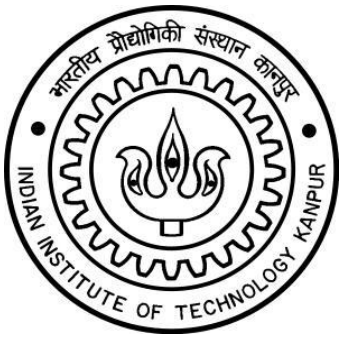
Analysis of the Advertising Media affecting Sales



MBA652A – Statistical Modelling for Business Analytics

Project Report

Submitted by:

Nitesh Sharma (20114013)

Declaration

This is to certify that the project report entitled ‘Analysis of the Advertising Media affecting Sales’ is based on our original research work. Our indebtedness to other works, studies and publications have been duly acknowledge at the relevant places.

INTRODUCTION

The Advertising data set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper. It is not possible for our client to directly increase sales of the product. On the other hand, they can control the advertising expenditure in each of the three media. Therefore, if we determine that there is an association between advertising and sales, then we can instruct our client to adjust advertising budgets, thereby indirectly increasing sales. In other words, our goal is to develop an accurate model that can be used to predict sales on the basis of the three media budgets.

DATA DESCRIPTION AND SOURCE

Dataset- Advertising.csv

In this setting, the advertising budgets are input variables while sales is an output variable.

Dependent variable

- sales(in thousands of units)

Independent variables

- TV (budget in thousands of dollars)
- radio (budget in thousands of dollars)
- newspaper (budget in thousands of dollars)

Below is sample of data set:

S.NO.	TV	Radio	newspaper	Sales
1	230.	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	152.	41.3	58.5	18.5
5	181.	10.8	58.4	12.9
6	8.7	48.9	75	7.2

http://www-bcf.usc.edu/~gareth/ISL/data.html/Advertising.csv

PURPOSE OF ANALYSIS

Here are few important questions that we might seek to address:-

a) Is there a relationship between advertising budget and sales?

b) How strong is the relationship between advertising budget and sales?

c) Which media contribute to sales?

d) How accurately can we estimate the effect of each medium on sales?

e) How accurately can we predict future sales?

f) Is the relationship linear?

ECONOMETRIC MODELS & ESTIMATION METHODS

Simple Regression: $Y \approx \beta 0 + \beta 1 \times X$

Model parameters

β0 : intercept

β1 : slope

Estimating the coefficients

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

MODEL – I

Applying Linear Regression Model by considering the independent variable TV (budget invested on advertising) & Number of sales. By the following formula:-

$$sales \approx \beta 0 + \beta 1 \times TV$$

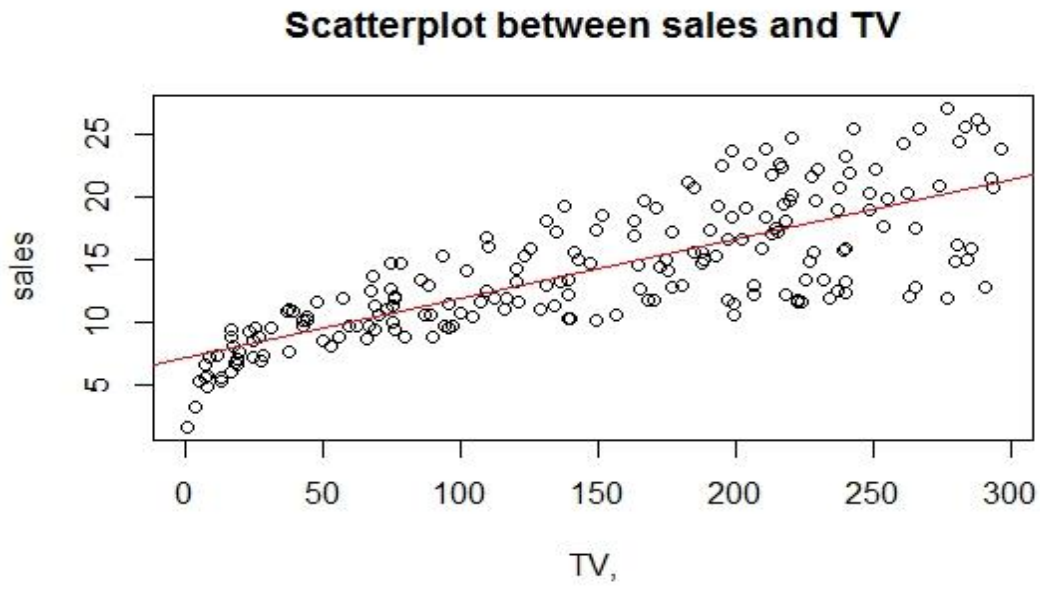
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.032594	0.457843	15.36	<2e-16 ***
TV	0.047537	0.002691	17.67	<2e-16 ***

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared: 0.6119 Adjusted R-squared: 0.6099

Confidence Interval:

	2.5 %	97.5 %
(Intercept)	6.12971927	7.93546783
TV	0.04223072	0.05284256



MODEL – II

Applying Linear Regression Model by considering the independent variable radio (budget invested on advertising) & Number of sales. By the following formula:-

$$sales \approx \beta 0 + \beta 1 \times radio$$

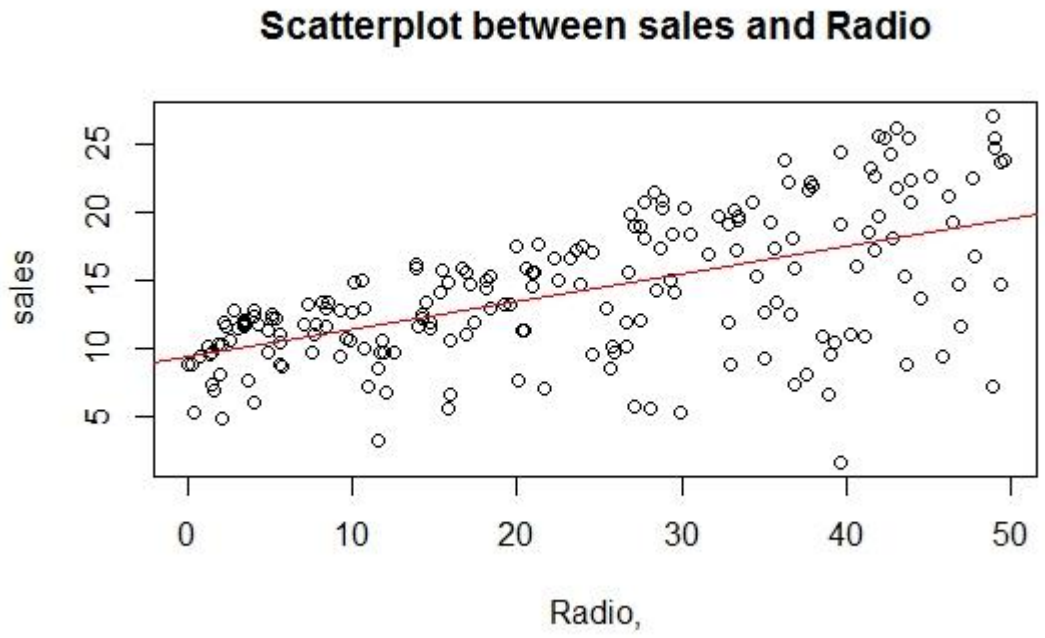
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	9.31164	0.56290	16.542	<2e-16	***
radio	0.20250	0.02041	9.921	<2e-16	***

Residual standard error: 4.275 on 198 degrees of freedom
Multiple R-squared: 0.332, Adjusted R-squared: 0.3287

Confidence interval:

	2.5 %	97.5 %
(Intercept)	8.2015885	10.4216877
radio	0.1622443	0.2427472



MODEL – III

Applying Linear Regression Model by considering the independent variable newspaper (budget invested on advertising) & Number of sales. By the following formula:-

$$sales \approx \beta 0 + \beta 1 \times newspaper$$

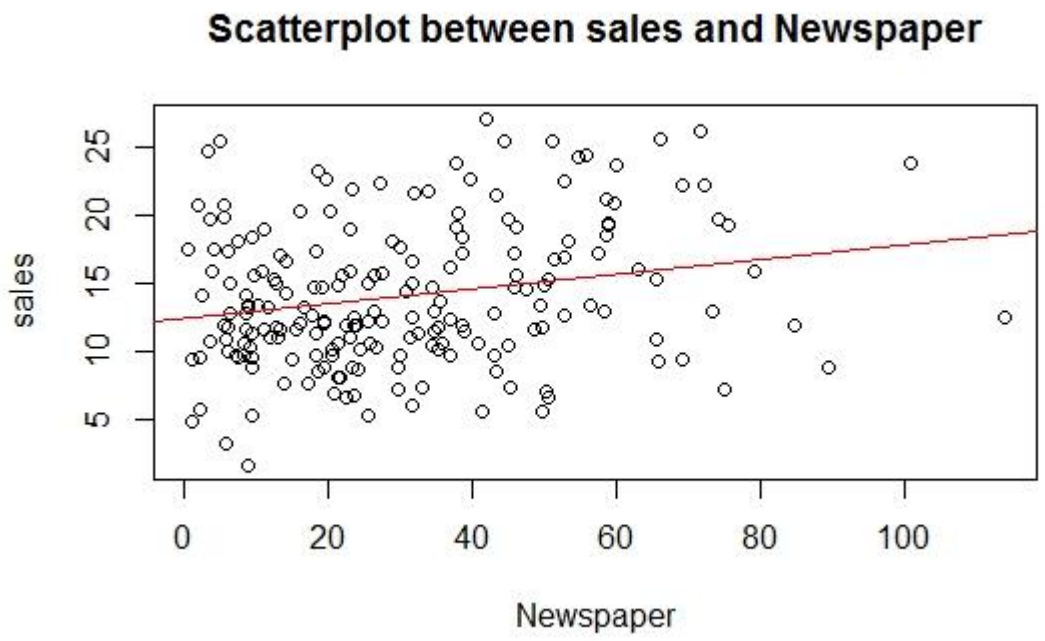
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	12.35141	0.62142	19.88	< 2e-16	***
newspaper	0.05469	0.01658	3.30	0.00115	**

Residual standard error: 5.092 on 198 degrees of freedom
Multiple R-squared: 0.05212, Adjusted R-squared: 0.04733

Confidence interval:

	2.5 %	97.5 %
(Intercept)	11.12595560	13.57685854
newspaper	0.02200549	0.08738071



MULTIPLE LINEAR REGRESSION

Correlation coefficient matrix

	TV	radio	newspaper	sales
TV	1.00000000	0.05480866	0.05664787	0.7822244
radio	0.05480866	1.00000000	0.35410375	0.5762226
newspaper	0.05664787	0.35410375	1.00000000	0.2282990
sales	0.78222442	0.57622257	0.22829903	1.0000000

Variance Inflation Factor

The Variance Inflation Factor (VIF) is always greater than or equal to 1. Values of VIF that exceed 10 are often regarded as indicating multicollinearity. In our model we can observe that vif value corresponding to each variables is nearly equal to 1.

TV	newspaper	radio
1.004611	1.145187	1.144952

MLR Model Assumptions:

- The mean of the response ,at each set of values of the predictors, is a **Linear function** of the predictors.
- The errors(residuals) are **Independent**.
- The errors(residuals)at each set of values of the predictors are **Normally distributed**.
- The errors, (residuals) at each set of values of the predictors , have **Equal variances** .

MODEL - IV

Applying Multiple Linear Regression Model by considering the independent variable TV, radio and newspaper (budget invested on advertising) & Number of sales. By the following formula:-

$$sales \approx \beta 0 + \beta 1 \times TV + \beta 2 \times radio + \beta 3 \times newspaper$$

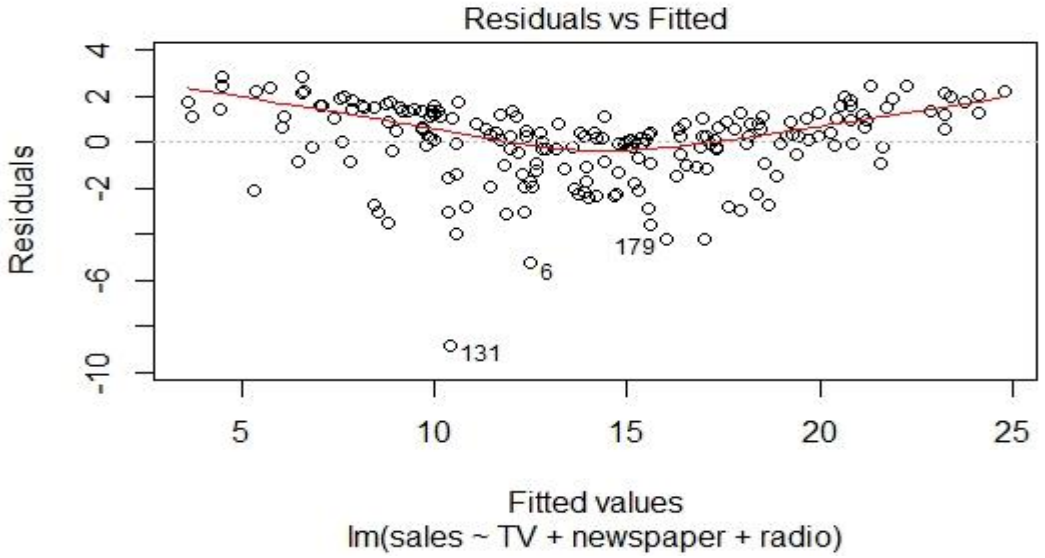
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.938889	0.311908	9.422	<2e-16 ***
TV	0.045765	0.001395	32.809	<2e-16 ***
newspaper	-0.001037	0.005871	-0.177	0.86
radio	0.188530	0.008611	21.893	<2e-16 ***

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared: 0.8972, Adjusted R-squared: 0.8956
F-statistic: 570.3 on 3 and 196 DF, p-value: < 2.2e-16

Confidence interval:

	2.5 %	97.5 %
(Intercept)	2.32376228	3.55401646
TV	0.04301371	0.04851558
newspaper	-0.01261595	0.01054097
radio	0.17154745	0.20551259



MODEL-V sales~ TV, radio

As we can observe from model-4 that variable newspaper is statistically insignificant. Hence we remove it in this model.

Residuals:

Min	1Q	Median	3Q	Max
-8.7977	-0.8752	0.2422	1.1708	2.8328

Coefficients:

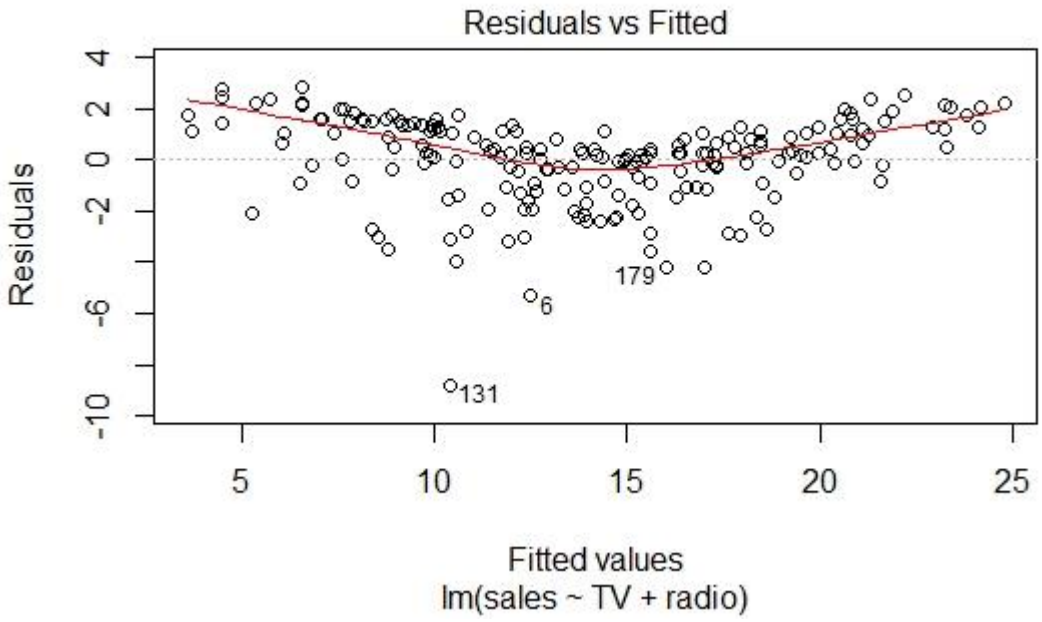
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.92110	0.29449	9.919	<2e-16	***
TV	0.04575	0.00139	32.909	<2e-16	***
radio	0.18799	0.00804	23.382	<2e-16	***

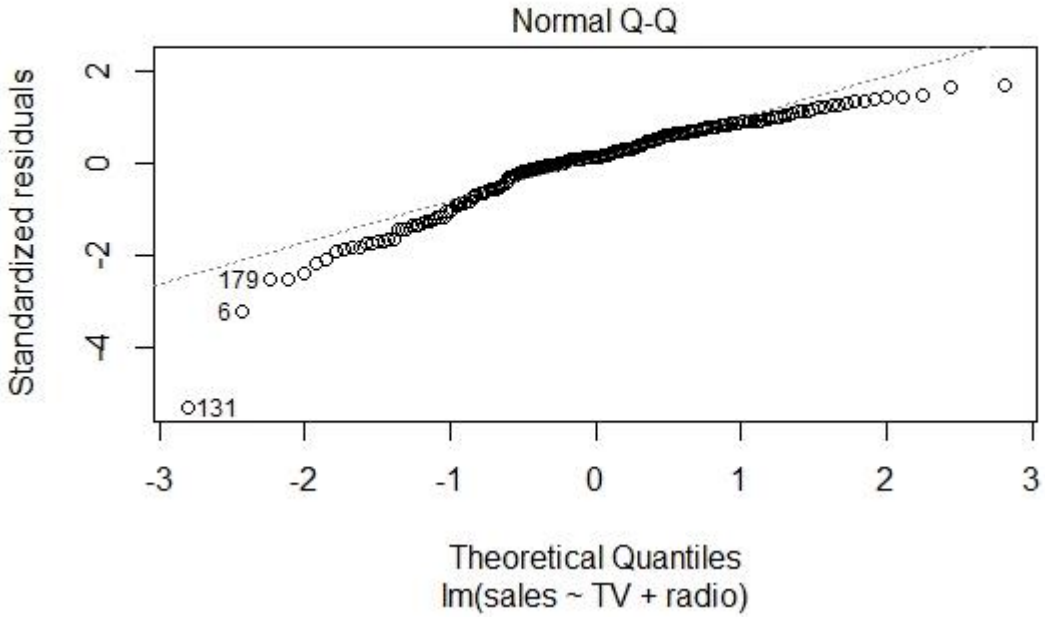
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.681 on 197 degrees of freedom
Multiple R-squared: 0.8972, Adjusted R-squared: 0.8962
F-statistic: 859.6 on 2 and 197 DF, p-value: < 2.2e-16

Coefficient interval:

	2.5 %	97.5 %
(Intercept)	2.34034299	3.50185683
TV	0.04301292	0.04849671
radio	0.17213877	0.20384969





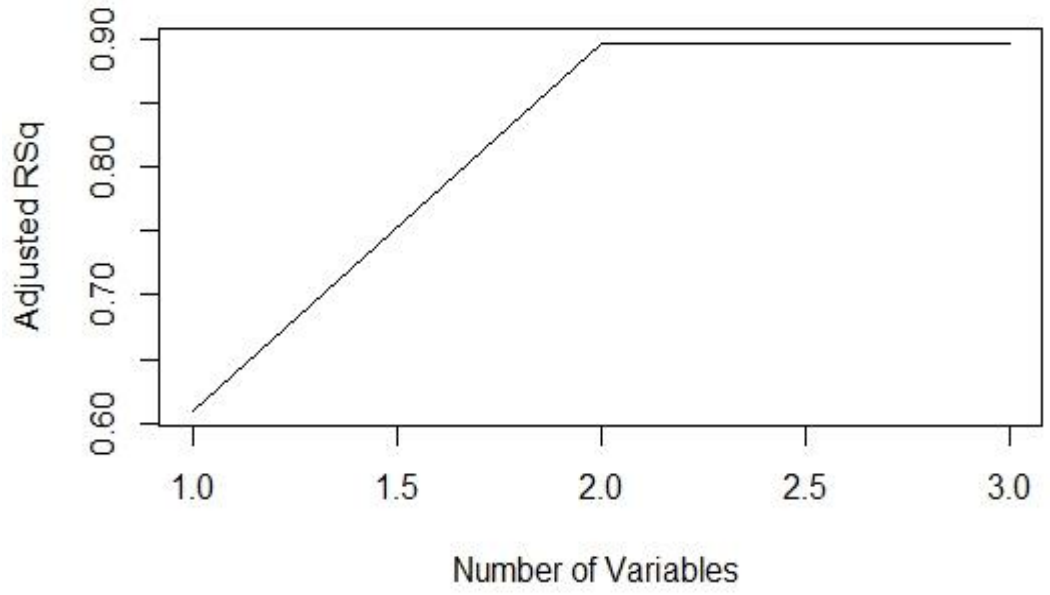
VARIABLE SELECTION

Subset Selection

Subset selection object
3

```
         TV  radio newspaper
1  ( 1 ) "*" " "   " "
2  ( 1 ) "*" "*"   " "
3  ( 1 ) "*" "*"   "*"
> reg.summary$rsq
[1] 0.6118751 0.8971943 0.8972106
> reg.summary$adjr2
[1] 0.6099148 0.8961505 0.8956373
> reg.summary$bic
[1] -178.6890 -439.0879 -433.8214
> coef(regfit.full,3)
```

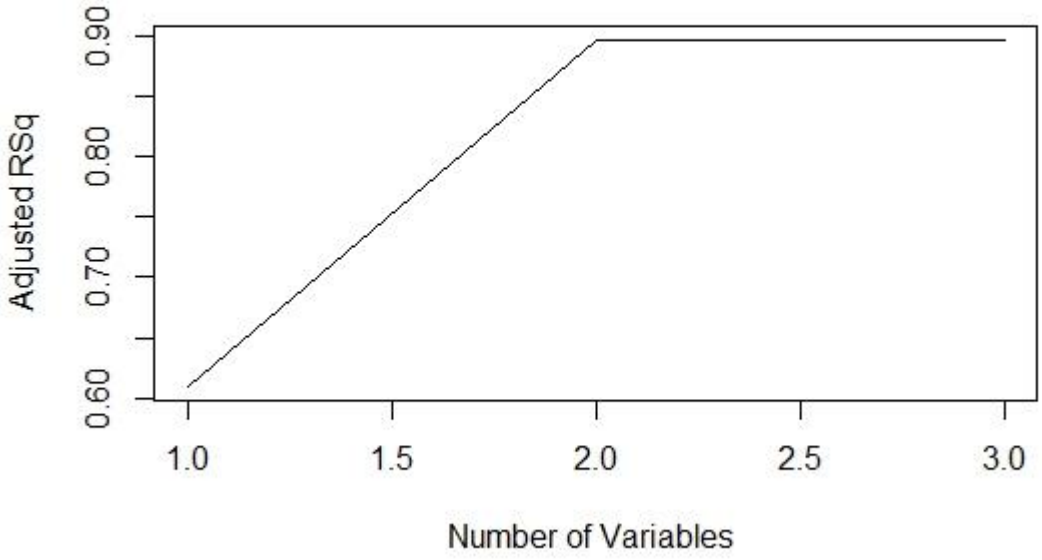
(Intercept)	TV	radio	newspaper
2.938889369	0.045764645	0.188530017	-0.001037493



Forward Selection

```
         TV  radio newspaper
1  ( 1 ) "*" " "   " "
2  ( 1 ) "*" "*"   " "
3  ( 1 ) "*" "*"   "*"
> reg.summary1$rsq
[1] 0.6118751 0.8971943 0.8972106
> reg.summary1$adjr2
[1] 0.6099148 0.8961505 0.8956373
> reg.summary1$bic
[1] -178.6890 -439.0879 -433.8214
> coef(regfit.fwd,3)
```

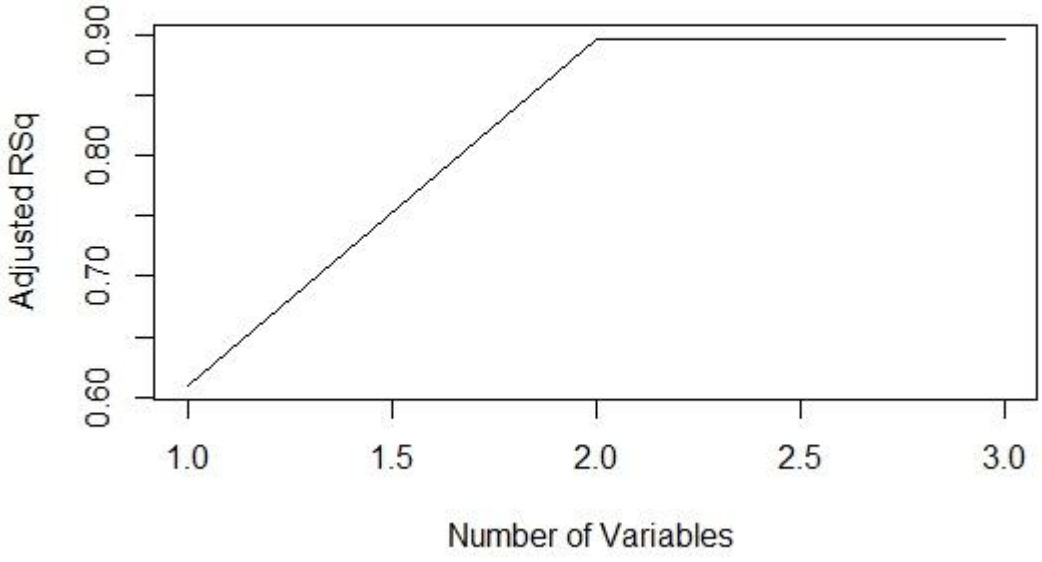
(Intercept)	TV	radio	newspaper
2.938889369	0.045764645	0.188530017	-0.001037493



Backward

```
         TV  radio newspaper
1  ( 1 ) "*" " "   " "
2  ( 1 ) "*" "*"   " "
3  ( 1 ) "*" "*"   "*"
> coef(regfit.bwd,3)
```

(Intercept)	TV	radio	newspaper
2.938889369	0.045764645	0.188530017	-0.001037493



MODEL-VI

Removing the Additive Assumption

```
lm(formula = sales ~ TV + radio + TV * radio, data = Advertising)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-6.3366	-0.4028	0.1831	0.5948	1.5246

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.750e+00	2.479e-01	27.233	<2e-16 ***
TV	1.910e-02	1.504e-03	12.699	<2e-16 ***
radio	2.886e-02	8.905e-03	3.241	0.0014 **
TV:radio	1.086e-03	5.242e-05	20.727	<2e-16 ***

```
Residual standard error: 0.9435 on 196 degrees of freedom
Multiple R-squared:  0.9678,   Adjusted R-squared:  0.9673
```

```
> confint(Advertising.lm5)
```

	2.5 %	97.5 %
(Intercept)	6.2613828568	7.239057549
TV	0.0161346865	0.022067461
radio	0.0112978842	0.046422796
TV:radio	0.0009831143	0.001189875

R CODES

```
summary(Advertising)

#Model 1 sales ~TV

plot(Advertising$TV,Advertising$sales,main = "Scatterplot between sales and TV",xlab =
"TV,",ylab="sales")

Advertising.lm1=lm(sales~TV,Advertising)

Advertising.lm1

summary(Advertising.lm1)

confint(Advertising.lm1)

abline(Advertising.lm1,col="red")

plot(Advertising.lm1)


#Model 2 sales ~ radio
```

```
plot(Advertising$radio,Advertising$sales,main = "Scatterplot between sales and Radio",xlab = "Radio,",ylab="sales")

Advertising.lm2=lm(sales~radio,Advertising)

Advertising.lm2

summary(Advertising.lm2)

confint(Advertising.lm2)

abline(Advertising.lm2,col="red")

plot(Advertising.lm2)


#Model 3 sales ~ newspaper

plot(Advertising$newspaper,Advertising$sales,main = "Scatterplot between sales and Newspaper",xlab = "Newspaper",ylab="sales")

Advertising.lm3=lm(sales~newspaper,Advertising)

Advertising.lm3

summary(Advertising.lm3)

confint(Advertising.lm3)

abline(Advertising.lm3,col="red")

plot(Advertising.lm3)


#Model 4 sales ~ TV, ,radio, ,newspaper

Advertising.lm4=lm(sales~TV+newspaper+radio,Advertising)

Advertising.lm4

summary(Advertising.lm4)

confint(Advertising.lm4)

plot(Advertising.lm4)


#Model 5 sales ~ TV, radio

Advertising.lm5=lm(sales~TV+radio,Advertising)

Advertising.lm5
```

```
summary(Advertising.lm5)
confint(Advertising.lm5)
plot(Advertising.lm5)
library(car)
vif(Advertising.lm4)
par(mfrow=c(2,2))
plot(Advertising.lm1)
#remove index column
Advertising=Advertising[c(2:5)]
View(Advertising)

# Correlation Plot
corr1=cor(Advertising, method="pearson",use="complete.obs")
corr1
library(corrplot)
corrplot(corr1,type="upper",order ="hclust",tl.col="black",tl.srt=45)
#Subset selection
library(leaps)
regfit.full=regsubsets (sales~.,Advertising)
reg.summary=summary(regfit.full)
reg.summary
reg.summary$rsq
reg.summary$adjr2
reg.summary$bic
coef(regfit.full,3)
par(mfrow=c(2,2))
plot(reg.summary$rss ,xlab="Number of Variables ",ylab="RSS",type="l")
plot(reg.summary$adjr2 ,xlab="Number of Variables ",ylab="Adjusted RSq",type="l")
```

```
regfit.fwd=regsubsets (sales~.,data=Advertising,method ="forward")
reg.summary1=summary (regfit.fwd)
reg.summary1
reg.summary1$rsq
reg.summary1$adjr2
reg.summary1$bic
coef(regfit.fwd,3)
par(mfrow=c(2,2))
plot(reg.summary1$rss ,xlab="Number of Variables ",ylab="RSS",type="l")
plot(reg.summary1$adjr2 ,xlab="Number of Variables ",ylab="Adjusted RSq",type="l")
regfit.bwd=regsubsets (sales~.,data=Advertising ,method ="backward")
reg.summary2=summary (regfit.fwd)
reg.summary2
reg.summary2$rsq
reg.summary2$adjr2
reg.summary2$bic
summary (regfit.bwd)
coef(regfit.bwd,3)
par(mfrow=c(2,2))
plot(reg.summary2$rss ,xlab="Number of Variables ",ylab="RSS",type="l")
plot(reg.summary2$adjr2 ,xlab="Number of Variables ",ylab="Adjusted RSq",type="l")

#Removing the Additive Assumption
Advertising.lm5=lm(sales~TV+radio+TV*radio,Advertising)
summary(Advertising.lm5)
summary(Advertising.lm5)
confint(Advertising.lm5)
plot(Advertising.lm5)
```

CONCLUSION

- Fitting a multiple regression model of sales onto TV, radio, and newspaper and testing null hypothesis ***H0 : βTV = βradio = βnewspaper = 0***,using F-statistics. There is a clear evidence of relationship between advertising budgets and sales.
- The predictors explain almost 90% of the variance in sales.
- In the multiple linear regression displayed the p-values for TV and radio are low, but the p-value for newspaper is not. This suggests that only TV and radio are related to sales.
- For the Advertising data, the 95% confidence intervals are as follows: (0.043, 0.049) for TV,(0.172, 0.206) for radio, and (−0.013, 0.011) for newspaper. The confidence intervals for TV and radio are narrow and far from zero, providing evidence that these media are related to sales. But the interval for newspaper includes zero, indicating that the variable is not statistically significant .
- We saw that residual plots can be used in order to identify non-linearity, heteroscedasticity and non-normality.

REFERENCES

1) An Introduction to Statistical Learning with Applications in R (Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani)
2) Introduction to Econometrics (James H. Stock and Mark Watson)