Exploring the Impact of Advertising on Sales: A Multiple Linear Regression Analysis

Ismah Ahmed

December 12, 2024

Introduction

Examining the Association of Advertising Spending on Sales

Does spending on different advertisement types have an effect on sales? Companies invest in different types of advertising, such as TV, radio, and newspapers advertisement in order to increase sales. This project examines whether spending on these advertisements are associated with a change in sales and identifies which type of spending is most effective. In this project, we will explore advertising data and use it to calculate a least squares regression equation that predicts Sales (in thousands) based on advertising spending on TV, newspapers, and radio. We will formally test whether the set of these predictors is associated with total sales at the $\alpha = 0.05$ significance level. Furthermore, we will analyze the significance of the model by summarizing the contribution of each type of advertising separately, again at the $\alpha = 0.05$ significance level.

Data Origin and Overview

The dataset used in this project, titled *Advertising Spend vs. Sales*, originates from Kaggle. Click *here* to be directed to the kaggle dataset. The dataset contains the following 4 numerical variables (all in thousands of dollars):

- **TV**: Total Spent on TV advertisements
- **Radio**: Total Spent on radio advertisements
- **Newspaper**: Total Spent on newspaper advertisements
- **Sales**: Total sales

Below, you will find the first few rows of the dataset:

Table 1: Top 5 rows from Advertising data

TV	radio	newspaper	sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

Data Cleaning and Examining

Firstly, we want to check if there are any missing values in our dataset. After running the following code, it appears there are no missing data.

```
colSums(is.na(data))
```

```
       TV     radio newspaper     sales 
        0         0         0         0 
```

Below, we are confirming the structure of the data set, making sure that the variables are all numeric.

```
str(data)
```

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

Before starting the multiple linear regression (MLR) model, some preliminary steps are required. Below are the summary statistics for the variables available in the dataset. TV has the highest mean spending among the advertisement categories, while Radio shows the smallest mean. The variation in TV spending is much wider compared to Radio and Newspaper, which have narrower ranges of values.

Variable	Mean	SD	Min	Q1	Q3	Max
TV	147.0425	85.854236	0.7	74.375	218.825	296.4
Radio	23.2640	14.846809	0.0	9.975	36.525	49.6
Newspaper	30.5540	21.778621	0.3	12.750	45.100	114.0
Sales	14.0225	5.217457	1.6	10.375	17.400	27.0

Table 2: Summary Statistics for Advertising Dataset

Next, let's examine the distribution of spending across our variables, including the response variable Sales. The distribution of TV and Radio spending appears to be relatively uniform, while newspaper shows a right skew, indicating that higher spending on newspaper ads is less common but can be high. This might impact our model as outliers or extreme values can disproportionately affect the results.

```
par(mfrow = c(2, 2), mar = c(4, 4, 2, 1))

hist(data$TV, main = "Distribution of TV Ad Costs", xlab = "Spending", col = "lightblue1")
hist(data$radio, main = "Distribution of Radio Ad Costs", xlab = "Spending", col = "lightblue1")
hist(data$newspaper, main = "Distribution of Newspaper Ad Costs", xlab = "Spending", col = "lightblue1"
hist(data$sales, main = "Distribution of Sales", xlab = "Sales", col = "lightblue1")
```

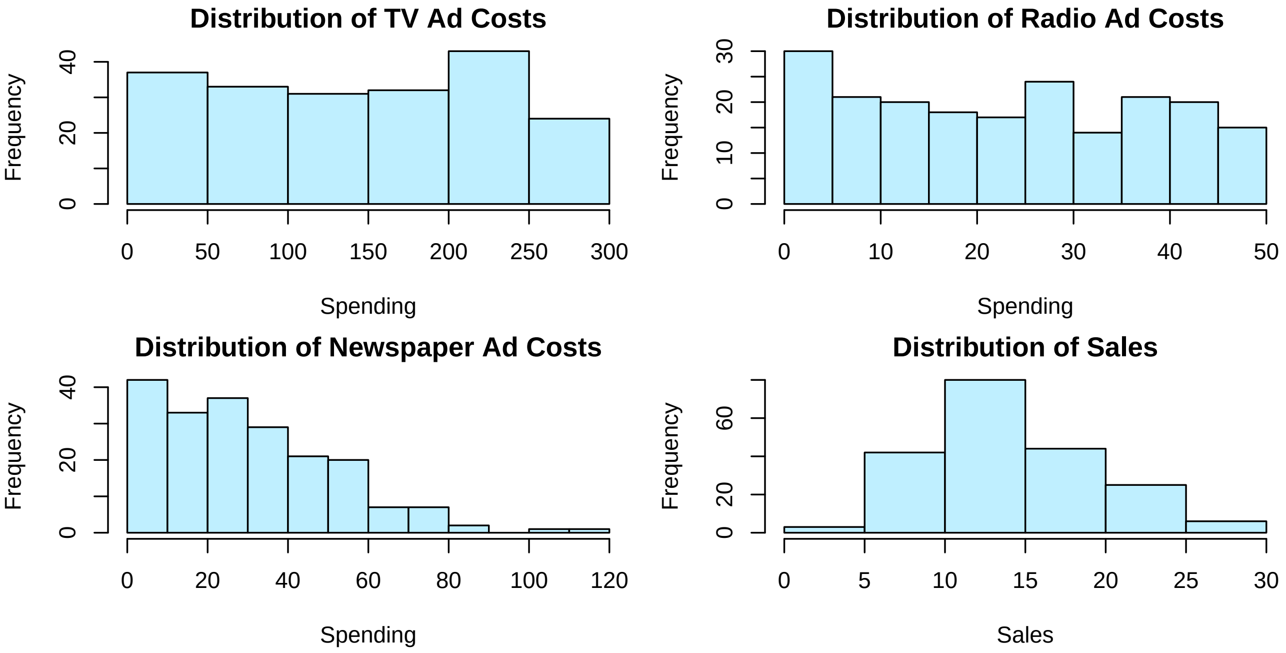


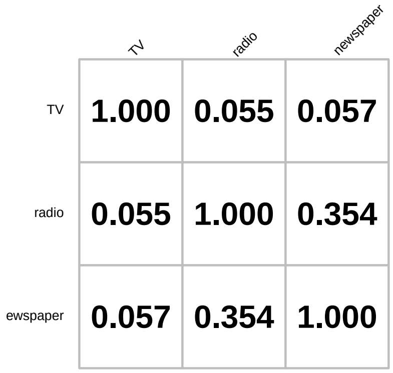
Figure 1: Distribution of TV, Radio, and Newspaper Spending and Total Sales in Thousands

The correlation matrix (labeled Figure 2) provides a measure of the linear relationship between pairs of variables.

- TV and Radio: Correlation is 0.055, very weak positive relationship. suggest no significant linear relationship
- TV and Newspaper: 0.057, very weak positive relationship, suggest no significant linear relationship
- Radio and Newspaper is 0.354, moderate positive relationship

These weak correlations suggest that multicollinearity is not a major concern in our analysis. Since the predictor variables (TV, Radio, and Newspaper) show very weak correlations with each other, we can proceed with the multiple linear regression model without the need for further adjustments.

Next, lets take a look at boxplots for each of the predictor variables



	TV	radio	newspaper
TV	1.000	0.055	0.057
radio	0.055	1.000	0.354
ewspaper	0.057	0.354	1.000

Figure 2: Correlation Matrix of TV, Radio, and Newspaper Spending

```
par(mfrow = c(1, 3))

boxplot(data$TV, main = "TV Advertising Costs", cex.main = 0.8)
boxplot(data$radio, main = "Radio Advertising Costs", cex.main = 0.8)
boxplot(data$newspaper, main = "Newspaper Advertising Costs", cex.main = 0.8)


par(mfrow = c(1, 1))
```

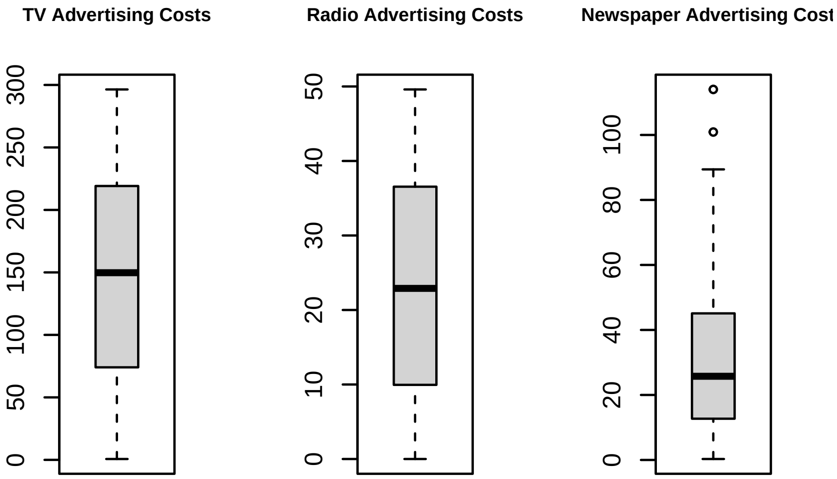


Figure 3: Boxplot of each predictor variable

At first glance, there appears to be 2 outliers in Newspaper Advertising however, after further investigation, I don't have any reason to believe these 2 data points were a mistake, therefore, I will be leaving them in.

Statistical Methods

This project utilizes several statistical methods to analyze the relationship between advertising variables and sales:

- Multiple Linear Regression (MLR): We will use multiple linear regression to calculate a least squares regression equation that predicts **Sales** based on advertising spending on TV, newspapers, and radio.
- F-Test: The F-test for multiple linear regression will serve as our decision rule to determine whether the predictors collectively explain variance in sales.
- Residual Analysis: A residual plot will be generated, showing the fitted values from the regression against the residuals. This will help assess the model's fit and detect any potential outliers.
- R^2: Represents the proportion (percentage) of the variation in the response variable, Sales, that is explained by the multiple regression model.

Performing our MLR

Step 1: Set up the hypotheses and select the alpha level

- H_0: $\beta_{\text{TV}} = \beta_{\text{radio}} = \beta_{\text{newspaper}} = 0$ *(These variables are not significant predictors of Sales)*
- H_1: At least one of $\beta_{\text{TV}}, \beta_{\text{radio}}, \beta_{\text{newspaper}} \neq 0$. *(At least one of these variables is a significant predictor of Sales)*
- $\alpha = 0.05$

Step 2: Select the appropriate test statistic

$F = \frac{\text{MS Red}}{\text{MS Res}}$, df = 3, n-k-1

Step 3: State the decision rule

- Using R, get the appropriate value from the F-distribution with 3, n - k - 1 = 200 - 3 - 1 = 196 degrees of freedom and associated with a right hand tail probability of $\alpha = 0.05$

```
qf(0.95, df1 = 3, df2 = 196)
```

```
[1] 2.650677
```

- Decision Rule: Reject H_0 if $F \geq 2.650677$
- Otherwise, do not reject H_0

Step 4: Compute the test statistic

```
mlr <- lm(data$sales ~ data$TV + data$radio + data$newspaper)
summary(mlr)$fstatistic[1]
```

```
   value
570.2707
```

Step 5: Conslusion

Reject H_0 since 570.2702 is greater than 2.650677. We have significant evidence at $\alpha = 0.05$ level that TV, radio and newspaper advertising spending when taken together are significant predictors of Sales. That is, there is evidence of a linear association between sales and TV, radio and newspaper advertising spending. The model reports a very small p value indicating that the overall model is highly statistically significant. This means that at least one of the predictors is significantly associated with Sales.

Lets take a closer look at the model output:

```
summary(mlr)
```

```

Call:
lm(formula = data$sales ~ data$TV + data$radio + data$newspaper)

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 ***
data$TV         0.045765   0.001395  32.809   <2e-16 ***
data$radio      0.188530   0.008611  21.893   <2e-16 ***
data$newspaper -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
```

- Intercept: The estimated Sales when all predictors (TV, radio, newspaper) are 0 is 2.94
- F-statistic: 570.2707037, lead us to reject null hypothesis
- Degrees of Freedom: 3 and 196
- P-value: <2.2e-16, very small indicating overall model is statistically significant
- R^2 value is 0.8972. The model explains about 89.72% of the variability in Sales
- Newspaper spending does not appear to contribute significantly, in fact, it has a high p value

Lets take a look at the contribution of each variable seperatly

```
coef(summary(mlr))
```

```
                   Estimate  Std. Error    t value     Pr(>|t|)
(Intercept)     2.938889369 0.311908236  9.4222884 1.267295e-17
data$TV         0.045764645 0.001394897 32.8086244 1.509960e-81
data$radio      0.188530017 0.008611234 21.8934961 1.505339e-54
data$newspaper -0.001037493 0.005871010 -0.1767146 8.599151e-01
```

- TV: The p value for TV is very small making it a significant predictor for Sales. The estimate is 0.0457 meaning that for every additional unit increase spent on TV advertising, sales is expected to go up that amount (in thousands of dollars). This is holding the other predictors constant

- Radio: The p value for Radio is very small also making it a significant predictor for Sales. For every additional unit increase spend on Radio advertising, sales are expected to go by by 0.1885 units (thousands of dollars)
- Newspaper: The p-value for newspaper is 0.8599 which suggests that spending money on newspaper ads does not have a significant association on Sales when TV and Radio predictors are included.

Below, we will generate a residual plot showing the fitted values from the regression against the residuals to determine if the fit of the model is reasonable.

```
residuals_mlr <- resid(mlr)
fitted_mlr <- fitted(mlr)

plot(fitted_mlr, residuals_mlr,
     main = "Fitted Values VS Residuals",
     xlab = "Fitted Values",
     ylab  = "Risiduals",
     pch = 20)
abline(0,0, col = "red")
```

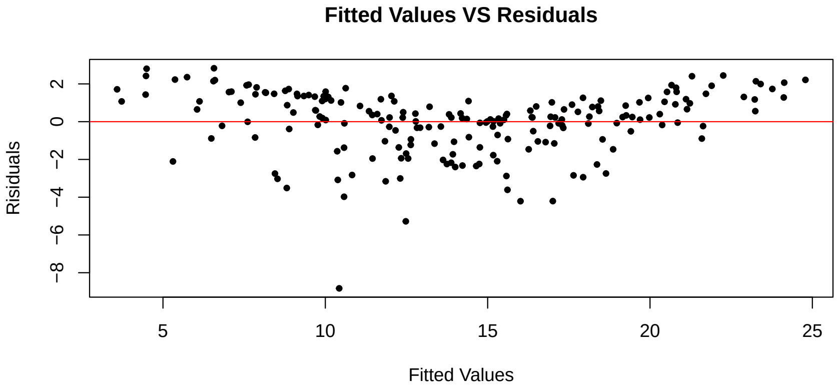


Figure 4: Plotting Residuals

In regression, variability of the response variable should be constant across the regression line and this assumption is checked using a residual plot. It is difficult to be certain just based on initial observation but it appears that there is a slight curve that is shown on the residual plot.

Conclusion

The results of our analysis show that TV and radio advertising are strong predictors of sales, while newspaper advertising does not seem to have a statistically significant impact. In other words, money spent on TV and radio ads is closely associated with total sales, but spending on newspaper ads does not show any significant change. However, after further analysis on residuals, I am more suspicious that the assumptions of the test does not hold.

Limitations

- One assumption of our multiple linear regression analysis is linearity. When we plotted the residuals (differences between predicted and actual sales), there is a slight curved pattern suggesting this assumption may not hold true.
- There is a clear outlier that is shown in the residual plot that may influence the overall model
- Our model only includes TV, radio, and newspaper advertising as predictors, however, there may be other factors (other types of advertising or differing conditions) that are not included but could be associated with sales.