Term_Project

Kokil Dhakal

2023-12-13

Describing research scenario and question

Research Scenario:

I am considering a marketing research scenario where a company wants to understand the effectiveness of its advertising strategies on sales. The company has historical data on advertising budgets for TV, Radio, and Newspaper, along with corresponding sales figures.

Research Question:

Is there a significant relationship between advertising budgets (TV, Radio, Newspaper) and sales? Or are the advertising budgets in tv and/or radio and/or newspaper help in predicting sales?

About data set: This Data set is obtained from the Kaggle.This data has four columns three of them are for the Money spent in the advertising in TV,Radio and Newspaper in thousands respectively while sales column is the revenue generated in millions.Here, expenses in advertisement is independent variables while sales revenue is dependent variable.There are 200 observations and 4 variables.

Loading the data

```
project.df <- read.csv("/Users/kokildhakal/Desktop/STUDY/BU/9.CS555/Term_project/Advertising Budget and
#first fews rows of dataset.
head(project.df)
```

```
##   X TV.Ad.Budget.... Radio.Ad.Budget.... Newspaper.Ad.Budget.... Sales....
## 1 1            230.1                37.8                    69.2      22.1
## 2 2             44.5                39.3                    45.1      10.4
## 3 3             17.2                45.9                    69.3       9.3
## 4 4            151.5                41.3                    58.5      18.5
## 5 5            180.8                10.8                    58.4      12.9
## 6 6              8.7                48.9                    75.0       7.2
```

Describing the data

Knowing some facts about the dataset:

Preprocessing the dataset:

It involves removing the extra index column, renaming the column names, and checking missing values

```
project.df <- project.df[-1] # removing first column
```

Since there is no use of this extra index column, I am removing it.

Renaming the column names:

```
colnames(project.df) <- c("tv_ad(K)","radio_ad(K)","newspaper_ad(K)","sales(M)")
head(project.df)
```

```
##   tv_ad(K) radio_ad(K) newspaper_ad(K) sales(M)
## 1    230.1        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    151.5        41.3            58.5     18.5
## 5    180.8        10.8            58.4     12.9
## 6      8.7        48.9            75.0      7.2
```

The reason behind renaming the columns is to make more convenient to work with dataset.

Checking if there is any missing values:

```
str(project.df)
```

```
## 'data.frame':    200 obs. of  4 variables:
##  $ tv_ad(K)       : num  230.1 44.5 17.2 151.5 180.8 ...
##  $ radio_ad(K)    : num  37.8 39.3 45.9 41.3 10.8 48.9 32.8 19.6 2.1 2.6 ...
##  $ newspaper_ad(K): num  69.2 45.1 69.3 58.5 58.4 75 23.5 11.6 1 21.2 ...
##  $ sales(M)       : num  22.1 10.4 9.3 18.5 12.9 7.2 11.8 13.2 4.8 10.6 ...
```

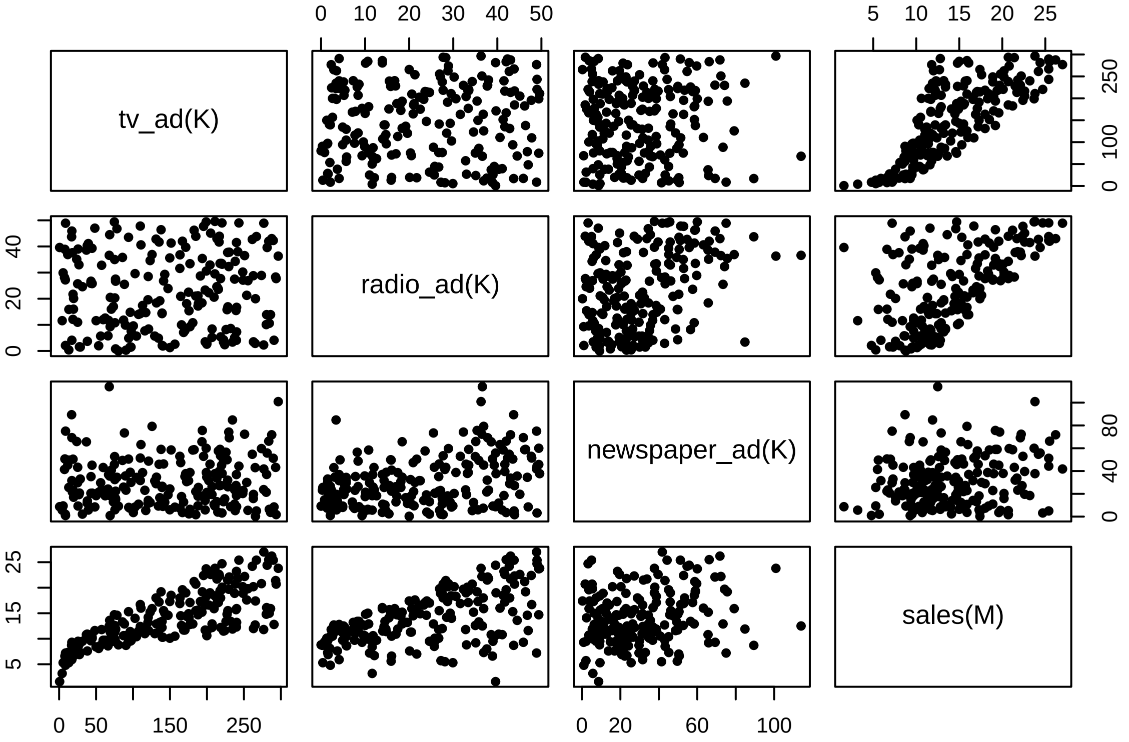
```
any(is.na(project.df))
```

```
## [1] FALSE
```

From above calculation there is no missing values and data set has 200 rows and 4 columns.By using str() method, it can be found that there are 200 rows and all variables are numerical.

Pairplotting:

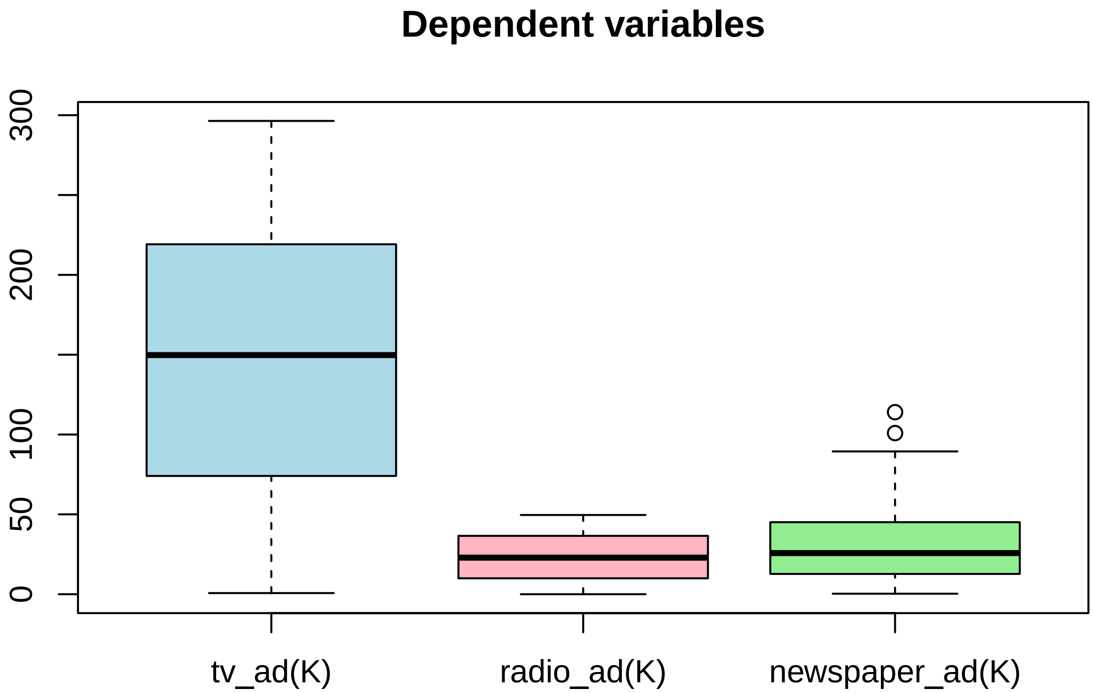
```
#pair plot
pairs(project.df,pch=16)
```



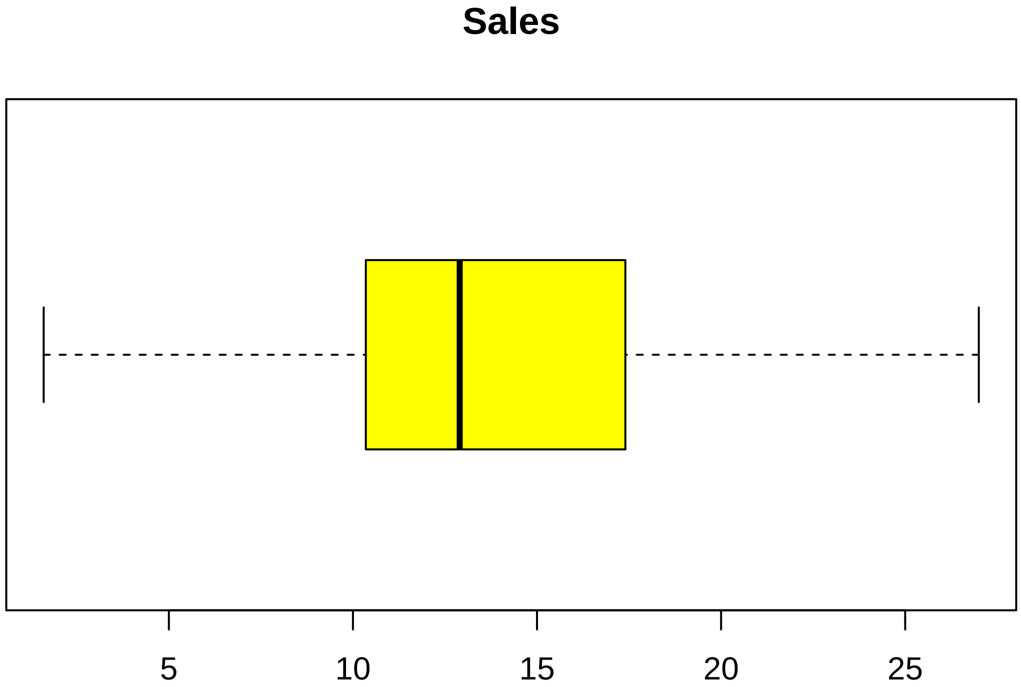
From above pair plot, it seems that there is linear and strong positive relationship between radio add and sales, tv_add and sales,however in case of newspaper_add and sales, there is some positive association,data points are more scattered but still have some linear association.As my responsive variable in this data set is sales, I am more concern about finding the association between sales and other variables.However association among explanatory variables is also important to know about the co-linearity.

Boxplot:

```
#plotting boxplot for dependent and independent variable seperately
boxplot.df <- project.df[1:3]
boxplot(boxplot.df,col = c("lightblue","lightpink","lightgreen"),main="Dependent variables")
```



```
boxplot(project.df[4],col = "yellow",main = "Sales",horizontal = T)
```



From above box plot, it can be seen that Median amount spent in tv_ad is higher than that newspaper ad and newspaper_ad.Similarly, More variability of data points can be seen in tv_ad followed by newspaper and radio_add.Also, there are two upper outlier points in case of newspaper while there are no outliers points in other two coulumns of data.

Calculating correlation:

```
# calculating correlations
cor.table <- cor(project.df)
kable(cor.table,align = "c",digits = 3)
```

	tv_ad(K)	radio_ad(K)	newspaper_ad(K)	sales(M)
tv_ad(K)	1.000	0.055	0.057	0.782
radio_ad(K)	0.055	1.000	0.354	0.576
newspaper_ad(K)	0.057	0.354	1.000	0.228
sales(M)	0.782	0.576	0.228	1.000

From above correlation table, the correlation between tv_ad and sales seems strong,positive. Similarly, relation between radio_ad and sales is also positive but moderately strong and, between newspaper_ad and sales is still positive and not so strong association.Also, we can see that there is some co-linearity between radio_ad and newpapter_add thought not so strong.

Making summary of datasets:

	min	max	mean	median	sd	Q1	Q3
tv_ad	0.0	296.4	147.04	149.75	85.85	74.38	218.82
radio_ad	0.0	49.6	23.26	22.90	14.85	9.97	36.52
newspaper_ad	0.3	114.0	30.55	25.75	21.78	12.75	45.10
sales	1.6	27.0	14.02	12.90	5.22	10.38	17.40

This summary tables gives the values that we described in boxplot section.we have two outliers in the newspaper_ad column.However I am not going to remove it from data set. Since I will be performing linear regression from this dataset.I will be checking if this data points really affect the overall model performance using R_squared value before and after incorporating potential influential points.I will just identify this outlier points just to know as follow.

```
upper.val <- Q3.news+1.5*(Q3.news-Q1.news)
upper.val
```

```
##    75%
## 93.625
```

```
outliers <- which(project.df$`newspaper_ad(K)` >upper.val)
outliers
```

```
## [1]  17 102
```

```
#nrow(project.df) #Checking total number of rows in original data set
#project.df <- project.df[-outliers,]
#nrow(project.df) # Checking total number of rows after removing outliers.
```

And, it seems that data points from row number 17 and 102 have values which are outside the threshold outlier value(93.625K) for newspaper_ad column.Usually those data points need to be re-evaluate about their validity to confirmed it is from real data and not put it in accidentally.

Method:

Since my data set contain all continuous values. I will be using simple linear and multiple linear regression to get the answer of my research question.

Simple Linear Regression:

For the simple linear regression, I will be use one independent variable(predictor) i.e tv_add and dependent variable sales.The reason behind choosing this variable is that it has higher correlation(0.782) with response variable.

```
#building the simple linear regression model
simple.tv.ad <- lm(data = project.df,`sales(M)`~`tv_ad(K)`)
summary(simple.tv.ad)
```

```
##
## Call:
## lm(formula = `sales(M)` ~ `tv_ad(K)`, data = project.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## `tv_ad(K)`  0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

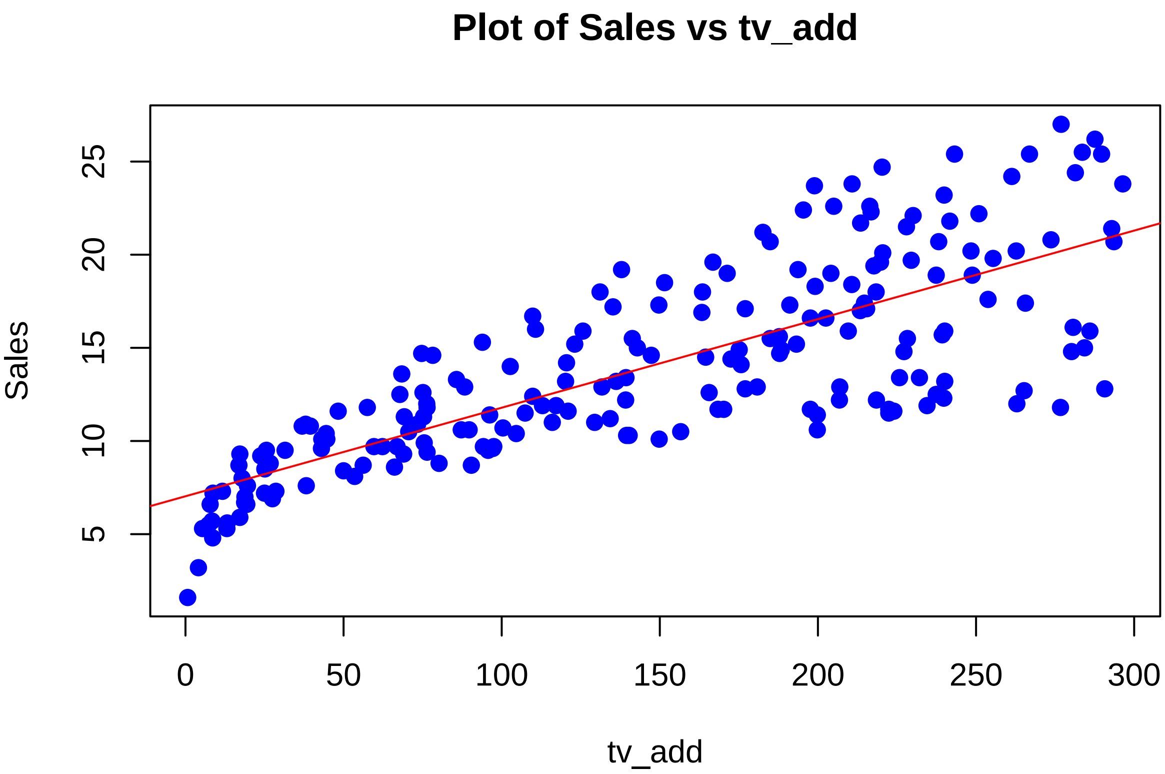
```
anova(simple.tv.ad)
```

```
## Analysis of Variance Table
##
## Response: sales(M)
##             Df Sum Sq Mean Sq F value    Pr(>F)
## `tv_ad(K)`   1 3314.6  3314.6  312.14 < 2.2e-16 ***
## Residuals  198 2102.5    10.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From above summary table, it can be interpreted as tv_ad is a good predictor of the sales as p-value is very low which means this predictor is significant at level of 0.05 in predicting the sales.It can also be interpreted as Beta(estimate) is not equal to zero.We have the tv_ad estimate(0.047381),which can be interpreted as each thousands spend in TV ad, sales is increased by an average of 047381 million or $47,381. The intercept is 7.03056 which can be interpreted as base level of sales without any advertisement keeping all variables constant.Also, from above summary table, we have the r-squared value of 0.6119 which can be interpreted as about 61.19% of variability in response variable(sales) is explained by the explanatory variable(tv_add)

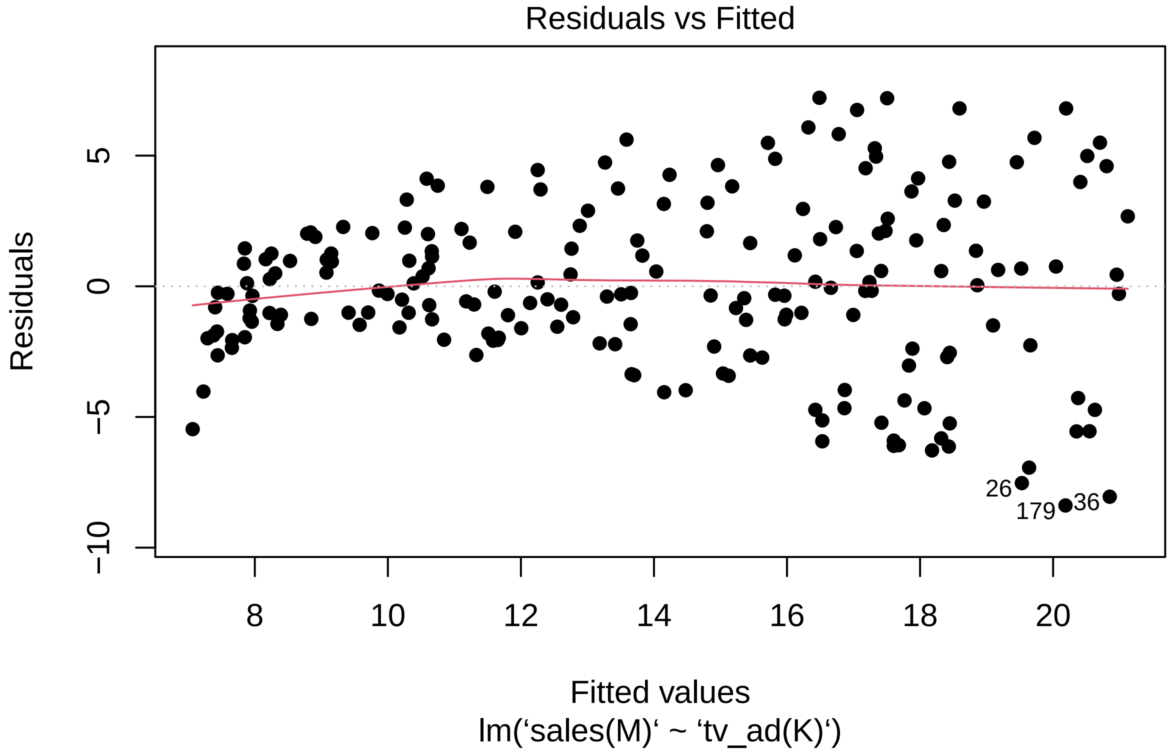
Now plotting the scatter plot between sales and tv_ad with the regression line as follows.

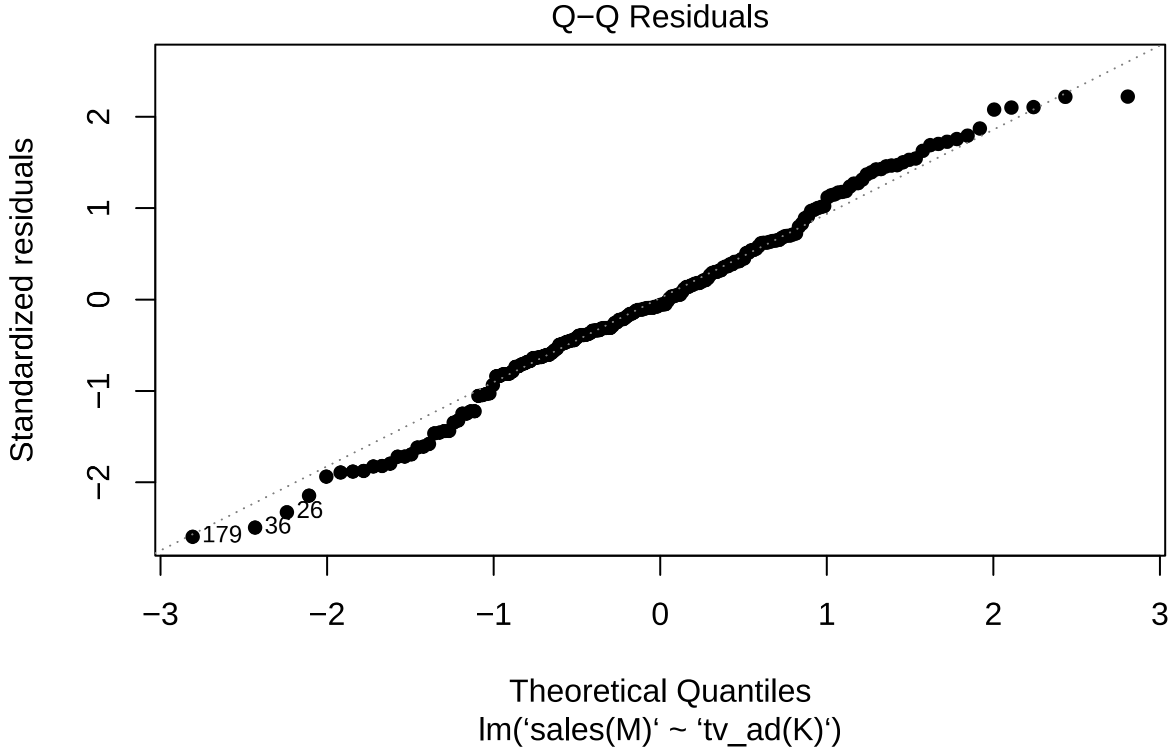
```
plot(x=project.df$tv_ad,y=project.df$sales,
     xlab = "tv_add",
     ylab = "Sales",
     main = "Plot of Sales vs tv_add",
     cex=1.2,
     col="blue",
     pch=16)
abline(simple.tv.ad,col="red")
```

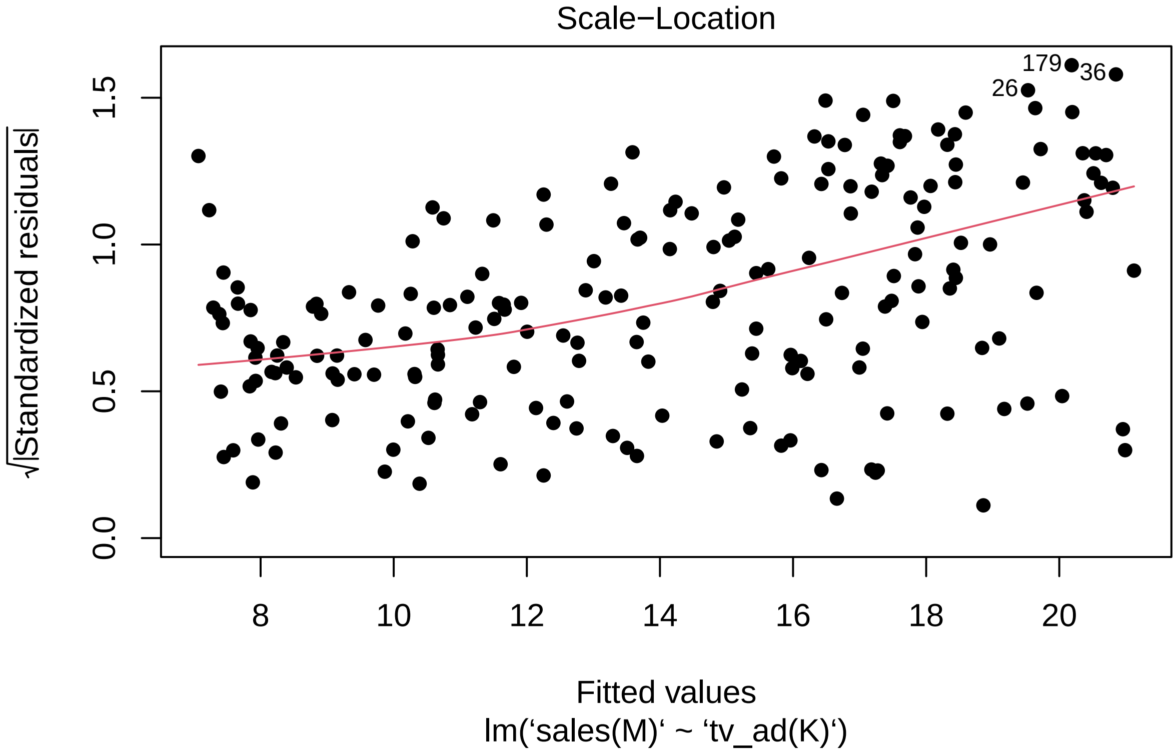


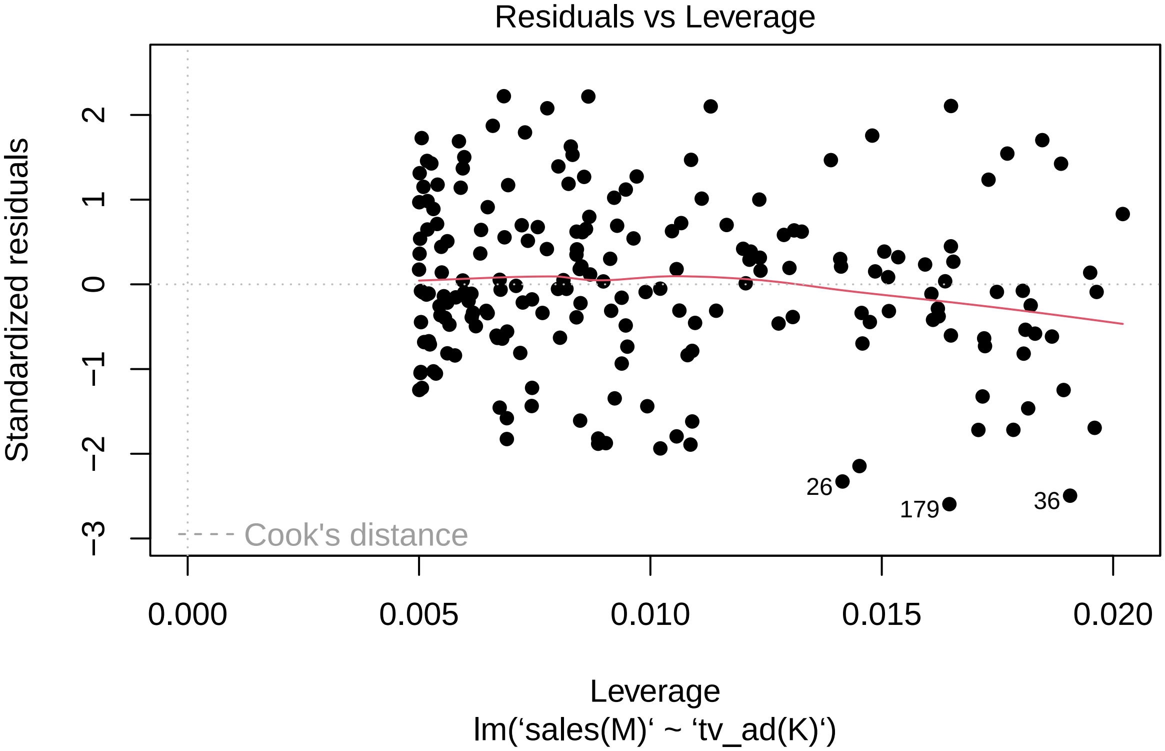
Now checking whether all assumptions of linear regression are met or not

```
plot(simple.tv.ad,pch=16)
```

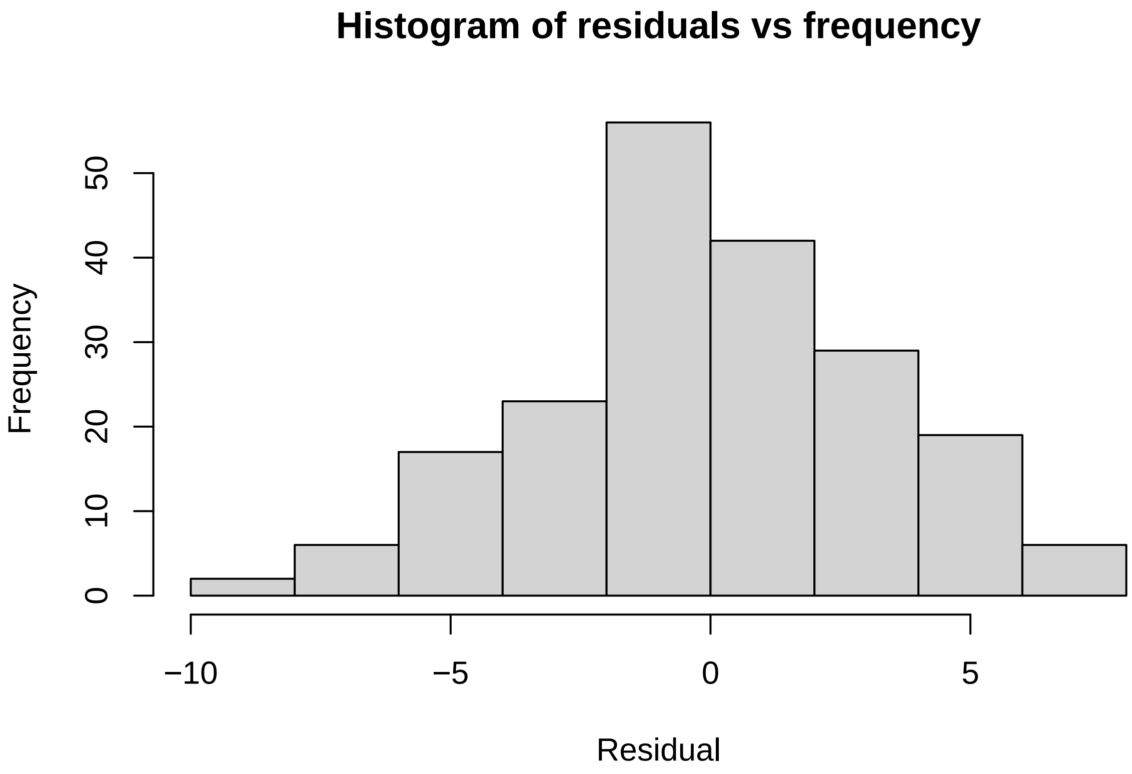

Residuals vs Fitted
Residuals
5
0
−5
−10
8
10
12
14
16
18
20
26
179
36
Fitted values
lm(‘sales(M)‘ ~ ‘tv_ad(K)‘)


Q−Q Residuals
Standardized residuals
2
1
0
−1
−2
−3
−2
−1
0
1
2
3
179
36
26
Theoretical Quantiles
lm(‘sales(M)‘ ~ ‘tv_ad(K)‘)





```r
hist(simple.tv.ad$residuals,main = "Histogram of residuals vs frequency",xlab="Residual")
```



From above plot it seems that residuals are some what normally distributed. we can also see that there is constant variance around the regression line from scale-location plot.Plot between tv_ad and sales is linearly associated and we assume the data points are independent as each point represent amount spent in tv_ad, radio_ad,newspaper_ad and I assume those are independent to each other.

Now checking influential/outliers

```
influential.points.simple <- cooks.distance(simple.tv.ad) >4/(nrow(project.df)-2)
which(influential.points.simple)
```

```
##  26  36  37 103 129 131 132 148 170 176 179 184 199
##  26  36  37 103 129 131 132 148 170 176 179 184 199
```

```
set.seed(123)
for (x in which(influential.points.simple)) {
 simple.data <- project.df #copying main data
 simple.data <- simple.data[-x,] #removing the row with potential influential point
 #building model after removing the potential influential point
 simple.model1 <- lm(data = simple.data,`sales(M)`~`tv_ad(K)`)

 #To get the slope of model
 slope.value1 <- summary(simple.model1)$coefficients[2,1]
 #to get r-squared from model
 rsquared1 <- summary(simple.model1)$r.squared
 cat("After removing row:",x,"Rsquard is ",rsquared1,"and slope is",slope.value1,"\n")
}
```

```
## After removing row: 26 Rsquard is  0.6222054 and slope is 0.04813994
## After removing row: 36 Rsquard is  0.6239703 and slope is 0.04834052
## After removing row: 37 Rsquard is  0.608518 and slope is 0.04706555
## After removing row: 103 Rsquard is  0.6176221 and slope is 0.04804945
## After removing row: 129 Rsquard is  0.6133365 and slope is 0.04717416
## After removing row: 131 Rsquard is  0.6062265 and slope is 0.04698042
## After removing row: 132 Rsquard is  0.6207721 and slope is 0.04810386
## After removing row: 148 Rsquard is  0.6111873 and slope is 0.04708534
## After removing row: 170 Rsquard is  0.617591 and slope is 0.04806516
## After removing row: 176 Rsquard is  0.6083276 and slope is 0.04692413
## After removing row: 179 Rsquard is  0.6247303 and slope is 0.04829031
## After removing row: 184 Rsquard is  0.6067362 and slope is 0.04700009
## After removing row: 199 Rsquard is  0.6069406 and slope is 0.04706408
```

```
cat("while origial r-squared is",summary(simple.tv.ad)$r.squared,"and slope",summary(simple.tv.ad)$coef
```

```
## while origial r-squared is 0.6118751 and slope 0.04753664
```

After finding potential influential points and then removing each points and building model to check influence on model.it is found that none of the points have high effect. So, none of them are actual influential points for our simple regression model. The r-squared value and slope of regression line after removing each potential point show there is not much change on those value before and after removing it.

Now calculating overall effect of all variables and then checking how each independent variable are associated with the response variable using multiple linear regression.

Building Multiple linear regression model:

```
set.seed(1234)
multi.linear.model <- lm(data = project.df,`sales(M)`~`tv_ad(K)`+`radio_ad(K)`+`newspaper_ad(K)`)
summary(multi.linear.model)
```

```
##
## Call:
## lm(formula = `sales(M)` ~ `tv_ad(K)` + `radio_ad(K)` + `newspaper_ad(K)`,
##     data = project.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.938889   0.311908   9.422   <2e-16 ***
## `tv_ad(K)`         0.045765   0.001395  32.809   <2e-16 ***
## `radio_ad(K)`      0.188530   0.008611  21.893   <2e-16 ***
## `newspaper_ad(K)` -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## 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
```

```r
anova(multi.linear.model)
```

```
## Analysis of Variance Table
##
## Response: sales(M)
##                      Df Sum Sq Mean Sq   F value Pr(>F)
## `tv_ad(K)`            1 3314.6  3314.6 1166.7308 <2e-16 ***
## `radio_ad(K)`         1 1545.6  1545.6  544.0501 <2e-16 ***
## `newspaper_ad(K)`     1    0.1     0.1    0.0312 0.8599
## Residuals           196  556.8     2.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now I am going to formally test (using the 5-step procedure) whether the set of these predictors are associated with sales at the = 0.05 level.The values from ANOVA and summary Table will be using to test the hypothesis.

Setting up the hypotheses 1.Selecting the alpha level and setting up the hypothesis.

- H0:Beta(tv_ad)=Beta(radio_ad)=Beta(newspaper_ad)=0 (tv_ad,radio_ad and newspaper_ad are not significant predictors of sales)
- H1:Beta(tv_ad) $\neq$ 0 and/or Beta(radio_ad) $\neq$ 0 and/or Beta(newspaper_ad) $\neq$ 0 (at least one of the slope coefficients is different than 0 tv_ad and/or radio_ad and/or newspaper_Ad are significant predictors/is a significant predictor of sales).
- $\alpha = 0.05$

2.Selecting the appropriate test statistics

- Selecting global F-stat from summary table with df1=3 and df2=196. Alternatively we can calculate the Res SS and Reg SS and calculate F-value.

3. Stating Decision rule:

-Determining the appropriate value from the F-distribution with 3 and 200-3-1=196 degrees of freedom and associated with a right hand tail probability of alpha= 0.05(F-distribution has only one-sided curve) -Using software Critical value is:

```r
qf(0.05,3,196,lower.tail = FALSE)
```

```
## [1] 2.650677
```

- Decision Rule : Reject H0 if F >=2.65
- Otherwise, do not reject H0

4. Compute the test statistics

it is found from above table that global F-statistics 553.5 with df1=3 and df2=196 and p_value <0.001.

5.Conclusion:

Since F-Statistics for set of predictors tv_ad,radio_ad and newspaper_ad are greater than critical value 2.65 and similarly p-values of these are way less than alpha value,we reject null hypothesis.This means at least one of the slopes coefficients is different than 0.Also,This can be interpreted at the alpha =0.05 level that tv_ad,radio_ad and newspaper_ad when taken together are significant predictors of sales.

Now checking association of each variable with dependent variable.Here, F-stat and p-values from summary table is obtained using following formula or can be obtained from table itself as follows.

From ANOVA table:

- F(tv_ad)= MS Reg(tv_ad)/MS Res =3314.6/2.8 = 1166.7308 i.e greater than critical value 2.699.Similarly p-value <0.001 -> Significant at level 0.05.

-F(radio_ad)= MS Reg(radio_ad)/MS Res=1545.6 /2.8 = 544.0501 i.e greater than critical value 2.699.Similarly p-value <0.001 -> Significant at level 0.05.

- F(newspaper_ad) = MS Reg(newspaper_ad)/MS Res = 0.1 /2.8 = 0.0312 i.e less than critical value 2.699.Similarly p-value=0.8599 >0.05 -> Not significant at level 0.05.

From above calculation, it can be interpreted as beta(tv_ad) and beta(radio_ad) are significant at alpha = 0.05. From above summary table beta(tv_ad) = 0.045765 and beta(radio_ad)=0.188530. Beta(tv_ad)=0.045765 can be interpreted as for each additional thousand dollars spent in tv_ad, sales is increase by an average of 0.045765 Million ,keeping other variables constant.Similarly, for each additional thousand dollars spent in radio_ad, sales is increase by 0.188530 million keeping other variables constant.Also, we have intercept of 2.938889 which can be interpreted as base sales of the company without any advertisements.Also, r-squared value is high i.e 0.8972 which can be interpreted as 89.72% of variability in response variable is explained by the independent variables.

```r
#Finding influential points
influential.points.multiple <- cooks.distance(multi.linear.model) >4/(nrow(project.df)-2)
which(influential.points.multiple)
```

```
##   3   6  26  36  76  79 127 129 131 132 133 136 159 166 170 179
##   3   6  26  36  76  79 127 129 131 132 133 136 159 166 170 179
```

```r
set.seed(123)
for (x in which(influential.points.multiple)) {
 multiple.data <- project.df #copying main data
 multiple.data <- multiple.data[-x,] #removing the row with potential influential point
 #building model after removing the potential influential point
 multiple.model1 <- lm(data = multiple.data,`sales(M)`~`tv_ad(K)`+`radio_ad(K)`+multiple.data$`newspape


 #to get r-squared from model
 rsquared1 <- summary(multiple.model1)$r.squared
 cat("After removing row:",x,"Rsquard is ",rsquared1,"\n")
}
```

```
## After removing row: 3 Rsquard is  0.8985289
## After removing row: 6 Rsquard is  0.9017618
## After removing row: 26 Rsquard is  0.8995997
## After removing row: 36 Rsquard is  0.9005542
## After removing row: 76 Rsquard is  0.8986306
## After removing row: 79 Rsquard is  0.8981135
## After removing row: 127 Rsquard is  0.8991809
## After removing row: 129 Rsquard is  0.896168
## After removing row: 131 Rsquard is  0.9095582
## After removing row: 132 Rsquard is  0.8987553
## After removing row: 133 Rsquard is  0.8973282
## After removing row: 136 Rsquard is  0.8982076
## After removing row: 159 Rsquard is  0.898155
## After removing row: 166 Rsquard is  0.8981964
## After removing row: 170 Rsquard is  0.8988337
## After removing row: 179 Rsquard is  0.9004865
```

```r
cat("while origial r-squared is",summary(multi.linear.model)$r.squared)
```

```
## while origial r-squared is 0.8972106
```

There were 16 potential influential points but none of them seems to have high impact on the model individually.Also,outliers seen in our boxplot(newspaper_ad) above have not that big impact on the model.

Calculating 95% confidence intervals for significant variables. Since tv_ad and radio_ad variables are significant at level 0.05 predicting the sales,I will be calculating 95% confidence interval for the slopes for those variables.

```r
confi.intv1 <- data.frame(confint(multi.linear.model,level = 0.95))[2:3,]
colnames(confi.intv1) <- c("Beta_Lower_value","Beta_Upper_value")
kable(confi.intv1,digits = 4,align = "c")
```

	Beta_Lower_value	Beta_Upper_value
‘tv_ad(K)‘	0.0430	0.0485
‘radio_ad(K)‘	0.1715	0.2055

Results:

1.Simple Linear Regression:

TV Advertising and Sales: There is a significant positive relationship between TV advertising spending and sales (Beta = 0.0474, p < 0.001). For every additional thousand dollars spent on TV advertising, sales increase by an average of $47,381. The model explains approximately 61.19% of the variability in sales (R-squared = 0.6119).

2.Multiple Linear Regression:

Combined Advertising Budgets and Sales: Considering TV, radio, and newspaper advertising budgets together, the set of predictors significantly predict sales (F = 570.3, p < 0.001). Both TV (Beta = 0.0458, p < 0.001) and radio (Beta = 0.1885, p < 0.001) advertising budgets have significant positive associations with sales.However newspaper_add not statistically significant in predicting sales(p=0.86) The overall model explains approximately 89.72% of the variability in sales (R-squared = 0.8972).

3.Confidence Intervals:

TV and Radio Advertising: The 95% confidence intervals for the slopes of TV and radio advertising,which are statistically significant predictors of sales, are [0.0430, 0.0485] and [0.1715, 0.2055], respectively.

Conclusions and limitations:

Conclusions:

The study offers valuable insights into the dynamics between advertising budgets and sales, revealing a strong and statistically significant relationship. Particularly, the Simple Linear Regression focusing on TV advertising indicates a substantial positive correlation.

Expanding the analysis to Multiple Linear Regression, encompassing TV, radio, and newspaper budgets, reinforces the significance of advertising expenditures. The combined effect of these channels significantly predicts sales, with both TV and radio budgets demonstrating meaningful positive associations. The model's high explanatory power, capturing approximately 89.72% of the variability in sales, highlights the collective impact of diverse advertising mediums on overall sales performance.

Limitations:

However, the robustness of these conclusions is contingent on several limitations that warrant consideration. Firstly, assumptions regarding linearity and model validity should be rigorously examined to ensure the reliability of the observed relationships. Any deviations from these assumptions could compromise the accuracy and generalizability of the results.

Secondly, the generalizability of the findings beyond the confines of the historical dataset may be limited. Market dynamics, consumer behavior, and economic conditions are subject to change, necessitating caution when applying the model to different time periods or economic contexts.

Furthermore, the model's exclusion of certain influential variables introduces potential limitations. Factors not considered in the analysis, such as online advertising or broader economic indicators, may play a role in influencing sales and should be acknowledged in a comprehensive understanding of the advertising-sales relationship.

Also, Though I did check whether all assumptions are met for linear regression and it did meet,I did not perform activities regarding checking whether all assumptions are met or not for multiple linear regression. We need to make sure all assumptions of regression are met before concluding and implementing results from the research project.

```
**The End**
```