Final_Project

Chimdi, Chisom and Oyefunmi

6/3/2023

```
library(Sleuth3)
library(ggplot2)
library(Rmisc)
```

```
## Loading required package: lattice

## Loading required package: plyr
```

```
library(graphics)
library(leaps)
```

The dataset is for predicting salary of workers in some parts of usa

```
df<- read.csv("project_anova.csv")
head(df)
```

```
##   Salary     Profession        Region
## 1 126411 Data Scientist San Francisco
## 2 108402 Data Scientist San Francisco
## 3  99399 Data Scientist San Francisco
## 4  91381 Data Scientist San Francisco
## 5 105023 Data Scientist San Francisco
## 6 108944 Data Scientist San Francisco
```

```
#View(df)
```

```
# Converting character columns to factor
for (i in seq_along(df)) {
  if (is.character(df[[i]])) {
    df[[i]] <- as.factor(df[[i]])
  }
}

head(df)
```

```
##   Salary     Profession        Region
## 1 126411 Data Scientist San Francisco
## 2 108402 Data Scientist San Francisco
## 3  99399 Data Scientist San Francisco
## 4  91381 Data Scientist San Francisco
## 5 105023 Data Scientist San Francisco
## 6 108944 Data Scientist San Francisco
```

```
table <- aggregate(Salary ~ Profession + Region, data = df, FUN = length)
xtabs(Salary ~ Profession + Region, table)
```

```
##                     Region
## Profession          New York San Francisco Seattle
##   BI Engineer             20            20      20
##   Data Scientist          20            20      20
##   Software Engineer       20            20      20
```

The data is balanced.

(b) Fitting the additive model for Salary of STEM employees vs their job location Based on the model fit, which factors seem to have a significant effect on the scores?

```
fit_add <- lm(Salary ~ Profession + Region, data = df)
summary(fit_add)
```

```
##
## Call:
## lm(formula = Salary ~ Profession + Region, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26593.4  -7950.0   -185.2   7285.4  30929.6
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    71758       2060  34.839  < 2e-16 ***
## ProfessionData Scientist       27608       2256  12.236  < 2e-16 ***
## ProfessionSoftware Engineer    18777       2256   8.322 2.38e-14 ***
## RegionSan Francisco            12215       2256   5.414 2.01e-07 ***
## RegionSeattle                   8724       2256   3.866 0.000156 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12360 on 175 degrees of freedom
## Multiple R-squared:  0.517,  Adjusted R-squared:  0.5059
## F-statistic: 46.82 on 4 and 175 DF,  p-value: < 2.2e-16
```

```
anova(fit_add)
```

```
## Analysis of Variance Table
##
## Response: Salary
##             Df     Sum Sq    Mean Sq F value    Pr(>F)
## Profession   2 2.3855e+10 1.1928e+10  78.099 < 2.2e-16 ***
## Region       2 4.7499e+09 2.3750e+09  15.551 6.078e-07 ***
## Residuals  175 2.6727e+10 1.5272e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Checking for the better model between additive and saturated

```
fit_sat <- lm(Salary ~ Profession * Region, data = df)
summary(fit_sat)
```

```
##
## Call:
## lm(formula = Salary ~ Profession * Region, data = df)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max 
## -23776  -8369  -1215   7426  36023 
## 
## Coefficients:
##                                                  Estimate Std. Error t value
## (Intercept)                                         77519       2632  29.454
## ProfessionData Scientist                            15093       3722   4.055
## ProfessionSoftware Engineer                         14011       3722   3.764
## RegionSan Francisco                                  1421       3722   0.382
## RegionSeattle                                        2236       3722   0.601
## ProfessionData Scientist:RegionSan Francisco        19866       5264   3.774
## ProfessionSoftware Engineer:RegionSan Francisco     12514       5264   2.377
## ProfessionData Scientist:RegionSeattle              17680       5264   3.359
## ProfessionSoftware Engineer:RegionSeattle            1783       5264   0.339
##                                                  Pr(>|t|)    
## (Intercept)                                       < 2e-16 ***
## ProfessionData Scientist                         7.61e-05 ***
## ProfessionSoftware Engineer                      0.000229 ***
## RegionSan Francisco                              0.703029    
## RegionSeattle                                    0.548786    
## ProfessionData Scientist:RegionSan Francisco     0.000221 ***
## ProfessionSoftware Engineer:RegionSan Francisco 0.018538 *  
## ProfessionData Scientist:RegionSeattle           0.000965 ***
## ProfessionSoftware Engineer:RegionSeattle        0.735213    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 11770 on 171 degrees of freedom
## Multiple R-squared:  0.5719, Adjusted R-squared:  0.5518 
## F-statistic: 28.55 on 8 and 171 DF,  p-value: < 2.2e-16
```

```
anova(fit_add, fit_sat)
```

```
## Analysis of Variance Table
## 
## Model 1: Salary ~ Profession + Region
## Model 2: Salary ~ Profession * Region
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)    
## 1    175 2.6727e+10                                   
## 2    171 2.3690e+10  4 3037177895 5.4809 0.0003555 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The anova output shows that the saturated model is more significant (p-value = 0.0003555)

showing the plot for additive model

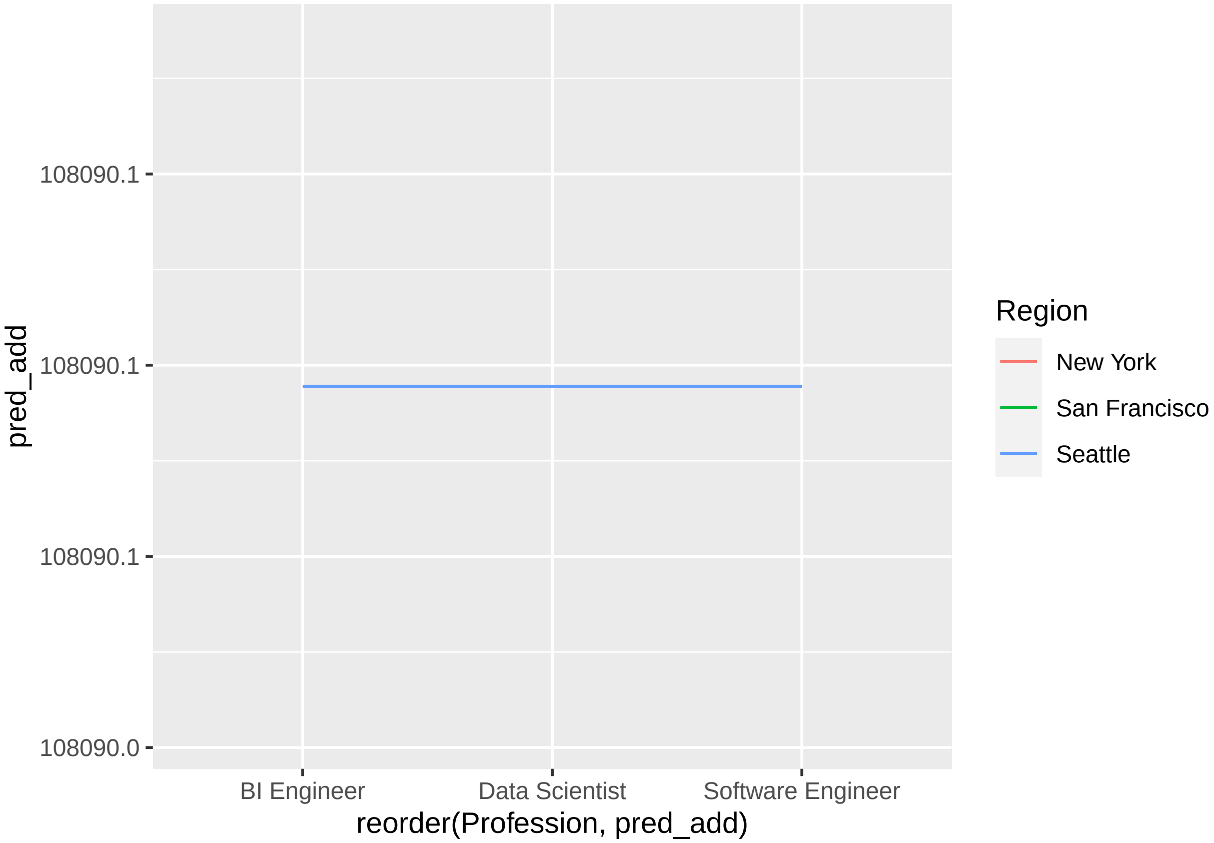
```
test_value <- data.frame(Profession = "Data Scientist", Region = "Seattle")
df$pred_add <- predict(fit_add, newdata = test_value)
```

```
round(xtabs(pred_add ~ Profession + Region, df), 2) # a table
```

```
##                      Region
## Profession            New York San Francisco Seattle
##   BI Engineer          2161801       2161801 2161801
##   Data Scientist       2161801       2161801 2161801
##   Software Engineer    2161801       2161801 2161801
```

```
qplot(reorder(Profession, pred_add), pred_add, data = df,
colour = Region, geom = "line", group = Region)
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



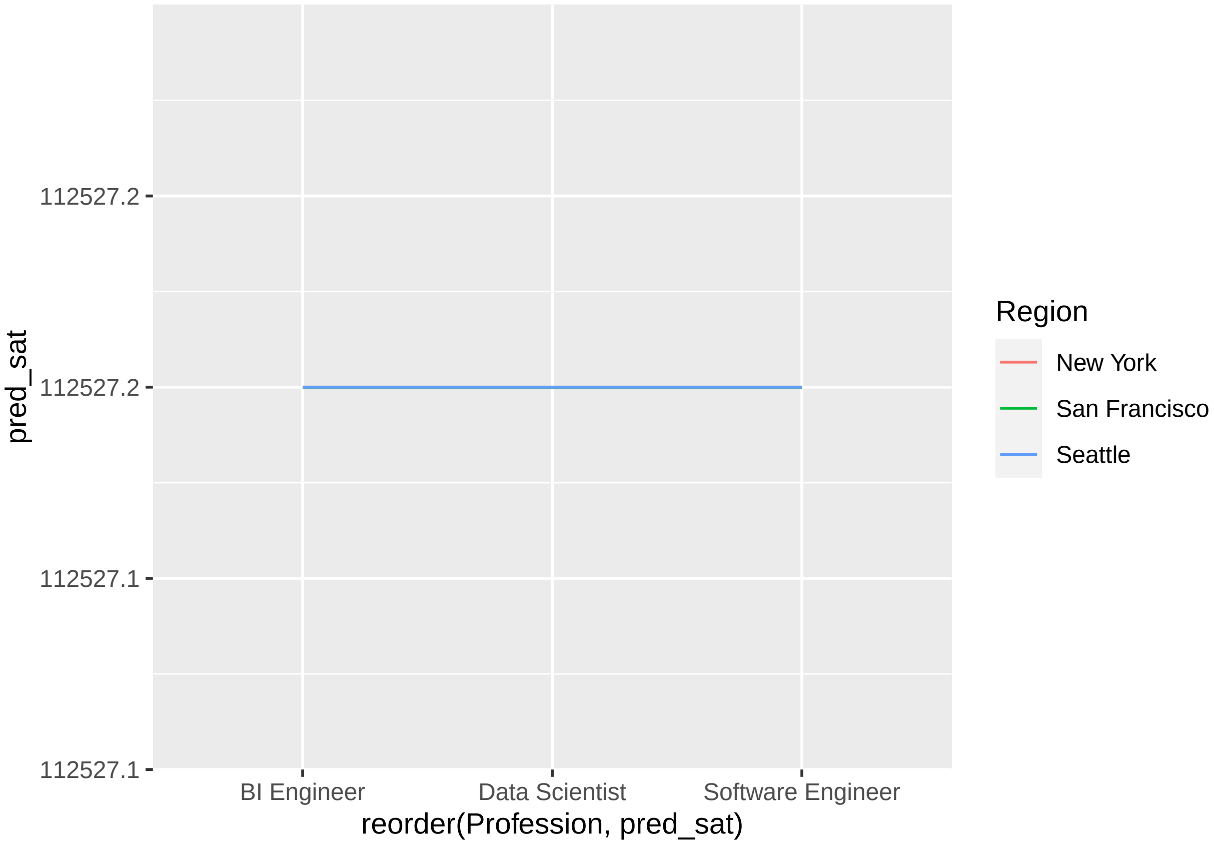
```
test_value <- data.frame(Profession = "Data Scientist", Region = "Seattle")
df$pred_sat <- predict(fit_sat, newdata = test_value)
```

showing the plot for saturated model

```
round(xtabs(pred_sat ~ Profession + Region, df), 2) # a table
```

```
##                      Region
## Profession            New York San Francisco Seattle
##   BI Engineer          2250543       2250543 2250543
##   Data Scientist       2250543       2250543 2250543
##   Software Engineer    2250543       2250543 2250543
```

```
qplot(reorder(Profession, pred_sat), pred_sat, data = df,
colour = Region, geom = "line", group = Region)
```



The plots are different, showing that the saturated model is better

```
test_value <- data.frame(Profession = "Data Scientist", Region = "Seattle")
pred_salary <- predict(fit_sat, newdata = test_value)
sprintf("predicted salary = %f", pred_salary)
```

```
## [1] "predicted salary = 112527.150000"
```

```
sprintf("Prediction interval")
```

```
## [1] "Prediction interval"
```

```
pred_int <-predict(fit_sat, test_value, interval ="prediction", level=0.95 )
pred_int
```

```
##        fit      lwr      upr
## 1 112527.1 88719.98 136334.3
```