Simple Regression Analysis

Author: Nichole Rethmeier

Abstract

This report will be reproducing a linear regression analysis found in the book, *An Introduction to Statistical Learning*, written by James, Witten, Tibshiri, and Hastie. The main results can be found in Chapter 3, "Simple Linear Regression".

Introduction

The purpose of this project is to gain insight as to how advertising budgets for a product affect sales of that product. Once we have examined the relationship between advertising and sales, we can create a model to predict the amount of sales based on various advertising budgets. We can use this relationship to optimize the amount of sales through advertising budget adjustments.

Data

The Advertising data set we will be using from *An Introduction to Statistical Learning*, consists of the following variables: sales, TV, radio, and newspaper. The sales variable tells you the amount of sales, in thousands of units, for the product throughout 200 different markets. TV, radio, and newspaper variables represent the advertising budgets for each of these medias in the same 200 markets. Therefore, the inputs are the budgets and the output is the amount in sales resulting from the budgets.

Methodology

Regression Coefficients

We choose to examine on particular media in this project, TV, and how that affects Sales. In order to do this, we apply a simple linear regression model. This is a method in order to predict a response, or change in Y, given a change in the predictor, or X. We write this relationship as follows:

$$Y \approx \beta_0 + \beta_1 X$$

We call this regressing Y onto X, where in this case the independent variable is TV and the dependent variable is Sales. Therefore, the formula for the regression would be:

$$Sales \approx \beta_0 + \beta_1 TV$$

The constants β_0 and β_1 are unknown and are the intercept and slope of our linear regression model. They can be estimated by plugging in values for X and Y. However, as these are estimates, theres stands to be some error involved. *Standard error* is a way of measure the statistical accuracy of these estimates. In general, we want the standard error of $\hat{\beta}_0$ and $\hat{\beta}_1$, denoted SE($\hat{\beta}_0$) and SE($\hat{\beta}_1$), to be as small as possible because that means our estimates are more accurate. Another important statistic is the *t-statistic*, which is used to test hypotheses. The t-statistic is given by:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

The t-statistic measures the size of difference in a hypothesized value relative to the variation in your data. The greater the t-statistic, the greater the evidence that there is no significant distance. The *p-value*, gives you the probability of observing a value equal to or larger than t. A small p-value means that there **is** a relationship between your X and Y.

Quality Indices

While those statistics speak to the regression itself, there are other important statistics that speak to the quality of the regression. The first we will discuss is the *Residual Standard Error*. This number tells us the average amount that the response, in other words the values of Y given by the regression model, will deviate from the true regression line. It is considered as a measure of the "lack of fit" of a model to the data. R^2 is another important statistic, which will always lie between 0 and 1. This measurement tells us the proportion of variance in Y which can that can be explained using variance in X. What this means is that the larger R^2 is the better the regression model is, because the closer to 1 it is, the more the variability in the response is justified. A number closer to 0 would tell us that the variability is not explained in the regression model, and the fit must be off. Lastly, the *F-statistic* is another indicator of the relationship between X and Y. When an F-statistic is larger than 1, it follows that X and Y are related.

Results

The results of the statistics discussed above are as follows. For the coefficient, using the Advertising data, we get the estimates $\hat{\beta}_0$=7.03 and $\hat{\beta}_1$=0.05. The standard errors were relatively small, suggesting that the model was a good fit. The values for these are as follows:

	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

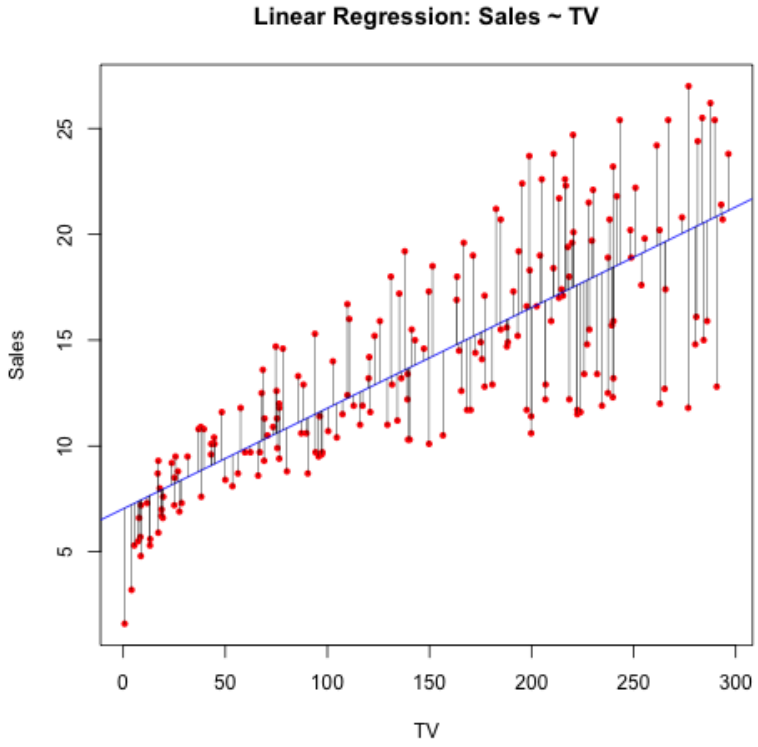
Table 1: Information about Regression Coefficients

Another indicator that there is a linear relationship between TV and *Sales* is the quality indices values. The R^2 is closer to 1 and also the F-statistic is relatively larger. The values for these are as follows:

	Value
Residual Std Error	3.26
R Squared	0.61
F statistic	312.14

Table 2: Regression Quality Indices

Lastly, we can visualize this relationship by looking at the scatterplot and regression line.



Conclusion

In conclusion, we can see that there clearly is a linear relationship between the TV and Sales variable. Generally, the larger the budget for TV is the greater amount of Sales will result.