# Simple Regression Analysis

*Minsu Kim*

*10/7/2016*

## Abstract

In this analysis, I explore the statistical relationship between advertising budget and its effectiveness. In particular, I focus on the linear relationship between the increase in the budgets spending on TV advertisement and the increase in the number of items sold by reproducing the results from the book An Introduction to Statistical Learning.

## Introduction

Nowadays, data analytics is often utilized in business sectors to effetively predict or forecast sales and earnings. In this analysis, I closely look at the relationship between advertising and sales, and implement a simple regression model to predict sales from budget spending on TV advertisement. Over the course of analysis, I first reproduce the results from the book An Introduction to Statistical Learning.

## Data

The advertising data set contains sales (in thousands of units) data as well as advertising budgets (in thousands of dollars) of a certain product in 200 different markets. There are three predictors namely TV, Radio and Newspaper as well as one response variable, Sales. All of the predictors are numerical values.

| | V1 | TV | Radio | Newspaper | Sales |
|---|---|---|---|---|---|
| 1 | Min. : 1.00 | Min. : 0.70 | Min. : 0.000 | Min. : 0.30 | Min. : 1.60 |
| 2 | 1st Qu.: 50.75 | 1st Qu.: 74.38 | 1st Qu.: 9.975 | 1st Qu.: 12.75 | 1st Qu.:10.38 |
| 3 | Median :100.50 | Median :149.75 | Median :22.900 | Median : 25.75 | Median :12.90 |
| 4 | Mean :100.50 | Mean :147.04 | Mean :23.264 | Mean : 30.55 | Mean :14.02 |
| 5 | 3rd Qu.:150.25 | 3rd Qu.:218.82 | 3rd Qu.:36.525 | 3rd Qu.: 45.10 | 3rd Qu.:17.40 |
| 6 | Max. :200.00 | Max. :296.40 | Max. :49.600 | Max. :114.00 | Max. :27.00 |

Table 1: Advertising data set shows the summary statistics for each predictor.

Table 1 shows the summary of the data set. The budget spending on TV is the highest on average. This intuitively makes sense because TV advertisement is often very expensive. Also, Newspaper seems to have a highly variance. So, it might be worth investigating what budget spending on newspaper has the high variance. However, in this analysis, I will mainly focus on TV and Sales columns.

## Methodology

In order to investigate the linear relationship between TV and Sales, I use a simple linear regression. It is a very straightforward approach and commonly used model for predicting a quantitative response Y on the basis of a single predictor variable X. It assumes that there is approximately a linear relationship between preditor and target variable. Mathematically, we can write this linear relationship as

$$Sales = \beta_0 + \beta_1 TV$$

In the context of this analysis, X represents TV advertising and Y represents sales. By fitting the model, I can regress sales onto TV.

Before fitting the model, I first look at the general distributions of both TV and Sales.

**Hisogram of TV**

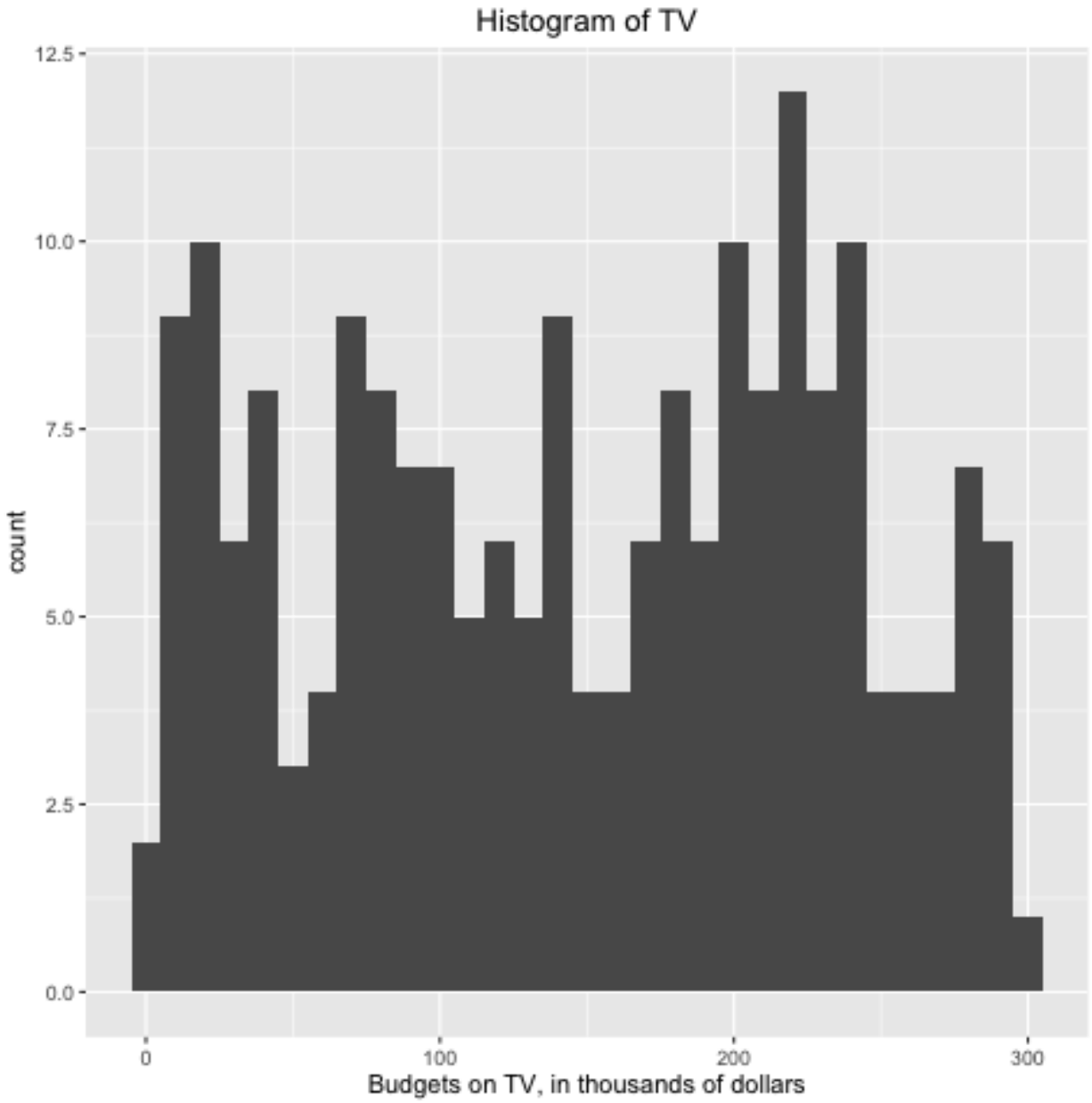


Figure 1. Histogram of Budgts on TV (in thousands of dollors). There is no distinct pattern on budge spending for TV.

**Hisogram of Sales**

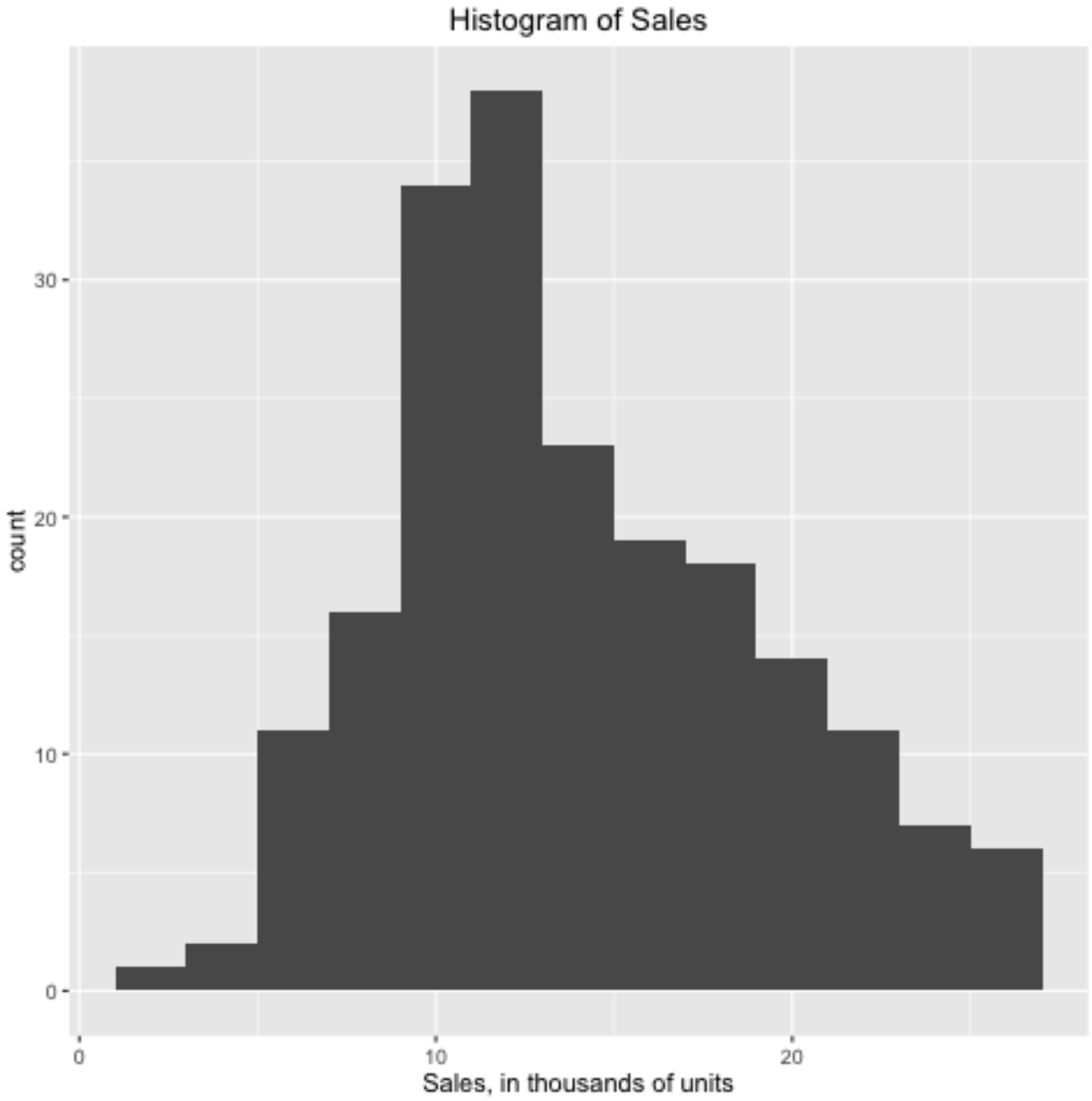


Figure 2. Histogram of Sales (in thousands of dollors). It indicates that sales is concentrated around 11 thousands. The distribution is quite skewed to the right.

## Results

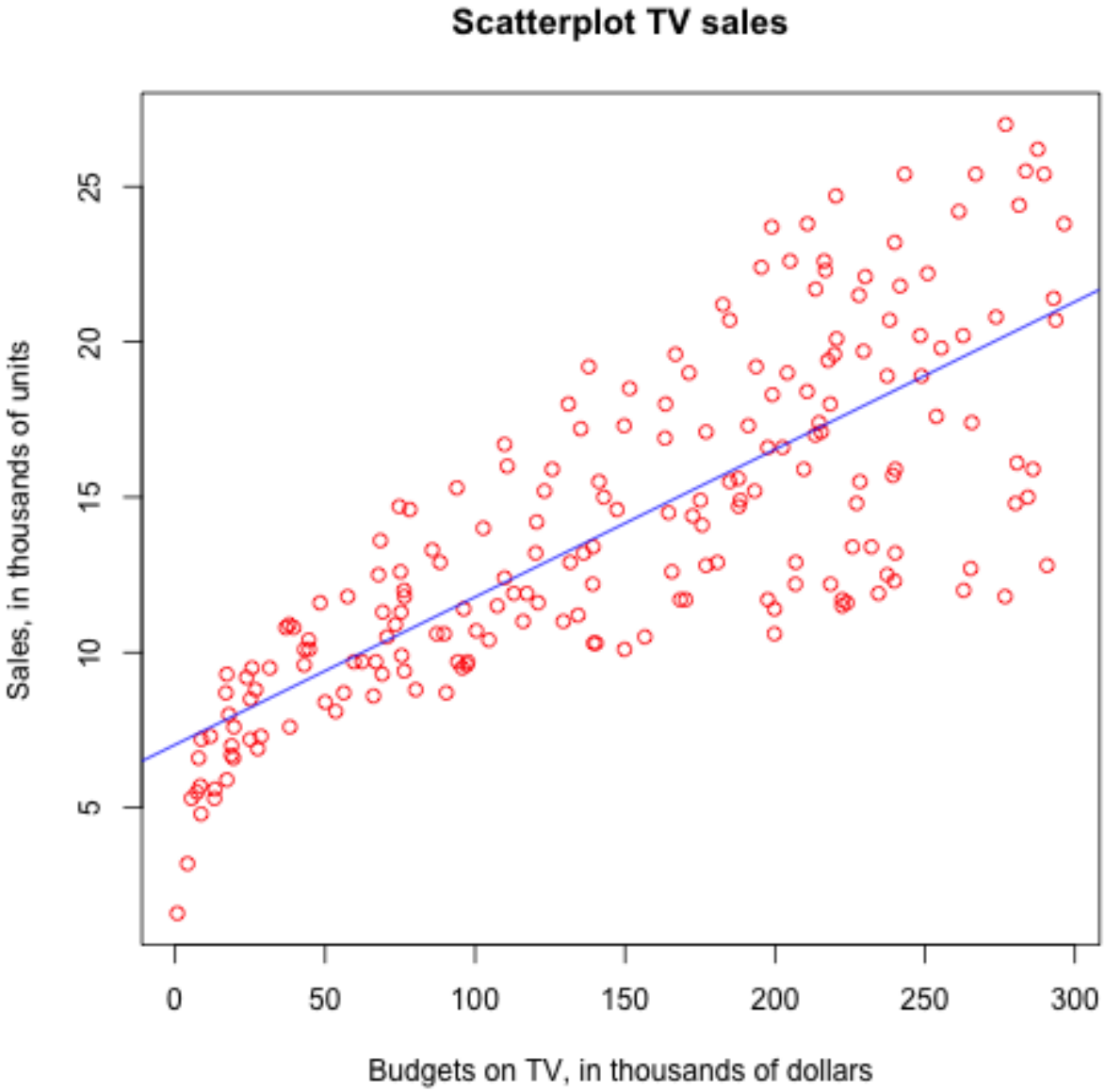


Figure 3. Scatterplot TV sales. This indciates the linear relationship between budget spending on TV and Sales. It also shows that the variane is getting bigger as budget spending on TV increases.

The scatterplot between TV advertising and Sales indicates that there is an approximately linear relationship between predictor and target variable. Since the variances from the regression line to points are getting bigger and bigger, it might suffer from heteroscadasticity. I could try to do a log-log transformation to alleviate this issue. However, overall, this plot indicates that there is a good chance that TV advertising affects sales.

Futhermore, I also look at regression coefficients shown in Table 2. The estimate for the slope (TV) is 0.0475. It means that as one unit of TV adverstising increases, the unit of 0.0475 in Sales increases. It is noteworthy that p value is very low. It implies that TV is a statistically significant variable in this model.

Lastly, I look at regression quality indices in Table 3 to determine the effectiveness of this model. R-squared indicates that this model can explain 61% of variability in this data set. RSS refers to Residual squared sum.

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 7.0326 | 0.4578 | 15.36 | 0.0000 |
| TV | 0.0475 | 0.0027 | 17.67 | 0.0000 |

Table 2: Information about Regression Coefficients. This cofficient table shows that one unit of TV adverstising increases, the unit of 0.0475 in Sales increases.

However, it does not directly provide interpretability because it is a relative measure. If I fit other models and try to compare the effectiveness of each model, I would use both RSS and R-sqaured.

| Quantity | Value |
|---|---|
| RSS | 2102.53 |
| R2 | 0.61 |
| F-stat | 312.14 |

Table 3: Regression Quality Indices. This statistics indicate the quality of the model.

# Conclusions

In this analysis, I explore advertising data set focusing on the relationship between TV advertising and Sales. To start with, I reproduce the results from the book Introduction to Statistical Learning. From the scatterplot, I notice the approximately linear relationship between predictor and target variable. Also, regression coefficient and R-squared indicate that this a simple linear is a quite effective to explain the variance in the data set. As a future work, I would add more predictors to improve accuracy and investigate interaction effect among predictors.