Marketing Data Regression Analysis

Woo Taek Kim

November 2, 2023

1 Introduction

This project is to do regression analysis on marketing data assuming I am making recommendation to the stakeholder. Yet, I will still be presenting the analysis process in detail along with code.

The data is from kaggle.
Link to the data is shown below:
Marketing data

2 Exploratory Data Analysis

Before getting straight into modeling, I started with exploring data. First I looked at basic statistics of the data. Below is the code and output from the code.

```
marketing_df <- read.csv("/Users/wootaekkim/Desktop/youtube
learning/marketing_sales_data.csv")

library(skimr)
skim(marketing_df)
```

```
> skim(marketing_df)
── Data Summary ────────────────────────
                           Values
Name                       marketing_df
Number of rows             572
Number of columns          5
_______________________
Column type frequency:
  character                2
  numeric                  3
________________________
Group variables            None

── Variable type: character ──────────────────────────────────────────────────
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 TV                    0             1   3   6     0        3          0
2 Influencer            0             1   4   5     0        4          0

── Variable type: numeric ────────────────────────────────────────────────────
  skim_variable n_missing complete_rate   mean    sd        p0    p25    p50    p75  p100 hist
1 Radio                 0             1  17.5   9.29 0.109       10.7   17.1   24.6  42.3
2 Social.Media          0             1   3.33  2.24 0.0000313   1.59   3.15   4.73  11.4
3 Sales                 0             1 189.   89.9 33.5        119.   184.   265.   358.
```

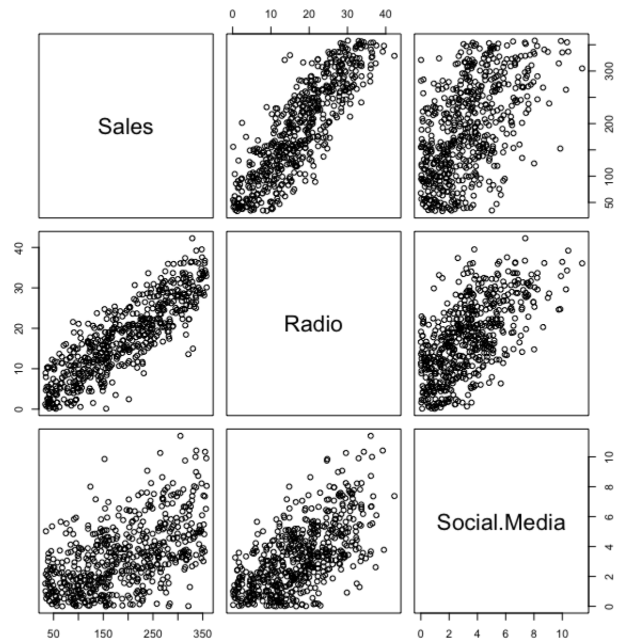
From above we can see that there are total of five variables of which 2 are categorical and 3 are numeric. Below is description of each variable from the source of the data.

Features
- TV promotional budget (in "Low," "Medium," and "High" categories)
- Social media promotional budget (in millions of dollars)
- Radio promotional budget (in millions of dollars)
- Sales (in millions of dollars)
- Influencer size (in "Mega," "Macro," "Micro," and "Nano" categories)

I had Sale as our target variables and remaining variables as predictor variabales. Our recommendations would be on how stakeholders should be allocating budget to maximize sales.

From the output of the code, we can also see that there is no missing value so that we do not need to deal with any missing values. Now we are going to look at the correlation between numerical variables. Below is the code and output from the code.

```
# check correlation coefficients
cor(subset(marketing_df, select = -c(TV,Influencer)))
attach(marketing_df)
pairs(Sales ~ Radio + Social.Media)

> cor(subset(marketing_df, select = -c(TV,Influencer)))
                 Radio Social.Media     Sales
Radio        1.0000000    0.6299406 0.8580363
Social.Media 0.6299406    1.0000000 0.5420478
Sales        0.8580363    0.5420478 1.0000000
```



From above, we can observe strong positive correlation between Sales and Radio budget and weak correlation between Sales and Social Media budget. Moreover. We can also see moderate correlation between Radio budget and Social Media budget. As we have Sales as out target variable and remaining variables as predictor variables, we might need to consider excluding one of them when modeling.

3 Modeling

After taking a look at basic features of the data, now we move on to modeling. Here we are going to use Linear Regression which also enables us better analyze relationship between Sales and other features.
In the beginning, I started with full model (including all the features.
Below is the code and output from the code:

```
m1 <- lm(Sales ~ Radio + Social.Media + TV + Influencer)
summary(m1)
```

```
> m1 <- lm(Sales ~ Radio + Social.Media + TV + Influencer)
> summary(m1)

Call:
lm(formula = Sales ~ Radio + Social.Media + TV + Influencer)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-61.570 -22.175   0.103  21.995  68.043

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      217.4784     6.5840  33.031   <2e-16 ***
Radio              2.9735     0.2352  12.644   <2e-16 ***
Social.Media      -0.1391     0.6761  -0.206    0.837
TVLow           -154.5736     4.9494 -31.231   <2e-16 ***
TVMedium         -75.5947     3.6473 -20.726   <2e-16 ***
InfluencerMega     2.4948     3.4620   0.721    0.471
InfluencerMicro    2.9391     3.3777   0.870    0.385
InfluencerNano     0.8015     3.3457   0.240    0.811
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.99 on 564 degrees of freedom
Multiple R-squared:  0.9042,Adjusted R-squared:  0.903
F-statistic: 760.4 on 7 and 564 DF,  p-value: < 2.2e-16
```

First thing we c notice is result of ANOVA test for utility of the model which is shown at the bottom. As p-value asscociated with F-value for ANOVA testing utility of the model is much less than 0.05, we can confidently say that this model is useful for predicting Sales.

Notice how p-values of associated with coefficients of Social Media and Influencer are bigger than 0.05. This implies that p-value associated with t-statistic for testing null hypothesis ,that says coefficients are zero, is greater than 0.05. Accordingly we cannot reject null hypothesis for coefficient of variables Social Media and Inlfuencer so that these coefficients are statistically insignificant.
Furthermore, as there were some correlation between Social Media and Radio, it is reasonable to remove Social Media variable.
Now try with variable Social Media removed.

```
m2 <- lm(Sales ~ Radio + TV + Influencer )
summary(m2)
```

```
> m2 <- lm(Sales ~ Radio + TV + Influencer )
> summary(m2)

Call:
lm(formula = Sales ~ Radio + TV + Influencer)

Residuals:
```

```
   Min     1Q Median     3Q    Max
-62.34 -22.18   0.22  21.81  67.78

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      217.3716     6.5579  33.146   <2e-16 ***
Radio              2.9529     0.2125  13.893   <2e-16 ***
TVLow           -154.5614     4.9449 -31.257   <2e-16 ***
TVMedium         -75.5800     3.6435 -20.744   <2e-16 ***
InfluencerMega     2.4652     3.4561   0.713    0.476
InfluencerMicro    2.9616     3.3731   0.878    0.380
InfluencerNano     0.7880     3.3422   0.236    0.814
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.97 on 565 degrees of freedom
Multiple R-squared:  0.9042,Adjusted R-squared:  0.9032
F-statistic: 888.6 on 6 and 565 DF,  p-value: < 2.2e-16
```

Compared to previous model we seem minimal increase in Adjusted R-squared value of 0.0002. With Social Media variable eliminated, we still have Influencer variables coefficients statistically insignificant. So, now try fitting model with both Influencer variables and Social Media variable.

```
m3 <- lm(Sales ~ Radio + TV)
summary(m3)
```

```
> m3 <- lm(Sales ~ Radio + TV)
> summary(m3)

Call:
lm(formula = Sales ~ Radio + TV)

Residuals:
    Min      1Q  Median      3Q     Max
-64.107 -21.985   0.677  21.878  67.207

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  218.5261     6.2612   34.90   <2e-16 ***
Radio          2.9669     0.2117   14.02   <2e-16 ***
TVLow       -154.2971     4.9291  -31.30   <2e-16 ***
TVMedium     -75.3120     3.6243  -20.78   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.92 on 568 degrees of freedom
```

```
Multiple R-squared:  0.904,Adjusted R-squared:  0.9035
F-statistic:  1783 on 3 and 568 DF,  p-value: < 2.2e-16
```

We can see that all the remaining coefficients of predictor variables are statistically significant. We also have minimally higher Adjusted R-squared compared to bothe first and second model.
Further confirm that this last model is better than last two models using partial ANOVA test.

```
anova(m2, m1)
anova(m3, m2)
```

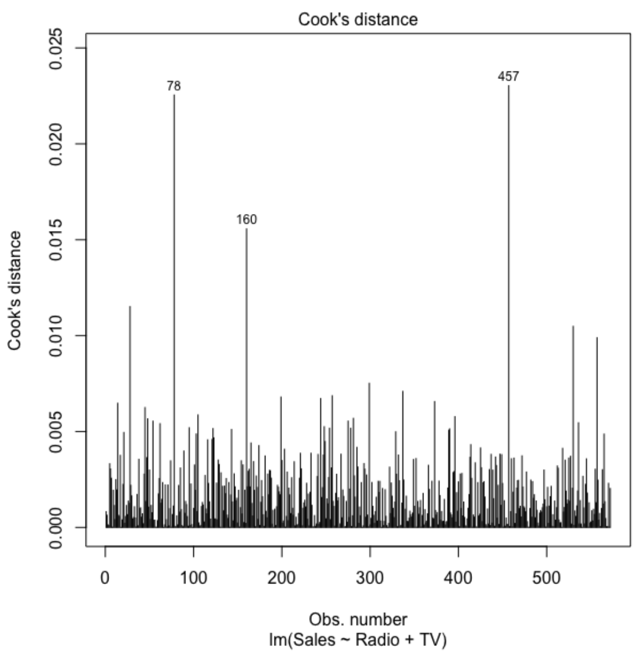
```
> anova(m2, m1)
Analysis of Variance Table

Model 1: Sales ~ Radio + TV + Influencer
Model 2: Sales ~ Radio + Social.Media + TV + Influencer
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    565 441899
2    564 441866  1    33.143 0.0423 0.8371
> # this implies that reduced model is better than full model
> anova(m3, m2)
Analysis of Variance Table

Model 1: Sales ~ Radio + TV
Model 2: Sales ~ Radio + TV + Influencer
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    568 442705
2    565 441899  3    805.62 0.3433  0.794
```

From first ANOVA test comparing between first and second model, notice how p-value associated with F-statistics is much larger than 0.05. According we fail to reject null hypothesis which says coefficient of Social Media is 0. Therefore, we can conclude that second model is better then first (full) model.

Then look at the ANOVA test comparing between second and third model. From the result of the test, we can see that p-value associated with F-statistics is much higher than 0.05. Therefore, we can conclude that further reduced model is better than second model.

Before finalizing model check for any leaverage points using Cook's distance plot

```
plot(m3, which=4)
```



From above we can see that none of the points are over 0.5. Accordingly, we can say that none of the point are influential enough to say that they are leverage point.

We have finalized the model. Following is the model equation for our final model

$$Sales = 218.53 + 2.97Radio + (-154.30)TV_{Low} + (-75.31)TV_{Medium} + \epsilon \quad (1)$$

Now check assumptions of linear regression models.

4 Model Assumption Check

Mulicolinearity

First check mulitcolinearity. We already have looked at correlation between numeric predictor variables and have eliminated Social Media variable. However, check once more use variance inflation factor. Following are code and output from the code.

```
library(car)
vif(m3)
```

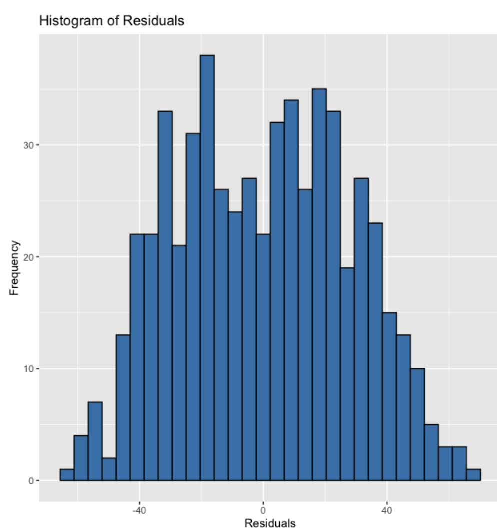
```
> vif(m3)
          GVIF Df GVIF^(1/(2*Df))
Radio 2.834074  1        1.683471
TV    2.834074  2        1.297486
```

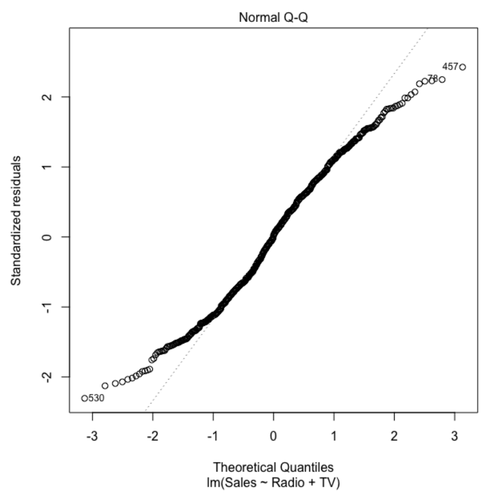
Notice how none of the variance inflation factors are greater than 5. Consequently, we can conclude that there is no mulitcolinearity issue with our model.

Normality of residual

Below are the code and output from the code for checking normality of residuals.

```
plot(m3, which=2)
ggplot(data = marketing_df, aes(x = m3$residuals)) +
  geom_histogram(fill = 'steelblue', color = 'black') +
  labs(title = 'Histogram of Residuals', x = 'Residuals', y = 'Frequency')
```



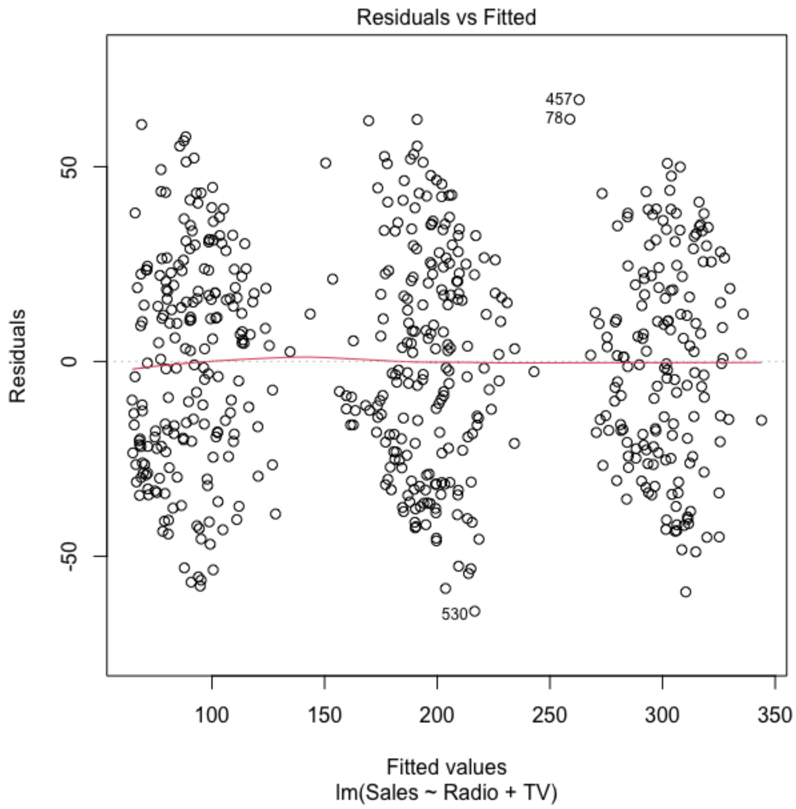


From above histogram of residuals, we can see that it roughly follows normal distribution. Moreover, as dots generally lives close to the $y = x$ line, it further support that residuals follow normal distribution.

Constant Variance

We can check constant vairance assumption using residual vs fitted values plot. Code and plot are shown below:

```
plot(m3, which=1)
```



From above, we can see random scatter around the mean 0 forming three scatters. Furthermore, we can also see constant variability around the mean 0. Accordingly we can confirm that we have residual with constant variance and TV has strong correlation with Sales as TV variables has three categories.

5 Interpretation of variables of Models

Below is the summary of our final model from above.

```
Call:
lm(formula = Sales ~ Radio + TV)

Residuals:
    Min      1Q  Median      3Q     Max 
-64.107 -21.985   0.677  21.878  67.207 

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  218.5261     6.2612   34.90   <2e-16 ***
Radio          2.9669     0.2117   14.02   <2e-16 ***
TVLow       -154.2971     4.9291  -31.30   <2e-16 ***
TVMedium     -75.3120     3.6243  -20.78   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.92 on 568 degrees of freedom
Multiple R-squared:  0.904,Adjusted R-squared:  0.9035
F-statistic:  1783 on 3 and 568 DF,  p-value: < 2.2e-16
```

Radio
- 95% confidence interval of the coefficient of Radio is $[2.551089, 3.382691]$
- 1 million dollars increase in Radio budget is associated with 2.9669 million dollars increase in sales.

TV
- 95% confidence interval of the coefficient of TVLow is $[-163.9785, -144.6156]$
- Having Low TV budget is associated with 154.2971 million dollars decrease in sales compared to having High TV budget and 78.9851 million dollars decrease in sales compared to having Medium TV budget.
- 95% confidence interval of the coefficient of TVMedium is $[-82.43062, -68.19335]$
- Having Medium TV budget is associated with 75.3120 million dollars decrease in sales compared to having High TV budget and 78.9851 million dollars increase in sales compared to having Low TV budget.

From above strong correlation with the Sales and huge influential to Sale, we could have given much more accurate and better analysis it we had numeric TV budget.

6 Recommendations based on Analysis

From above Analysis we can confidently say that it is best of company's interest to maintain High TV budget as reducing to Medium TV budget causes $75.3120 Million (CI 95% [68.19335,82.43062]) loss in sales and reducing to Low TV budget causes $154.2971 Million (CI 95% [144.6156,163.9785]) in sales compared to maintaining High TV budget.

Having decent amount of Radio marketing budget as increasing $ 1 Million is associated $2.9669 Million (CI 95% [2.551089,3.382691]) increase in sales.

After all, it is recommended to focus most having High proportion of TV marketing budget and allocate some budget to Radio when possible