Multiple Regression Analysis

Priscilla Hartono

10/14/2016

Source file

Abstract

This paper is the third homework of STAT 159, Reproducible and Collaborative Statistical Data Science, taught by Professor Gaston Sanchez in the fall of 2016 at UC Berkeley.

In this paper we reproduce the analysis in section 3.2, *Multiple Linear Regression*, of the book **An Introduction to Statistical Learning** (by James et al) using dataset, **Advertising.csv**.

Introduction

In the real world, when predicting a response, we want to take into account all possible predictors. So far we know how to make a simple regression analysis of a single predictor variable. Now, we want to make more meaningful analysis by including multiple predictors. We can do this by multiple linear regression, which accomodates multiple predictors by giving each a separate slope coefficient in a single model.

Data

The dataset we used in this paper is Advertising.csv. It contains 200 entries of different markets' marketing budget through each media medium (TV, Radio, Newspaper) and how many units were sold.

Sales in thousands of units. Budgets in thousands of dollars

For this paper, we used all the variables: TV, Radio, Newspaper, and Sales. The numbers under TV, Radio, and Newspaper refers to how much budget was put into advertising by each, and the number under Sales reflects the number of units sold in that particular market.

Before we start doing the regressions, we want to look at how our data look like through histograms and simple linear regression.

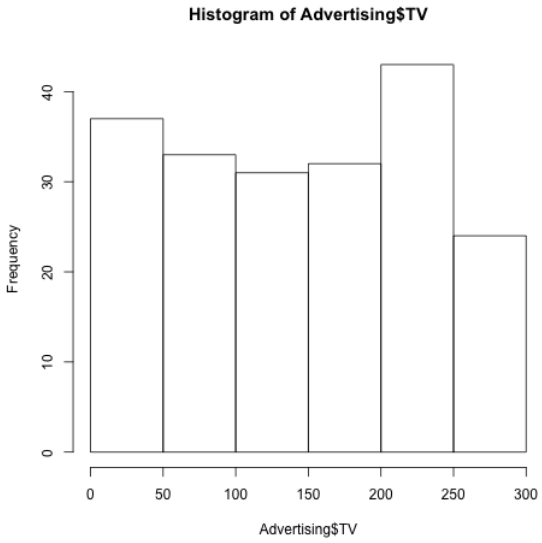

Figure 1: Histogram of TV

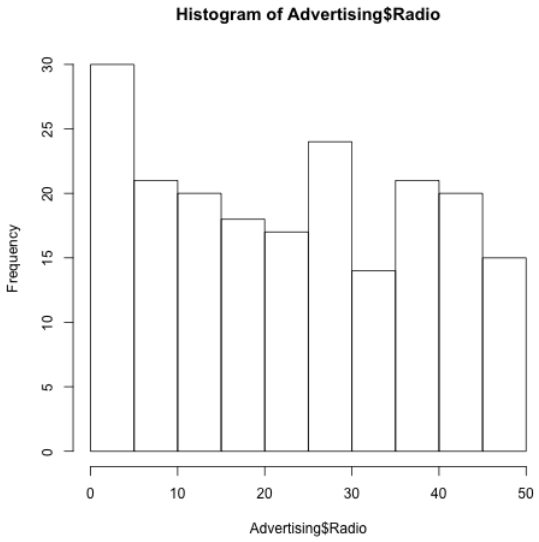

Figure 2: Histogram of Radio

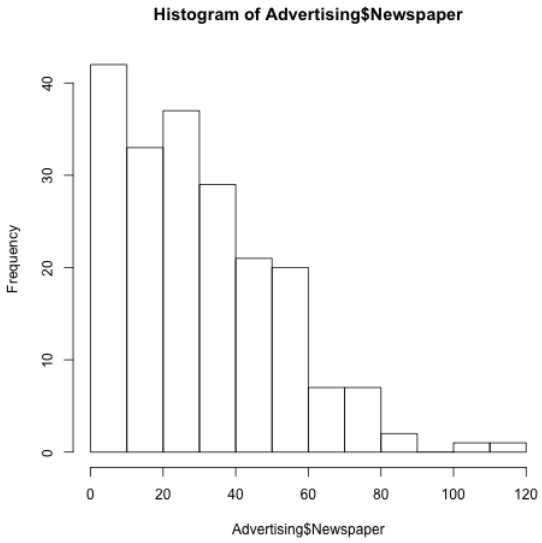

Figure 3: Histogram of Newspaper

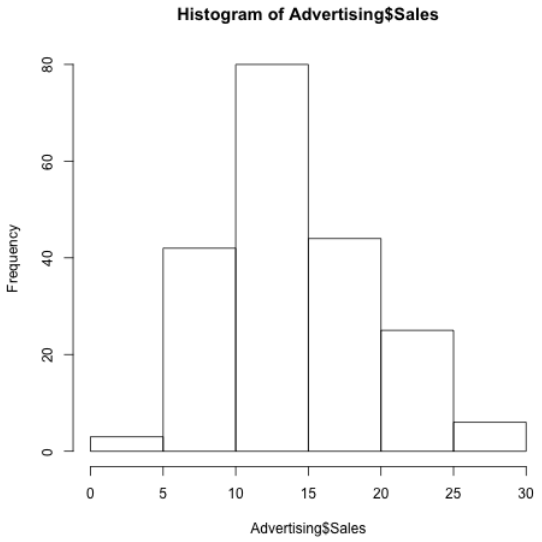

Figure 4: Histogram of Sales

	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 1: Regression of TV on Sales

	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: Regression of Radio on Sales

	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

Table 3: Regression of Newspaper on Sales

Methodology

The multiple linear regression model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$$

As in the case of simple linear regression, we want to get estimates of $\beta_0, \beta_1, ..., \beta_p$. $\hat{\beta}_i$ is the value we are interested in. We need it to determine whether there is a positive or negative correlation. When $\hat{\beta}_i$ comes out to be positive, there is positive correlation, meaning higher spending of advertising through this media medium results in more sales.

Knowing which predictors produces results, we want to know its significance. Here, we need the RSS, TSS, RSE, R2, and F-statistic.

Results

To have an idea of how much sales increase with an additional $1000 budget in each media medium we need the multiple regression coefficient estimates.

	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 4: Least squares coefficient estimates of the multiple linear regression of number of units sold on TV, radio, and newspaper advertising budgets

For example, Table 3 shows approximately 189 units were sold following an increase of $1000 budget in radio advertising.

Continuing from here, to see more clearly the relationship between each predictor and target, we can study the scatterplot matrix and each predictor's scatterplot.

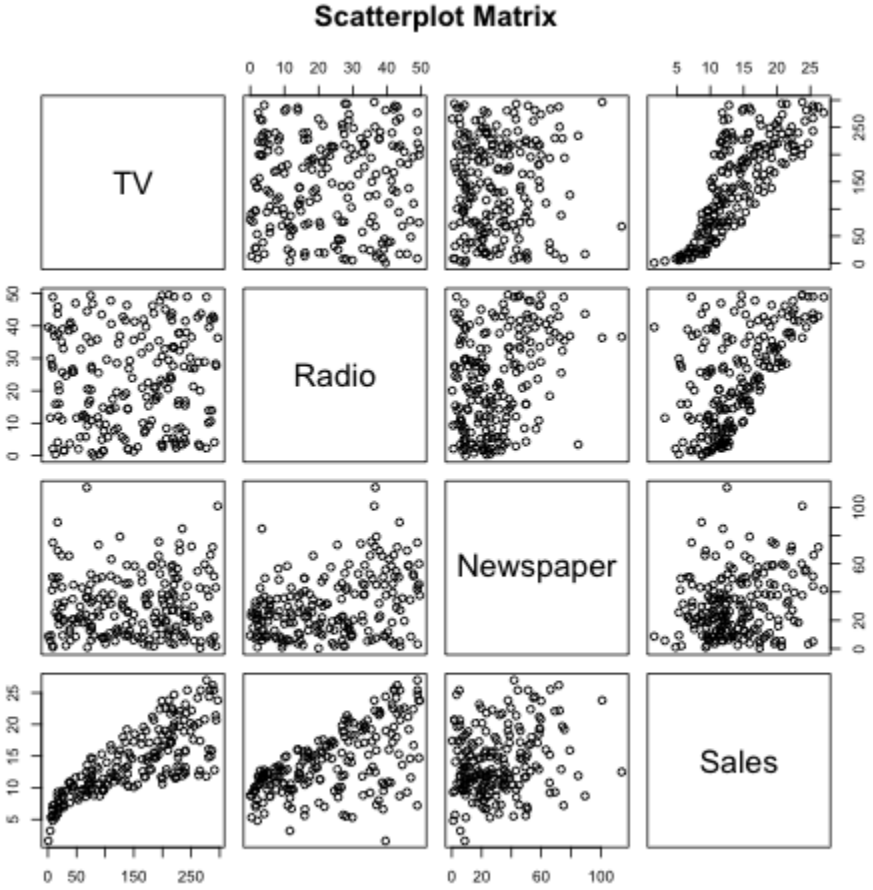

Figure 5: Scatterplot Matrix

Table 5: Correlation Matrix

	X	TV	Radio	Newspaper	Sales
X	1	0.01771	-0.1107	-0.1549	-0.05162
TV	0.01771	1	0.05481	0.05665	0.7822
Radio	-0.1107	0.05481	1	0.3541	0.5762
Newspaper	-0.1549	0.05665	0.3541	1	0.2283
Sales	-0.05162	0.7822	0.5762	0.2283	1

Notice that the correlation between radio and newspaper is 0.35, revealing that there is a tendancy to spend more on newspaper advertising in market where more is spent on radio advertising. Although newspaper advertising does not really affect sales, it is a surrogate for radio advertising, meaning newspaper gets "credit" for the effect of radio on sales.

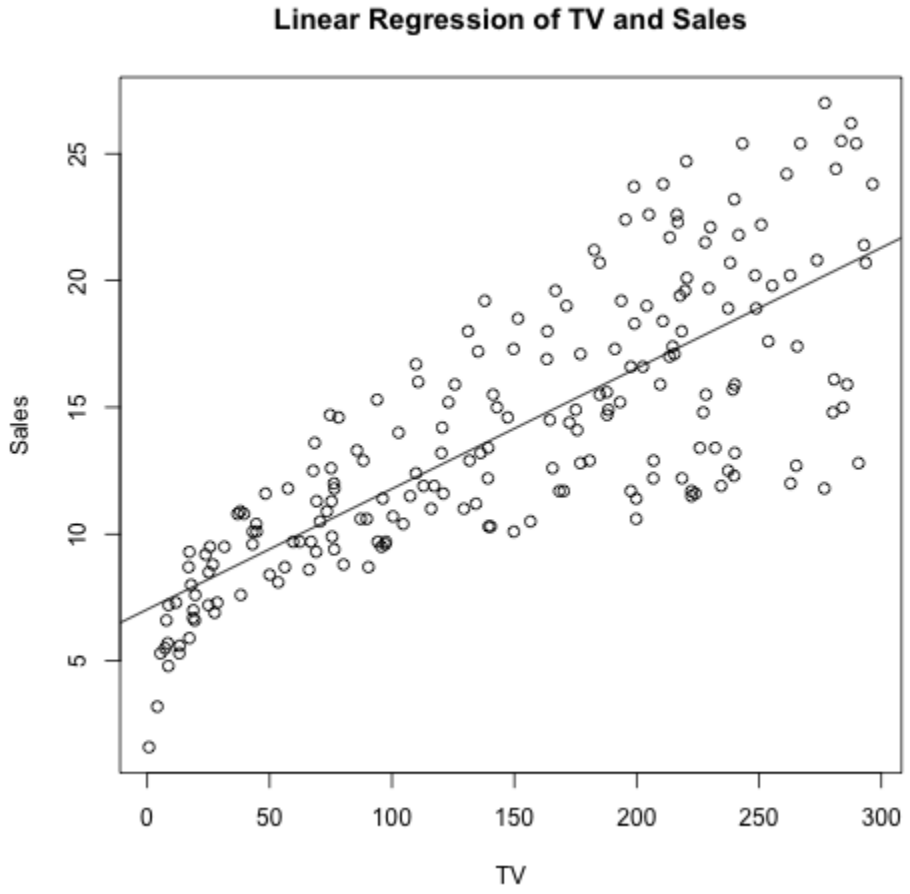

Figure 6: Scatterplot TV and Sales

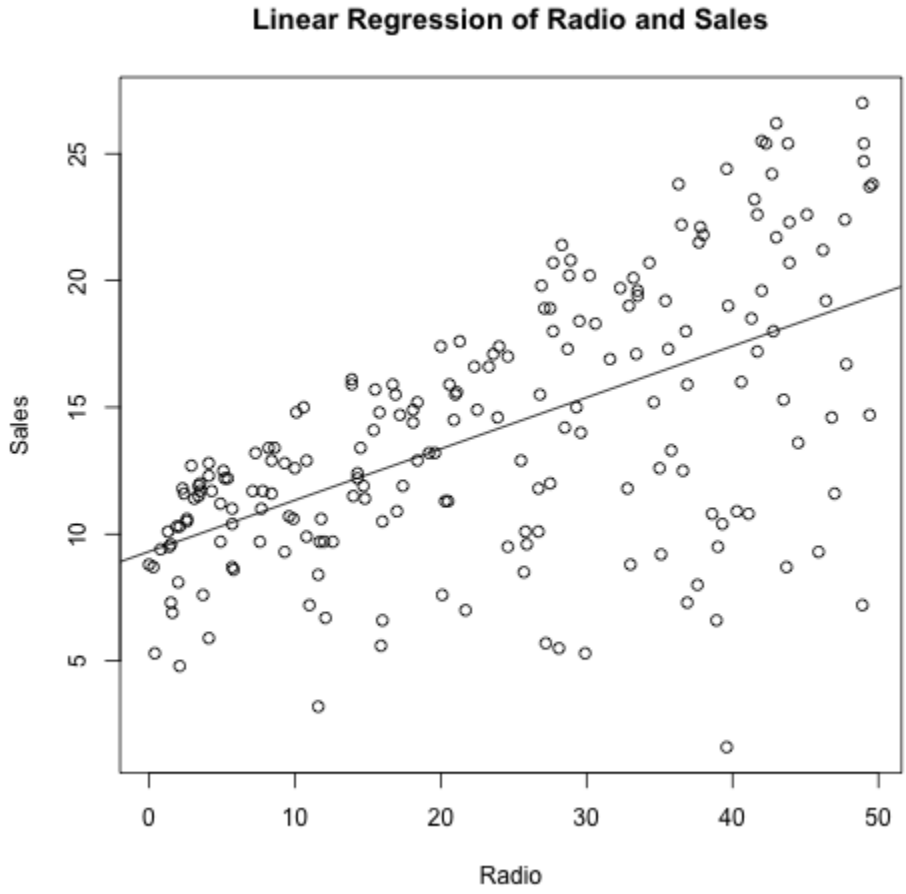

Figure 7: Scatterplot Radio and Sales

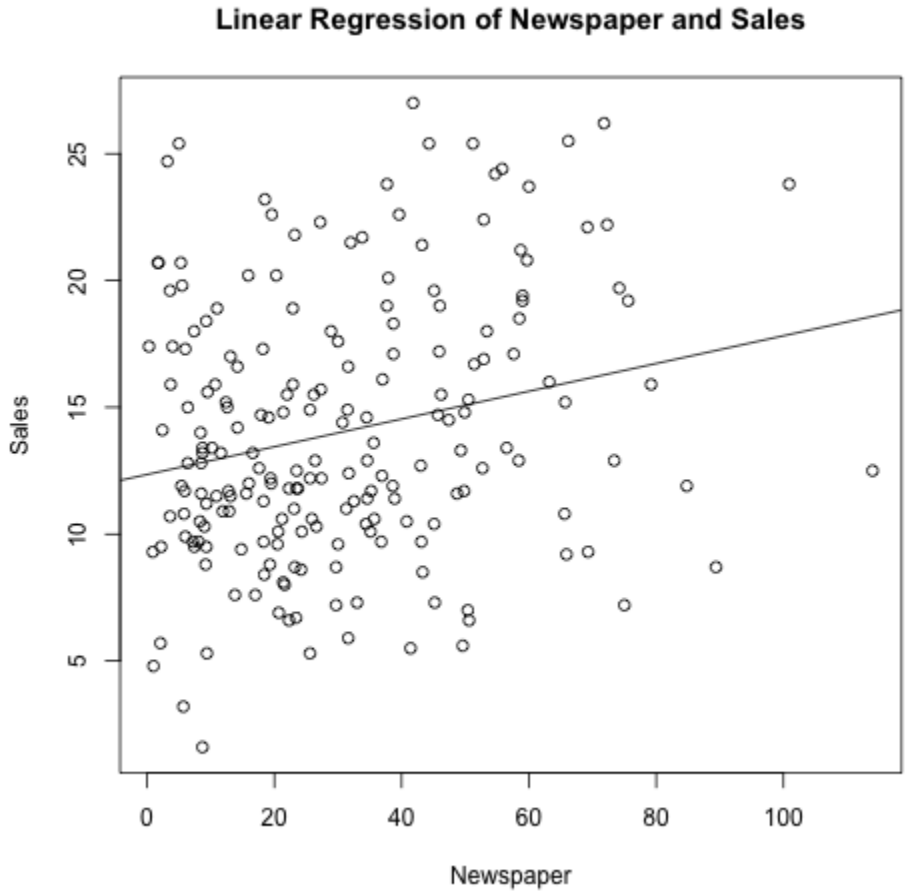

Figure 8: Scatterplot Newspaper and Sales

Finding out the significance of each predictor is key to producing a meaningful analysis. Table 6 gives us this information.

	Quality	Value
1	RSE	1.69
2	R2	0.90
3	F-statistic	570.27

Table 6: More information about the least squares model for the regression of number of units sold on TV, radio, and newspaper advertising budgets in the Advertising data.

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

In order to determine if there is a relationship between the response and the predictor we want to check whether $\beta_i = 0$. We do this by testing the null versus alternative hypothesis. If the H_0 is true, we would expect the F-statistic to be close to 1 when there is no relationship and if H_1 is true, we would expect the F-statistic to be greater than 1.

For this dataset's case, since F-statistic is greater than 1, it suggests that at least one of the advertising media must be related to sales.

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

It is possible that all of the predictors are associated with the response, but it is more often the case that the response is only related to a subset of the predictors. CONTINUE

3. How well does the model fit the data?

To determine how well the model fit the data, we need the RSE and R^2. An R^2 value close to 1 indicates that the model explains a large portion of the variance in the response variable.

For this dataset's case, the model uses all 3 advertising media to predict sales has an R^2 value of 0.8972. On the other hand, the model that uses only TV and Radio predicts sales has R^2 value of 0.89719, meaning there is a small increase in R^2 value if we include newspaper advertising (advertising in newspaper is not significant). In contrast, the model containing only TV as predictor has R^2 value of 0.61. Adding radio to the model leads to a substantial imporvement.

As for RSE, when the model contains TV and radio, it has an RSE of 1.681, and when the model contains all 3 predictors, it has an RSE of 1.686. Meanwhile, the model that contains only TV has an RSE of 3.26. This shows that using a model that uses TV and radio expenditures to predict sales is more accurate than using a model that uses TV expenditure only.

4. How accurate is the prediction?

To check for accuracy, we use confidence interval. CONTINUE

Conclusions