Multiple Linear Regression Analysis

Andrea Widjaja

10/14/16

Stat 159 Homework 3

Abstract

This paper consists of the reproduced results from *Section 3.2 Multiple Linear Regression* from the book **An Introduction to Statistical Learning** *by Gareth James, Daniel Witten, Trevor Hastie, and Robert Tibshirani.* This paper will discuss and examine topics of multiple linear regression analysis such as estimating the coefficients, finding a relationship between response and predictors, deciding on important variables, model fit, and predictions.

Introduction

Simpe linear regression is a useful tool to predict a response on the basis of a single predictor variable. Simple linear regressions are usually in the form

$$Y = \beta_0 + \beta_1 X$$

When taking into account to error, the simple linear regression will be:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

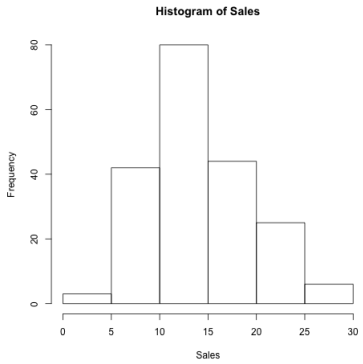
The single predictor variable X are independet variables, whereas the response variable Y are dependent variables. This method of simple linear regression is a very simple approach to predict a quantitative response. In reality, however, there are often more than one predictor. If we approach this problem by fitting a seperate simple linear regression model for each predictor, the estimates generated might be misleading and inaccurate. This is why it better to utilize Multiple Linear Regression.

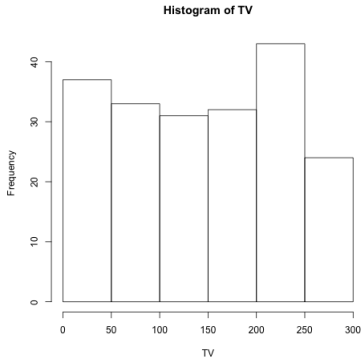
Multiple Linear Regression is a predictive analysis tool that can interpret relationships between one dependent variable and several independent variables. Hence, it gives us the ability to predict the value of a variable based on the value of several variables.

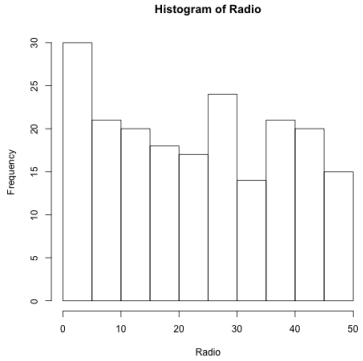
Data

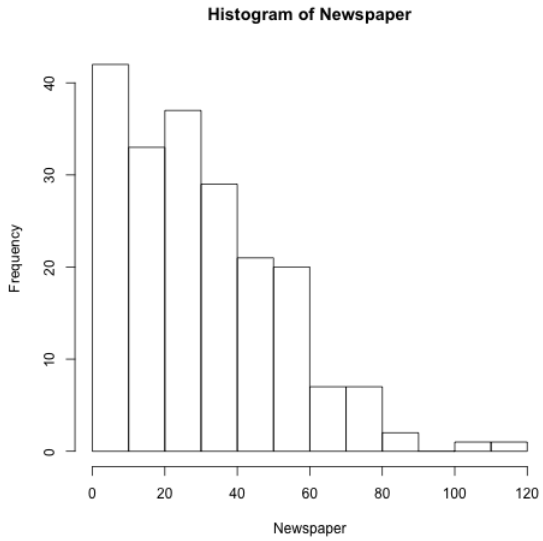
The data set given is *Advertising Data Set*, downloaded from `http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv` called **Advertising.csv**. This data set contains the sales of a product (in thousands of units) across `200` different markets and the advertising budget (in thousands of dollars) for the product in each market. The advertising data contains budgets for three different advertising medias: `TV`, `radio` and `newspaper`. It has `200` records and `5 variables(index, TV, Radio, Newspaper, Sales)`. Each record contains the information of advertising budget for `TV`, `Radio`, and `Newspaper`, including the sales progress for each market.

To give a rough picture on what data we are dealing with, these are the histograms for Sales and the advertising medias TV, Radio and Newspaper.


Histogram of Sales
Frequency
0
20
40
60
80
0
5
10
15
20
25
30
Sales


Histogram of TV
Frequency
0
10
20
30
40
0
50
100
150
200
250
300
TV


Histogram of Radio
Frequency
0
5
10
15
20
25
30
0
10
20
30
40
50
Radio



Methodology

Suppose we know from above that a simple linear regression, taking into account for error, takes the form

$$Y = \beta_0 + \beta_1 X + \epsilon$$

We know that multiple regression have several predictor X variables. Say that we have p numbers of predictors, then the multiple linear regression model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_p X_p + \epsilon$$

where X_i represents the ith predictor, β_i indicates the correlation, β_0 as the intersection, and ϵ as the error. To make it more understandable, β_i represents the average increase in Y as X_i increases by `1 unit`, holding all other predictors fixed.

For our case, since we want to estimate the effect of advertising budgets of three different medias `TV`, `Newspaper`, and `Radio` onto `Sales`, then our multiple regression model can be represented like this

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

Estimating The Coefficients

Just like simple linear regression, we have to estimate the values of the regression coefficients $\beta_0, \beta_1, \beta_2, ..., \beta_p$, because they are unknown. Using `lm()` in R, we can directly get the estimated values $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2 and \hat{\beta}_3$. Additionaly, we can also use the `summary()` function to summarize the regression object, to get more statistical details. (Table 4)

F-Statistic

Having the same concept with simple linear regression, it is important to know if there is actually a relationship between the response variable and the predictor variable. To determine wheter there is a relationship, one should conduct the null hypothesis to test if $\beta_0 = \beta_1 = \beta_2 = ... = \beta_p = 0$.

Null hypothesis :

$$H_0 : There is no relationship between X i and Y$$

$$H_0 : \beta_i = 0$$

versus

Alternative hypothesis:

$$H_a : There is some relationship between X i and Y$$

$$H_a : \beta_i \neq 0$$

The hypothesis test is performed by computing the *F-statistic*,

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

where

$$TSS = \sum (y_i - \bar{y})^2$$

is the total sum of squares, and

$$RSS = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i 1 - \hat{\beta}_2 x_i 2 - ... - \hat{\beta}_n x_i p)^2$$

is used to estimate the least square fit for the regression.

If the Null Hypothesis is true, then

$$E(TSS - RSS)/p = \sigma^2$$

, showing that when there is no relationship between the response and predictors the *F-statistc* take on a value close to 1. If the Alternative Hypothesis is true, then

$$E(TSS - RSS)/p > \sigma^2$$

, meaning that F is expected to be greater than 1 when there is some kind of relationship between the reponse and predictors.

p-value

The *p-value* is widely used in null-hypothesis testing. When the null hypothesis is true, the *p-value* is the porbability of obtaining a result equal to or "more extreme" than what was actually observed. This can be used to determine whether or not to reject the null hypothesis H_0. Typically, a small *p-value* rejects the null-hypothesis, concluding that there is a relationship between the variables.

Results

We run three seperate simple linear regression, each of which uses a different advertising medium as a predictor.

Table 1: Regression of TV on Sales (note: Sales are in thousands of dollars, and TV is in thousands of units

	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

From the estimate prediction above, we can see that for a `$1,000` increase in TV advertising, there will be an increase in sales by around `50 units`.

Table 2: Regression of Radio on Sales (note: Sales are in thousands of dollars, and Radio is in thousands of units

	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 3: Regression of Newspaper on Sales(note: Sales are in thousands of dollars, and Newspaper is in thousands of units

	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

From the estimate prediction above (table 2), we can see that for a `$1,000` increase in TV advertising, there will be an increase in sales by around `203 units`.

From the estimate prediction above (table 3), we can see that for a `$1,000` increase in TV advertising, there will be an increase in sales by around `55 units`.

When doing three seperate simple linear regression, this ignores the other two media in forming estimates for the regression coefficients, which leades to misleading estimates of the individual media effects on sales. Therefore, we have to do multiple linear regression to get better results.

Table 4: Multiple Linear Regression Model

	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
Newspaper	-0.0010	0.0059	-0.1767	0.8599

This table displays the multiple regression coefficient estimates when all three predictor variables (TV, radio, and newspaper advertising budgets) are used to predict product sales. We can see that the results in this table, differ from the predicted coefficents using seperate simple linear regression. The estimated coefficeint for TV and radio are pretty similar, however there are significant differences for the newspaper regression.

We can see that the newspaper regression coefficient estimate is `non-zero` from the simple regression (Table 4), but close to zero in the multiple regression model. Additionaly, notice that the p-value is no longer significant, with a value around `0.86`.

Analysis Questions

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

Goal: F=statistic > 1, p-value=0

F-statistic:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

In our case, the F-statistic for the multiple linear regression model is 570. When F-statistic is way greater than 1, then this suggests that at least one of the advertising media is related to sales. In other words, this rejects the null hypothesis H_0.

An F-statistic that is jsut a little greater to 1, might still provide evidence agianst the null hypothesis. This, however, depends on the values of n and p. If n is large, then an F-statistic just a little greater than 1 might

provide evidence against the null hypothesis. Thus, a larger F-statistic is needed to reject the Null hypothesis H_0 if n is small.

Since we are give the value of n and p, we can compute the p-value associated with the F-statistic. Since the p-value associated with the F-statistic is almost equals to 0, then there is an extremely strong evidence that at least one of the media is associated with increasing sales.

Table 5: Correlation Matrix for The 3 Medias

	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

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

Goal: finding the fitted linear model containing variables with low p-value

From the question above, computing the F-statistic and examining the associated p-value, we can determine if there are at least one predictor related to the response. Now, we want to know which of these predictors are useful. One way of approaching this question is to take a look at the individual p-values (table 4). However, if p is large, then we are likely to make some false discoveries. It is possible to have all of the predictors associated with the response, however, most of the time, the response is only related to a subset of the predictors. Another way of selecting the "best" subset of predictors is by variable selection. We perform variable selection by trying out different models with different subset of predictors. In our case, we have `3` predictor variables X_1, X_2, and X_3.

$$Sales = \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 Newspaper + \epsilon$$
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

We can consider the following 8 models: 1. A model containing no variables, β_0 only 2. A model containing X_1 only 3. A model containing X_2 only 4. A model containing X_3 only 5. A model containing X_1 and X_2 6. A model containing X_2 and X_3 7. A model containing X_1 and X_3 8. A model containing X_1 X_2 and X_3

To approach this, there are 3 classical methods: * Forward selection - Starting with no variables in the model with only the intercept , we then proceed by adding a variable with the lowest RSS. This is continued until some stopping rule is satisfied. * Backward selection - Starting with all variables in the model, we proceed to the next step by removing the variable with the largest p-value. We refit the model, and remove the variable with the largest p-value again. We stop when all p-values are below a certain threshold. * Mixed selection - Combining Forward selection and Backward selection, Mixed selection start with no variables in the model, and add variables that provide the best fit one-by-one. As new predictors are added, the p-value for one of the variables in the model can get larger. If it rises above a certain treshold, then we should remove that variable. Continue to do this until all variables in the model have a low p-value, and all variables outside the model have a large p-value if added to the model.

The fitted linear model maximizes the correlation among all possible linear models. For our case, only the subset of predictors TV and Radio are useful.

3. How well does the model fit the data?

Goal : high R^2,

After finding the fitted linear model, we should the accuracy of the model. The quality of a linear regression fit is assessed using two related quantities: *Residual Standard Error* and $R^2 Statistic$.

Table 6: Regression Quality Indices

	Quantity	Value
1	RSE	1.6855
2	R^2	0.8972
3	F-statistic	570.2707

Residual Standard Error(RSE)

The RSE is an estimate of the standard deviation of ϵ. It is the average amount that the response will deviate from the true regression line. In simple words, it is a measure of the lack of fit of the linear model to the data.

$$RSE = \sqrt{(\frac{1}{n-p-1}) * RSS}$$

$$RSE = \sqrt{(\frac{1}{n-p-1}) * (\sum_{i=1}^{n}(y_i - \hat{y}_i)^2)}$$

When the RSE is small, then we can assume that for $i = 1, ..., n$, $\hat{y}_i = y_i$. This means that we can conclude that the model fits the data very well. On the contrary, if $\hat{y}_i$ is very far from y_i, for one or more of the observations, then the RSE will be large, indicating that the model does not fit the data very well.

For our case, the Residual Standard Error is `1.69`, and cosidered small, then our model fits the data well.

R^2 Statistic

The $R^2 Statistic$ provides an alternative measure of good fit to the RSE. It is a measure of the linear relationship between X_i and Y, with a range of [0,1].

Recall that in simple linear regretssion, R^2 is identical to the squared correlation. In mulitple linear regression, it is equals to

$$Cor(Y, \hat{y})^2$$

the square of the correlation between the response and the fitted linear model.

The formula of R^2 is given by:

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

where,

$$TSS = \sum (y_i - \hat{y}_i)^2$$

is the total sum of squares.

An R^2 value close to 1 shows that the model explains a large portion of the variance in the response variable. R^2 always increase as we add variables to the model. Generally, the higher the R^2 the better the model fits the data.

If R^2 is near 0, then our linear model does not fit the data well. On the contrary, if R^2 is near 1, the our linear model fits the data well.

In our case, the R^2 is `0.90`, which indicates that our model does fit the data well.

4. How accurate is the prediction?

Goal: calculate Confidence Intervals, and Prediction Intervals

Suppose $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p$ are coefficient estimates for $\beta_0, \beta_1, \beta_2, ..., \beta_p$, then we have the least squares plane

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_p X_p$$

is an extimate for the true population regression plane

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

From this, we can compute *confidence intervals* in order to determine how close $\hat{Y}$ will be to $f(X)$. Additionally, we can also compute *prediction intervals* to compute how much Y will vary from $\hat{Y}$.

For our case, *confidence interval* can be used to quantify the uncertainty surrpounding average `sales` over a large number of cities. *prediction intervals* can be used to quantify the uncertainty surrounding `sales` of a *particular* city.

Conclusions

This paper contains indepth information regardin Multple Linear Regression. By modelling multiple linear regression, I have come to a better understanding of the effect of each predictor variables onto the response variable. For this case, the predictor variables that are useful are Radio and TV. Hence, to gain a fitted model with satisfying RSE and R^2 values, we have to remove the Newspaper variable.