Multiple Linear Regression Analysis

Lauren Hanlon

October 7, 2016

Abstract

For this project I will refer to the book *An Introduction to Statistical Learning* (by James et al).

My goal for this project was to conduct a multiple regression analysis using the Advertising.csv dataset to look at sales across various products as a function of advertising budget, specifically referring to Newspaper, Radio and TV media spend.

In particular, we were answering the questions:

1. Is at least one of the predicotrs 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?

Introduction

In order to answer the questions regarding predictors and the response, we need to not only look at the specific relations between said predictors and the response, but also at the interaction between the predictors and the response. To analyze this interaction, I will use a *multivariate linear regression model*, which predicts a quantitative repsonse Y on the basis of multiple predictor variables *X1*, *X2*, *X3*, etc.

For this analysis looking at Newspaper, Radio and TV media spend versus sales, our predictor varialbes will be Newspaper, Radio and TV media spend, and our Y varialbe will be sales.

`sales` $= \beta_0 + \beta_1$ x `newspaper` $+ \beta_2$ x `radio` $+ \beta_3$ x `TV`

Data

The Advertising data set is 200 x 5 in dimensions. There are 200 rows, each row being a unique item and there are 5 columns:

- `X`: index
- `newspaper`: Advertising budget on newspaper (in thousands $)
- `radio`: Advertising budget on radio (in thousands $)
- `TV`: Advertising budget on TV (in thousands $)
- `sales`: Product sales (in thousands $)

The table contains `sales` in thousands of units for a particular product as a function of advertising budgets (in thousands of dollars).

Methodology

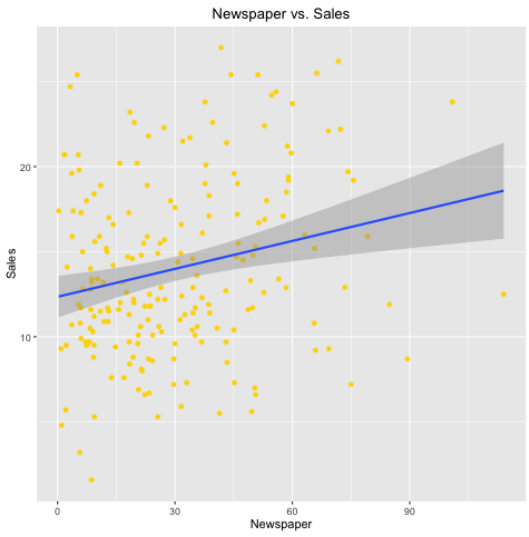
Coefficient estimates of simple regression models

First, I looked at each media source and it's individual relationship with sales.

Newspaper

Table 1: Simple regression of sales on newspaper

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.351	0.621	19.876	0
Newspaper	0.055	0.017	3.300	0.001



Equation for `sales` ~ `newspaper`:

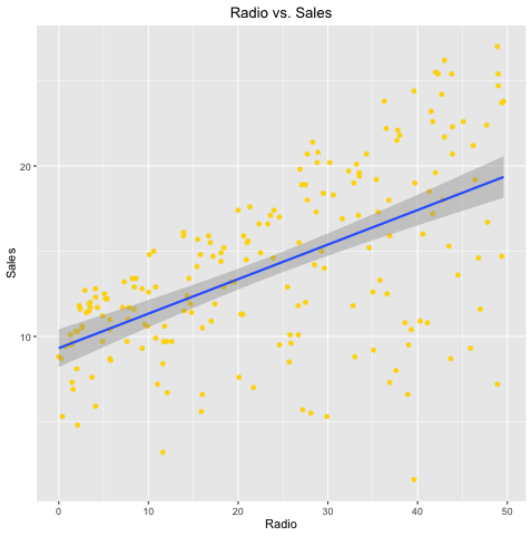
`sales` = 12.351 + 0.055 x `newspaper`

The graph as well as the coefficient of 0.055 indicates a positive correlation between sales and newspaper media spend. From the scatterplot we can tell that there is a higher density of points towards the left of the graph, with few points past 90 and a much higher standard deviation as newspaper media spend increases. Relative to the other predictors, we will soon see that newspaper has a higher standard error than radio and TV. Also note that the y-intercept is 3 points above any of the other predictors, which means that for $0 newspaper media spend, we're still seeing $12,351 in sales being reported.

Radio

Table 2: Simple regression of sales on radio

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.312	0.563	16.542	0
Radio	0.202	0.020	9.921	0



Equation for `sales` ~ `radio`:

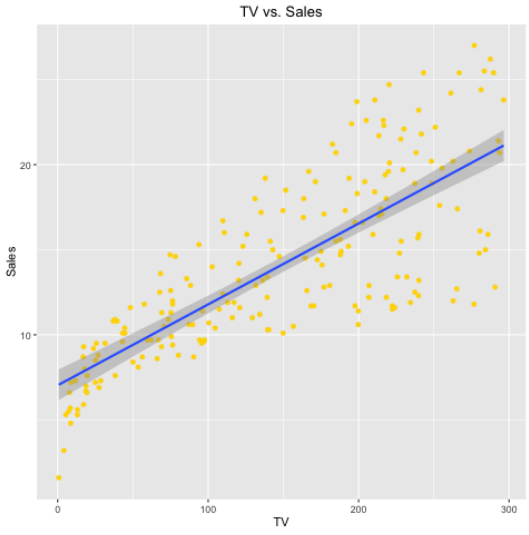
`sales` = 9.312 + 0.202 x `radio`

The correlation coefficient is highest for radio at 0.202 in relation to the other predictors, with a low standard error, indicating a strong correlation between sales and radio media spend.

TV

Table 3: Simple regression of sales on TV

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.033	0.458	15.360	0
TV	0.048	0.003	17.668	0



Equation for `sales` ~ `TV`

`sales` = 7.033 + 0.048 x `TV`

The linear model for TV has the lowest standard error amonst each of the predictors, as well as the lowest y-intercept, indicating that although it's correlation coefficient isn't the highest, that it might have a stronger effect on sales than if we just looked at the coefficients individually. We will be investigating this in the rest of the report.

Coefficient estimates of the least squares model

Table 4: Coefficient estimates

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.939	0.312	9.422	0
TV	0.046	0.001	32.809	0
Newspaper	-0.001	0.006	-0.177	0.860
Radio	0.189	0.009	21.893	0

This table shows the least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper media spend. How we should interpret this is that the coefficients for each of these represent the average effect of increasing that particular predictor, while holding all other predictors constant; e.g. the coefficient for `TV` is the average effect of increasing TV media spend by $1,000 while holding `newspaper` and `radio` fixed.

What we note here is that the coefficient for `newspaper` is significantly close to zero, and actually has a negative effect on sales when compared against the other predictors

The main takeaway from this table should be that the coefficients can vary when you're looking at single versus multiple linear regressions, and that when dealing with multiple variables, a multi regression should be taken into account.

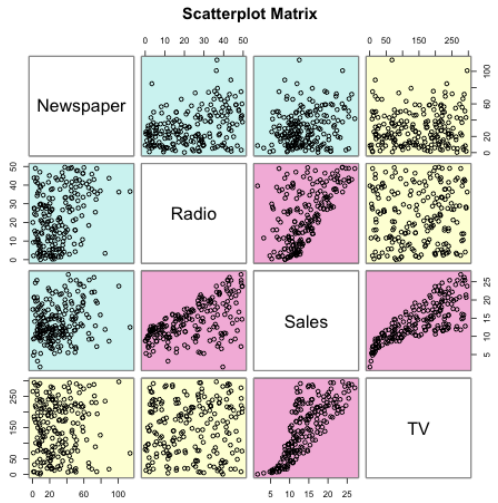
Correlation matrix

Table 5: Correlation matrix

	TV	Sales	Radio	Newspaper
TV	1	0.782	0.055	0.057
Sales	0.782	1	0.576	0.228
Radio	0.055	0.576	1	0.354
Newspaper	0.057	0.228	0.354	1

In this correlation matrix table we can clearly tell that `TV` has the strongest correlation with `sales` (0.782) while `radio` (0.576) and `newspaper` (0.228) are weaker. We can also note the fact that there is a trend for an increased newspaper media spend where more is spent on radio media. The bottom line is that, if we were to look at simple linear regressions alone, we would not have noticed such a weak relationship between newspaper media spend on sales.

These relations are represented visually in the scatterplot matrix graph below. The pink scatterplots represent those with the highest correlations - namely `TV` ~ `sales` and `radio` ~ `sales`, whereas the relationship between `newspaper` and `sales` is represented as blue, indicating a very weak correlation.



Residual Sum of Squares

The residual sum of squares measures the amount of variability that is left unexplained after performing the regression. To calculate this, you sum up the residual coefficients and predictor coefficients, square each of these sums then add them all together.

```
residual_sum_squares(reg)
```

```
## [1] 556.8253
```

R^2

The R^2 statistic takes the form of a *proportion*-the proportion of variance explained as a value between 0 and 1, which is independent of the scale of Y. Essentially it measures the *proportion of variability in Y that can be explained using X*. A higher number indicates that a higher proportion of the variability can be explained by the regression. In this case with a $R^2 > 0.8$, we can assume that the regression does a fairly accurate job of fitting the data.

$R^2 = 1$ - RSS/TSS

```
r_squared(reg)
```

```
## [1] 0.8972106
```

F-statistic

If there is no relationship between the predictors and the response, then the F-statistic would take a value closer to 1. Since we have a very large F-statistic of over 500, we can assume that at least one of the predictors is related to our response; in other words at least one of `TV`, `radio`, or `newspaper` is related to `sales` (which we already knew).

F-statistic = (TSS - RSS)/p / RSS/(n - p - 1)

```
f_statistic(reg)
```

```
## [1] 564.4516
```

We can see all of these results organized in the table below.

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Oct 13, 2016 - 15:20:17

Table 6: Summary of Multiple Regression Model

	Dependent variable:
	Sales
TV	0.046***
	(0.001)
Newspaper	−0.001
	(0.006)
Radio	0.189***
	(0.009)
Constant	2.939***
	(0.312)
Observations	200
R^2	0.897
Adjusted R^2	0.896
Residual Std. Error	1.686 (df = 196)
F Statistic	570.271*** (df = 3; 196)
Note:	*p<0.1; **p<0.05; ***p<0.01

Results

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

   Because of our *F-statistic* is so high, we can conclude that at least one of the predicotrs is useful in predicting the response.

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

   Looking at each of the predictors individually, we can see that they exude a positive correlation, which leads us to believe that all of them are useful in explaining the response. Looking at the data more closely, we found that there exists a very strong correlation between TV media spend and sales. With an estimate of 0.046 and a standard error of 0.001, we can conclude that TV is useful in explaining the response.

3. *How well does the model fit the data?*

   We look at the R^2 statistic to assess whether the model fits the data. The closer to 1 the R^2 is, the better job the model does at fitting the data. We saw an R^2 statistic of 0.897, which indicates that the model does an accurate job of fitting the data (but there are room for improvements:)

4. *How accurate is the prediction?*

   We can see the p-values associated with each variable in Table 6. We can see that TV and Radio have statistically strong p-values of less than 0.01, whereas newspaper gives us a p-value of 0.860, indicating that it is not an accurate predictor. Holding all variables constant we see a p-value of less than 0.01, which indicates that the model is accurate in it's prediction.

Conslusions

In conclusion, we learned two things. The first is that James' *An Introduction to Statistical Learning* is reproducible, since we were able to obtain all of the same results (yay!). The second is that when you're performing data analysis, it is important to look at all of the factors and their relations versus focusing on a subset of variables that are relevant to your particular project. For example, had we only looked at the relationship between newspaper and sales we might have incorrectly concluded that newspaper had a strong effect on sales. However, when you look at all the data in a multiple regression model, we were able to determine the right relationships between each of the predicotrs and the response.