# Simple Regression Analysis

*Erica Wong*

*October 13, 2016*

## Abstract

The purpose of this report is to apply the computational tools that we have learned about in class to reproduced a multiple regression analysis. Specifically, we are trying to reproduce the analysis that was done in section 3.2 of *An Introduction of Statistical Learning* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. You can find a link to their book and data set by clicking on this sentence.

## Introduction

The goal of this study is to see how different types of advertising affects sales so that we can provide advice on what medium would be the best to improve sales of a specific product. In my report, I am specifcally looking to see if there is some association between TV, radio, newspaper, and sales. If there is an association between the response and explainatory variables, then I want to have a model that will be able to predict sales based on these 3 variables.

## Data

In this report, we are using data from Advertising.csv. In the data set we have four different columns. There is the TV advertising budget, radio advertising budget, and newspaper advertising budget. These budgets are given in thousands of dollars. Additionally, there is a sales column, which is given in thousands of units. In my report, we will be focusing on the TV and sales column of the Advertising.csv.

## Methodology

We look at the data set and study the relationship between TV, radio, newspaper, and sales. This statement helps us to set up our null and alternate hypothesis. The null hypothesis is $H_0$: $\beta_1 = \beta_2 = \cdots = \beta_p$. The alternative hypothesis is $H_1$: There exists at least one $\beta_j$ that does not equal 0. To test this, we want to use a multiple regression model, where the equation is $Sales = \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 Newspaper + error$.In order to solve for $\beta_0$, $\beta_1$, $\beta_2$, and $\beta_3$ we need to fit a linear model to our data, this will give us $\hat{\beta_0}$, $\hat{\beta_1}$, $\hat{\beta_2}$, and $\hat{\beta_3}$, which are estimates of $\beta_0$, $\beta_1$, $\beta_2$, and $\beta_3$. By solving for these objects,we get a line as close as possible to the 200 data points we have. In this report, we will fit a multiple linear regression model using the least squares criterion to our data.

## Results

Here are the tables of the simple linear regression of sales vs each of the three variables: Table 1 shows simple regression of sales on TV.

Table 1: Simple Regression of Sales on TV

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 7.0326 | 0.4578 | 15.3603 | 0.0000 |
| TV | 0.0475 | 0.0027 | 17.6676 | 0.0000 |

Table 2: Simple Regression of Sales on Radio

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 9.3116 | 0.5629 | 16.5422 | 0.0000 |
| radio | 0.2025 | 0.0204 | 9.9208 | 0.0000 |

Table 2 shows simple regression of sales on radio.

Table 3 shows simple regression of sales on newspaper.

Table 3: Simple Regression of Sales on Newspaper

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 12.3514 | 0.6214 | 19.8761 | 0.0000 |
| newspaper | 0.0547 | 0.0166 | 3.2996 | 0.0011 |

When we did multiple regression for our model, this these are the results that we obtained.

Table 4: Multiple Regression Table

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 2.9389 | 0.3119 | 9.4223 | 0.0000 |
| TV | 0.0458 | 0.0014 | 32.8086 | 0.0000 |
| Radio | 0.1885 | 0.0086 | 21.8935 | 0.0000 |
| newspaper | -0.0010 | 0.0059 | -0.1767 | 0.8599 |

Additionally, we have included the correlation matrix between the different variables. The correlation matrix is shown in table 5.

In this report, we want to answer 4 main questions. Those questions are:

1. Is at least one of the predictors usefil in predicting the response?
2. Do all predictors help to explain the response, or is it only a subset of the predictors useful?
3. How well does the model fit the data?
4. How accurate is the prediction?

To answer these questions, we will be using the informations from the tables produced above (in addition to other tables) and draw conclusions from those.

**1. Is at least one of the predictors useful in predicting the response?**

When asking this question, we are setting up the null hypothesis to be: $H_0$: $\beta_1 = \beta_2 = \cdots = \beta_p$. The alternative hypothesis in this case is $H_1$: There exists at least one $\beta_j$ that does not equal 0. In order to see if there is a relationship we can look at the F-statistic. If there is no relationship between the predictors and the response variable, in our case TV, radio, and newspaper with sales, then the F-statistic will be close to 1. However, if the F-statistic is greater than 1, then there is probably a relationship between the predictors and the response and we will reject the null hypothesis. In table 6, there is a chart containing additional information, and one of the things that is shown in this table is the F-statistic.

From this table, we can see that the F-statistic = 570.271. This is a very large number, which is much greater than 1. Thus, we would reject the null hypothesis due to the large number and we have enough reason to believe that there is at least one predictor variable has a relationship with the response variables.

Table 5: Correlation Matrix

| | TV | Radio | Newspaper | Sales |
|---|---|---|---|---|
| TV | 1 | 0.0548 | 0.0566 | 0.7822 |
| Radio | | 1 | 0.3541 | 0.5762 |
| Newspaper | | | 1 | 0.2283 |
| Sales | | | | 1 |

Table 6: More Info on Multiple Least Squares Model

| Quantity | Value |
|---|---|
| RSE | 1.686 |
| R2 | 0.896 |
| F-stat | 570.271 |

**2. Do all predictors help to explain the response, or is it only a subset of the predictors useful?**

When deciding on which predictors help to explain the reponse, we can look at the p-values. When looking at the p-value, if the value is small, then there is reason to believe that the predictors help to explain the response. However, if the p-value is large, then we have reason to believe that there is no relationship between that predictor and the response variable. The p-value (rounded to the thousandth) of TV in relation to sales is 0 and the p-value of radio (rounded to the thousandth) is 0. Both of these p-values are fairly small, so there is good reason to believe that these predictors help to explain the reponse variable. However, the p-value of newspaper is 0.8599151. This is a very large p-value, so there is reason to believe that there is no relationship between newspaper budget and sales.

**3. How well does the model fit the data?**

To know how well our model fits the data, we can look at the RSE and Rˆ2. RSE stands for residual standard error. RSE is the estimate of the standard deviation of error. It can also be described as the lack of fit in the model. So, when RSE is small, that means that the model fits the data well. In this multiple regression model, the RSE = 1.686. Since RSE is measured in absolute terms, we can tell that our RSE is small enough to say that this is a good fit.

We can also look at $R^2$, which tells us how much of the dependent variable can be explained by the independent variables. $R^2$ is the square of the correlation of the response and the variable, so the closer that is to 1, the better our model fits the data. $R^2$ can only take on values between 0 and 1. Looking at the model that uses all 3 variables, our $R^2 = 0.897$. This is very close to 1. So, that means that a lot of our response variable, sales, can be explained by the explainatory variables.

So, looking at these 2 criteria, we can see that our multiple regression model with all 3 variables fits very well overall.

**4. How accurate is the prediction?**

As explained in the previous question, we know that our prediction is fairly accurate because our $R^2$ is fairly high. So, this contributes to how accurate our model is. Also, we can look at the stanard deviations of each of our explainatory variables. The standard deviation related to each of predictor variables are all fairly small For TV, radio, and newspaper, they are 0.0013949, 32.8086244, approximately 0. respectively. So, that means that our predictions of $\beta_1$, $\beta_2$, and $\beta_3$ are very close to the true model and thus we have an accurate fitted model so the predictions are fairly accurate.

## Conclusion

So, in this project, we were able to recreate the study that was done in chapter 3.2 of *An Introduction of Statistical Learning* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. We also found that there is some relationship between TV, radio, and newspaper budget with sales. This is shown through the rejection of our null hypothesis. Our null hypothesis is $H_0$: $\beta_1 = \beta_2 = \cdots = \beta_p$. Additionally, we know that our model fits well because $R^2$ is close to 1 and RSE is small by comparative standards.