

Multiple Linear Regression

Author : Rose Ellison

I will be building a multiple linear regression model based off the '50_Startups' data and using the forward elimination model. In this dataset there are five columns *Profit*, *R. D. Spend*, *Administration*, *Marketing*, and *State*. Profit is our dependent variable while the other four are our independent variables. We want to determine if there are any correlations between the profit and expenditures, such as r&d, admin, and marketing. Additionally, is there any correlation between profit and which state the company is operating?

MultipleLinearRegressionFormula :

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4$$

Preparing the data

For this step, I need to read in the csv file, deal with categorical data, and then split the data into training and test sets.

```
# Importing the dataset
startups <- read.csv('../data/50_Startups.csv')

# Dealing with categorical data - 'State'
# Encoding
startups$State <- factor(startups$State,
                          levels <- c('New York', 'California', 'Florida'),
                          labels <- c(1, 2, 3))

# Splitting the data into test and training sets
set.seed(1)
split <- sample.split(startups$Profit, SplitRatio = .8)
training.set <- subset(startups, split == TRUE)
test.set <- subset(startups, split == FALSE)
```

Fitting Multiple Linear Regression to the Training Set

The Independent variable, Profit, is going to be a linear combination of all the dependent variables.

```
# Regressor with all dependent variables
regressor <- lm(Profit ~ ., training.set)
summary(regressor)
```

```
##
## Call:
## lm(formula = Profit ~ ., data = training.set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30230.1  -3255.0   606.6   6683.7  13424.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.203e+04  7.535e+03   6.905  5.9e-08 ***
## R.D.Spend      8.612e-01  5.468e-02  15.749 < 2e-16 ***
## Administration -7.261e-02  5.923e-02  -1.226   0.229
## Marketing.Spend 1.893e-02  1.926e-02   0.983   0.332
## State2         5.173e+02  3.688e+03   0.140   0.889
## State3        -1.967e+02  3.728e+03  -0.053   0.958
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9365 on 34 degrees of freedom
## Multiple R-squared:  0.955, Adjusted R-squared:  0.9483
## F-statistic: 144.2 on 5 and 34 DF,  p-value: < 2.2e-16
```

According to the data, the only strong predictor in profit is the r&d spend. Due to this, we could rewrite our regressor with only one dependent variable and we should still get the same results.

```
# Regressor with only the R.D.Spend dependent variable
regressor <- lm(Profit ~ R.D.Spend, training.set)
summary(regressor)
```

```
##
## Call:
## lm(formula = Profit ~ R.D.Spend, data = training.set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -31194.8  -4500.5    58.8   5638.2  17478.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.588e+04  2.823e+03  16.25  <2e-16 ***
## R.D.Spend    8.836e-01  3.283e-02   26.91  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9318 on 38 degrees of freedom
## Multiple R-squared:  0.9502, Adjusted R-squared:  0.9488
## F-statistic: 724.4 on 1 and 38 DF,  p-value: < 2.2e-16
```

Although we changed the regressor to only using the R.D.Spend variable, the p-value remains the same. This is another indication that R.D.Spend is the only strong dependent variable predictor.

Predicting the Test Set Results

```
y.pred <- predict(regressor, newdata = test.set)
```

Visualizing the training set

```
# Visualizing the training set results
ggplot() +
  geom_point(aes(x = training.set$R.D.Spend, y = training.set$Profit), col = 'lightblue') +
  geom_line(aes(x = training.set$R.D.Spend, y = predict(regressor, newdata = training.set)), col = 'darkblue') +
  theme_bw() +
  ggtitle('Profit VS R.D.Spend (Training Set)') +
  xlab('Research and Development Spend') +
  ylab('Profit')
```



Visualizing the test set results

```
ggplot() +  
  geom_point(aes(x = test.set$R.D.Spend, y = test.set$Profit), col = 'lightblue') +  
  geom_line(aes(x = training.set$R.D.Spend, y = predict(regressor, newdata = training.set)), col = 'darkblue') +  
  theme_bw() +  
  ggtitle('Profit VS R.D.Spend (Test Set)') +  
  xlab('Research and Development Spend') +  
  ylab('Profit')
```

Profit VS R.D.Spend (Test Set)

