Linear Regression Model for DVD sales

Reading the data from Excel File and preparing the dataframe.

library(readxl) #this loads the package readxl

Sales\_dataset <- read\_excel("C:/Users/AcadgildDesktop/Sales\_dataset.xlsx")

View(Sales\_dataset)

Fitting the Linear Model

lmfit=lm(sales~advertise+plays+attractiveness,df)

summary(lmfit)

Call:

lm(formula = sales ~ advertise + plays + attractiveness, data = df)

Residuals:

Min 1Q Median 3Q Max

-122.728 -28.760 1.476 29.422 142.960

Coefficients:

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

(Intercept) -28.140377 17.373604 -1.62 0.107

advertise 0.084642 0.006908 12.25 < 2e-16 \*\*\*

plays 3.385493 0.277723 12.19 < 2e-16 \*\*\*

attractiveness 11.333342 2.437340 4.65 6.1e-06 \*\*\*

---

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

Residual standard error: 47.1 on 196 degrees of freedom

Multiple R-squared: 0.6645, Adjusted R-squared: 0.6593

F-statistic: 129.4 on 3 and 196 DF, p-value: < 2.2e-16

Output analysis:

Residuals are differences in actual dependent variable and predicted variable.The summary output shows you the minimum residual and maximum residual. A The coefficients in this output define the regression line which can be fitted . From this output, we have determined that the intercept is -28.140377 and the coefficient of advertise is 0.084642, for plays it is 3.385493 and for attractiveness is 11.333342.

model<- -28.140377 + (0.084642\* advertise)+(3.385493\*plays)+(11.333342\*attractiveness)

This equation tells us that the predicted number of DVD sales will increase by 0.084642 for every one percent increase in the advertise, will increase by 3.385493 for every one percent increase in the plays and will increase by 11.333342 for every one percent increase in the attractiveness.

The Adjusted R-squared value is 0.6593 which means that the model can explain about 66% of the variance of the Sales variable.

The standard error for each coefficient is given as 0.006908, 0.277723 and 2.43730 respectively. It is root of mean square error which is the sum of square errors divided by degrees of freedom.

The p-values of the attributes attractiveness and number of plays are below the significance level .The significance level is given by t-values shown above. Since the p-values shown above are lower than the significance level these are statistically significant.

The output shows that F = 129.4 (p < 2.2e-16), indicating that we should clearly reject the null hypothesis that the variables advertise,number of plays and attractiveness collectively have no effect on sales

Fitted Values and Residuals

df['Predicted']<-NULL

> fitted<-data.frame(df, fitted.value=fitted(lmfit),residual=resid(lmfit))

> fitted

advertise sales plays attractiveness fitted.value residual

1 10.256 330 43 10 231.63734 98.3626573

2 985.685 120 28 7 229.41738 -109.4173751

3 1445.563 360 35 7 292.04091 67.9590854

4 1188.193 270 33 7 263.48556 6.5144353

5 574.513 220 44 5 226.11608 -6.1160825

6 568.954 170 19 5 141.00822 28.9917772

7 471.814 70 20 1 90.83821 -20.8382065

8 537.352 210 22 9 193.82321 16.1767937

9 514.068 200 21 7 165.80022 34.1997792

10 174.093 300 40 7 201.34836 98.6516378

11 1720.806 290 32 7 305.18161 -15.1816078

12 611.479 70 20 2 113.99310 -43.9931010

13 251.192 150 24 8 165.03964 -15.0396392

14 97.972 190 38 6 176.80098 13.1990150

15 406.814 240 24 7 166.87849 73.1215138

16 265.398 100 25 5 135.62753 -35.6275350

17 1323.287 250 35 5 259.02452 -9.0245217

18 196.650 210 36 8 201.04900 8.9509955

19 1326.598 280 27 8 266.22085 13.7791501

20 1380.689 230 33 8 291.11219 -61.1121913

21 792.345 210 33 7 229.98012 -19.9801187

22 957.167 230 28 6 215.67021 14.3297928

23 1789.659 320 30 9 326.90517 -6.9051737

24 656.137 210 34 7 221.83667 -11.8366672

25 613.697 230 49 7 269.02685 -39.0268524

26 313.362 250 40 8 224.46974 25.5302615

27 336.510 60 20 4 113.38580 -53.3858028

28 1544.899 330 42 7 324.14739 5.8526143

29 68.954 150 35 8 186.85504 -36.8550407

30 785.692 150 8 6 133.44632 16.5536811

31 125.628 180 49 7 227.71562 -47.7156181

32 377.925 80 19 8 158.83913 -78.8391325

33 217.994 180 42 6 200.50188 -20.5018846

34 759.862 130 6 7 135.82237 -5.8223657

35 1163.444 320 36 6 260.21389 59.7861067

36 842.957 280 32 7 230.87854 49.1214635

37 125.179 200 28 6 145.24891 54.7510882

38 236.598 130 25 8 167.18986 -37.1898643

39 669.811 190 34 8 234.32741 -44.3274062

40 612.234 150 21 6 162.77587 -12.7758655

Predictions:

predict(lmfit,interval = "confidence")

fit lwr upr

1 231.63734 211.59237 251.68232

2 229.41738 221.05928 237.77547

3 292.04091 278.60949 305.47234

4 263.48556 253.01375 273.95738

5 226.11608 211.12239 241.10977

6 141.00822 129.93634 152.08011

7 90.83821 62.81115 118.86527

8 193.82321 180.40252 207.24389

9 165.80022 158.08818 173.51226

10 201.34836 189.89791 212.79881

11 305.18161 288.72130 321.64192

12 113.99310 90.48527 137.50093

13 165.03964 154.49909 175.58019

14 176.80098 164.56742 189.03455

15 166.87849 159.40874 174.34823

16 135.62753 124.24014 147.01493

17 259.02452 243.46792 274.58112

18 201.04900 189.66473 212.43328

19 266.22085 253.24155 279.20015

20 291.11219 277.72086 304.50352

21 229.98012 222.43290 237.52734

22 215.67021 206.63959 224.70082

23 326.90517 307.13596 346.67439

24 221.83667 214.37988 229.29346

25 269.02685 255.67406 282.37964

26 224.46974 212.88928 236.05019

27 113.38580 98.42846 128.34314

28 324.14739 308.38315 339.91162

29 186.85504 174.59467 199.11541

30 133.44632 120.48155 146.41109

31 227.71562 212.36944 243.06180

32 158.83913 147.93194 169.74633

33 200.50188 187.71542 213.28835

34 135.82237 121.83368 149.81105

35 260.21389 248.42678 272.00100

36 230.87854 223.26524 238.49183

37 145.24891 135.33705 155.16077

38 167.18986 156.67001 177.70971

39 234.32741 225.21715 243.43766

40 162.77587 154.71870 170.83303

41 244.34150 235.92648 252.75653

42 268.71120 246.38525 291.03716

43 325.42771 305.47320 345.38221

44 225.04551 215.86707 234.22396

45 225.97479 215.11042 236.83916

46 304.09434 289.69468 318.49399

47 155.84527 144.30039 167.39015

48 156.53093 147.83779 165.22406

49 282.33633 270.88005 293.79262

50 265.29048 253.36191 277.21906

51 228.11554 218.65852 237.57256

52 92.17628 76.29433 108.05823

53 84.51415 66.44234 102.58596

54 212.10532 203.32618 220.88446

55 304.79376 289.69586 319.89166

56 240.14874 230.64633 249.65115

57 146.25170 133.82346 158.67995

58 250.63066 241.60560 259.65572

59 101.56978 86.25803 116.88152

60 175.10310 162.39368 187.81253

61 201.79282 195.01019 208.57545

62 318.13840 303.07207 333.20473

63 198.91147 189.11338 208.70955

64 119.90039 109.11978 130.68100

65 181.49905 167.74210 195.25600

66 138.70907 129.43561 147.98253

67 124.13441 114.54928 133.71953

68 181.27199 169.72370 192.82027

69 138.28791 125.95203 150.62380

70 218.81203 209.86937 227.75470

71 125.23074 112.82860 137.63287

72 274.49947 256.02614 292.97280

73 151.79277 141.80844 161.77710

74 213.93562 206.90391 220.96734

75 259.40816 245.31328 273.50303

76 173.80408 161.91240 185.69577

77 228.16395 219.25243 237.07546

78 119.85352 108.87546 130.83159

79 253.50022 239.91987 267.08056

80 104.30331 91.29609 117.31053

81 116.96940 102.47773 131.46108

82 42.32256 23.16164 61.48348

83 45.14657 28.12815 62.16499

84 175.76113 162.78081 188.74145

85 116.97063 102.67473 131.26652

86 275.18790 264.35701 286.01879

87 315.04344 292.65683 337.43005

88 319.47746 298.03404 340.92089

89 137.18402 127.50898 146.85907

90 174.05438 163.69340 184.41536

91 227.47501 218.06199 236.88802

92 238.47618 224.06647 252.88589

93 290.53906 272.10873 308.96939

94 77.31323 60.83738 93.78909

95 267.42833 257.00784 277.84882

96 211.00647 201.78074 220.23220

97 94.14331 82.02651 106.26012

98 218.11156 209.31829 226.90483

99 333.28435 316.03685 350.53185

100 152.76268 137.27655 168.24881

101 161.39045 150.06885 172.71205

102 313.49606 297.78090 329.21123

103 123.69517 110.10913 137.28120

104 148.70379 140.46359 156.94400

105 248.51487 230.85462 266.17512

106 134.41548 124.81986 144.01109

107 101.44419 85.83382 117.05456

108 90.01726 77.63384 102.40068

109 167.13717 155.10132 179.17303

110 134.35545 124.38038 144.33052

111 179.96665 169.23022 190.70309

112 223.91355 213.70424 234.12286

113 79.30803 59.15342 99.46263

114 77.71026 61.59547 93.82506

115 190.79360 176.43566 205.15155

116 99.47501 79.52535 119.42467

117 143.12374 133.02773 153.21974

118 149.09847 138.12255 160.07439

119 310.05602 291.43804 328.67400

120 145.13294 133.41021 156.85566

121 217.25107 209.46943 225.03271

122 143.19192 132.41022 153.97361

123 124.95752 114.56077 135.35427

124 288.06202 272.81267 303.31137

125 59.40464 43.89468 74.91460

126 270.71969 255.46154 285.97784

127 219.08074 207.14473 231.01676

128 259.70977 246.71782 272.70171

129 206.56339 199.76523 213.36154

130 253.57666 243.77422 263.37910

131 179.98139 168.57475 191.38804

132 247.90418 239.20785 256.60051

133 202.15729 195.37937 208.93520

134 180.13174 169.32954 190.93393

135 158.20946 150.25271 166.16621

136 260.20878 249.47046 270.94711

137 108.08803 95.64395 120.53211

138 80.56202 51.35011 109.77392

139 203.57442 192.63909 214.50975

140 213.69204 206.57265 220.81143

141 196.90998 184.61035 209.20960

142 203.49803 194.91806 212.07801

143 188.05853 178.02408 198.09297

144 162.66628 149.86565 175.46692

145 142.14273 133.08866 151.19680

146 173.47134 158.98965 187.95304

147 116.10857 104.89467 127.32247

148 289.37715 277.03746 301.71683

149 231.72460 222.11536 241.33384

150 207.28038 200.19746 214.36329

151 75.29845 60.98285 89.61405

152 161.46218 145.92582 176.99854

153 136.31207 126.01309 146.61105

154 207.14326 192.81590 221.47063

155 108.26921 93.55441 122.98401

156 151.17711 142.06492 160.28930

157 244.36339 235.68180 253.04499

158 217.35387 206.12722 228.58051

159 82.82727 66.64421 99.01033

160 165.27499 152.15929 178.39069

161 172.44492 164.88968 180.00017

162 106.46050 95.24420 117.67679

163 214.02914 206.31679 221.74148

164 242.72809 224.42049 261.03568

165 169.47189 158.07698 180.86680

166 160.34071 151.33478 169.34663

167 133.76722 124.54736 142.98708

168 169.47848 158.95702 179.99994

169 217.03971 203.73919 230.34023

170 180.86756 170.10229 191.63283

171 226.80326 217.09171 236.51480

172 222.72454 209.74163 235.70744

173 207.87605 197.74685 218.00526

174 107.73896 95.88210 119.59581

175 288.27511 273.03944 303.51078

176 164.59789 156.61112 172.58466

177 188.15594 176.70617 199.60571

178 238.70938 229.59152 247.82724

179 267.72895 252.45714 283.00076

180 223.89586 208.01768 239.77403

181 62.67345 34.38579 90.96112

182 166.26692 155.93138 176.60245

183 285.10238 266.07897 304.12579

184 325.77185 300.22870 351.31501

185 290.59352 278.60503 302.58201

186 161.86173 153.81332 169.91014

187 165.57155 153.47538 177.66773

188 186.57565 178.59890 194.55240

189 270.65059 259.38796 281.91321

190 147.41697 137.33686 157.49707

191 154.07266 144.47137 163.67395

192 156.39667 145.12519 167.66815

193 239.17955 230.06646 248.29263

194 147.94952 139.36068 156.53836

195 251.68014 239.08643 264.27384

196 216.31227 208.41436 224.21018

197 162.46701 151.39049 173.54354

198 222.73226 213.75891 231.70561

199 226.73008 208.38577 245.07439

200 208.07243 194.10573 222.03914

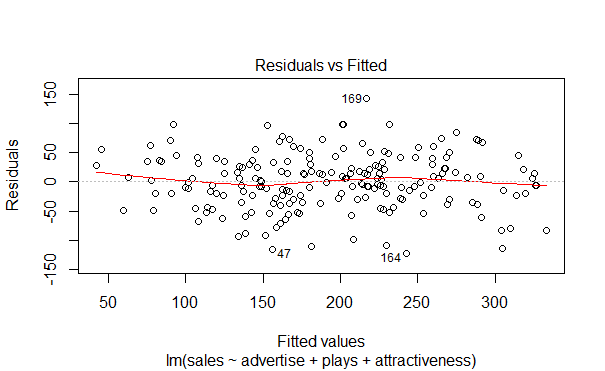
plot(lmfit)

coef(lmfit)

confint(lmfit)

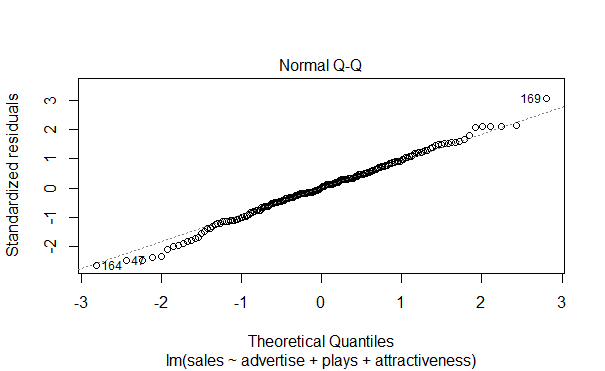
Residual Diagnostics:

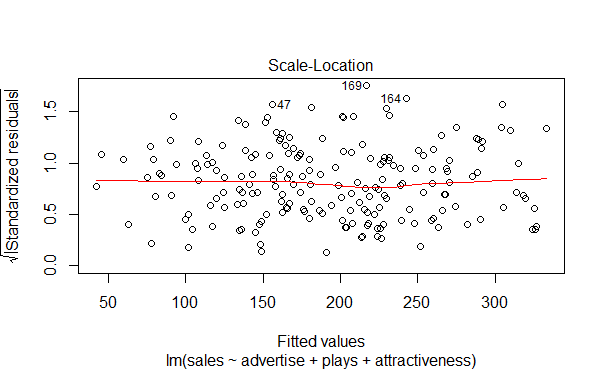
plot(lmfit)



Since the residuals are evenly spread above and below the red line the model.Due to this uniform distribution the inference can be made that the

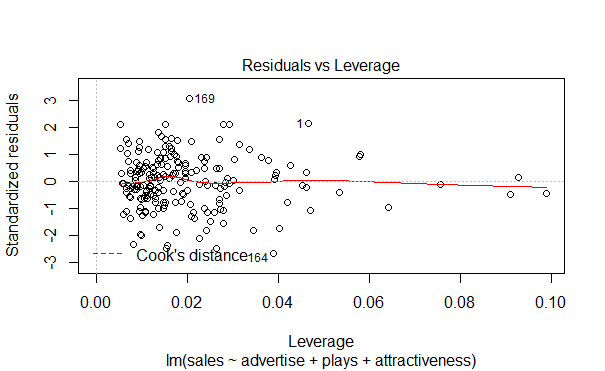
Model is not non-linear in nature.Thus shows a good enough linearity and thus it is correct to apply linear regression for the model.

This plot shows if residuals are normally distributed. Deviations from a straight line could mean that the errors which do not follow a normal distribution. Here you can a few outliers.



It’s also called Spread-Location plot. This plot shows if residuals are spread equally along the ranges of predictors. This is how you can check the assumption of equal variance (homoscedasticity). It’s good if you see a horizontal line with equally (randomly) spread points.

In the above plot the line is clearly horizontal and the points are also equally spread above and below the line.



This plot tells us the cases of influential outliers. These points on inclusion or exclusion can affect the regression to a great extent. When cases are outside of the Cook’s distance (meaning they have high Cook’s distance scores), the cases are influential to the regression results.