 3rd Qu.:708.0 3rd Qu.:12.000 3rd Qu.:3.000

Max. :10.000 Max. :920.0 Max. :28.000 Max. :3.000

week SalesInThousands

Min. :1.00 Min. :17.34

1st Qu.:1.75 1st Qu.:42.55

Median :2.50 Median :50.20

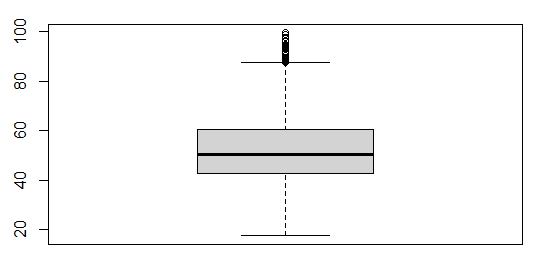
Mean :2.50 Mean :53.47

3rd Qu.:3.25 3rd Qu.:60.48

Max. :4.00 Max. :99.65

> #displaying the dependent variable through boxplot; to see the outliers

> boxplot(data$SalesInThousands)



> quantile(data$SalesInThousands, c(0,0.05,0.1,0.25,0.5,0.75,0.90,0.95,0.99,0.995,1))

0% 5% 10% 25% 50% 75% 90% 95% 99%

17.34000 30.62150 35.41200 42.54500 50.20000 60.47750 82.26600 88.45800 94.67380

99.5% 100%

96.77945 99.65000

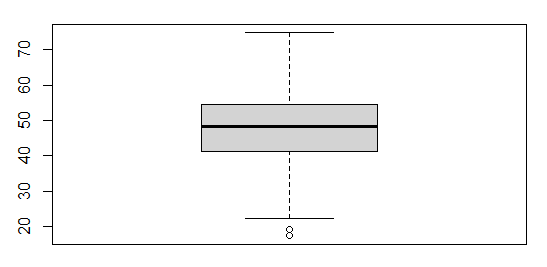
> #removing the outliers

> data2 <- data[data$SalesInThousands <75, ]

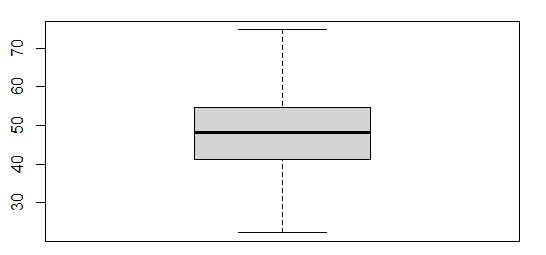
> data3 <- data2[data2$SalesInThousands >20, ]

> #displaying the dependent variable through box plot; to see the outliers

> boxplot(data2$SalesInThousands)



> boxplot(data3$SalesInThousands)



> sapply(data3, function(x) sum(is.na(x)))

MarketID MarketSize LocationID AgeOfStore Promotion

0 0 0 0 0

week SalesInThousands

0 0

> #removing null values if any

> data3 <- na.omit(data3)

> data <- data3

> nrow(data)

[1] 462

> #fitting into a linear model

> fit<- lm(SalesInThousands ~ MarketID + MarketSize + LocationID + AgeOfStore + Promotion +

+ week, data=data)

> summary(fit)

Call:

lm(formula = SalesInThousands ~ MarketID + MarketSize + LocationID +

AgeOfStore + Promotion + week, data = data)

Residuals:

Min 1Q Median 3Q Max

-22.0821 -5.2167 0.2316 5.2486 30.3590

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 55.87151 7.77178 7.189 2.70e-12 \*\*\*

MarketID -12.75937 8.60251 -1.483 0.13871

MarketSizeMedium -4.75154 1.20754 -3.935 9.63e-05 \*\*\*

MarketSizeSmall 12.23986 1.83157 6.683 6.89e-11 \*\*\*

LocationID 0.14121 0.08614 1.639 0.10183

AgeOfStore 0.15460 0.05508 2.807 0.00522 \*\*

Promotion -1.29111 0.46136 -2.798 0.00535 \*\*

week -0.22435 0.32504 -0.690 0.49040

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.807 on 454 degrees of freedom

Multiple R-squared: 0.4042, Adjusted R-squared: 0.395

F-statistic: 44.01 on 7 and 454 DF, p-value: < 2.2e-16

> #keeping only significant codes

> final\_fit<- lm(SalesInThousands ~ MarketSize + AgeOfStore + Promotion, data=data)

> summary(final\_fit)

Call:

lm(formula = SalesInThousands ~ MarketSize + AgeOfStore + Promotion,

data = data)

Residuals:

Min 1Q Median 3Q Max

-21.0658 -5.3319 0.0962 5.9235 22.0938

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 55.83929 1.32449 42.159 < 2e-16 \*\*\*

MarketSizeMedium -10.44672 1.03417 -10.102 < 2e-16 \*\*\*

MarketSizeSmall 2.50650 1.43045 1.752 0.08040 .

AgeOfStore 0.15618 0.05871 2.660 0.00809 \*\*

Promotion -1.26928 0.48845 -2.599 0.00966 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.344 on 457 degrees of freedom

Multiple R-squared: 0.315, Adjusted R-squared: 0.309

F-statistic: 52.54 on 4 and 457 DF, p-value: < 2.2e-16

|  |
| --- |
| > ##Assumption Diagnostic Tests  > #Multicollinearity Test  > #checking the VIF score; vif>2 means presence of multicollinearity  > vif(final\_fit)  GVIF Df GVIF^(1/(2\*Df))  MarketSize 1.035316 2 1.008715  AgeOfStore 1.026271 1 1.013050  Promotion 1.017141 1 1.008534 |
|  |
| |  | | --- | | > | |

> ## MAPE

> data$pred <- fitted(final\_fit)

> #Calculating MAPE

> attach(data)

> (sum((abs(data$SalesInThousands-data$pred))/data$SalesInThousands))/nrow(data)

[1] 0.1573574

> dwt(final\_fit)

lag Autocorrelation D-W Statistic p-value

1 0.5181597 0.9602602 0

Alternative hypothesis: rho != 0

> bptest(final\_fit)

studentized Breusch-Pagan test

data: final\_fit

BP = 10.016, df = 4, p-value = 0.04016

> resids <- final\_fit$residuals

> #get Anderson-Darling test for normality

> ad.test(resids)

Anderson-Darling normality test

data: resids

A = 0.61108, p-value = 0.1115

> #writing in excel

> write.csv(data,"linear\_output.csv")

Summary:

Based on the above Assumption Diagnostic Test, we generate the below insights of the linear model:

1. **Normality Test**: We start with Hypothesis as below

* Null Hypothesis: The errors are normally distributed. p-value should be more than 0.05.
* Alternative Hypothesis: The errors are not normally distributed. p-value is less than 0.05.

Based on the Anderson- Daring Test, we find the p-value is 0.1115. Since the p-value is high than 0.05, we accept the null hypothesis and reject the alternative hypothesis. Finally, we conclude that the errors are normally distributed. We have passed the Normality Test for this linear model.

1. **Homoscedasticity Test**: The assumption means that the variance around the regression line is same for all values of the predictor variable (X). For this test, we use Breusch-Pagan Test.

* Null Hypothesis: The error variances are all equal. p-value should be greater than 0.05.
* Alternative Hypothesis: The error variances are not equal. p-value should be less than 0.05

Based on the Breusch-Pagan Test, we find the p-value is 0.04016. Since the p-value is less than 0.05, we reject the null hypothesis which says the variance is scattered similarly. We accept the alternative hypothesis and the regression model is heteroscedasticity – the variances are differently scattered.

1. **Multicollinearity Test**: Here, we check if the independent variables have relationship between them, i.e, the correlation between them. The correlation between independent or explanatory variables are called multicollinearity. We use VIF (Variance Inflation Factor) for this test.

* The value of VIF should be lower than 1.7.

Based on the test, we find the VIF score for the variables are less than 1.7. As such, we conclude that there is no multicollinearity on the linear model i.e, no correlation between independent variables in the model.

1. **Auto – Correlation Test**: The good model should not have autocorrelation in the data. Auto- Correlation occurs when the residuals are not independent from each other. For this, we use Durbin-Watson’s d test.

* Null Hypothesis: Residuals are not linearly auto correlated. Values 1.5 < d < 2.5 shows there is no auto correlation in data.
* Alternative Hypothesis: The residuals are linearly auto correlated. Values 1.5 > d > 2.5 shows there is linear auto correlations between data.

Based on the Durbin-Watson Test, we find that the values are less than 2.5, as such there is no linear auto correlation between data.

Finally, we pass all the assumption diagnostic test except the homoscedasticity test.