Linear Regression-1

MGinorio

11/3/2021

library(tidyverse)  
library(moderndive)  
library(skimr)  
library(ISLR)  
library(tinytex)

## Dataset

**evals dataset**

Researchers at the University of Texas in Austin, Texas (UT Austin) tried to answer the following research question: what factors explain differences in instructor teaching evaluation scores?

To this end, they collected instructor and course information on 463 courses. A full description of the study can be found at openintro.org.

glimpse(evals)

## Rows: 463  
## Columns: 14  
## $ ID <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17~  
## $ prof\_ID <int> 1, 1, 1, 1, 2, 2, 2, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, ~  
## $ score <dbl> 4.7, 4.1, 3.9, 4.8, 4.6, 4.3, 2.8, 4.1, 3.4, 4.5, 3.8, 4.~  
## $ age <int> 36, 36, 36, 36, 59, 59, 59, 51, 51, 40, 40, 40, 40, 40, 4~  
## $ bty\_avg <dbl> 5.000, 5.000, 5.000, 5.000, 3.000, 3.000, 3.000, 3.333, 3~  
## $ gender <fct> female, female, female, female, male, male, male, male, m~  
## $ ethnicity <fct> minority, minority, minority, minority, not minority, not~  
## $ language <fct> english, english, english, english, english, english, eng~  
## $ rank <fct> tenure track, tenure track, tenure track, tenure track, t~  
## $ pic\_outfit <fct> not formal, not formal, not formal, not formal, not forma~  
## $ pic\_color <fct> color, color, color, color, color, color, color, color, c~  
## $ cls\_did\_eval <int> 24, 86, 76, 77, 17, 35, 39, 55, 111, 40, 24, 24, 17, 14, ~  
## $ cls\_students <int> 43, 125, 125, 123, 20, 40, 44, 55, 195, 46, 27, 25, 20, 2~  
## $ cls\_level <fct> upper, upper, upper, upper, upper, upper, upper, upper, u~

### Question

Can we explain differences in teaching evaluation score based on various teacher attributes?

### Variables

Average teaching based on students evaluations

Attributes like gender, ethnicity, bty\_avg

Investigate correlation between this variables

* ID
* Score
* Age
* Gender
* Rank
* Cls\_students

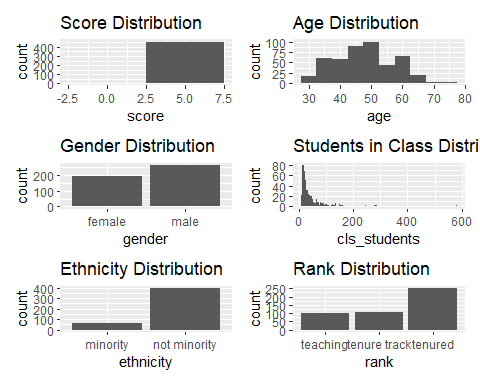
evals\_disc12 <- evals %>%   
 select(ID, score, gender, age, rank, cls\_students, ethnicity)  
  
evals\_disc12

## # A tibble: 463 x 7  
## ID score gender age rank cls\_students ethnicity   
## <int> <dbl> <fct> <int> <fct> <int> <fct>   
## 1 1 4.7 female 36 tenure track 43 minority   
## 2 2 4.1 female 36 tenure track 125 minority   
## 3 3 3.9 female 36 tenure track 125 minority   
## 4 4 4.8 female 36 tenure track 123 minority   
## 5 5 4.6 male 59 tenured 20 not minority  
## 6 6 4.3 male 59 tenured 40 not minority  
## 7 7 2.8 male 59 tenured 44 not minority  
## 8 8 4.1 male 51 tenured 55 not minority  
## 9 9 3.4 male 51 tenured 195 not minority  
## 10 10 4.5 female 40 tenured 46 not minority  
## # ... with 453 more rows

### EDA

Understanding Each Variable

We will perform an exploratory analysis of the selected variables before any formal modeling



### Statistics

evals\_disc12 %>%   
 select(score, gender, age, rank, cls\_students, ethnicity) %>% skim()

Data summary

|  |  |
| --- | --- |
| Name | Piped data |
| Number of rows | 463 |
| Number of columns | 6 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 3 |
| numeric | 3 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

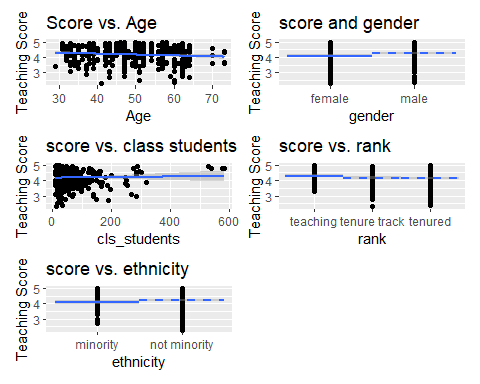
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| gender | 0 | 1 | FALSE | 2 | mal: 268, fem: 195 |
| rank | 0 | 1 | FALSE | 3 | ten: 253, ten: 108, tea: 102 |
| ethnicity | 0 | 1 | FALSE | 2 | not: 399, min: 64 |

**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| score | 0 | 1 | 4.17 | 0.54 | 2.3 | 3.8 | 4.3 | 4.6 | 5 | ▁▁▅▇▇ |
| age | 0 | 1 | 48.37 | 9.80 | 29.0 | 42.0 | 48.0 | 57.0 | 73 | ▅▆▇▆▁ |
| cls\_students | 0 | 1 | 55.18 | 75.07 | 8.0 | 19.0 | 29.0 | 60.0 | 581 | ▇▁▁▁▁ |

**Visual Relationship**

Following the scatterplots indicate the relationship of the independent variables witht the dependent variable (score)



* Score =
* Age =
* Cls\_students =

CATEGORICAL VARS -

When using a categorical predictor variable, the intercept corresponds to the mean for the baseline group, while coefficients for the non-baseline groups are offsets from this baseline. Thus in the visualization the baseline for comparison group’s median is marked with a solid line, whereas all offset groups’ medians are marked with dashed lines

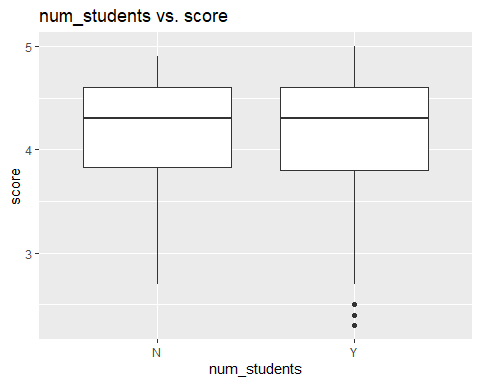
* Gender =
* Rank =
* Ethnicity =

### Dichotomous Var

We can see in the scatter plot above that the amount of students in the classrooms goes from 0-600. In which the vast majority is distributed between 0-200. Having that in mind I will create a Dichotomous variable for num\_students in a classroom less than 100 for “Y” and more than 100 to “N”

evals\_disc12$num\_students <- ifelse(evals\_disc12$cls\_students <= 100, "Y", "N")  
evals\_disc12$num\_students\_n <- ifelse(evals\_disc12$num\_students == "Y", 1,0)

evals\_disc12 %>%   
 ggplot(aes(x = num\_students, y = score)) +  
 geom\_boxplot() +   
 labs(title = "num\_students vs. score")

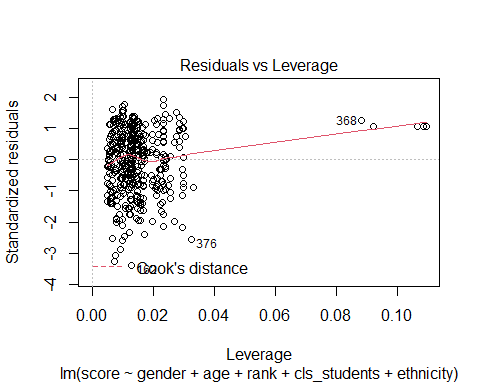
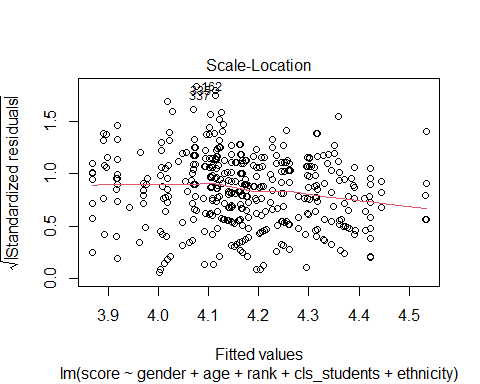
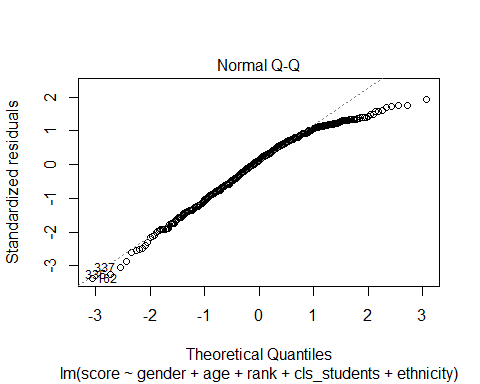
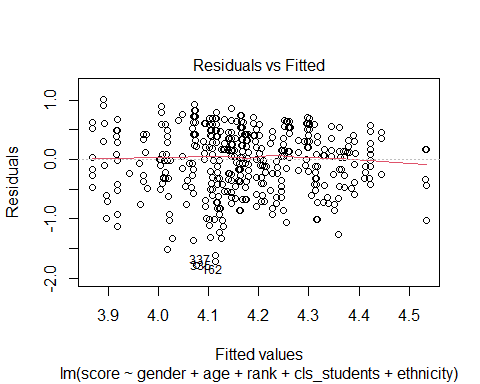
 num\_students boxplot indicates a slight difference in score when teacher has more than 100 students

### Create Model

**Score Model**

##   
## Call:  
## lm(formula = score ~ gender + age + rank + cls\_students + ethnicity,   
## data = evals\_disc12)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.7769 -0.3589 0.0773 0.4233 1.0083   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.649e+00 1.759e-01 26.429 < 2e-16 \*\*\*  
## gendermale 1.942e-01 5.328e-02 3.645 0.000298 \*\*\*  
## age -1.098e-02 3.111e-03 -3.531 0.000457 \*\*\*  
## ranktenure track -2.165e-01 8.246e-02 -2.626 0.008941 \*\*   
## ranktenured -1.659e-01 6.395e-02 -2.594 0.009794 \*\*   
## cls\_students 8.954e-05 3.380e-04 0.265 0.791216   
## ethnicitynot minority 9.342e-02 7.318e-02 1.277 0.202406   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5299 on 456 degrees of freedom  
## Multiple R-squared: 0.06308, Adjusted R-squared: 0.05076   
## F-statistic: 5.117 on 6 and 456 DF, p-value: 4.21e-05

plot(score\_model)



### Transform Variables

**Quadratic Term**

I decided to create quadratic term for class students

evals\_disc12$cls\_students\_sq2 <- evals\_disc12$cls\_students^2

**Dichonomous by quatitative**

to create a dichotomous by quatitative variable. I will multiply age by a Dichotomous variable ‘num\_students’

num\_students was created earlier based on the scatterplot

evals\_disc12$cls\_by\_age <- evals\_disc12$num\_students\_n \* evals\_disc12$age

### Second Model

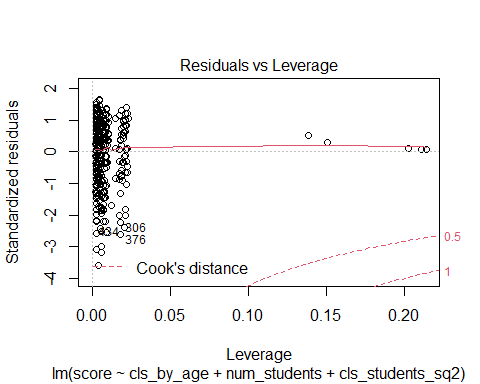
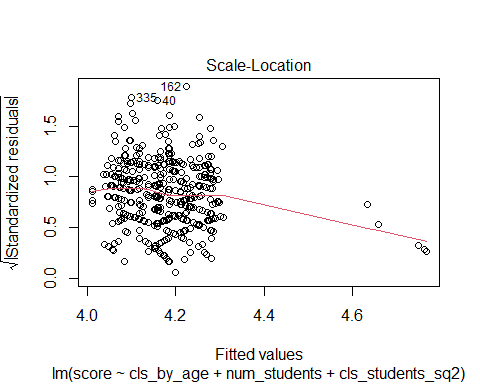
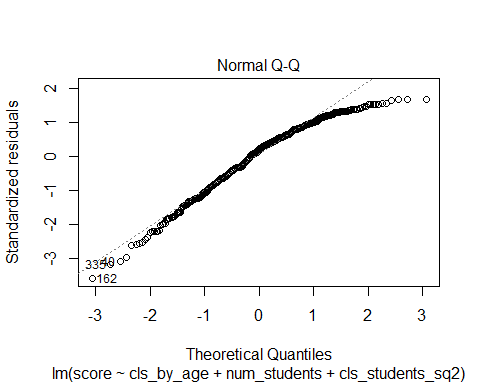
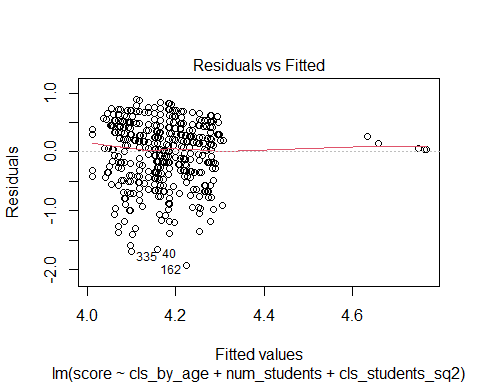
**New Variables**

* cls\_by\_age
* num\_students\_n
* num\_students

score\_model\_t <- lm(score ~ cls\_by\_age + num\_students + cls\_students\_sq2, data = evals\_disc12)  
  
summary(score\_model\_t)

##   
## Call:  
## lm(formula = score ~ cls\_by\_age + num\_students + cls\_students\_sq2,   
## data = evals\_disc12)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.9245 -0.3462 0.1093 0.4183 0.8889   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.022e+00 8.643e-02 46.530 <2e-16 \*\*\*  
## cls\_by\_age -6.674e-03 2.777e-03 -2.403 0.0166 \*   
## num\_studentsY 4.761e-01 1.612e-01 2.954 0.0033 \*\*   
## cls\_students\_sq2 2.205e-06 8.547e-07 2.580 0.0102 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5382 on 459 degrees of freedom  
## Multiple R-squared: 0.02692, Adjusted R-squared: 0.02056   
## F-statistic: 4.232 on 3 and 459 DF, p-value: 0.005752

plot(score\_model\_t)



## Conclusion

Overall, Adjusted R-squared: 0.009235 wich means the model only accounts for 0.9% of the variability in the data.

the selection of class size (cls\_students) given the influence of the outliers.

this model does not meet the level of the baseline model.