Multiple Regression Analysis

Author: Nichole Rethmeier

Abstract

This report will be reproducing a multiple 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, "Multiple 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 budgets 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 examine three types of media in this regression, Newspaper, TV and Radio, and examine how the budgets for those variable affects Sales. In order to do this, we apply a multiple linear regression model. This is a method in order to predict a response, or change in Y, given a change in a set of predictors, $X_1, X_2, ..., X_n$. We write this relationship as follows:

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_1 X_2 + ... + \beta_n X_n$$

We call this regressing Y onto $X_1, X_2, ..., X_n$, where in this case the independent variables are the budgets for TV, Radio, and Newspaper and the dependent variable is Sales. Therefore, the formula for this particular regression would be:

$$Sales \approx \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 Newspaper$$

The constants $\beta_0, \beta_1, \beta_2, \beta_3$ are unknown and are the coefficients for our regression model. They can be estimated by plugging in values for X_1, X_2, X_3 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 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$, denoted SE($\hat{\beta}_i$), 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 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 predictors 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 $\hat{\beta}_0$=2.94, $\hat{\beta}_1$=0.05, $\hat{\beta}_2$=0.19,$\hat{\beta}_3$=0. Due to the the very low standard errors, we can infer that there is a correlation between the predictors and Sales. However, as the coefficient of Newspaper is 0, this variable does not factor in. Therefore, it is *not* helpful in predicting a response, and the only predictors that are relevant are TV and Radio budgets.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.94	0.31	9.42	0.00
TV	0.05	0.00	32.81	0.00
Radio	0.19	0.01	21.89	0.00
Newspaper	-0.00	0.01	-0.18	0.86

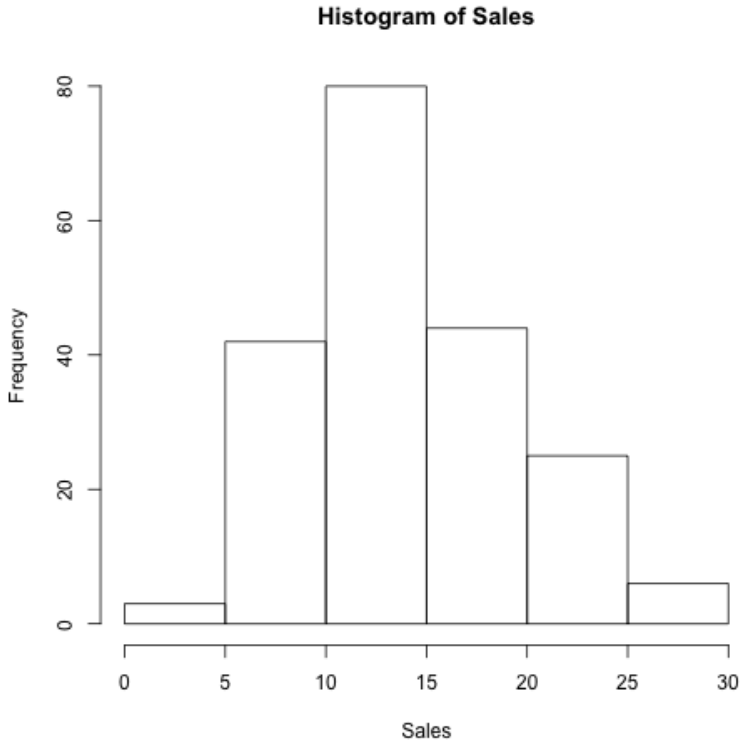
Table 1: Information about Regression Coefficients

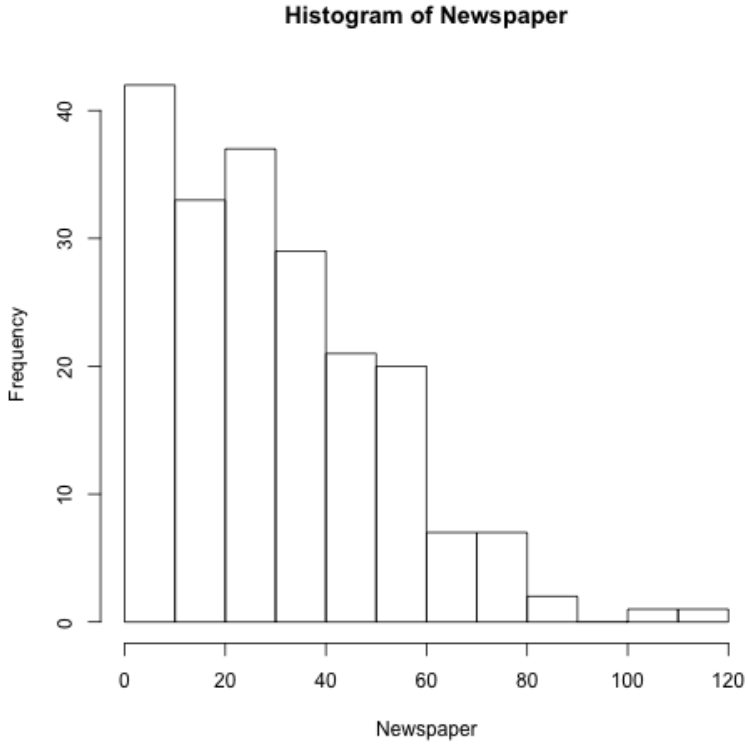
Now, we will examine how good of a fit the multiple regression model is. The R-squared of 0.90 tells us that this model is an excellent fit, as the closer to 1 this number is the more accurate the model. Also as the residual standard error is only 1.69, the predictions created by the model are very close to the actual values:

	Value
Residual Std Error	1.69
R Squared	0.90
F statistic	570.27

Table 2: Regression Quality Indices

We can also see that the Y values, or Sales, seem to follow a normal distribution. It's also why the Newspaper variable is not an accurate predictor in this model as it does not resemble a normal distribution at all.


Histogram of Sales
Frequency
0
20
40
60
80
0
5
10
15
20
25
30
Sales


Histogram of Newspaper
Frequency
0
10
20
30
40
0
20
40
60
80
100
120
Newspaper

Conclusion

In conclusion, we can see that there clearly is a relationship between the budgets for TV, Radio and Sales. Generally, the larger the media budgets are the greater the increase in Sales will be. However, the Newspaper variable is not an accurate predictor here, quite possibly because the importance of Newspapers dwindles as technology advances.