Multiple Least Squares Regression Report

Joseph Francia
Professor Gaston Sanchez
Stats 159
14 October 2016

Abstract

In this report, I'm going to first examine datasets corresponding to the following variables: sales, newspaper advertising, radio advertising, and tv advertising. Afterwards, I am going to show the relationship between these various modes of advertising and sales by running a multiple regression of sales on tv advertising, radio advertising, and newspaper advertising. I will then discuss the relevant regression statistics from this regression I ran.

Introduction

All the internet products that I use are free. **Facebook**, **Gmail**, and **Google**, amongst many other technology products, are all free to use. The only source of revenue for these companies is essentially advertising revenue. So how effective is it to advertise? Moreover, how would one go about testing the effectiveness of advertising? In this report, I am going to examine a dataset that contains data on sales and different forms of advertising.

Data

Before we have a discussion on how advertising (our explanatory variables) interacts with sales (our response variable) through linear regression, lets first take a look at a correlation matrix between all four of our variables. An entry in a correlation matrix details the level of correlation between two variables. It turns out that all of our variables are positively correlated with each other to varying degrees. The strongest correlation (.782) exists between sales and television advertising. Meanwhile, the weakest correlation (.054) exists between television and newspaper advertising. These results seem to suggest that television advertising could be quite useful in predicting the response. The correlation matrix also seems to suggest that newspaper advertising might not help much in explaining our response variable. I've shown the correlation matrix below:

```
## [1] "Correlation Matrix"
```

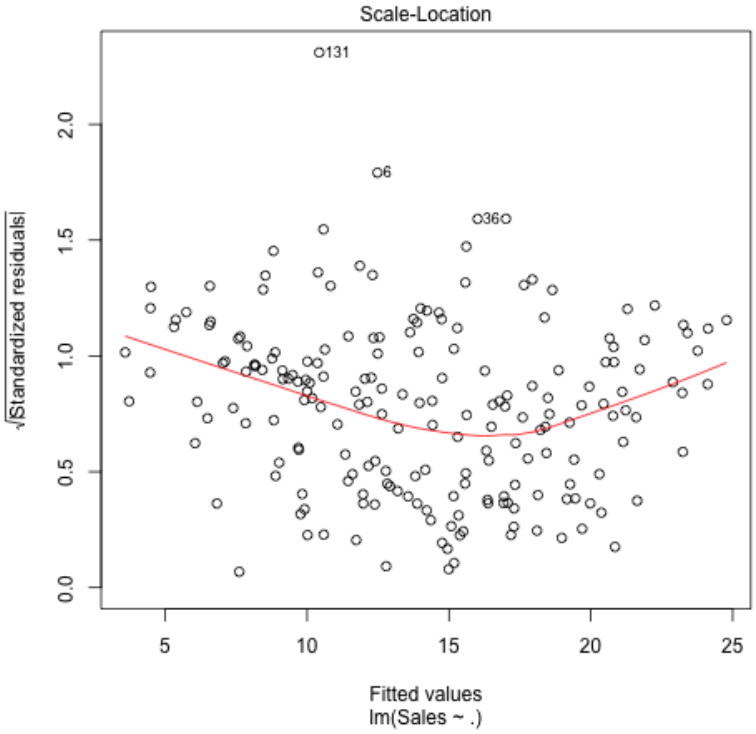
```
##                    TV      Radio  Newspaper     Sales
## TV         1.00000000 0.05480866 0.05664787 0.7822244
## Radio      0.05480866 1.00000000 0.35410375 0.5762226
## Newspaper 0.05664787 0.35410375 1.00000000 0.2282990
## Sales      0.78222442 0.57622257 0.22829903 1.0000000
```

Methodology

In order to analyze the relationship between advertising and revenue, I first had to designate which variable was explanatory and which one was the response. In this assignment, the specification was for advertising spending to be the explanatory variable with revenues as the response. In order to measure the

relationship between tv/radio/newspaper advertising spending and revenues, I ran a multiple least squares regression. Least squares regression outputs a beta vector, which quantifies the relationship between the different forms of advertising spending and revenues. For instance, if the beta coefficient for newspaper advertising took on the value of .35, this would indicate that an increase in newspaper advertising by 1 dollar causes an increase in sales by .35 units (holding all other variables constant). But how do we know that we're using multiple least squares regression correctly?

One of the key assumptions in linear regression is that the residuals are on average zero for all fitted values. In the graph below, I plotted a scatterplot, showing the relationship between the regression's standardized residuals and their corresponding fitted values.



Results

So what did our multiple least squares regression say about the relationship between ad spending and revenues? According to our least squares regression, for every $1000 increase in tv ad spending, sales go up by 45.77 units. Meanwhile, every $1000 increase in radio ad spending led to an increase in sales by 188.53 units. Finally, a $1000 dollar increase in newspaper ad spending actually *decreases* sales by 1.03 units. The table below details the coefficient estimates for the multiple least squares model.

```
## [1] "Summary of Multiple Least Squares Regression"
```

```
## $coefficients
##                Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  2.938889369 0.311908236  9.4222884 1.267295e-17
## TV           0.045764645 0.001394897 32.8086244 1.509960e-81
## Radio        0.188530017 0.008611234 21.8934961 1.505339e-54
## Newspaper   -0.001037493 0.005871010 -0.1767146 8.599151e-01
```

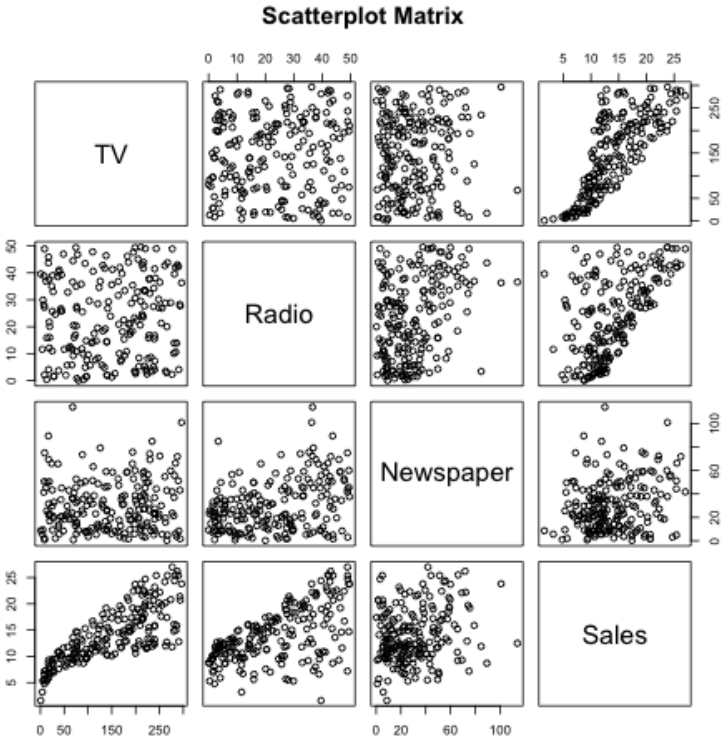
I've also attached details about the F-Squared, R-Squared, and RSE. The F-Squared statistic tells us how statistically significant our regression results are when tested against the null hypothesis that all of the variables are useless. A larger F Statistic means that more statistical significance. R-Squared, meanwhile, measures how well our explanatory variables explain the response variable. An R-Squared value close to one indicates that our regression line is doing a good job of capturing our response values. RSE (Residual Standard Error), meanwhile, tells us the average deviation of our fitted values from our true response values.

```
## [1] "More Statistics For Our Multiple Least Squares Regression"
```

```
##      Quantity       Value
## 1         RSE   1.6855104
## 2   R-Squared   0.8972106
## 3 F Statistic 570.2707037
```

So what do these values tell us? For one, the R-Squared of .8972 is fairly close to 1, and this tells us our explanatory variables do a fairly good job of explaining the variation in our response. In other words, our model fits the data pretty well. Meanwhile, the RSE value of 1.685 is quite small and indicates that our predictions are fairly accurate and do not deviate very much from our response values.

Because this multiple least squares regression has more than two explanatory variables, there isn't a way to show the regression line in a two or three dimensional space. As a result, I've shown below a pairwise scatterplot below that demonstrates all of the possible two dimensional scatterplots that could have been constructed from the dataset.



In addition, I've attached coefficient estimates of these simple least squares regression model.

```
## [1] "Summary of TV Advertising Regression"
```

```
## $coefficients
##                Estimate  Std. Error  t value    Pr(>|t|)
## (Intercept) 7.03259355 0.457842940 15.36028 1.40630e-35
## TV          0.04753664 0.002690607 17.66763 1.46739e-42
```

```
## [1] "Summary of Radio Advertising Regression"
```

```
## $coefficients
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 9.3116381 0.56290050 16.542245 3.561071e-39
## Radio       0.2024958 0.02041131  9.920765 4.354966e-19
```

```
## [1] "Summary of Newspaper Advertising Regression"
```

```
## $coefficients
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 12.3514071 0.62142019 19.876096 4.713507e-49
## Newspaper    0.0546931 0.01657572  3.299591 1.148196e-03
```

The above visual shows a pretty clear positive relationship between all advertising spending and sales. The scatterplot that shows the strong positive relationship between sales and newspaper advertising, however, seems to contradict our findings from our multiple least squares regression, which finds a negative relationship between the two variables. Surprisingly, it is fairly common for beta coefficients of variables to change when new variables are added into the regression. For instance, newspaper advertising could be strongly correlated with both tv and radio advertising. As a result, in our Sales~Newspaper regression, the positive relationship between the two variables may actually just come from radio and tv advertising, which do not show up in the regression.

This is actually a strong advantage of multiple least squares regression. Including more variables into a regression can allow us to get a more precise estimate for the relationship between two variables.

Conclusions

Least squares regression tells us how variables move and interact with each other. It doesn't, however, tell us which direction the causation runs. For instance, it could very well be that increasing revenues causes an increase in ad spending. Maybe an increase in revenues allows a company's advertising budget to increase, thus causing more ad spending. This is called reverse causality. It could also very well be that increasing ad spending is merely positively correlated with another variable that actually causes revenue to increase. This is called omitted variable bias.

In short, the results of least squares regression should always be taken with a grain of salt. Demonstrating a mere relationship between two variables does not imply any sort of causation between variables in either direction.