BIS623 Homework 6

Due on 11/07/2019 before the lecture

04 November, 2019

Please use the data `salarygov` in the R library `alr4`.

```
if(!require("alr4")){
  install.packages("alr4")
}
```

```
## Loading required package: alr4
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: effects
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
library(alr4)

dt = salarygov

if(!require("data.table")){
  install.packages("data.table")
}
```

```
## Loading required package: data.table
```

```
library(data.table)
setDT(dt)
```

The data file gives the maximum monthly salary for 495 nonunionized job classes in a midwestern governmental unit in 1986. The variables are described below

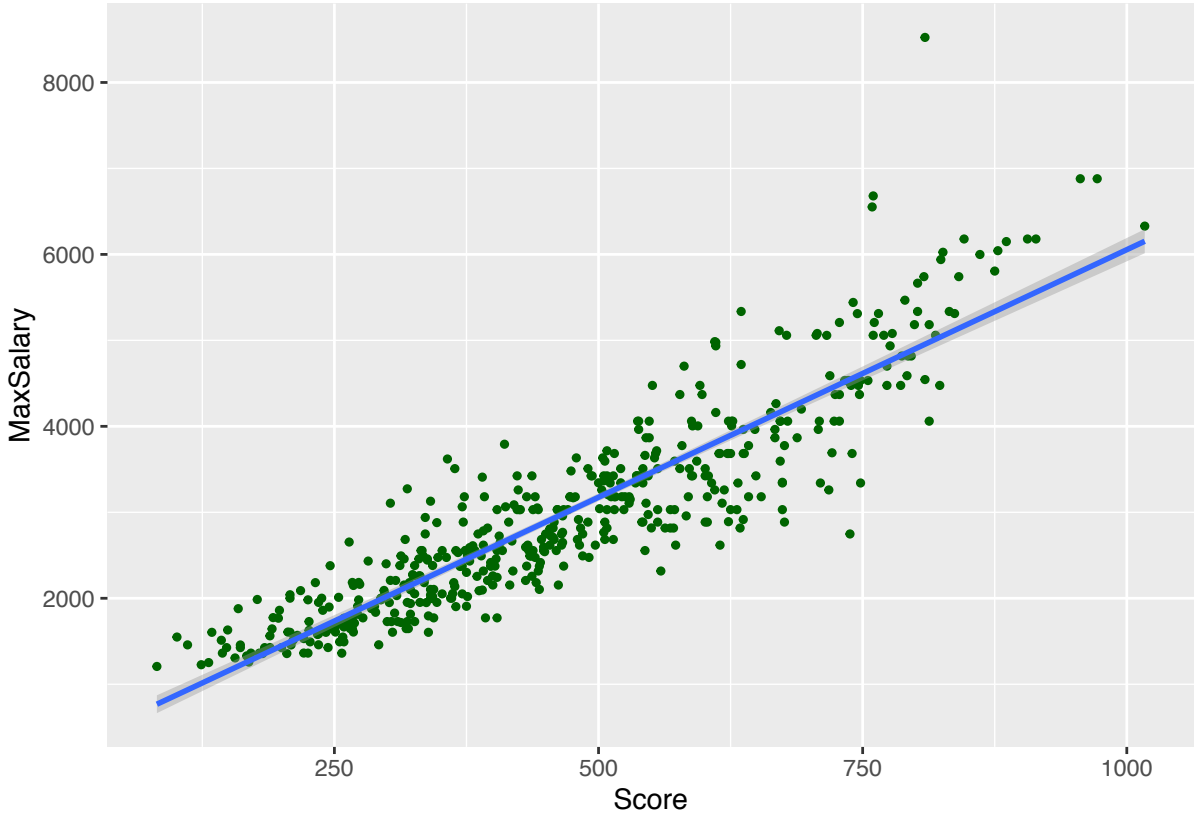
- MaxSalary: Maximum salary in dollars for employees in this job class, which is the response variable
- NE: Total number of employees currently employed in this job class
- NW: Number of women employees in the job class
- Sore: Score for job class based on difficulty, skill level, training requirements and level of responsibility as determined by a consultant to the governmental unit. This value for these data is in the range between 82 and 1017.
- JobClass: Name of the job class; a few names were illegible or partly illegible

Question 1:

Examine the scatterplot of MaxSalary versus Score. Will simple linear regression (SLR) provides a good fit? SLR provides a good fit by looking at R^2 and p-value in the summary(slr) and looking at the graph directly.

But the fit can be improved.

```
library(ggplot2)
p = ggplot(data = dt, aes(x = Score, y = MaxSalary)) +
    geom_point(color = "darkgreen", size = 1) + #Plotting the scatter point
    stat_smooth(method = lm)
p
```




```
slr <- lm(MaxSalary ~ Score, data = dt)
summary(slr)
```

```
##
## Call:
## lm(formula = MaxSalary ~ Score, data = dt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1797.9  -284.1   -42.0   248.7  3569.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  295.274     62.012   4.762 2.53e-06 ***
## Score          5.760      0.123  46.844  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 507.2 on 493 degrees of freedom
## Multiple R-squared:  0.8165, Adjusted R-squared:  0.8162
```

```
## F-statistic:  2194 on 1 and 493 DF,  p-value: < 2.2e-16
```

Question 2:

According to Minnesota statutes, and probably laws in other states as well, a job class is considered to be female dominated if 70% of the employees or more in the job class are female. Create a factor with two levels that divides the job classes into female dominated or not.

```
#female_dominated is 1 if female-dominated and 0 if not
dt$women_prop=dt$NW/dt$NE
dt[,female_dominated:=ifelse(women_prop>=0.7, 1, 0)]
head(dt)
```

```
##                      JobClass NW NE Score MaxSalary women_prop
## 1:              Account_clerk 52 68   258      1549 0.76470588
## 2: Account_clerk_Intermediate 26 29   269      1712 0.89655172
## 3:    Account_clerk_Principal 10 13   321      2182 0.76923077
## 4:       Account_clerk_Senior 16 24   273      1982 0.66666667
## 5:                 Accountant  1 12   352      2555 0.08333333
## 6:           Accountant_Chief  0  5   709      4060 0.00000000
##    female_dominated
## 1:                1
## 2:                1
## 3:                1
## 4:                0
## 5:                0
## 6:                0
```

Question 3:

Fit a model for MaxSalary on Score, the newly created variable in Question 2, as well as their interaction. Interprete each of the estimated coefficient.

```
m=lm(MaxSalary~Score + female_dominated + Score:female_dominated, data=dt)
summary(m)
```

```
##
## Call:
## lm(formula = MaxSalary ~ Score + female_dominated + Score:female_dominated,
##     data = dt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1910.9  -271.5   -40.4   214.5  3450.7
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            351.7125    75.2637   4.673 3.84e-06 ***
## Score                    5.8363     0.1394  41.856  < 2e-16 ***
## female_dominated       206.2744   127.1966   1.622    0.106
## Score:female_dominated  -1.2508     0.2804  -4.461 1.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 477.8 on 491 degrees of freedom
## Multiple R-squared:  0.8378, Adjusted R-squared:  0.8368
## F-statistic: 845.5 on 3 and 491 DF,  p-value: < 2.2e-16
```

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

where:

$$\begin{aligned} X_{i1} &= \text{Score} \\ X_{i2} &= \begin{cases} 1 & \text{if the job class is female-dominated} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Then the response function for this model is

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

For non-female-dominated job class, $X_2 = 0$ and hence $X_1 X_2 = 0$. Response function therefore becomes for non-female-dominated job class:

$$E\{Y\} = \beta_0 + \beta_1 X_1 \text{ non-female-dominated job class}$$

β_0 is the intercept represents the mean response $E\{Y\}$ when the score is 0 in the case of non-female-dominated job class.

β_1 indicates the change in the mean response $E\{Y\}$ per unit increase in the score when in the case of non-female-dominated job class.

For female-dominated job class, $X_2 = 1$ and hence $X_1 X_2 = X_1$. Response function therefore becomes

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2(1) + \beta_3 X_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 \text{ female-dominated job class}$$

in this senario, β_2 here indicates how much greater (smaller) is the Y intercept ofthe response function for the class coded 1 than that for the class coded O. Similarly, β_3 indicates how much greater (smaller) is the slope of the response function for the class coded 1 than that for the class coded O.

Question 4:

For Question 1, if SLR is not a good choice, what other model you plan to use? Fit the alternative model and draw the fitted line.

```
library(splines)
bs_model=lm(MaxSalary~bs(Score,df=3),data=dt)
summary(bs_model)
```

```
##
## Call:
## lm(formula = MaxSalary ~ bs(Score, df = 3), data = dt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1842.8  -257.8   -45.7   245.0  3343.5
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1098.6      141.2   7.781 4.28e-14 ***
```

```
## bs(Score, df = 3)1   1435.6      374.8   3.830 0.000145 ***
## bs(Score, df = 3)2   2398.6      233.9  10.257  < 2e-16 ***
## bs(Score, df = 3)3   6272.2      325.6  19.261  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 484 on 491 degrees of freedom
## Multiple R-squared:  0.8336, Adjusted R-squared:  0.8325
## F-statistic: 819.7 on 3 and 491 DF,  p-value: < 2.2e-16
```

```
p2 = ggplot(data = dt, aes(x = Score, y = MaxSalary)) +
    geom_point(color = "darkgreen", size = 1) + #Plotting the scatter point
    stat_smooth(method = "lm", se = FALSE, formula= y ~ bs(x,df=3))
p2
```