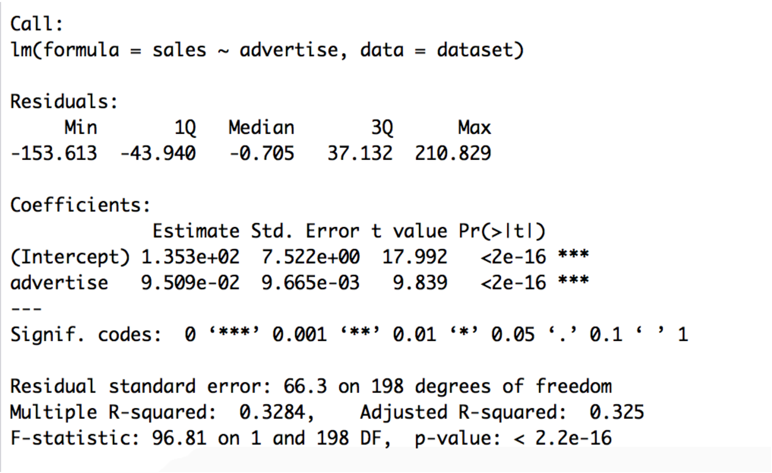
Project - 2

Create a Linear Regression Model for DVD sales data set

```
data <- read.csv("~/Desktop/Sales_dataset.csv", header =T, sep = ",")
colnames(data)
is.na(data)
newdata <-na.omit(data)
fit = lm(sales~advertise , data= dataset)
summary(fit)
```

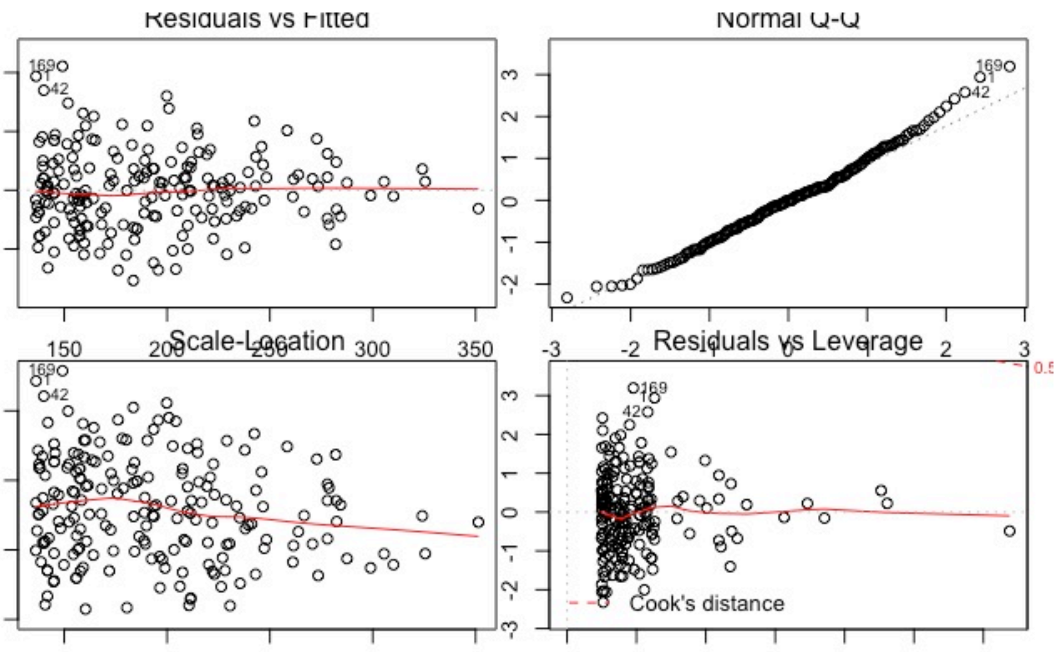


```
Call:
lm(formula = sales ~ advertise, data = dataset)

Residuals:
     Min       1Q   Median       3Q      Max 
-153.613  -43.940   -0.705   37.132  210.829 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.353e+02  7.522e+00  17.992   <2e-16 ***
advertise   9.509e-02  9.665e-03   9.839   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 66.3 on 198 degrees of freedom
Multiple R-squared:  0.3284,	Adjusted R-squared:  0.325 
F-statistic: 96.81 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
plot(fit)
```

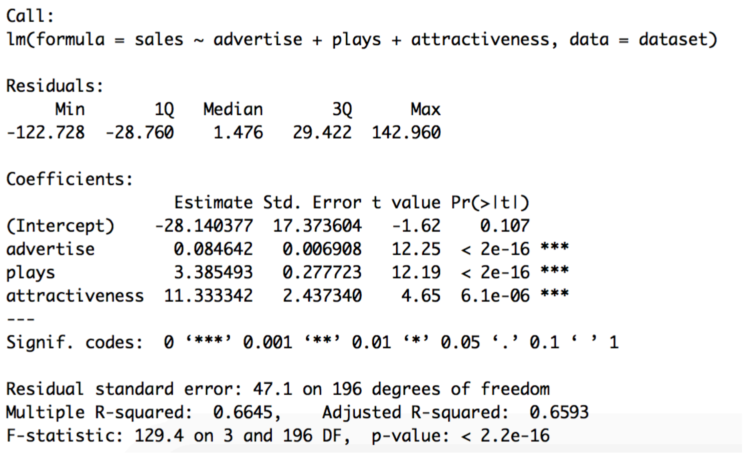
The summary contains the model statistics, parameter estimates, their standard errors, and p-values to determine if the coefficients are different from 0.

The R^2 value is 0.325 , indicating the model is not very efficient at minimising residual error. We can now predict the test values based on the model using the prediction function



Model 2 -

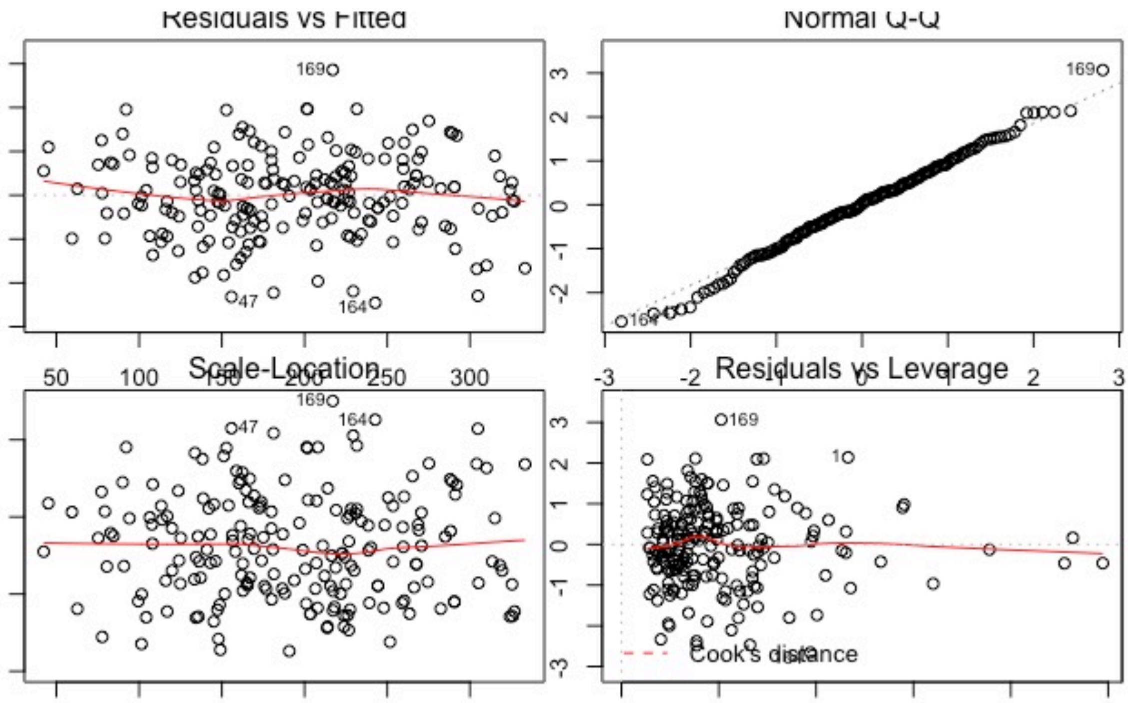
fit1 = lm(sales ~ advertise + plays + attractiveness, data= dataset)
summary(fit1)



```
Call:
lm(formula = sales ~ advertise + plays + attractiveness, data = dataset)

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
```

We observe that in the Summary Report advertising, number of times song being played and attractiveness provides the highest significance value for prediction (see *** symbol). But even this value is statistically no-significant in the multiple linear regression model of data.
The coefficient of determination of the multiple linear regression model for the data set 'data' is at 66% and Adjusted R-squared at 67% which is fairly good.

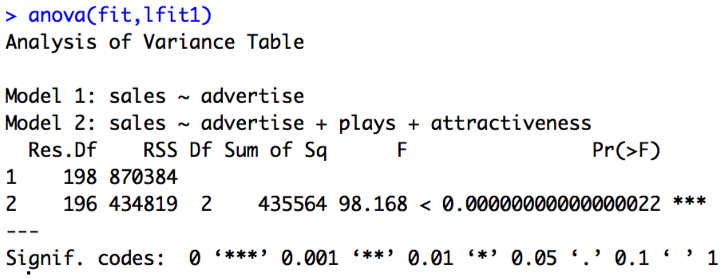
plot(fit1)



Deviance residuals - If we look at the residuals it is quite symmetrical. Intercept value is -28.140 which means when advertise spend is zero we expect to gain sales by -28.14. Positive values of Attractiveness and number of times song has been played contributing positively to the DVD sales .

Slope - Increase of 1000 in advertise will increase the dvd sales by .084642. Similar if we increase the advertising spend by 10 times , dvd sales will increase by 8.4642. As the p-value is much less than 0.05, we reject the null hypothesis that β = 0. Hence there is a significant relationship between the variables in the linear regression model of the data set faithful. The model can predict the sales with 66% confidence interval with acceptable margin of error, given the advertise, number of times song played and attractiveness score of the song.

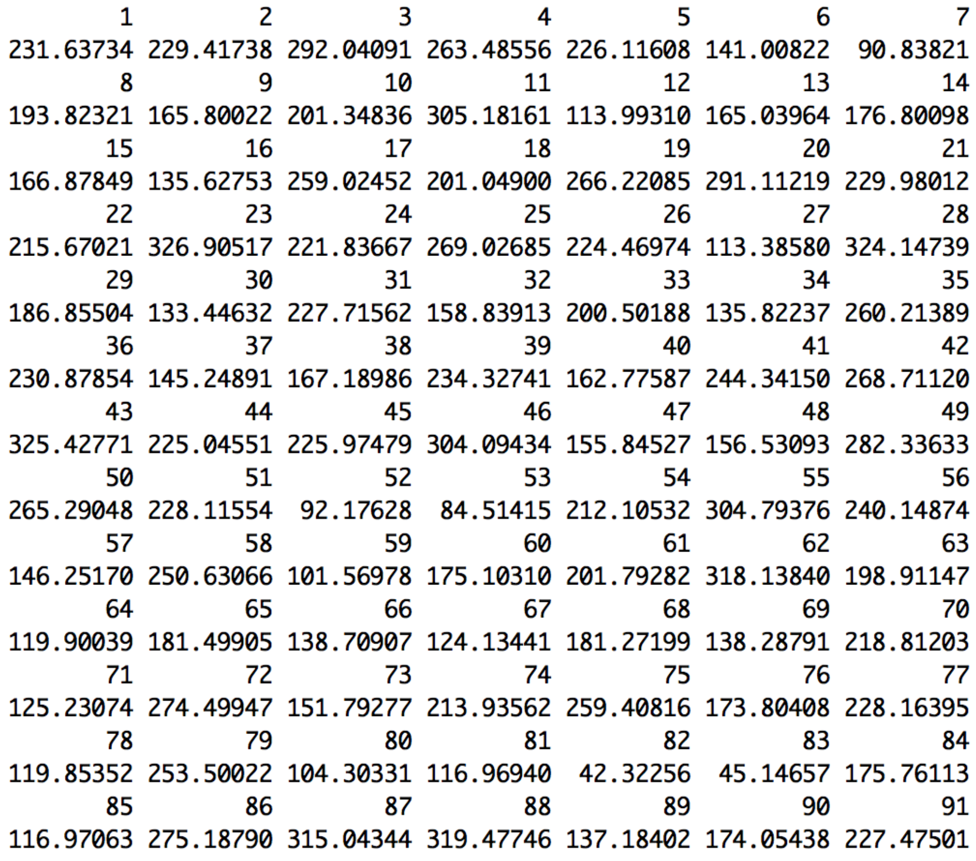
Comparison of two models -



```
> anova(fit,lfit1)
Analysis of Variance Table

Model 1: sales ~ advertise
Model 2: sales ~ advertise + plays + attractiveness
  Res.Df    RSS Df Sum of Sq      F                Pr(>F)    
1    198 870384                                              
2    196 434819  2    435564 98.168 < 0.00000000000000022 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Predict the values -

res <- signif(residuals(fit1), 5)
pre <- predict(fit1)



```
        1         2         3         4         5         6         7 
231.63734 229.41738 292.04091 263.48556 226.11608 141.00822  90.83821 
        8         9        10        11        12        13        14 
193.82321 165.80022 201.34836 305.18161 113.99310 165.03964 176.80098 
       15        16        17        18        19        20        21 
166.87849 135.62753 259.02452 201.04900 266.22085 291.11219 229.98012 
       22        23        24        25        26        27        28 
215.67021 326.90517 221.83667 269.02685 224.46974 113.38580 324.14739 
       29        30        31        32        33        34        35 
186.85504 133.44632 227.71562 158.83913 200.50188 135.82237 260.21389 
       36        37        38        39        40        41        42 
230.87854 145.24891 167.18986 234.32741 162.77587 244.34150 268.71120 
       43        44        45        46        47        48        49 
325.42771 225.04551 225.97479 304.09434 155.84527 156.53093 282.33633 
       50        51        52        53        54        55        56 
265.29048 228.11554  92.17628  84.51415 212.10532 304.79376 240.14874 
       57        58        59        60        61        62        63 
146.25170 250.63066 101.56978 175.10310 201.79282 318.13840 198.91147 
       64        65        66        67        68        69        70 
119.90039 181.49905 138.70907 124.13441 181.27199 138.28791 218.81203 
       71        72        73        74        75        76        77 
125.23074 274.49947 151.79277 213.93562 259.40816 173.80408 228.16395 
       78        79        80        81        82        83        84 
119.85352 253.50022 104.30331 116.96940  42.32256  45.14657 175.76113 
       85        86        87        88        89        90        91 
116.97063 275.18790 315.04344 319.47746 137.18402 174.05438 227.47501 
```

So if we look at the summary of both model we can see it very clearly that Model 2 is quite good in predicting sales. Sales of the dvds depend upon advertise spending, number of times song played and attractiveness of the score.