Class Activity 9

Courtney Hodge

2024-07-31

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.3     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.4     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
startups <- read.csv("C:\\Users\\hodge\\Desktop\\UVA_Coding_Folder\\Statistics-6021\\Startups.csv")
```

1

```
model1 <- lm(Profit~ State, data = startups)
summary(model1)
```

```
##
## Call:
## lm(formula = Profit ~ State, data = startups)
##
## 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
```

> The model is:

$\hat{Profit} = 103905 + (14869 * StateFlorida) + (9851 * StateNewYork)$

2

```
model2 <- lm(Profit~State + R.D.Spend, data = startups)
summary(model2)
```

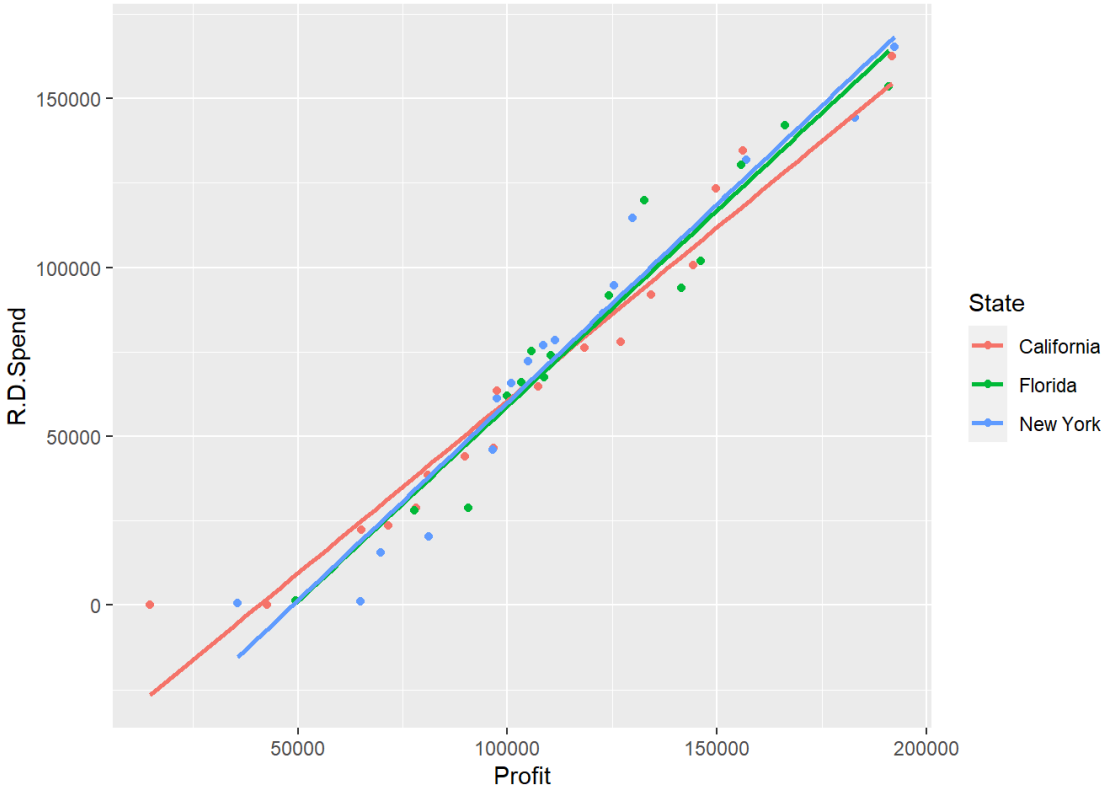
```
##
## Call:
## lm(formula = Profit ~ State + R.D.Spend, data = startups)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -34069  -4302   -555   6554  16343
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.875e+04  3.040e+03  16.036   <2e-16 ***
## 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
```

2a

- Explore the three variables in a visualization. Also, superimpose a linear regression line predicting Profit based on R.D.Spend. for each State.

```
ggplot(startups, aes(x = Profit, y = R.D.Spend, color = State)) + geom_jitter() + geom_smooth(method = "lm", model.
extract(model2), se = F)
```

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



2b

- write down the model

> The model is:

$ = 4.875e+04 + (1.164e+03 * StateFlorida) + (9.597e+00 * StateNew York) + (8.530e-01*R.D.Spend) $

2c

The coefficient of StateFlorida in this context is 1.164e+03 more than the California Baseline of 4.875e+04 when R.D.Spend is included in the model.

3

```
model3 <- lm(Profit~State * R.D.Spend, data = startups)
summary(model3)
```

```
##
## Call:
## lm(formula = Profit ~ State * R.D.Spend, data = startups)
##
## 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
```

3a

$$\hat{Profit} = 4.388e+04 + (9.242e+03 * StateFloridax - 1.151e-01) + (7.921e+03 * StateNewYorkx - 1.153e-01)$$

3b

When the company is in Florida, the effect of R.D. spend on Profit is reduced 1.151e-01 times on average, and when the company is in New York, the effect of R.D.Spend on Profit is reduced 1.153e-01 times on average.

4

```
startups$State2<-factor(startups$State, levels = c("New York", "California", "Florida"))

mod4 <- lm(Profit~State2, data = startups)
summary(mod4)
```

```
##
## Call:
## lm(formula = Profit ~ State2, data = startups)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -89224 -22673  -6835  26283  87887
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        113756       9862  11.535 2.62e-15 ***
## State2California    -9851      13946  -0.706    0.483
## State2Florida        5018      14163   0.354    0.725
## ---
## 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
```

```
startups$State3<-factor(startups$State, levels = c("Florida", "California", "New York"))

mod5 <- lm(Profit~State3, data = startups)
summary(mod5)
```

```
##
## Call:
## lm(formula = Profit ~ State3, data = startups)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -89224 -22673  -6835  26283  87887
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        118774      10165  11.684 1.67e-15 ***
## State3California   -14869      14163  -1.050    0.299
## State3New York      -5018      14163  -0.354    0.725
## ---
## 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
```

State is not a useful predictor based on the p-values. When we change the reference category for state, it still remains not significant, therefore, the predictor category for State is not useful.