Simple Regression Analysis

Dakota Lim

October 5, 2016

Abstract

In the following report we wil preform a basic linear regression to explain how well one data set predicts a corresponding set. In particular, we will reproduce the work done in sections 3.1 of *Simple Linear Regression}* using the exact same data sets and producing some of the same graphics.

Introduction

The goal of this paper is to advise retailers on how to spend their advertising money amongst the available media outlets. We will do this by regressing the relative ammount of money spent on advertising on the total sales, determining if there is some relationship, and determine the strength of the relationship. Using this information we can provide statistically supported advise to the retailer.

Data

The data in question is the Advertising data set found at `http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv`. The data has a total of 200 observation in four variables: **Tv**, **Radio**, **Newspaper**, and **Sales**. The **Sales** data is given in thousands of units, and all other data is given in thousands of dollars spent.

Methodology

In this paper we will only consider the **Tv** data and its influence on **Sales**. The methods used can easilly be extended to other variables if desired. To describe the relationship we first consider a basic scatterplot of **Tv** against **Sales**:

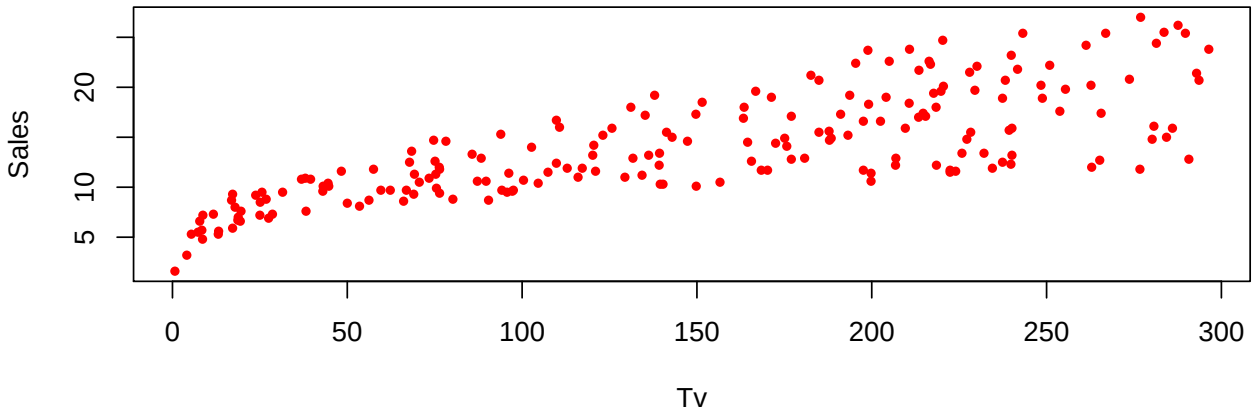


Figure 1: Tv vs Sales Scatterplot

There appears to be a linear relationship between **Tv** and **Sales**, as **Tv** increases **Sales** tends to as well. Thus, we decide on the following linear model:

$$Y = X * \beta$$

With $Y \coloneqq$ **Sales** and $X \coloneqq$ **Tv**. Upon fitting the model to our data we well be able to solve for the optimal β. Recall that in linear regression the objective function is to minimize the following following L2 norm over all possible β values:

$$\|Y - X\beta\|_2$$

Some basic matrix algebra gives the following optimal value of β given some Y and X data matricies:

$$\hat{\beta}_{OLS} \coloneqq (X^T X)^{-1} X^T Y$$

Next we consider the Residual Standard Error, or RSE. The RSE is considered a measure of "lack of fit", meaning it measures how poorly a model fits the data. We define RSE as follows:

$$RSE \coloneqq \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)}$$

Thus, we may think of RSE as the average deviation from the OLS prediction.

We will also consider the R^2 statistic to help determine the strength of the correlation between our variables, R^2 is given by:

$$R^2 \coloneqq \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$RSS \coloneqq \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$TSS \coloneqq \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Thus, we may think of R^2 as a proportion of explained variance : total variance.

Finally we consider the F-statistic, which can be used to determine if our model has excess regressors. The F test uses the following hypothesis:

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

Results

Once we loaded and prepared the data we observed the following values relating to $\hat{\beta}$ in Table 1:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.0326	0.4578	15.36	0.0000
Tv	0.0475	0.0027	17.67	0.0000

Table 1: Regressor Analytics

and recorded statistics relating to the overall model in Table 2 on the following page.

We note that Pr(>|t|) is so small that it is rounded down to 0, implying the **intercept** and **Tv** regressors both contribute significantly in the prediction of **Sales**, which in turn tells us to keep both in the model. To further support this we note the large t-scores for each regressor, implying they are significant to the model.

Next we look at the actual values and interpret them. The intercept ($\hat{\beta}_0$) is 7.0326, this can be interpreted as the total **Sales** if no money is spent on **Tv**. The Tv value ($\hat{\beta}_1$) is .0475, this can be interpreted as the unit of **Sales** increase per one unit increase in **Tv**. Note Table 1 also provides standard errors, allowing us to (if we wish) produce $(100 * \alpha)\%$ confidence intervals of the form:

$$\hat{\beta}_i \pm t_{n-2}(\frac{\alpha}{2}) * Std.\ Error_i$$

Now we consider the information in Table 2.

	Statistics	Values
1	RSS	2102.53
2	MSE	10.51
3	RSE	3.26
4	R2	0.61
5	FStat	312.14

Table 2: Model Analytics

The RSS gave a value of 2102.53, which is not too helpful in judging the models performance. The MSE was only 10.51, which has a direct interpretation as the average of the squares of the errors given by our model.

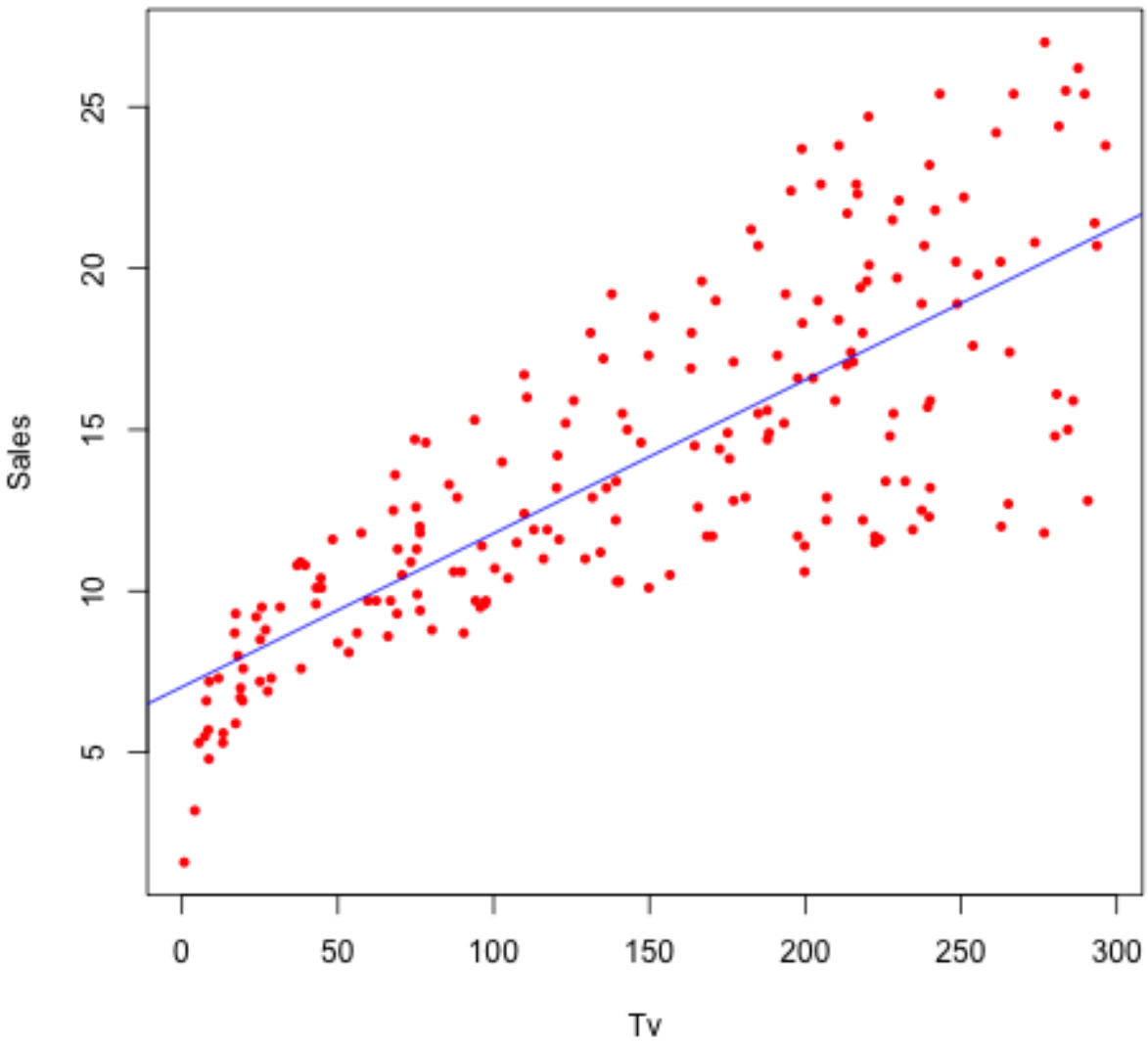
The RSE gave a value of 3.26, as described in the **Methodology** section this is interpreted as a measure of 'lack of fit'. A relatively small RSE, like 3.26, implies that our model is fitting the data relatively well.

The R^2 statistic is described in the **Methodology** section as well. Noting that the value is .61 for this particular model, it is safe to say **Tv** has a positive correlation with **Sales**, albeit not a perfect correlation but a clear correlation none the les.

Recall the FStat tells us the probability our model contains excess regressors. In our model the we have an F-Statistic with (1, 198) degrees of freedom and a value of 312.14. Using this information we can use the F-Statistic table (an online tool can be found here: `http://stattrek.com/online-calculator/f-distribution.aspx`) to see that we would reject $H_0 \ \forall \ \alpha > 1.525879e-05$.

Conclusions

Using the information discussed in **Results** and **Methodology**, we have sufficient evidence to support the conclusion that spending more money on **Tv** results in an increase in **Sales**. By viewing the scatterplot shown in Figure 1 with the Least Squares line superimposed on it we can clearly see that this claim holds:



In conclusion, we advise retailers to use the following equation to estimate **Sales**: **Sales** = 7.0326 + .0475***Tv**

To further elaborate, we expect a increase of .0475 thousand units sold for every 1 thousand dollars spend on Tv advertising. Note these results were calculated without considering any kinds of external influences, one off events, or other regressors.

Similarly, we would advise retailers to be aware of the RSE. The RSE is informing us that, on average, our predictions for **Sales** are 3.26 thousand units sold off their true value.