Stat159 HW03 Multiple Regression Analysis

Siyu Chen

10/14/2016

Abstract

The purpose of this assignment is to extend the scope of the previous HW. Not only we will look into relationships among variables, but also we will write funtions to compute some important statistics. In addition, we need to write unit tests to test our written functions.

Introduction

The overall goal is to examine the effect of TV, Newspaper and Radio advertising budgets on Sales by calculations and plotting graphs. If we do find a relationship, then we want to build a good linear model that can be used for Sales prediction based on TV, Newspaper or Radio advertising budgets.

Data

The advertising dataset consists of Sales(in thousands of units) of a particular product in 200 different markets, along with advertising budgets (in thousands of dollars) for the product in each of those markets for three different media (TV, Newspaper and Radio).

Methodology

We consider Sales vs TV, Newspaper and Radio advertising budgets in our dataset and try to fit them in a multiple linear regression model:

$$Sales = \beta_0 + \beta_1 TV + \beta_2 Newspaper + \beta_3 Radio$$

In order to estimate four coefficients β_0, β_1, β_2 and β_3 we fit the linear regression model via the least square criterion.

Results

Simple linear regression

Before analysis about multiple linear regression, it's good to examine the relationships between Sales and each of TV, Newspaper and Radio advertising budgets.

	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: Regressing Sales on TV advertising budgets

As we can see, the p-value of predictor is pretty small, which is 0. So we can infer that there is an association between the predictor(TV advertising budgets) and the response(Sales).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.3514	0.6214	19.88	0.0000
Newspaper	0.0547	0.0166	3.30	0.0011

Table 2: Regressing Sales on Newspaper advertising budgets

As we can see, the p-value of predictor is pretty small, which is 0.0011. So we can infer that there is an association between the predictor(Newspaper advertising budgets) and the response(Sales).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.3116	0.5629	16.54	0.0000
Radio	0.2025	0.0204	9.92	0.0000

Table 3: Regressing Sales on Radio advertising budgets

As we can see, the p-value of predictor is pretty small, which is 0. So we can infer that there is an association between the predictor(Radio advertising budgets) and the response(Sales).

According to simple linear regression, Sales has a linear relationship with each of TV, Newspaper or Radio advertising budgets independently

Multiple linear regression

Now, let's do some multiple linear regression analysis .

First, we should examine least squares coefficient estimates.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.9389	0.3119	9.42	0.0000
TV	0.0458	0.0014	32.81	0.0000
Radio	0.1885	0.0086	21.89	0.0000
Newspaper	-0.0010	0.0059	-0.18	0.8599

Table 4: Coefficient estimats of the least squares model

As we can see, the coefficient estimate for **Newspaper** in the multiple regression model is close to zero, and the corresponding p-value is no longer significant, which is 0.86. This is different from the result of simple linear regression.

In fact, it makes sense that multiple regression suggest no relationship between **Sales** and **Newspaper** while the simple linear regression implies the opposite.

Let's look at the following correlation matrix. The correlation between **radio** and **newspaper** is 0.35. This reveals a tendency to spend more on newspaper advertising in markets where more is spent on radio advertising. In a simple linear regression which only examines **Sales** versus **Newspaper**, we observe that higher values of **Newspaper** tend to be associated with higher values of **Sales**, which is because **Newspaper** is a surrogate for **Radio** advertising; **Newspaper** gets "credit" for the effect of **Radio** on **Sales**.

	TV	Radio	Newspaper	Sales
TV	1.00	0.05	0.06	0.78
Radio	0.05	1.00	0.35	0.58
Newspaper	0.06	0.35	1.00	0.23
Sales	0.78	0.58	0.23	1.00

Table 5: Correlation matrix

Now, let's answer following questions.

	Quantity	Value
1	Residual standard error	1.69
2	R squared	0.90
3	F-statistic	570.27

Table 6: RSE, R-squared and F-statistic of the least squares model

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

The F-statistic for the multiple linear regression model obtained by regressing **Sales** onto **TV**, **Newspaper**, **Radio** is 570.27. Since this is far larger than 1, it suggests that at least one of the advertising media must be related to **Sales**.

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

As we discussed before, if the significance level is 0.05, the p-value of **Newspaper** is **Table4** is larger than our significance level. So only a subset of the predictors (**TV** and **Radio**) are useful.

How well does the model fit the data?

Two of the most common numerical measures of model fit are the RSE and R-squared. As we can see, in **Table 6**, the R-squared is 0.9, which is very close to 1. So the model explains a large portion of the variance in the response variable and fits well.

The RSE is 1.69, which is less than the RSE of model that contains only **TV**, 3.26. This corroborates our previous conclusion that a model uses TV and Radio advertising budgets to predict Sales is much more accurate.

How accurate is the prediction?

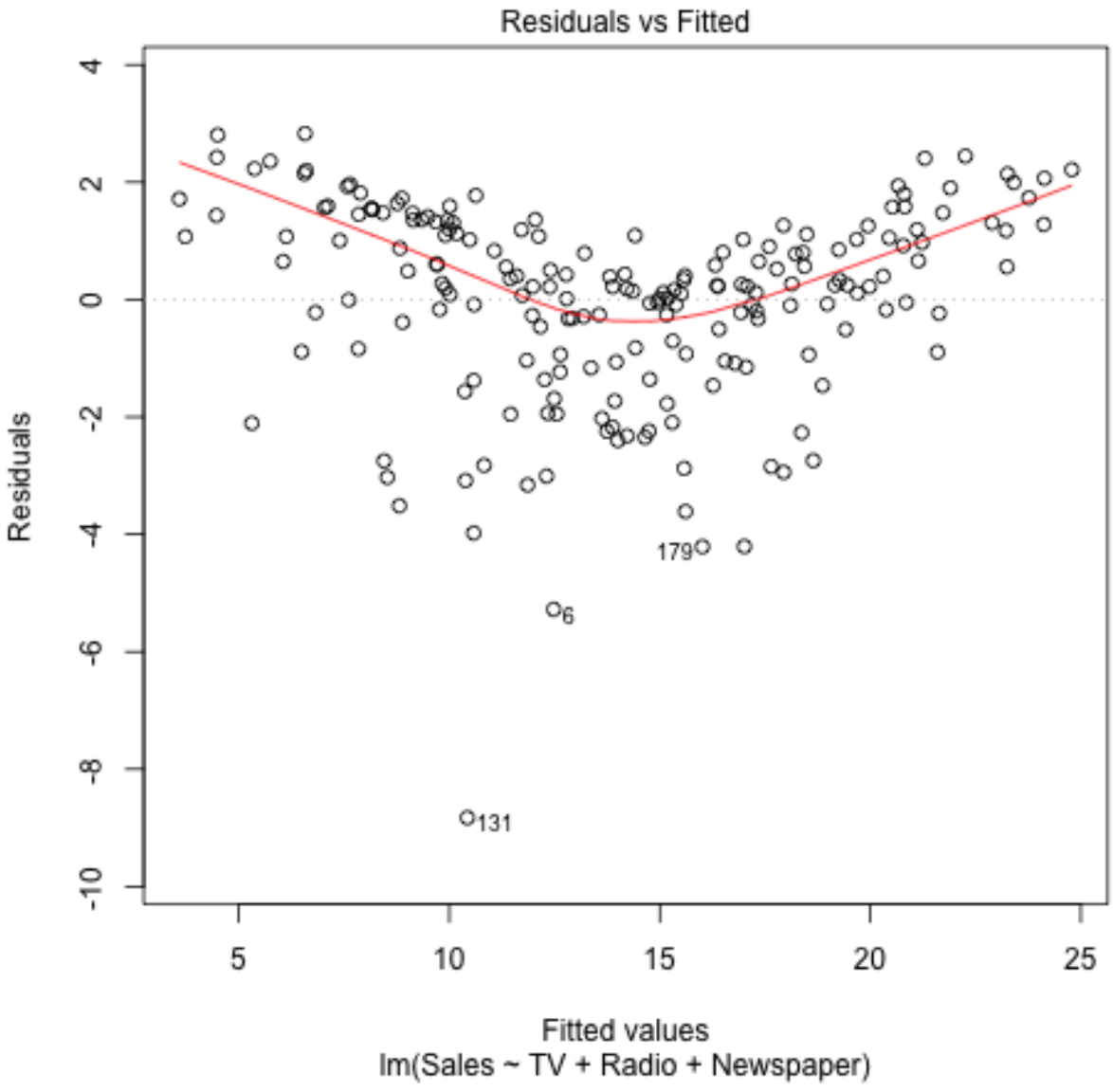

Figure 1: residual-plot

According to residual plot, we can say that our prediction is pretty accurate when the expenditure of TV, Newspaper and Radio is between 10 thousands of dollars and 20 thousands of dollars. But it becomes less accurate when the expenditure is below 10 thousands of dollars or above 20 thousands of dollars.

Conclusion

Through comparison between simple linear regression and multiple linear regression, we find that they may lead to different conclusions. With correlation of **Newspaper** and **Radio**, we can conclude that Newspaper does not help predicting Sales. Also, by examing R-squared and RSE, we are confident that our model fits pretty well. At last, we make a estimation that our model is accurate when the expenditure is between 10 thousands of dollars and 20 thousands of dollars.