Regression analysis

Mohammad Waleed

2024-08-23

Contents

1 Experiment 1

To check the effect and its significance on sales from three types of advertisement, TV, Radio and Newspaper using linear regression model.

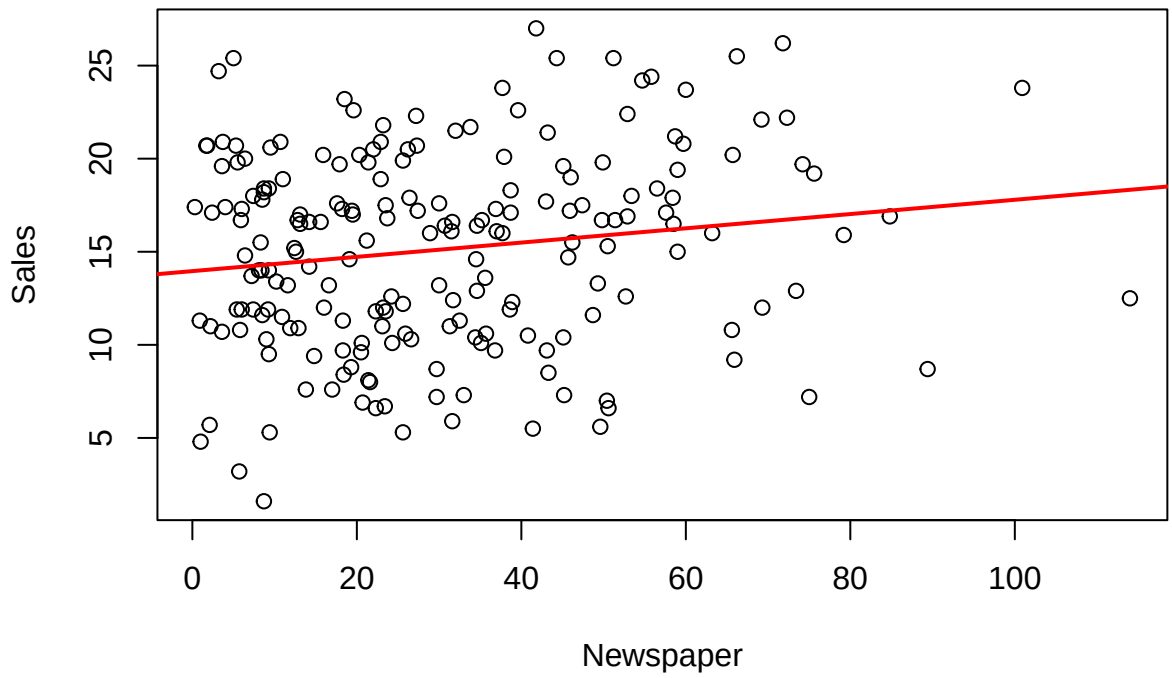
1.1 Experiment 1(a)

(a) Appliying linear regression model onto sales and newspaper. We have linear model $Sales = \beta_0 + \beta_1 * Newspaper + \epsilon$.

```
advert = read.csv("advertising.csv")
cor(Newspaper,Sales)
```

```
## [1] 0.15796
```

```
plot( Newspaper,Sales)
M1 = lm(Sales~Newspaper, data = advert)
abline(M1, col = "red", lwd = 2)
```



Hypothesis testing $H_o : \beta_i = 0$ for all i's i.e., the model is insignificant. $H1 : \beta_i \neq 0$ for some i, i.e., the model is significant.

1.1.1 Checking accuracy of the model

```
summary.lm(M1)
```

```
##
## Call:
## lm(formula = Sales ~ Newspaper, data = advert)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.6930  -3.8807   0.6591   3.9083  11.4385
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.95955    0.63829  21.870   <2e-16 ***
## Newspaper    0.03832    0.01703   2.251   0.0255 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.231 on 198 degrees of freedom
## Multiple R-squared:  0.02495,    Adjusted R-squared:  0.02003
## F-statistic: 5.067 on 1 and 198 DF,  p-value: 0.02549
```

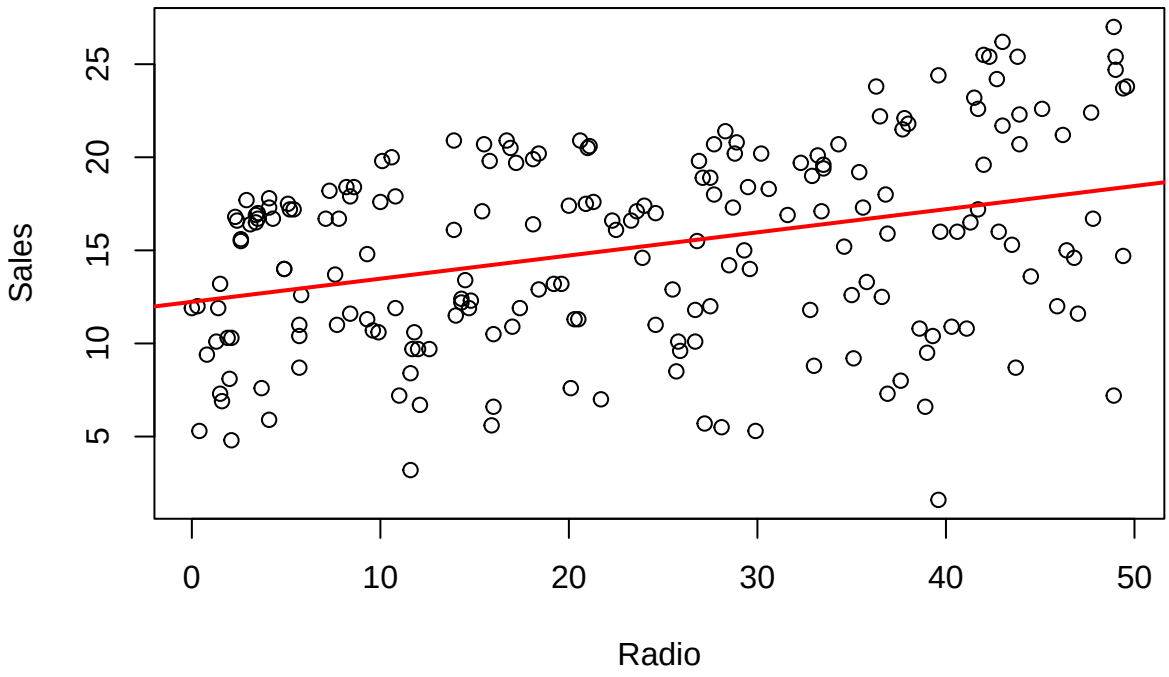
1.2 Experiment 1(b)

(b) Appliying linear regression model onto sales and Radio. We have linear model $Sales = \beta_0 + \beta_1 * Radio + \epsilon$

```
advert = read.csv("advertising.csv")
cor(Radio,Sales)
```

```
## [1] 0.3496311
```

```
plot(Radio,Sales)
M2 = lm(Sales~Radio, data = advert)
abline(M2, col = "red", lwd = 2)
```



Hypothesis testing $H_o : \beta_i = 0$ for all i's i.e., the model is insignificant. $H1 : \beta_i \neq 0$ for some i, i.e., the model is significant.

1.2.1 Checking accuracy of the model

```
summary.lm(M2)
```

```
##
## Call:
## lm(formula = Sales ~ Radio, data = advert)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5632  -3.5293   0.6714   4.2504   8.6796
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.2357     0.6535  18.724  < 2e-16 ***
## Radio         0.1244     0.0237   5.251 3.88e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.963 on 198 degrees of freedom
## Multiple R-squared:  0.1222, Adjusted R-squared:  0.1178
## F-statistic: 27.57 on 1 and 198 DF,  p-value: 3.883e-07
```

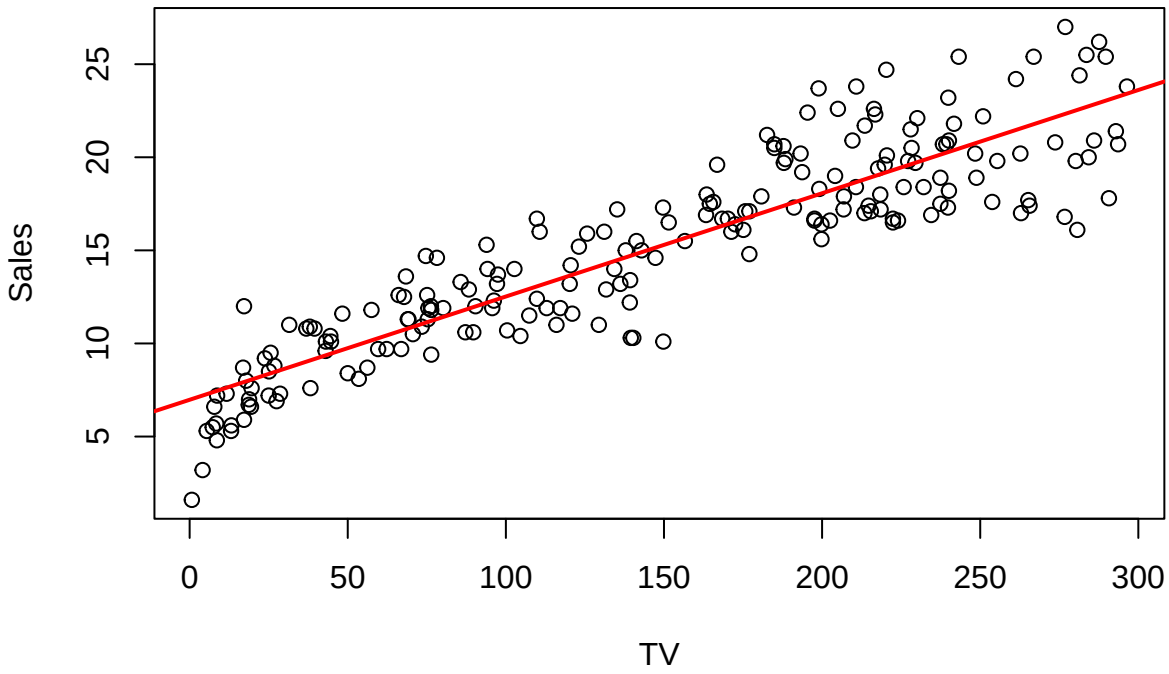
1.3 Experiment 1(c)

(c) Appliying linear regression model onto sales and TV. We have linear model $Sales = \beta_0 + \beta_1 * TV + \epsilon$

```
advert = read.csv("advertising.csv")
cor(Sales, TV)
```

```
## [1] 0.9012079
```

```
M3 = lm(Sales~TV, data = advert)
plot(TV,Sales)
abline(M3, col = "red", lwd = 2)
```



Hypothesis testing H_o : $\beta_i = 0$ for all i's i.e., each coeffcient is zero and the model is insignificant. $H1$: $\beta_i \neq 0$ for some i, i.e., atleast one is not zero and the model is significant.

1.3.1 Checking accuracy of the model

```
summary.lm(M3)
```

```
##
## Call:
## lm(formula = Sales ~ TV, data = advert)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4438 -1.4857  0.0218  1.5042  5.6932
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.974821   0.322553   21.62   <2e-16 ***
## TV          0.055465   0.001896   29.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.296 on 198 degrees of freedom
## Multiple R-squared:  0.8122, Adjusted R-squared:  0.8112
## F-statistic: 856.2 on 1 and 198 DF,  p-value: < 2.2e-16
```

1.4 Result

1(a) Since for the calculated p-value = 0.02549 < 0.05, our model is significant enough to reject the null hypothesis that all $\beta_i's$ are zero. The positive value of t-value(= 2.251) indicates positive correlation and the adjusted R-squared close to 0.02 indicates the 2% accuracy of the model.

1(b) Since for the calculated p-value $<$ 2.2e-16 < 0.05, our model is significant enough to reject the null hypothesis that all $\beta_i's$ are zero. The positive value of t-value(= 29.26) indicates positive correlation and the adjusted R-squared close to 0.81 indicates the 81% accuracy of the model.

1(c) Since for the calculated p-value $<$ 3.88e-07 < 0.05, our model is significant enough to reject the null hypothesis that all $\beta_i's$ are zero. The positive value of t-value(= 5.251) indicates positive correlation and the adjusted R-squared close to 0.12 indicates the 12% accuracy of the model.

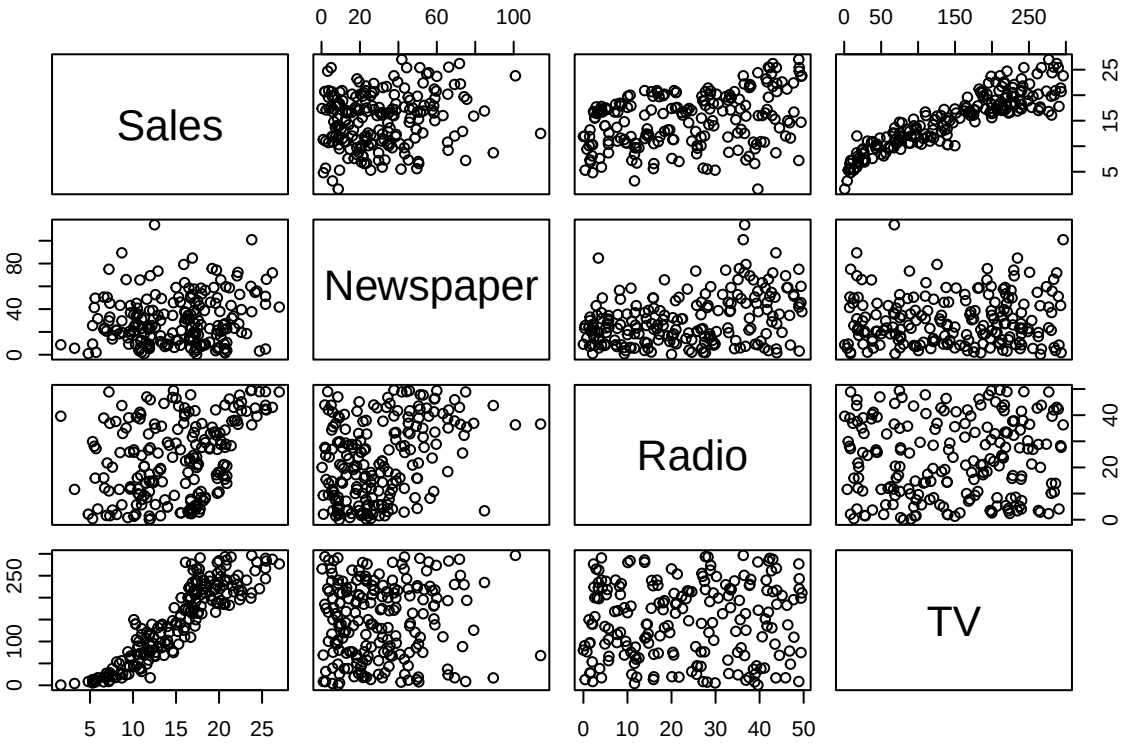
2 Experiment 2.

To check the effect of all three regressors on sales using multiple linear regression model.

```
advert = read.csv("advertising.csv")
Mlm1 = lm(Sales ~ Radio + Newspaper + TV, data = advert)
Mlm1
```

```
##
## Call:
## lm(formula = Sales ~ Radio + Newspaper + TV, data = advert)
##
## Coefficients:
## (Intercept)        Radio    Newspaper           TV
##   4.6251241    0.1070012    0.0003357    0.0544458
```

```
plot_all = data.frame(Sales, Newspaper, Radio, TV)
pairs(plot_all)
```



2.1 Checking accuracy of the model

```
advert = read.csv("advertising.csv")
Mlm1 = lm(Sales~Radio + Newspaper + TV, data = advert)
summary.lm(Mlm1)
```

```
##
## Call:
## lm(formula = Sales ~ Radio + Newspaper + TV, data = advert)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3034 -0.8244 -0.0008  0.8976  3.7473
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.6251241  0.3075012  15.041   <2e-16 ***
## Radio       0.1070012  0.0084896  12.604   <2e-16 ***
## Newspaper   0.0003357  0.0057881   0.058    0.954
## TV          0.0544458  0.0013752  39.592   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.662 on 196 degrees of freedom
## Multiple R-squared:  0.9026, Adjusted R-squared:  0.9011
## F-statistic: 605.4 on 3 and 196 DF,  p-value: < 2.2e-16
```

```
Mlm1
```

```
##
## Call:
## lm(formula = Sales ~ Radio + Newspaper + TV, data = advert)
##
## Coefficients:
## (Intercept)        Radio    Newspaper           TV
##   4.6251241    0.1070012    0.0003357    0.0544458
```

2.2 Analysis without insignificant regressor.

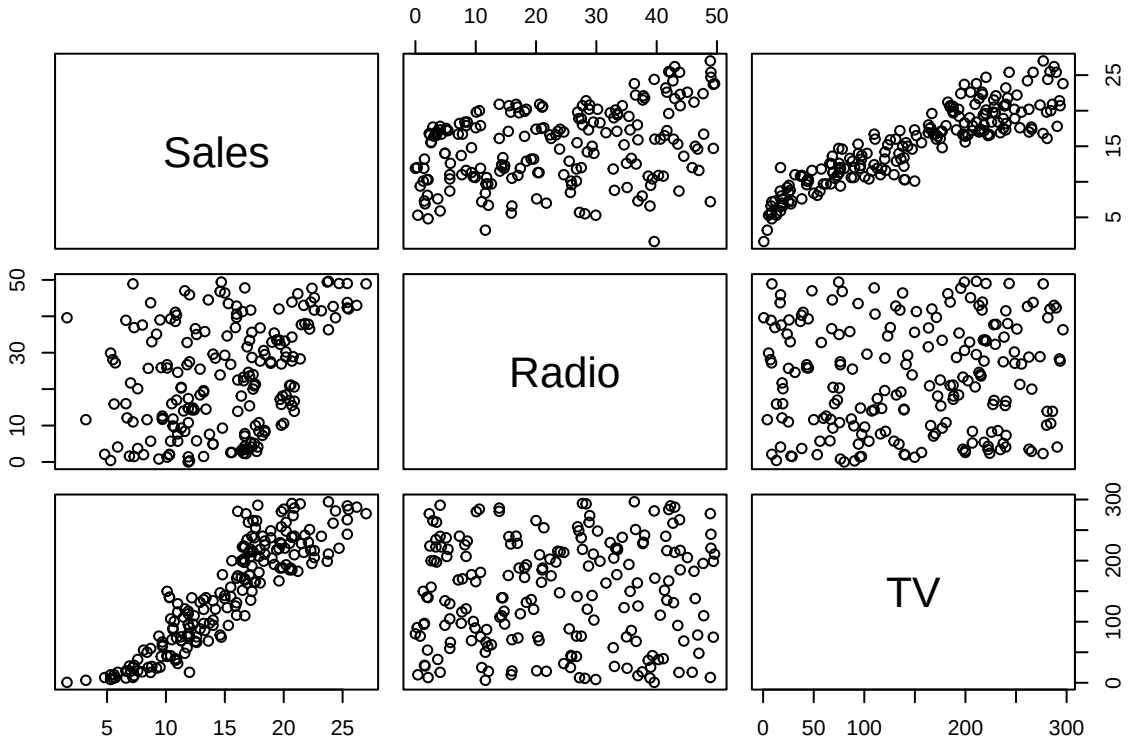
Since P value for newspaper is greater than 0.05, the regressor is insignificant for sales.

```
advert = read.csv("advertising.csv")

Mlm2 = lm(Sales ~ Radio + TV, data = advert)
Mlm2
```

```
##
## Call:
## lm(formula = Sales ~ Radio + TV, data = advert)
##
## Coefficients:
## (Intercept)        Radio           TV
##     4.63088      0.10717      0.05445
```

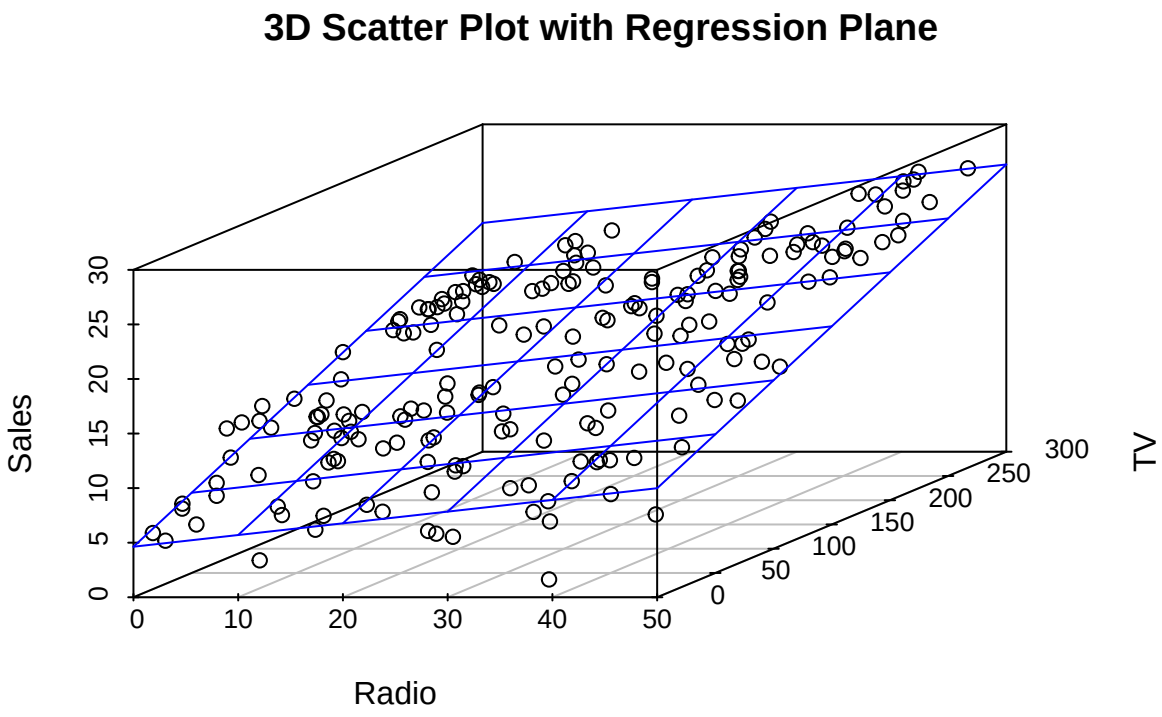
```r
plot_all = data.frame(Sales, Radio, TV)
pairs(plot_all)
```




```r
# Create the scatterplot3d object

plot3d = scatterplot3d(Radio, TV, Sales, main = "3D Scatter Plot with Regression Plane", xlab = "Radio"

# Add the regression plane
plot3d$plane3d(Mlm2, lty = "solid", col = "blue")
```



2.3 Checking accuracy of the model

```
advert = read.csv("advertising.csv")
Mlm2 = lm(Sales~Radio + TV, data = advert)
summary.lm(Mlm2)
```

```
##
## Call:
## lm(formula = Sales ~ Radio + TV, data = advert)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3131 -0.8269  0.0095  0.9022  3.7484
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.630879   0.290308   15.95   <2e-16 ***
## Radio       0.107175   0.007926   13.52   <2e-16 ***
## TV          0.054449   0.001371   39.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.657 on 197 degrees of freedom
```

```
## Multiple R-squared:  0.9026, Adjusted R-squared:  0.9016
## F-statistic: 912.7 on 2 and 197 DF,  p-value: < 2.2e-16
```

```
Mlm2
```

```
##
## Call:
## lm(formula = Sales ~ Radio + TV, data = advert)
##
## Coefficients:
## (Intercept)        Radio           TV
##     4.63088      0.10717      0.05445
```

2.4 Result

(1) While performing multiple linear regression model on the advertising data set we noticed the p-value for newspaper is $0.954 > 0.05$ which makes it insignificant compared to TV and Radio.

(2) After eliminating the insignificant regressor we get our linear model with adjusted R-squared $= 0.9016$, which indicates the accuracy of model to be 90.16%.