Regression Analysis on the Relationship Between TV Advertising Budgets and Product Sales

Shannon Chang

October 6, 2016

1 Abstract

In this report, I will reproduce the scatterplot and fitted regression line shown in **Figure 3.1** (page 62) of *An Introduction to Statistical Learning* by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. In addition, I will also reproduce the summary regression coefficients shown in **Table 3.1** (page 68) and the quality indices shows in **Table 3.2** (page 69). These results are based on the `Advertising.csv` dataset that is paired with the textbook, and contains data on product sales in over two hundred different markets along with the advertising budgets for the product in each market by different mediums: `TV`, `Radio`, and `Newspaper`.

2 Introduction

Suppose a company wants advice on how to increase sales for one of its products. There is, of course, no concrete way to insure increased sales, but we can influence greater sales through advertising. Imagine that we are statistical consultants hired for this project. To convince the company to invest in advertising campaigns, we must first prove to the company that there is a relationship between advertising and sales. From here, we can then advise the company on appropriate advertising budgets to better reach sales targets. Thus, the goal for this analysis is to determine whether there is a relationship between advertising and sales and, if so, construct an accurate model that can be utilized to predict sales based on the size of advertising budget. For the purposes of this paper, we suppose that the company is only surveying TV advertising; therefore I will only construct analyses for TV advertising budgets and sales. This analyses can, however, be extended and compared to all remaining mediums in the dataset (radio and newspaper).

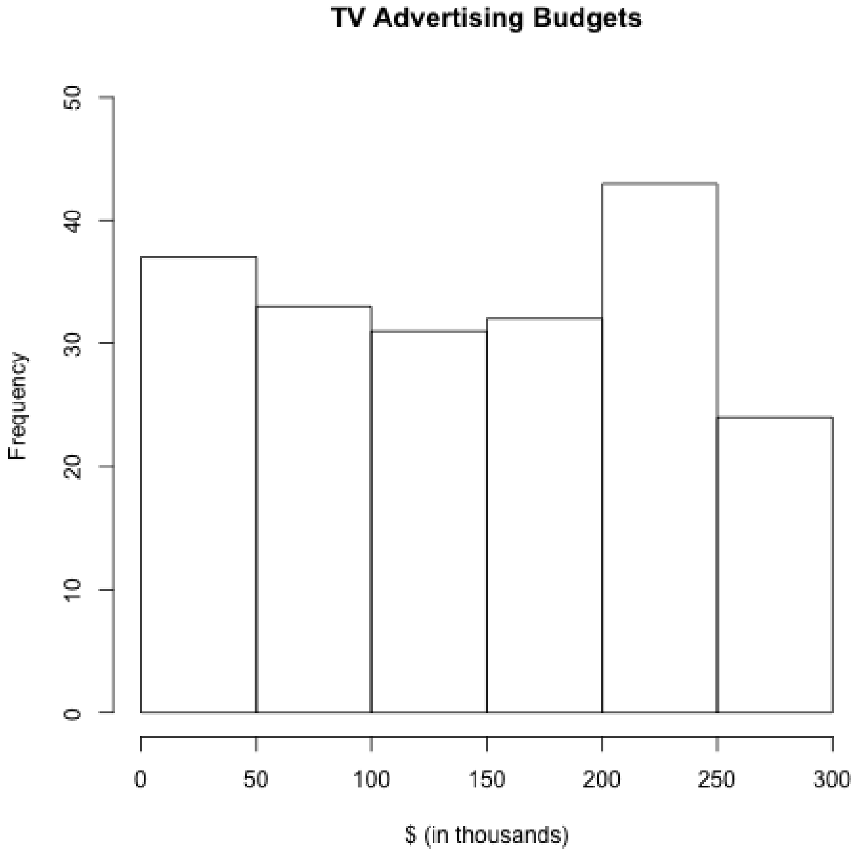

Figure 1: Histogram of TV Advertising Budgets

3 Data

The `Advertising.csv` dataset consists of advertising budgets (in thousands of dollars) by medium: `TV`, `Radio`, and `Newspaper`. Product sales (in thousands of units) are listed under `Sales`. There are 200 rows of data, indicating 200 different markets. I will only be using data in the `TV` and `Sales` columns. Histograms for the two columns are shown below:

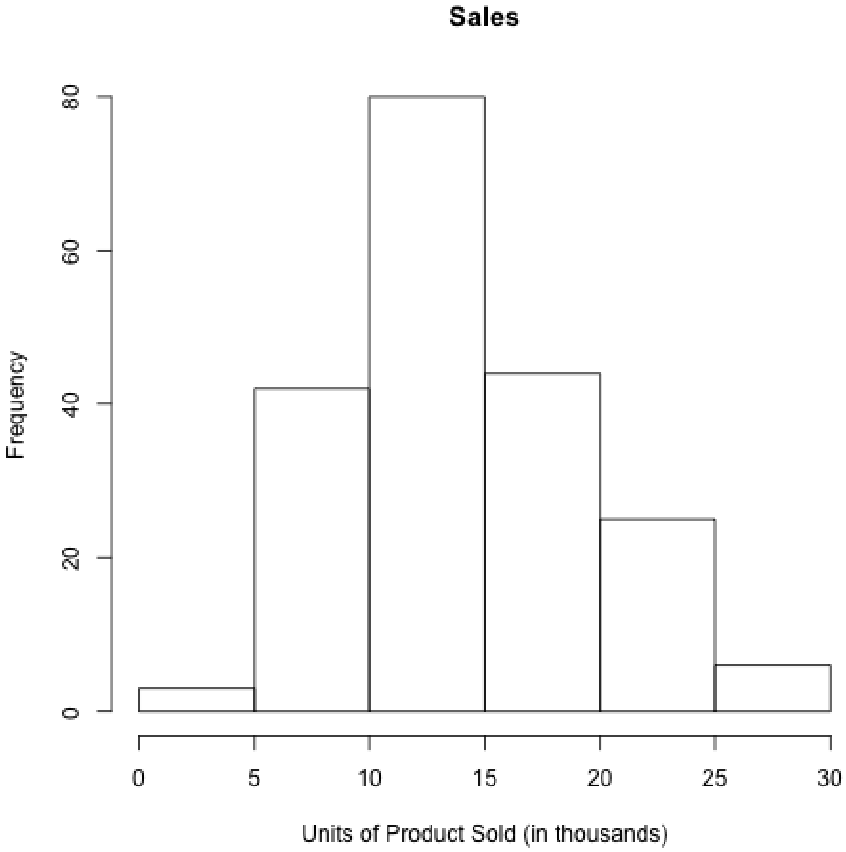

Figure 2: Histogram of Product Sales

4 Methodology

4.1 Setting Up a Model

To examine the association between `TV` and `Sales`, we model their relationship by *simple linear regression.* This method involves predicting a quantitative response Y based on the predictor variable X, assuming that there is a linear relationship between the two. For `TV` and `Sales`, we model their relationship as:

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

Here, β_0 represents the intercept of the linear model while β_1 represents the slope. Since β_0 and β_1 are unknown, we would need to calculate estimates for the two coefficients instead. In a visual manner of speaking, we would want to graph all the data for 'TV' and 'Sales' and fit a line $Sales = \beta_0 + \beta_1 TV$ as close as possible to our 200 data points. We can optimize the fit of this line using the *least squares criterion.* This involves minimizing the sum of squared errors (distance between each data point and its predicted value from the linear model). The line/linear model that we fit would be based on an average of the squares. From a computational perspective, we can start fitting the line by using sample means as estimates for β_0 and β_1, since the average of sample means over a large number of datasets will be very close to the actual/population mean. To evaluate the accuracy of these estimates, we start by calculating standard errors of the standard means. We can then use these standard errors to perform *hypothesis tests* on the estimates. Thus, we would be testing the *null hypothesis* that

$$H_0 : There\, is\, no\, relationship\, between\, TV\, and\, Sales$$

versus the *alternative hypothesis* that

$$H_0 : There\, is\, some\, relationship\, between\, TV\, and\, Sales$$

Numerically, we would be testing

$$H_0 : \beta_1 = 0$$

versus

$$H_0 : \beta_1 \neq 0$$

To do so, we would calculate a *t-statistic*:

$$t = \frac{\hat{\beta_1} - 0}{SE(\hat{\beta_1})}$$

This measures the number of standard deviations that our estimate for β_1 is away from 0. From this, we can calculate a *p-value*, which is the probability of observing any value greater than or equal to t. A small p-value would indicate that it is unlikely to observe a meaningful association between the predictor (`TV`) and the response (`Sales`) purely by chance without some true relationship between the two. Thus, a small p-value would allow us to *reject the null hypothesis* and determine that there is a relationship between `TV` and `Sales`. In general, 5% or 1% are used as p-value benchmarks.

4.2 Evaluating Accuracy of the Model

After conducting the hypothesis test, we will want to examine the extent to which the model fits the data. There are two quantities we can look at to assess this: *residual standard error* and the R^2 statistic.

4.3 Residual standard error (RSE)

The RSE is an estimate of the standard deviation of errors, the distances from each data point to its predicted value based on the linear model we fit. In other words, it is the average amount that the response (`Sales`) will differ from the true regression line and is given by the formula:

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum (y_i - \hat{y_i})^2}$$

4.4 R^2 statistic

The R^2 statistic is, technically speaking, the proportion of variance explained by our fitted model. Specifically, it measures the *proportion of variability in the response (`Sales`) that can be explained using the predictor (`TV`).* The closer R^2 is to 1, the greater the proportion of variability that is explained. Its formula is given by:

$$R^2 = \frac{(TSS - RSS)}{TSS} = 1 - \frac{RSS}{TSS}$$

Here, the *total sum of squares*, TSS $= \sum (y_i - \bar{y})^2$ measures the total variance in the response Y, and can be thought of as the amount of variability that already exists in the response, even before we perform any regression analysis. Thus, the R^2 value is a ratio of variability in Y that can be explained by our model to the variability that exists inherently in Y.

5 Results

Using data collected in *Advertising.csv* for TV advertising budgets and their corresponding product sales, I was able to generate the following calculations for the regression coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.03	0.46	15.36	0.00
TV	0.05	0.00	17.67	0.00

A visualization of the coefficient estimates in relation to the observed data can be seen in this scatterplot:

Calculations for the quality indices yield the following:

Note that the F-statistic is included here as well. This is usually utilized for multiple linear regression and thus, will not be discussed in this report.

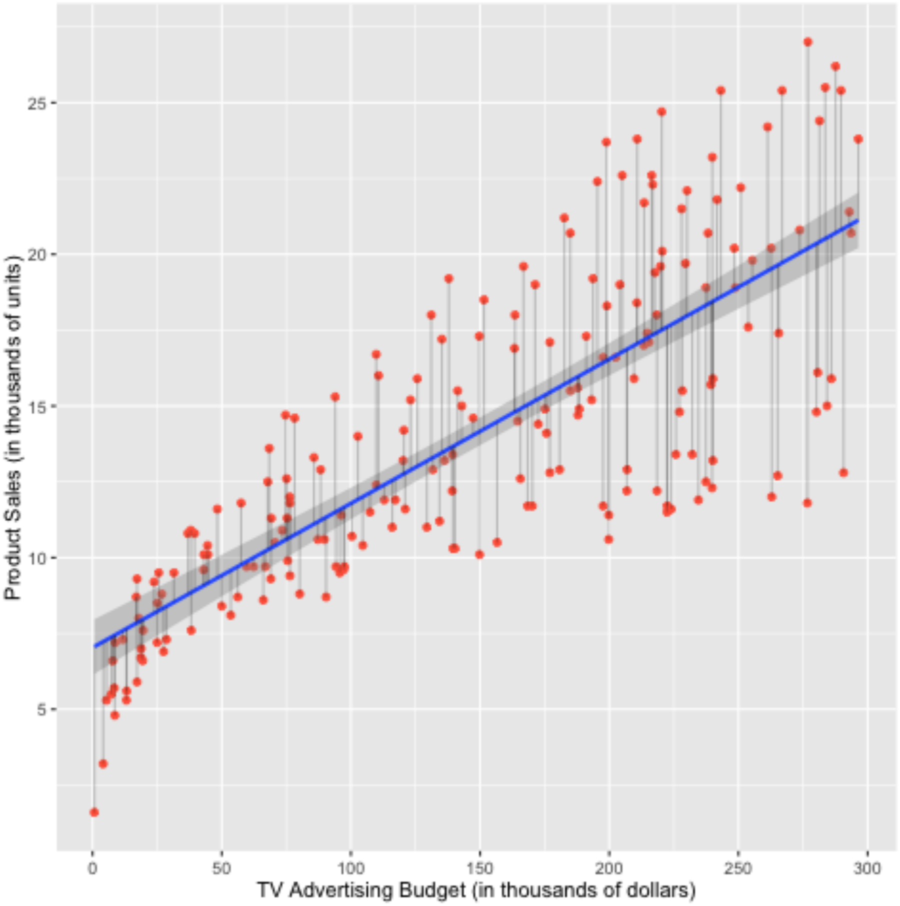

Figure 3: Predicting Product Sales Based on TV Advertising Budgets

Table 1: Quality Indices for the Simple Linear Regression of Sales on TV

Quantity	Value
Residual standard error	3.26
R^2	0.61
F-statistic	312.14

6 Conclusions

Since the p-value for the estimate of `TV` is essentially 0, we can reject the null hypothesis and infer that there is indeed a relationship between `TV` and `Sales`. From the quality indices, we can see that on average, the observed data deviates from its predicted value by 3.26, meaning $3260. The R^2 value tells us that 61% of the variability in `Sales` can be explained by `TV`. The fact that only about half the variability is explained by the model suggests that this simple linear regression may not be the best model fit for the relationship between `TV` and `Sales`.