Report

Jamie Stankiewiz

October 13, 2016

Abstract

The goal is to look at a resopnse variable that is dependent on more than one predictor. We will look at the advertising.csv data set provided by James et al. in the book "An Introduction to Statistical Learning". In this case, our response variable we are looking at is Sales, and how that is determined the multiple predictor variables: TV, Radio and Newspaper.

Introduction

We will look at how each variable relates pairwise (in a simple linear regression) as well as how all 3 predictors come together to create the response, Sales. We will take a close look at the data by first comparing Sales on each individual predictor, using simple linear regression. Then we will look at pairwise correlation to see correlation coefficients between two variables.

Data

```
ad <- read.csv('../data/Advertising.csv')
```

The below tables are the coefficients of simple linear regression models of pairwise variables from the multiple regression model. This tells us if the individual pairs of variables are related. Linear regression models with p-values of less than 1% are considered related to each other.

Simple regression of Sales on Radio

```
coefficients(summary(lm(Sales ~ Radio, data= ad)))
```

```
##              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
```

Simple regression of Sales on Newspaper

```
coefficients(summary(lm(Sales ~ Newspaper, data=ad)))
```

```
##               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
```

Simple regression of Sales on TV

```
coefficients(summary(lm(Sales ~ TV, data = ad)))
```

```
##                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
```

The correlation matrix produced describes the pairwise correlation between the two variables. A correlation value close to 1 indicates a strong relationship between the two variables, where a correlation value close to 0 tells that the two variables are independent from one another, or not related.

Correlation Matrix:

```
corrmat <- round(cor(ad[2:5]),4)
corrmat[lower.tri(corrmat)] <- NA
corrmat
```

```
##           TV  Radio Newspaper  Sales
## TV         1 0.0548    0.0566 0.7822
## Radio     NA 1.0000    0.3541 0.5762
## Newspaper NA     NA    1.0000 0.2283
## Sales     NA     NA        NA 1.0000
```

Coefficient estimates of the least squares model

```
table4 <- round(summary(lm(Sales ~ TV+Radio+Newspaper , data=ad))$coefficients, 3)
table4
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.939      0.312   9.422     0.00
## TV             0.046      0.001  32.809     0.00
## Radio          0.189      0.009  21.893     0.00
## Newspaper     -0.001      0.006  -0.177     0.86
```

These are the coefficients of the least squares model for multiple regression. Since TV and Radio have a p-value of <1%, they both significantly contribute to sales. Their β_0 value is non-zero. While newspaper doesn't seem to correlated to sales in multiple regression. Its β_0 value is closer to 0.

Methodology

We first look at our data set, advertising.csv to look at the variables used. TV, Radio, Newspaper and Sales are the variables we get by looking at the column names of our data sheet. We then wish to look at the relationship between Sales and all of the variables individually, as well as all together. To do so, we first generate a scatterplot of the 2 against each other. Next, we create the regression line using the lm() function of the data. You can find these images in the images/ directory. To get these estimates of the coefficients, we look at the outputs of the lm() function to see the statistics of the regression line (as shown below).

```
regression <- lm(ad)
regression
```

```
##
## Call:
## lm(formula = ad)
##
## Coefficients:
## (Intercept)           TV        Radio    Newspaper        Sales
##   116.36108      0.05031     -0.12609     -0.35569     -0.67448
```

After looking at the scatterplots, we ran the data through R to see if individual predictors relate to Sales (response). This data was previously show in the Data section of this paper. After looking at each of the variables individually, we look at them all together.

Results

This table refers to the least squares model statistics for the multiple regression.

```
regdata <- summary(lm(Sales ~ TV+Radio+Newspaper , data=ad))
values <- c(round(regdata$sigma,3), round(regdata$r.squared, 3), round(regdata$fstatistic, 3)[1])
names <- c('Residual standard error', 'R-squared', 'F-statistic')
matvals <- matrix(c(names,values), nrow=3, ncol=2)
df <- as.data.frame(matvals, row.names=NA, col.names=c('Quantity', 'Value'))
colnames(df) <- c('Quantity', 'Values')
df
```

```
##                   Quantity  Values
## 1 Residual standard error   1.686
## 2               R-squared   0.897
## 3             F-statistic 570.271
```

The ANOVA test compares all of the β_0 values to check if they all are equal to zero. This lets us check to see if there's a difference between any given mean in a pairwise fashion. Since our F-statistic corresponds to a very small p-value, we can reject the null hypothesis (H_0). Therefore, we can conclude that at least one of the predictors is useful in predicting the response.

From the table, we can see that the variables,

```
for(i in 1:nrow(table4))
  if(table4[i,4]<.01){
  print(rownames(table4)[i])
  }
```

```
## [1] "(Intercept)"
## [1] "TV"
## [1] "Radio"
```

have a p-value of less than 1%. Therefore, these are the variables that are significant to the response.

To see how well the model fits the data, we look at the R^2 test statistic. R^2 tells us how well the data fits the regression model. Here, our R^2 value is

```
summary(lm(Sales ~ TV+Radio+Newspaper , data=ad))$r.squared
```

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

You can see that is value is closer to 1 than it is to 0, meaning that the regression model is explained very much so by the predictors.

To see how accurate our prediction is, we look at the confidence intervals or prediction intervals. This is shown more in (James et al.) "An Introduction to Statistical Learning".

Conclusions

From this data, we can conclude that in multiple regression, TV and Radio had an impact on Sales, while Newspaper did not. We also looked at our r^2 value to see how well the predictors as a whole determined the response variable. We also see that individual effects on Sales had a much different result than putting all of the variables together.