Forecasting, Homework 1

Robby Jeffries

1/26/2022

Set working directory

```
setwd("~/MSEA2022/Spring 2022/ECON 5753, Forecasting")
```

Import packages and install them if necessary

```
list.of.packages <- c("tidyverse", "caTools", "pastecs", "ggplot2")
new.packages <- list.of.packages[!(list.of.packages %in% installed.packages()[,"Package"])]
if(length(new.packages)) install.packages(new.packages)
library(tidyverse)
library(caTools)
library(pastecs)
library(ggplot2)
```

1) Import data

```
df = read.csv("Data/50_Startups.csv")
```

2) Show descriptive statistics

```
options(scipen = 100)
options(digits = 2)
pastecs::stat.desc(df)
```

```
##                 R.D.Spend Administration Marketing.Spend State        Profit
## nbr.val             48.00          50.00            47.0    NA         50.00
## nbr.null             0.00           0.00             0.0    NA          0.00
## nbr.na               2.00           0.00             3.0    NA          0.00
## min                542.05       51283.14          1903.9    NA      14681.40
## max             165349.20      182645.56        471784.1    NA     192261.83
## range           164807.15      131362.42        469880.2    NA     177580.43
## sum            3686080.78     6067231.98      10551254.9    NA    5600631.96
## median           74661.71      122699.79        229161.0    NA     107978.19
```

```
## mean              76793.35      121344.64        224494.8    NA     112012.64
## SE.mean            6383.20        3962.32         16528.9    NA       5700.15
## CI.mean.0.95      12841.34        7962.57         33271.0    NA      11454.89
## var          1955769803.42   784997271.25   12840630064.4    NA 1624588173.41
## std.dev           44224.09       28017.80        113316.5    NA      40306.18
## coef.var              0.58           0.23             0.5    NA          0.36
```

3) Replace missing data with the mean of each variable

```
df$R.D.Spend = ifelse(is.na(df$R.D.Spend),
                      mean(df$R.D.Spend, na.rm=TRUE),
                        df$R.D.Spend)

df$Marketing.Spend = ifelse(is.na(df$Marketing.Spend),
                      mean(df$Marketing.Spend, na.rm=TRUE),
                      df$Marketing.Spend)
```

4) Simple Linear Regression: Dep. variable-Profit, Ind. variable-R&D Spend

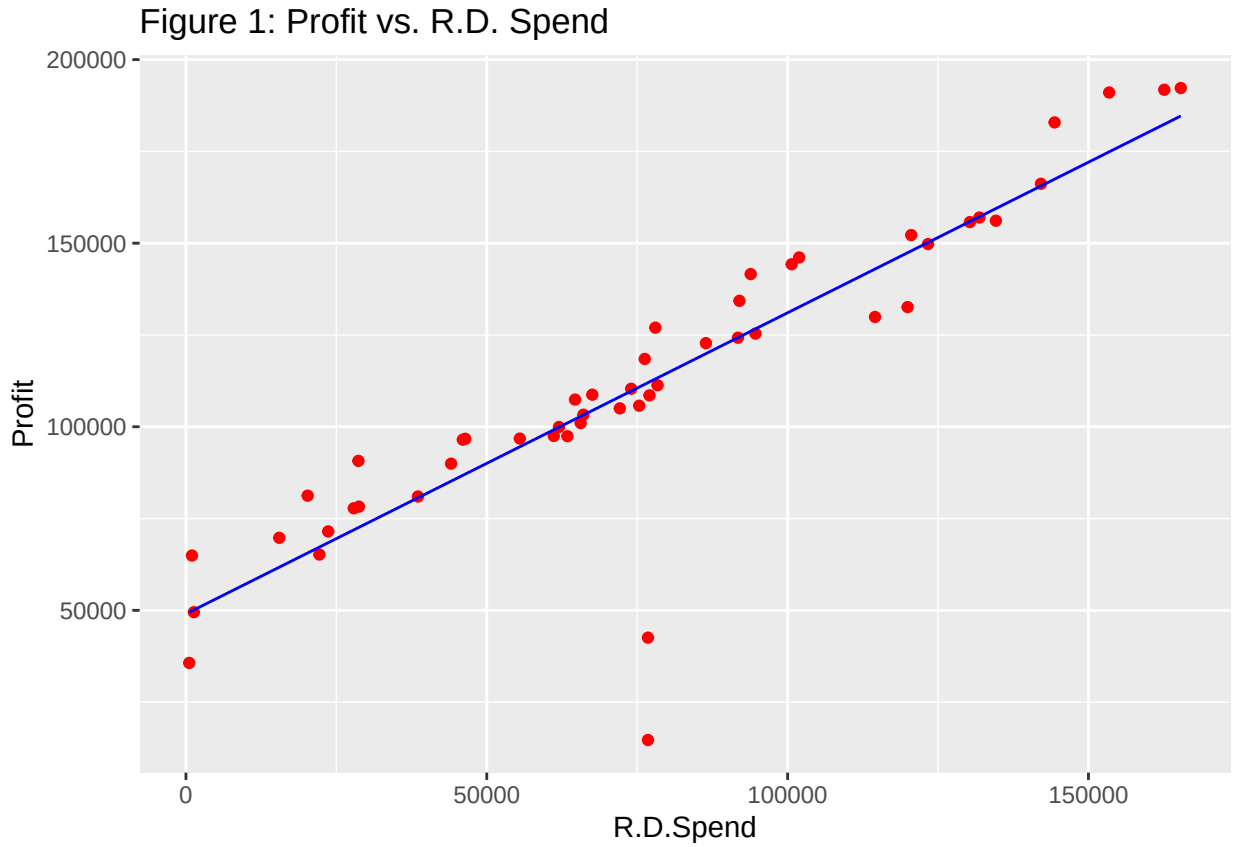
```
reg = lm(formula = Profit ~ R.D.Spend, data = df)
summary(reg)
```

```
##
## Call:
## lm(formula = Profit ~ R.D.Spend, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -97331  -1796   1435   9056  18171
##
## Coefficients:
##               Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 49027.8566  5581.5152    8.78       0.000000000015 ***
## R.D.Spend       0.8202     0.0635   12.92 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19200 on 48 degrees of freedom
## Multiple R-squared:  0.777,  Adjusted R-squared:  0.772
## F-statistic:  167 on 1 and 48 DF,  p-value: <0.0000000000000002
```

```
y_hat = predict(reg, newdata = df)
```

5) Visualize the data

```
ggplot() +
  geom_point(aes(x = df$R.D.Spend,
                 y = df$Profit),
             colour = 'red') +
  geom_line(aes(x = df$R.D.Spend,
                y = y_hat),
            colour = 'blue') +
  ggtitle('Figure 1: Profit vs. R.D. Spend') +
  xlab('R.D.Spend') +
  ylab('Profit')
```



6) Generate quadratic form of ind. variable R&D Spend

```
quadraticModel = lm(formula = Profit ~ df$R.D.Spend + I(df$R.D.Spend^2), data = df)
```

7) Print out the quadratic regression result

```
summary(quadraticModel)
```

```
##
```

```
## Call:
## lm(formula = Profit ~ df$R.D.Spend + I(df$R.D.Spend^2), data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -93600  -1151   3148   8371  18975
##
## Coefficients:
##                         Estimate     Std. Error t value     Pr(>|t|)
## (Intercept)       58450.15678175  8075.62228632    7.24 0.0000000036 ***
## df$R.D.Spend          0.49305467     0.21476587    2.30        0.026 *
## I(df$R.D.Spend^2)     0.00000203     0.00000127    1.59        0.118
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18900 on 47 degrees of freedom
## Multiple R-squared:  0.788,  Adjusted R-squared:  0.779
## F-statistic: 87.5 on 2 and 47 DF,  p-value: <0.0000000000000002
```

```
y_hat2 = predict(quadraticModel, newdata = df)
```

8) Visualize the data

```
ggplot(df, aes(R.D.Spend, Profit)) +
  geom_point(aes(x = R.D.Spend,
                 y = Profit),
             colour = 'firebrick') +
  geom_line(aes(x = R.D.Spend,
                y = y_hat),
            colour = 'gray30',
            size = 1) +
  geom_smooth(method = lm,
              formula = y ~ x + I(x^2),
              se = FALSE,
              colour = 'seagreen',
              size = 1) +
  ggtitle('Figure 2: Profit vs. R.D. Spend') +
  xlab('R.D.Spend') +
  ylab('Profit')
```

Figure 2: Profit vs. R.D. Spend