Simple Linear Regression

Nura Kawa

October 7, 2016

Abstract

In this report we reproduce the main results displayed in section 3.1 "Simple Linear Regression" in the book *An Introduction to Statistical Learning*.

Introduction

Our goal is to provide advice on how to improve sales of the particular product. To do this, we look at the impact of advertising on sales and model the association between them. We then develop a simple linear regression model to predict sales on the basis of advertising via three media.

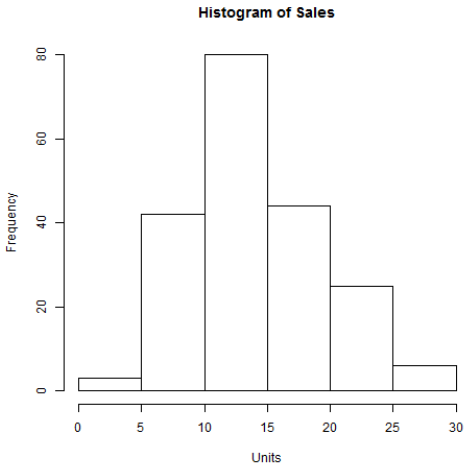
Data

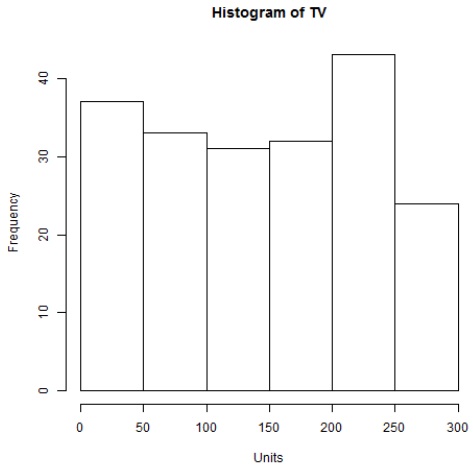
The Advertising dataset consists of *Sales* (in thousands of units) of a particular product in 200 different markets, along with advertising budgets (in thousands of dollars) for the product in each of those markets for three different media: *TV*, *Radio*, and *Newspaper*.

Below is the first five rows of the Advertising Dataset:

```
##      TV Radio Newspaper Sales
## 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
```

The *Sales* and *TV* variables reveal the following distribution:





Methodology

We use Simple Linear Regression to model the association between **Sales** and **TV**. Our method predicts **Sales** using **TV**, using the following linear model:

$$Y = \beta_0 + \beta_1 X$$

Here, our **Y** is **Sales** and our **X** is **TV**.

The parameters β_0 and β_1 are respectively the intercept and slope of our linear model (also called a regression line) fitted to our data. These are estimated as follows:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

where $\bar{x}$ is the mean of **TV** and $\bar{y}$ is the mean of **Sales**.

The *Least Squares Model* minimizes the Residual Sum of Squares, which is defined as:

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + ... + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

Figure 1 displays the simple linear regression fit to the data, where $\beta_0 = 7.03$ and $\beta_1 = 0.0475$.

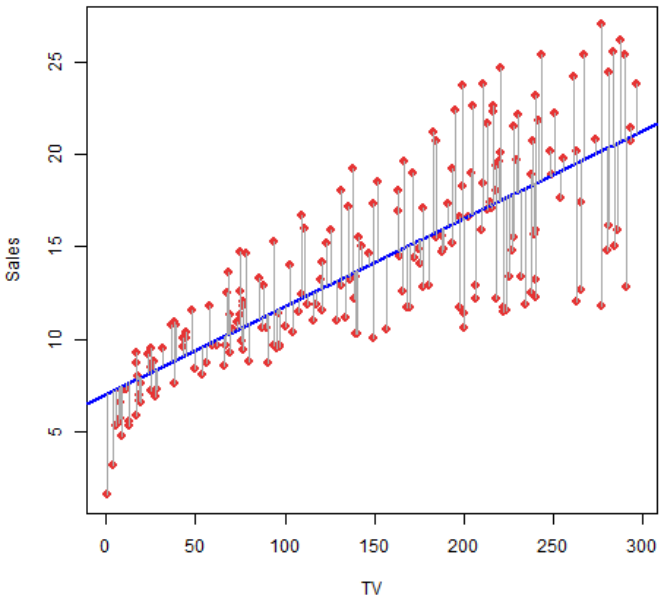


Figure 1: Scatterplot of TV regressed onto Sales

Hypothesis Testing

To assess the existance and strength of an association, we use Hypothesis testing. Our *null hypothesis*, H_0, is that there does not exist an association between **TV** and **Sales**. Our *alternative hypothesis*, H_1, is the opposite - that TV advtertising does impact sales. More specifically, we have:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

We first assume the null hypothesis. To test the null hypothesis, we need to determine whether $\hat{\beta_1}$ is sufficiently far from zero that we can be confident that our true β_1 is non-zero. We thus compute a *t-statistic* defined as,

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

,

which measures the number of standard deviations that $\hat{\beta_1}$ is away from zero. Keeping our assumption of the null hypothesis, we then compute the probability of observing any value equal to $\mid t \mid$ or larger. This is our *p-value*. Typically we look for a p-value of 0.05 or less, meaning that we are 95% confident that our alternative hypothesis is correct. For our purpose we will reject the null hypothesis if our p-value is less than or equal to 0.05.

More about the Linear Model

There exist several regression coefficients that allow us to asses at the accuracy of our linear model. Specifically, these coefficients give us an idea of how well our particular model allows us to predict **Sales** from **TV**. These are the Residual Standard Error (RSE) and the R^2.

Residual Standard Error (RSE)

The RSE is an estimate of the standard deviation of the error in our model. This shows how far our data will deviate from the generated regression line. The equation for RSE from page 69 of *Introduction to Statistical Learning* is:

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

where y_i is a data point of our response variable **Sales** *as predicted by our model*, and $\hat{y_i}$ is our actual data point. Our goal always is to generate a model with a minimal RSE.

R-Squared

The R^2 measures the goodness of fit of our model. It measures the proportion of variance explained by our model. Thus it has the range [0,1] and is independent of Y's scale.

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

Where TSS measures the total sum of squares; the total variance inherent in the response before performing our regression. RSS meausres the amount of variability explained by our regression. Thus the above formula measures the proportion of variability in Y that is explained using X.

If R^2 is a value close to 1, a large proportion of variability in the response explained by the regression; i.e. our model fits the data well.

If R^2 is a value near 0, the regression did not explain much variability in the response; i.e. our regressin performed poorly.

Results

The following tables show the results of our linear regression. We used R to create the model:

```
# reading in advertising data
advertising <- read.csv("data/Advertising.csv", stringsAsFactors = FALSE)

# computing regression object via lm()
regression_object <- lm(Sales ~ TV, data = advertising)

# summary of regression object
regression_summary <- summary(regression_object)
```

and obtained these resulting coefficients:

	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

Quantity	Value
Residual Standard Error	3.26
R^2	0.61
F-Statistic	312.14

The tables show a very small p-value (it is not actually zero, but very close to it); hence, if we see a small p-value, we can infer that there is an association between the predictor and the response. We rejet the null hypothesis - that is, we declare a relationship to exist between TV and Sales, and thus a correlation.

However, looking at our model assessment, we see that a linear model is not a very close fit, given our R^2 is 0.61, and not very close to 1. Thus, we can declare an association but must take into account the possibility of an ill-fitting model. It is better to try different types of models, including non-linear ones, to predit Sales.

Conclusions

The extremely small p-value indicates that there is indeed a correlation between TV and Sales, but our R^2 value indicates that we should try fitting alternative models. Alternative methods include using more predictors and running multiple regression, or fitting a non-linear model.