Report

Jamie Stankiewiz

October 4, 2016

Abstract

In this report, we will investigate the findings of comparing the relationship of TV vs. Sales. We will mostly look at the data through R, specifically looking at the scatterplot of TV and Sales to see if there is a trend between the 2, then try to fit a regression line. We will calculate the data of the regression line through R and see what we can conclude from the simple linear model.

Introduction

We examine a simple linear regression model. We will look at the data from the book "An Introduction to Statistical Learning" by James et al. We will specifically look at the plot of TV vs. Sales and take a look at the regression line. We are particularly interested in the data for the regression residuals, including intercept and slope, more commonly known as the coefficients of the regression line, as we look at the strength of the line. In addition, we will also examine the R^2 statistic as a measure of the linear relationship between TV and Sales, as well as the F-statistic, which is part of the ANOVA hypothesis test.

Figure 1: This is a result of the least squares regression fit of Sales on TV from the Advertising data set.

Data

The data set, Advertising.csv contains data for TV, Radio, Newspaper, and Sales.

This summary tables details out the estimated coefficients of the linear least squares line. Complete with a standard error of the estimates as well as the t- and p- values. From this, we can conclude that $\hat{\beta}_0$ and $\hat{\beta}_1$ are not equal to 0.

Table 1: Coefficients of the regression line of the least squares model.

	Coefficient	Std. Error	t-statistic	p-value
Intercept	7.03	0.46	15.36	0.00
TV	0.05	0.00	17.67	0.00

This marks the code of the actual summary output in R. These are the details of the regression line.

The content in this table displays 3 things. First, the residual standard error, or RSE which is the estimated standard error of errors, ϵ, describing the so-called "lack of fit" of the model. The second term, R^2 explains how much of the variability was due to the regression. This is on a scale of 0 to 1. The last term is the F-statistic, is the test statistic of running an ANOVA hypothesis test to find if the means of the two categories are the same or not. In this case, our F-statistic corresponds to a p-value of $<<0.000001$, so therefore you can safely reject the null hypothesis, H_0.

Table 2: More information of the least squares model.

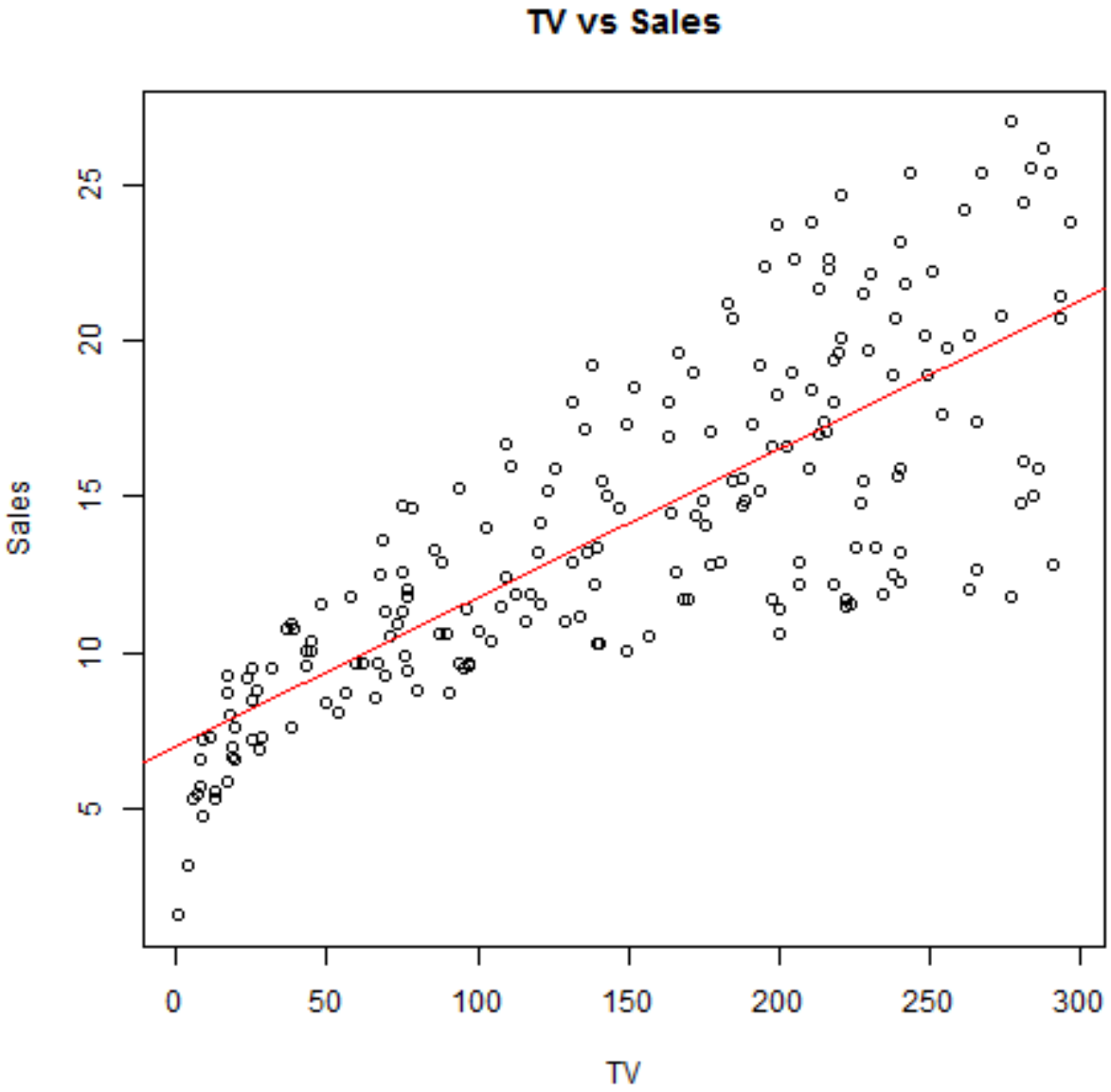

Figure 1:

	Quantity	Values
A	Residual standard error	3.259
B	R^2	0.612
C	F-statistic	312.145

Methodology

We will first look at our data set, advertising.csv to look at the varibales used. TV, Radio, Newspaper and Sales are the variables we get by looking at the column names of our data sheet. We then wish to look at the relationship between TV and Sales. To do so, we first generate a scatterplot of the 2 against each other. Next, we create the regression line using the lm() function of the data. To get these estimates of the coefficients, we look at the outputs of the lm() function to see the statistics of the regression line (as shown below).

```
##
## Call:
## lm(formula = Sales ~ TV, data = ad)
##
## 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 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 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
```

Results

By looking at the data table of the coefficients of the regression line, we see that estimated coefficient, $\hat{\beta}_0$ is an estimate with a standard error of .46. From the t-test, we find that the both t-statistics have p-values of less than <<0.00001. Therefore, we can conclude that both $\hat{\beta}_0$ and $\hat{\beta}_0$ are not equal to 0. This means that there is an actual relationship between the 2 categories. Looking at the second table, we find the value of the standard error of the residuals is 3.259. The R^2 statistic, is known as the 'adjusted R-squared' value, represents the percentage of the linear regression line is actually due to the 2 predictors. The F-statistic, we find is 312.145, is generated from the ANOVA test to determine whether the means of the 2 categories are the same. From the more interesting p-value, 2.2×10^-16 concludes that the means of TV and Sales are differnet.

Conclusions

From the data, we can conclude that there is a linear relationship between the 2 data categories, TV and Sales from the data set given. This was shown the the regression line and the values of the estimated coefficients.