M12 Activity

Hannah Valenty

2024-07-31

```
st <- read.csv('data/Startups.csv')
head(st)
```

```
##   R.D.Spend Administration Marketing.Spend      State   Profit
## 1  165349.2      136897.80        471784.1   New York 192261.8
## 2  162597.7      151377.59        443898.5 California 191792.1
## 3  153441.5      101145.55        407934.5    Florida 191050.4
## 4  144372.4      118671.85        383199.6   New York 182902.0
## 5  142107.3       91391.77        366168.4    Florida 166187.9
## 6  131876.9       99814.71        362861.4   New York 156991.1
```

Task 1

```
mod_state <- lm(Profit~State, data = st)
summary(mod_state)
```

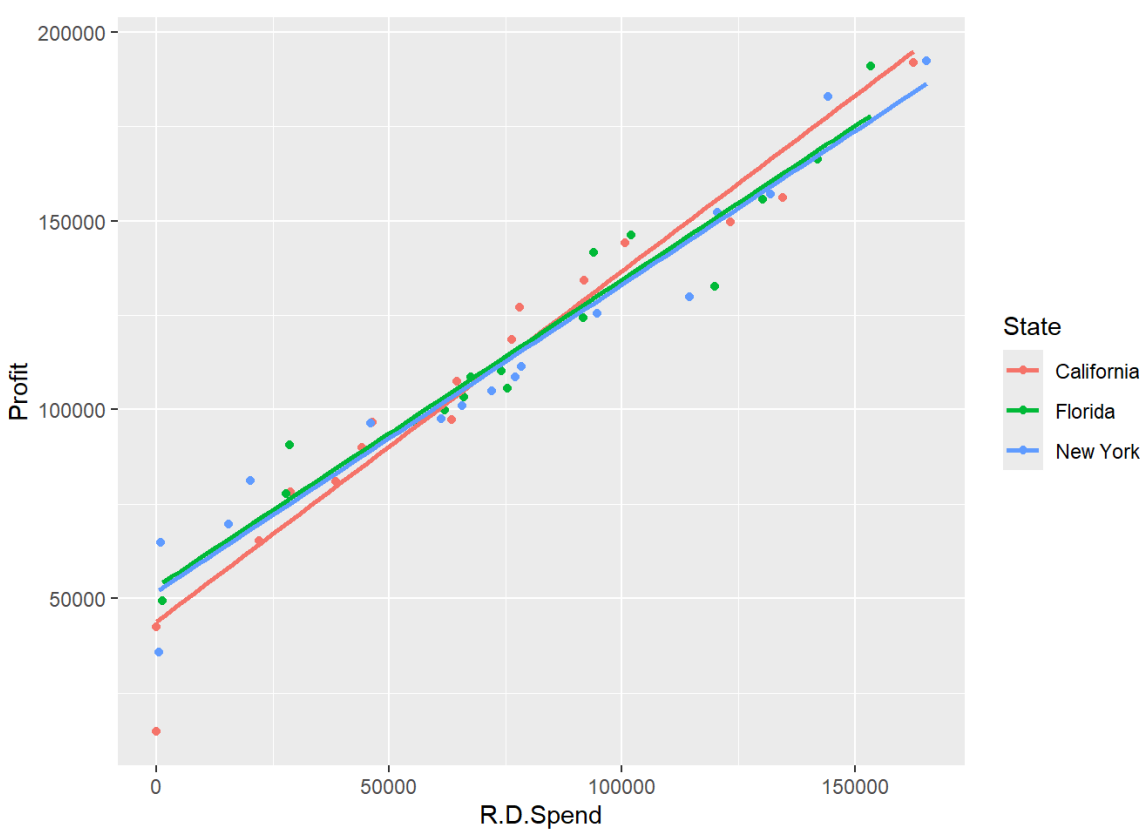
```
##
## Call:
## lm(formula = Profit ~ State, data = st)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -89224 -22673  -6835  26283  87887
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     103905       9862  10.536 5.77e-14 ***
## StateFlorida     14869      14163   1.050    0.299
## StateNew York     9851      13946   0.706    0.483
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40660 on 47 degrees of freedom
## Multiple R-squared:  0.02388,    Adjusted R-squared:  -0.01766
## F-statistic: 0.5748 on 2 and 47 DF,  p-value: 0.5667
```

Model Equation: $\hat{Profit} = 103905 + 14869 * I_{Florida} + 9851 * I_{NewYork}$

Task 2.A – No Interaction

```
library(ggplot2)
ggplot(st, aes(x=R.D.Spend, y=Profit, colour = State))+
  geom_jitter()+
  geom_smooth(method='lm', aes(group=State), se=F)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Task 2.B

```
mod_state_rd <- lm(Profit~State+R.D.Spend, data = st)
summary(mod_state_rd)
```

```
##
## Call:
## lm(formula = Profit ~ State + R.D.Spend, data = st)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -34069  -4302   -555   6554  16343
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.875e+04  3.040e+03  16.036   <2e-16 ***
## StateFlorida  1.164e+03  3.380e+03   0.344    0.732
## StateNew York 9.597e+00  3.312e+03   0.003    0.998
## R.D.Spend     8.530e-01  3.022e-02  28.226   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9603 on 46 degrees of freedom
## Multiple R-squared:  0.9467, Adjusted R-squared:  0.9432
## F-statistic: 272.4 on 3 and 46 DF,  p-value: < 2.2e-16
```

Model Equation: $\hat{Profit} = 48750 + 0.853 * R.D.Spend + 1164 * I_{Florida} + 9.597 * I_{NewYork}$

Task 2.C

The coefficient of the State of Florida can be interpreted as, Florida start ups making $1164 more than start ups in California when spending nothing on R& D.

Task 3.A – With Interaction

```
mod_inter <- lm(Profit~State*R.D.Spend, data = st)
summary(mod_inter)
```

```
##
## Call:
## lm(formula = Profit ~ State * R.D.Spend, data = st)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -29194  -4112   -313   5924  14278
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               4.388e+04  4.000e+03  10.969 3.58e-14 ***
## StateFlorida              9.242e+03  6.569e+03   1.407    0.167
## StateNew York             7.921e+03  5.880e+03   1.347    0.185
## R.D.Spend                 9.284e-01  5.067e-02  18.322  < 2e-16 ***
## StateFlorida:R.D.Spend   -1.151e-01  7.666e-02  -1.501    0.140
## StateNew York:R.D.Spend  -1.153e-01  6.972e-02  -1.653    0.105
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9461 on 44 degrees of freedom
## Multiple R-squared:  0.9505, Adjusted R-squared:  0.9449
## F-statistic: 169.1 on 5 and 44 DF,  p-value: < 2.2e-16
```

Model Equation:

$\hat{Profit} = 43880 + 0.9284 * R.D.Spend + 9242 * I_{Florida} + 7921 * I_{NewYork} - 0.1151 * R.D.Spend * I_{Florida} - 0.1153 * R.D.Spend * I_{NewYork}$

Task 3.B

- Florida:R.D.Spend coefficient: Every change in R&D for start ups in Florida has a higher average Profit from start ups in California by 0.9284 - 0.1151 (0.8133).
- New York:R.D.Spend coefficient: Every change in R&D for start ups in Florida has a higher average Profit from start ups in California by 0.9284 - 0.1153 (0.8131).

Task 4

```
st$State <- factor(st$State, level = c('Florida','New York','California'))
# Florida reference
mod_inter <- lm(Profit~State*R.D.Spend, data = st)
summary(mod_inter)
```

```
##
## Call:
## lm(formula = Profit ~ State * R.D.Spend, data = st)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -29194  -4112   -313   5924  14278
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                5.312e+04  5.211e+03  10.193  3.7e-13 ***
## StateNew York             -1.321e+03  6.763e+03  -0.195    0.846
## StateCalifornia           -9.242e+03  6.569e+03  -1.407    0.167
## R.D.Spend                  8.134e-01  5.752e-02  14.139  < 2e-16 ***
## StateNew York:R.D.Spend   -1.757e-04  7.485e-02  -0.002    0.998
## StateCalifornia:R.D.Spend  1.151e-01  7.666e-02   1.501    0.140
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9461 on 44 degrees of freedom
## Multiple R-squared:  0.9505, Adjusted R-squared:  0.9449
## F-statistic: 169.1 on 5 and 44 DF,  p-value: < 2.2e-16
```

```
# New York reference
st$State <- factor(st$State, level = c('New York','Florida','California'))
mod_inter <- lm(Profit~State*R.D.Spend, data = st)
summary(mod_inter)
```

```
##
## Call:
## lm(formula = Profit ~ State * R.D.Spend, data = st)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -29194  -4112   -313   5924  14278
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                5.180e+04  4.310e+03  12.017 1.72e-15 ***
## StateFlorida               1.321e+03  6.763e+03   0.195    0.846
## StateCalifornia           -7.921e+03  5.880e+03  -1.347    0.185
## R.D.Spend                  8.132e-01  4.788e-02  16.983  < 2e-16 ***
## StateFlorida:R.D.Spend     1.757e-04  7.485e-02   0.002    0.998
## StateCalifornia:R.D.Spend  1.153e-01  6.972e-02   1.653    0.105
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9461 on 44 degrees of freedom
## Multiple R-squared:  0.9505, Adjusted R-squared:  0.9449
## F-statistic: 169.1 on 5 and 44 DF,  p-value: < 2.2e-16
```

State does not seem to be a useful predictor of Profit. For each iteration of State reference variables the state p-values all fall above 0.05. It would be best to remove state from the model.