Stat 159 Hw2 Report

Ellen Hwang

October 31, 2016

Abstract

In this report, I am performing a simple linear regression analysis. I am using *the Advertising Dataset*, taken from the webpage for the text, *An Introduction to Statistical Learning*, to perform this simple linear regression. This report will include a description of the data, methodology, and results for the linear regression.

1 Introduction

One of the basic models data scientists should understand is simple linear regression. Regression, in its most simple terms, is a statistical process that estimates the relationship between a dependent variable and one or more independent variables. In this report, we will only be examing the relationship between one dependent variable and one independent variable to understand regression. This report will specifically examine the relationship between TV advertising budget and number of sales using linear regression.

2 Data

We will be working with the *Advertising Dataset.* This dataset holds information for 200 different markes for 4 different variables: *Sales*, *TV*, *Newspaper* and *Radio*. *Sales* represents the amount of units sold (in thousands). *TV*, *Newspaper*, and *Radio* each represent the advertising budget spent on those platforms.

3 Methodology

For running a regression, we using the ‘lm()‘ function. More specially, we regress *Sales* on *TV* using this code: ‘lm(Sales TV, data = advertising)‘. The immediate output of this code is the estimated coefficient and the constant value. Second, I use the ‘summary‘function on the regression object to see further coefficient information, residuals, and other statistics.

4 Results

Based on the regression of Sales on TV, we found that TV is a statistically significant variable because of a p-value far below zero. The R squared value is at .6099 meaning about 61% of the variance from the population regression line is explained by TV. We can also see that the coefficient on TV is a postive value, meaning TV and Sales are positively correlated.

Simple Regression

```
> load('regression.RData')
> simp_reg

Call:
lm(formula = Sales ~ TV, data = advertising)

Coefficients:
(Intercept)           TV
    7.03259      0.04754
```

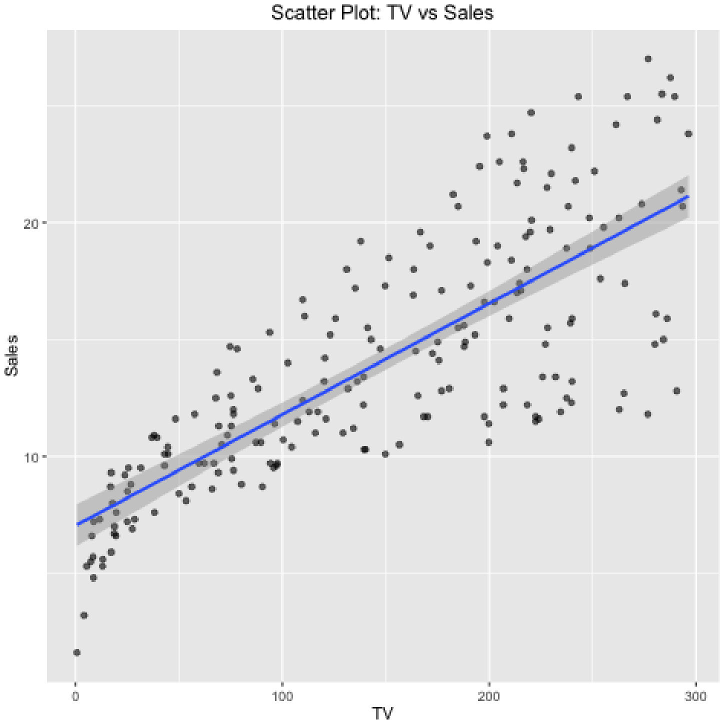
Summary of Simple Regression

```
> sum_simp_reg

Call:
lm(formula = Sales ~ TV, data = advertising)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3860 -1.9545 -0.1913  2.0671  7.2124

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
TV          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,        Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

5 Conclusion

Ultimately, we can see there is a relationship between Sales and TV. We see that the regression output gives us several summary statistics to understand the

Figure 1: Scatter Plot with Regression of Sales on TV



relationship between Sales and TV. For instance, the high F-statistic and low p-value tell us that TV is a significant explanatory variable to Sales. The positive coefficient on TV tells us that TV and Sales are positively correlated. Although the statistics tell us a good story, a visualization of the relationship through a scatter plot can help us better comprehend what is going on. Overall, this regression analysis has shown us several strategies for using regression analysis to understand the data we have at hand.