Multiple Linear Regression Analysis

PHILHOON OH

October 11, 2016

Abstract

In this report, I will reproduce the main results displayed in section 3.1 *Simple Linear Regression* (chapter3) of the book **An introduction to Statistical Learning.** descirbed below :

- Table 3.3 (page 72): Coefficient estimates of simple regression models: Sales on TV, Sales on Radio, and Sales on Newspaper. The book only shows two tables (those of Radio and Newspaper) but also include the table for TV.
- Table 3.4 (page 74): Coefficient estimates of the least squares model.
- Table 3.5 (page 75): Correlation matrix.
- Table 3.6 (page 76): RSE, R^2 and F-statistic of the least squares model.

Introduction

The goal of this homework is to reproduce the results of Explanatory Data Analysis of Sales onto TV, Radio, Newspaper advertising budget. In order to see assoication among those variables, by applying multiple regression analysis We will answer questions described below :

1. Is at least one of the predictors useful in predicting the response?
2. Do all predictors help to explain the response, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. How accurate is the prediction?

Data

Data set consists of the Sales(in thousand of units) of a certain products in 200 different markets, and the advertising budgets of 3 differnet media: TV, Radio, and Newspaper (4 variabales and 200 observations.) It could be redownloaded; however, the data set must stays in the data directory. Other data set in data directory are produced by runing eda-script.R, regression-script.R.

Methodology

First, a simple linear regression model will be applied to see the linear relationship between sales and each of TV, Radio, Newspaper:

$$Sales = \beta_0 + \beta_1 TV$$

$$Sales = \beta_0 + \beta_1 RADIO$$

$$Sales = \beta_0 + \beta_1 NEWSPAPER$$

To estimate each of the coefficients β_0 and β_1, we fit a regression model via the least squares criterion (Best Linear Prediction method).

Second, a multiple linear regression model will be applied to see the linear relationship among sales, TV, Radio, Newspaper at once:

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

To estimate the coefficients β_0, β_1, β_2, β_3 we fit a multiple regression model via the least squares criterion (Best Linear Prediction method).

Lastly, by comparing these two methods and anaylzing its statistics, we are going to conclude which method is more appropriate to analyze the given data.

Results

1. The simple regression coefficients are discribed below in a Table 1-3:

	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: Simple regression of sales on tv

	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: Simple regression of sales on radio

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.3514	0.6214	19.8761	0.0000
news	0.0547	0.0166	3.2996	0.0011

Table 3: Simple regression of sales on newspaper

- In the Table 1, sales on tv, P-value of intercept is significant to reject the null. The estimate of the intercept is 7.0325935 with sd of 0.4578429. P-value of slope is significant to reject the null. The estimate of the slope is 0.0475366 with sd of 0.0026906
- In the Table 2, sales on radio, P-value of intercept is significant to reject the null. The estimate of the intercept is 9.3116381 with sd of 0.5629005. P-value of slope is significant to reject the null. The estimate of the slope is 0.2024958 with sd of 0.0204113
- In the Table 3, sales on newspaper, P-value of intercept is significant to reject the null. The estimate of the intercept is 12.3514071 with sd of 0.6214202. P-value of slope is significant to reject the null. The estimate of the slope is 0.0546931 with sd of 0.0165757

2. The multiple regression coefficients are discribed below in a Table 4:

Table 4 displays the multiple regression coefficient estimates. The interpretation is as follow: spending an additional $1,000 on tv advertising for a fixed amount of budgets of radio and newspaper, will lead to an increase in sales by approximately 45 units.

Now, compare this coefficient to those displayd in Table 1-3. The coefficient estimates of TV and Radio of multiple regression are similar to those of the simple regression; however, the estimate coefficient of

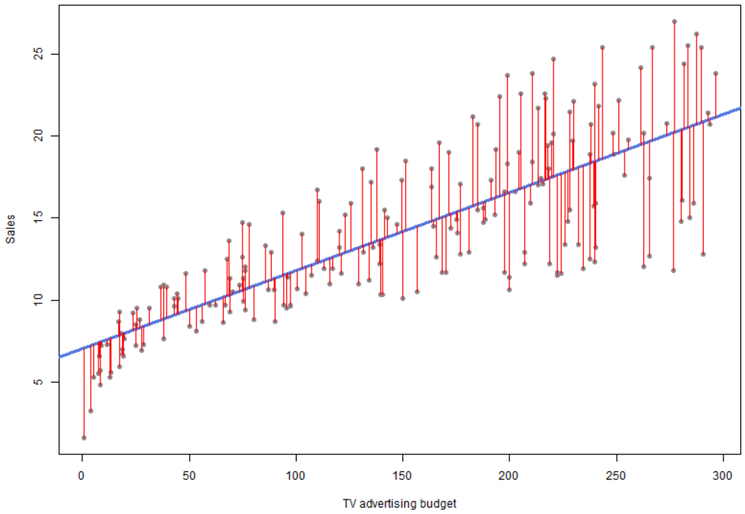
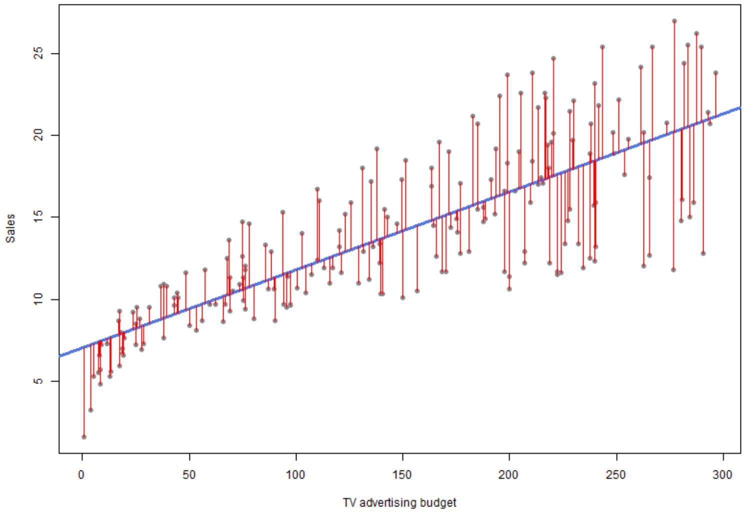


Figure 1: Scatterplot with fitted regression line

Figure 2: Simple linear regression sales onto tv



Newspaper in multiple regression is close to zero and pretty lower than that of single regression. Moreover the corresponding p-value, 0.8599151, is no longer significant.

Since the p-value is no longer significant, we do not reject the Null hypthosis, the slope of Newspaper is now considered to be zero. This is caused by the difference between simple and multiple regression method. The previous one only takes one varaible into consideration estimating sales while the latter one takes three variables into consideration to estimate sales.

Figure 3: Simple linear regression sales onto radio

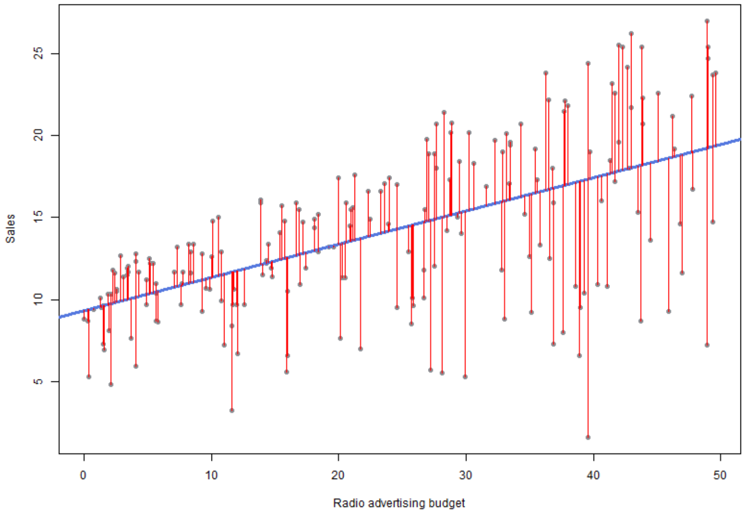
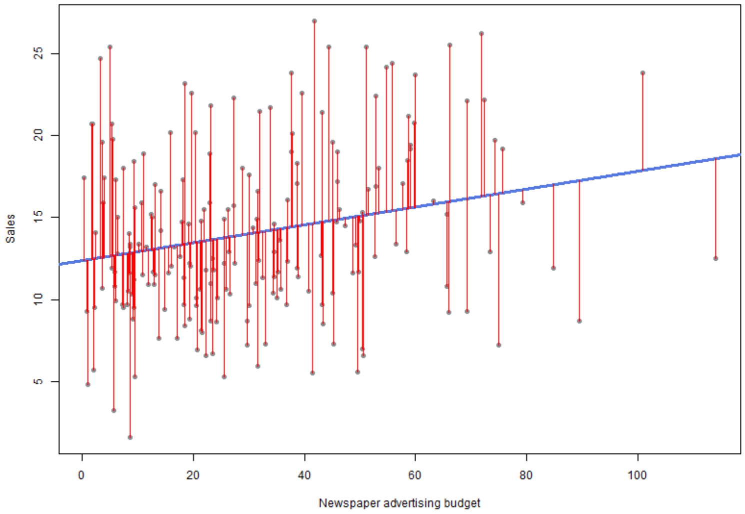


Figure 4: Simple linear regression sales onto newspaper



3. Correlation matrix for TV, radio, newspaper, and sales for the Advertising data are discribed below in a Table 5:

As mentioned previously, multiple regression shows that there is no relationship between sales and newspaper while simple regression does not. In this case, the multiple regression is a better method to explain the relationship for a given data set. Table 4 suggests that there is strong correlation between radio and newspaper, 0.3541 and it means that markets where more money is spent on radio advertising tend to spend more on newspaper advertising. In simple linear regression, newspaper advertising might look like contributing to sales increasement but the underlying fact that actually affects sale increasement is radio advertising.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.9389	0.3119	9.4223	0.0000
tv	0.0458	0.0014	32.8086	0.0000
radio	0.1885	0.0086	21.8935	0.0000
news	-0.0010	0.0059	-0.1767	0.8599

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

Figure 5: residual-plot of multiple regression

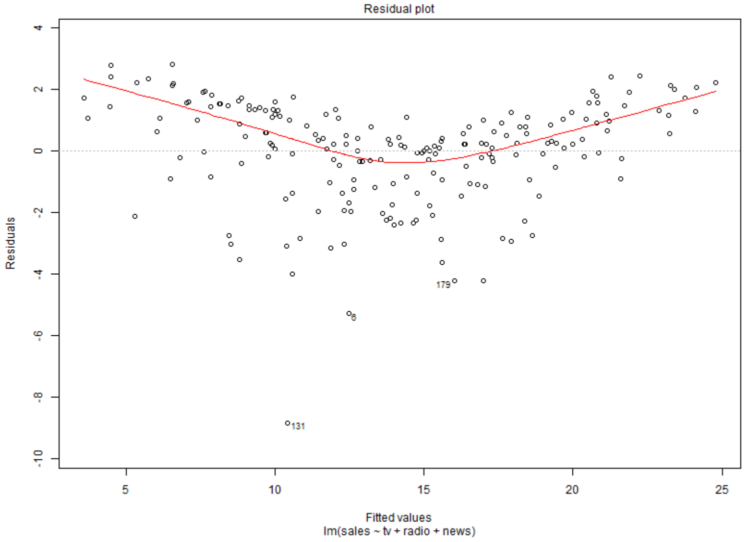
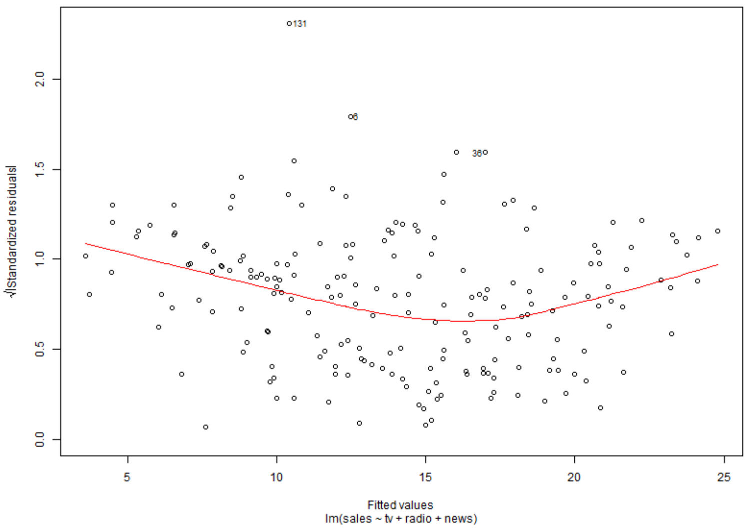


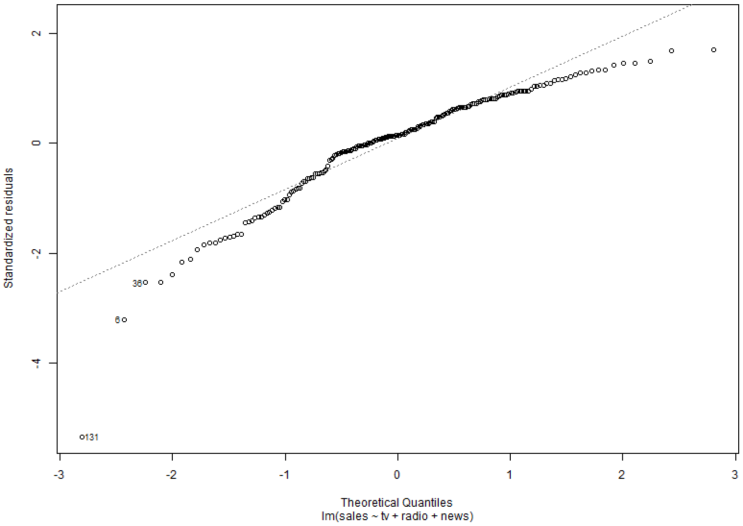
Figure 6: scale-location-plot of multiple regression



4. Quality indices RSE, R square, and F-statistic of multiple regression are given in the Table 6 below:

- R sqaure is the percent of variation that can be explained by the multiple regression equation. Here, it means that about 0.8972106 precent of variations in sales can be explained by the multiple regression

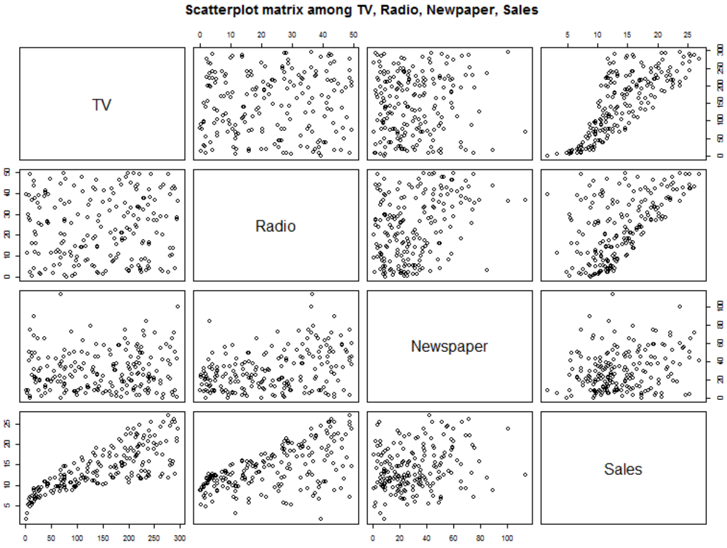
Figure 7: normal-qq-plot of multiple regression



	TV	radio	newspaper	sales
TV	1.00000	0.05481	0.05665	0.78222
radio		1.00000	0.35410	0.57622
newspaper			1.00000	0.22830
sales				1.00000

Table 5: Correlation matrix for TV, radio, newspaper, and sales for the Advertising data.

Figure 8: scatterplot-matrix



equation. When we fit the multiple regression model using just tv and radio advertising on sales, then R squared would be 0.8971943 which is higher than 0.8972106. Therefore, it would be better to exclude the newspaper to fit the multiple regression model to the given data.

- Residual Standard Error refers the estimated standard error of residuals. It measures the distance

	Quantity	Value
1	R2	0.8972
2	RSE	570.2707
3	F-stat	1.6855

Table 6: Multiple Regression Quality Indices

between the data point and the regression equation. Here, RSE is 1.6855104. When we fit the multiple regression model using just tv and radio advertising on sales, then Residual Standard Error would be 1.6813609 which is lower than 1.6855104. This suggests that this model would be more accurate than using all tv, radio, and newspaper variables. There is no need to include newspaper.

- Here F-statistics is obtained by applying multiple linear regression to sales onto radio, tv and newspaper. The value is 570.2707037, which is much larger than 1. Therefore, we reject the null hypothesis of all the regression coefficients, β_0, β_1, β_2, β_3, are zero. This suggests that at least one advertising media must be related to sales.

Conclusions

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

Yes. According to F-statistics, at least one of the predictors should be related to sales.

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

No. As mentioned previously, R-squared is higher and Residual Standard Error is lower when applying TV and Radio predictors to sales. This suggests that this model would be much accurate in predicting response than using all three media. Also, Table 4 suggests that the regression coefficient of newspaper is not significant. Therefore, based on these observations, it is safe to exclude the newspaper media. So the subset of the predictors, TV and Radio, is useful.

How well does the model fit the data?

Residaul Standard Error and the R square are the most common numerical figures of model fit. If R square is close to 1, the model explains that much of varaiation in the response variable. The smaller Residaul Standard Error indicates the better model. So, multiple regression using TV, Radio, and Newspaper on Sales is better model than using simple linear regression on each variables; however, it is not as good as just using TV and Radio as predictor variables on Sales, the response variable.

How accurate is the prediction?

Typically, prediction interval is larger than the confidence interval. Confidence interval is used to get the expected mean given predictor values whereas prediction is using one sample to create regression equation that would predict a value within a particular population. Thus, prediction includes the error in the estimate for true population regression plane and the uncertainty resulting from the distance between individual point from the population regression plane as well.