Multiple Linear Regression

Nura Kawa

October 14, 2016

Abstract

This paper reproduces Section 3.2 of *Introduction to Statistical Learning* by Hastie, Tibshirani, James, and Witten. This section discusses the Multiple Linear Regression model and fits one to the *Advertising* dataset. The model uses data on television, newspaper, and radio advertising to predict units of sales.

Introduction

We return to the *Advertising* dataset, where we want to build a model from data on three advertising media to predict Sales. A past paper discusses the Simple Linear Regression model; so, one solution is fit three Simple Linear Regression models - one for each predictor. However, this is not the only option - we can simply tweak the linear model to allow for more predictors, using a **Multiple Linear Regression Model**.

A Multiple Linear Regression model extends the Simple Linear Regression model to accomodate more predictors for a single response variable. Recall the Linear Regression model we have previously encountered:

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

,

where Y is our response variable, X is our predictor, ϵ is random noise, and the beta coefficients are the intercept and slope of our linear model, respectively.

The Multiple Linear Regression model is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon$$

,

where X_j represents the jth predictor and β_j quantifies the association between that variable and the response.

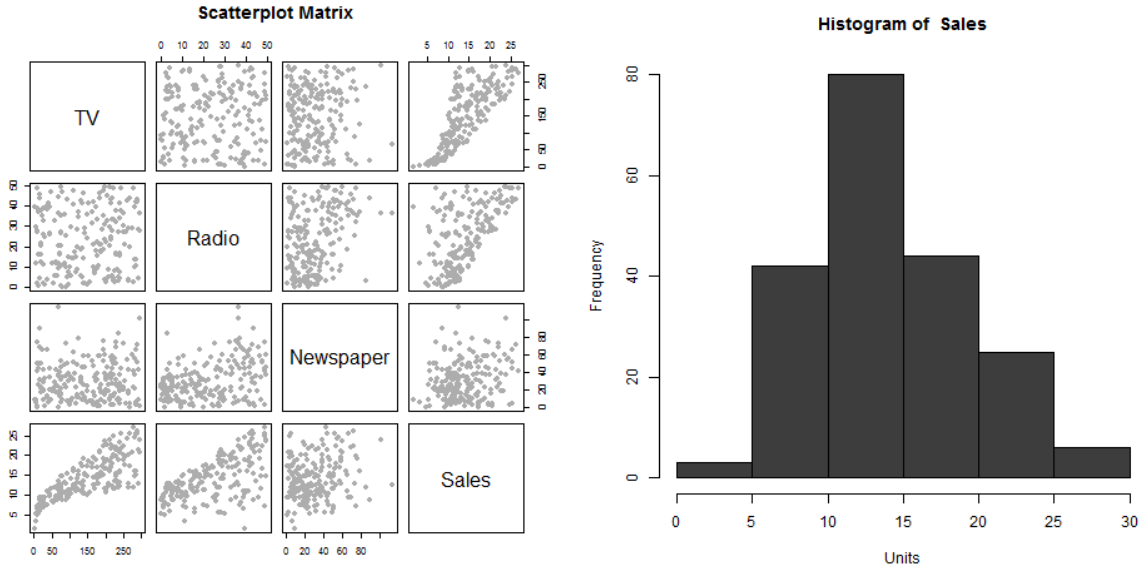
Data

Before fitting any model it is imperative that we discuss our 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*, *Radio*, and *Newspaper*.

Below is the first five rows of the Advertising Dataset:

```
##      TV Radio Newspaper Sales
## 1 230.1  37.8      69.2  22.1
## 2  44.5  39.3      45.1  10.4
## 3  17.2  45.9      69.3   9.3
## 4 151.5  41.3      58.5  18.5
## 5 180.8  10.8      58.4  12.9
```

It is also a good idea to look at some data visualizatoins:



Finally, because we are using multiple predictors, it is important to see their level of correlation. Predictors that are too highly correlated may cause problems in making predictions, as one predictor can be easily interchanged with another.

Figure 2 Correlation matrix for TV, radio, newspaper, and sales for the *Advertising* dataset.

	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

Methodology

We discuss how to specifically fit a multiple linear regression model and compare the outcome to those of three simple linear models.

Fitting Simple Linear Regression Models

We use the methodology discussed in the paper *Simple Linear Regression* by Nura Kawa (reproduced from Section 3.1 of *Introduction to Statistical Learning*) to fit a simple linear model for each predictor with Sales. The results are as follows:

Figure 1 Simple Regression Models for the *Advertising* dataset. The three tables below show coefficients of the linear regression models of Sales regressed onto Radio, Newspaper, and TV, respectively.

	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 1: Simple Regression of Sales on Radio

	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: Simple Regression of Sales on Newspaper

	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 3: Simple Regression of Sales on TV

This solution to our question of prediction leaves much to be desired, as it becomes difficult to make one estimate using three completely different linear models. Furthermore, each model leaves out a possible influence from a separate predictor. Thus, we will fit a Multiple Linear Regression Model.

Fitting a Multiple Linear Regression Model

Specifically, our Multiple Linear Regression model is the following:

$$\textbf{sales} = \beta_0 + \beta_1\textbf{TV} + \beta_2\textbf{radio} + \beta_3\textbf{newspaper} + \epsilon$$

.

As with simple linear models, we first test hypotheses, namely:

Hypothesis Testing

Our *null hypothesis* H_0 is: H_0: $\beta_1 = \beta_2 = ... = \beta_p = 0$

Our *alternative hypothesis* H_a is: H_a: at least one β_j is non-zero

Simply put, we assume the null hypothesis of each predictor having no influence on Sales, and test to see if at least one predictor has an association with Sales. We set our p-value to be 0.05, meaning that anything below this would allow us to reject our null hypothesis.

Computing Coefficients

As with a simple linear model, we will compute the following coefficients as follows:

Residual Sum of Squares (RSS)
Residuals are the difference between the observed value of the dependent variable y and the predicted value, $\hat{(y)}$. The Residual Sum of Squares is simply the summation of all residuals, squared:

$$RSS = \sum(y_1 - \hat{y_i})^2$$

Total Sum of Squares (TSS)
The total sum of squares explains the variability already present in the dataset: it is the sum of the difference between each value of y (Sales, in our case) and its mean, squared:

$$TSS = \sum(y_1 - \bar{y})^2$$

R-Squared
The R^2 measures the goodness of fit of our model. It measures the proportion of variance explained by our model. Thus it has the range [0,1] and is independent of Y's scale.

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

Residual Standard Error
The RSE is an estimate of the standard deviation of the error in our model. This shows how far our data will deviate from the generated regression line.

$$RSE = \sqrt{\frac{1}{n-2}\sum (y_i - \hat{y_i})^2}$$

F-Statistic
This is our test statistic: Allowing p to be the number of predictors and n to be the number of observations, we have:

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

Model Selection

Now that we have discussed how to fit the model, we also need to pay attention to model selection. In cases where we have many predictors, using all predictors may not always be ideal. Our goal of a good model is to **minimize our p-value** and to **minimize RSS**. The following three methods allow us to select a model that approaches our goals.

Forward selection

We begin with a model that has only intercept terms. We iteratively fit all p (where p is number of predictors) possible Simple Linear Regression models, measure our RSS, and then add the variable that minimizes the RSS, until we reach our stopping point.

Backward selection

This is the same as Forward Selection, but we begin with *all* predictors and iteratively remove the predictors that yield the largest p-value.

Mixed selection

This method combines the previous two. We first forward select, then, if the p-value becomes too large due to the addition of one variable, we remove that variable from our model. This iterative method has the goal of dually minimizing RSS and the p-value. This may not result in the optimum of either, but allows one to balance out the negatives of both forward and backward selection.

To assess **model accuracy**, we can also look at the R^2 and the RSE coefficients, which are explained previously.

Results

We fit a Multiple Linear Regression model on the *Advertising* data, using TV, Radio, and Newspaper to predict Sales. The results are shown below:

Figure 3 Multiple Linear Regression Coefficients___

	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: Multiple Linear Regression of Sales onto Radio, TV, and Newspaper

Figure 4 More information about the least squares model for the regression of number of units sold on TV, newspaper, and radio advertising budgets in the *Advertising* dataset.

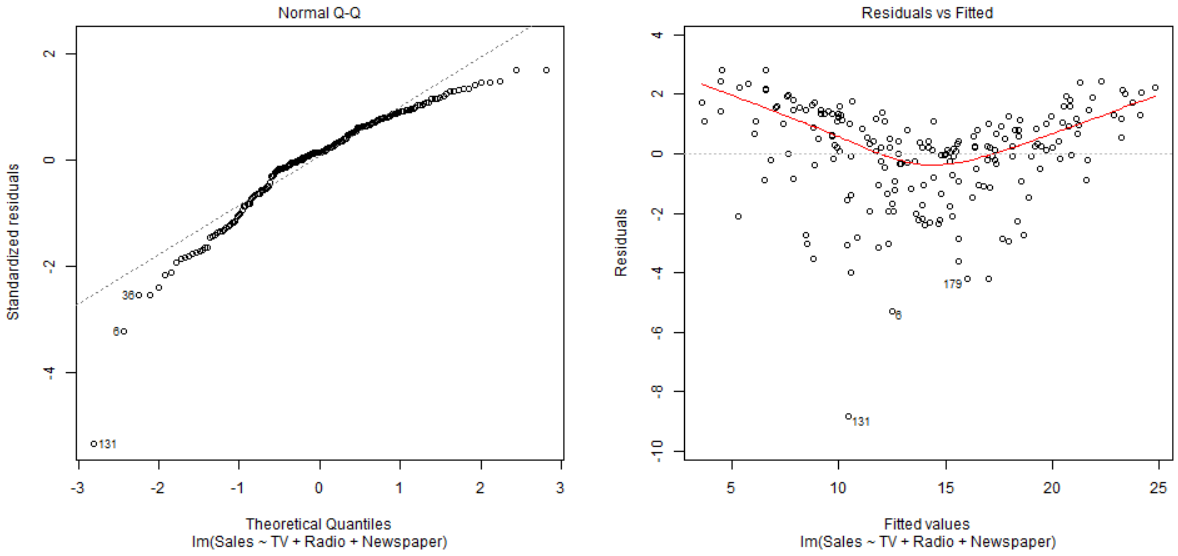
Quantity	Value
Residual Standard Error	1.68
R-Squared	0.90
F-Statistic	570.27

Table 5: Multiple Linear Regression Coefficients

Checking Model Assumptions

The Multiple Linear Regression Model assumes **independent predictors**, **normally distributed variables**, and **constant variance**.

The images below are part of an evaluation of our Model. The **QQ-Plot** checks the assumption of normality. We see that our valus are indeed normally distributed, as they fit the theoretical quantiles. The *residuals*, or the difference between the actual values of y and our predicted values, $\hat{y}$, are interesting to look at. Namely, **it is important that we check whether there exists a non-linear relationship in our data**. We look at the *heteroschedasticity* of our residuals - whether or not they display a pattern.



Our residuls show a non-linear (nearly quadratic) pattern by the curve of the red line, meaning that perhaps a different model could better explain the data.

Checking our Model Fit

Like in Simple Linear Regression, we look at the R^2. Our value of 0.9 is very good and indicates a much better-fitting model than the Simple Linear Model with TV used in a previous paper. To interpret an **F-Statistic** we look at its size - since this is a ratio, we hope for a value larger than 1. We have 570.27) reveals that this model is useful because a relationship definitely exists between our dependent and independent variables. Our **RSE** is not too large to discredit the model, meaning that it is a good fit.

Now we look at **p-values**. Ourp-values are very low for all predictors but Newspaper, meaning that we could leave out Newspaper and perhaps get a better fit. Thus we can **reject the null hypothesis**: at least one predictor has an association with our response, Sales.

Conclusion

We fit a Multiple Linear Regression model and found that for all predictors but Newspaper very small p-values, allowing us to reject our null hypothesis: at least one predictor has a positive association. Specfically, TV and Radio have positive association with Sales. However, this model is not perfect; we see that our residuals are heteroschedastic, meaning that variance of the residuals is not constant. There appears to be an underlying non-linear relationship in our Sales predictor that cannot be explained by a linear model. Thus, while we have progressed in fitting a model to our *Advertising* dataset, we must continue trying different methods to get an optimal prediction.