Linear Regression

MGinorio

11/3/2021

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

## 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~

### Subset data

evals\_disc11 <- evals %>%   
 select(ID, score, bty\_avg, age)  
  
evals\_disc11

## # A tibble: 463 x 4  
## ID score bty\_avg age  
## <int> <dbl> <dbl> <int>  
## 1 1 4.7 5 36  
## 2 2 4.1 5 36  
## 3 3 3.9 5 36  
## 4 4 4.8 5 36  
## 5 5 4.6 3 59  
## 6 6 4.3 3 59  
## 7 7 2.8 3 59  
## 8 8 4.1 3.33 51  
## 9 9 3.4 3.33 51  
## 10 10 4.5 3.17 40  
## # ... with 453 more rows

## EDA

Random sample of 10 out of the 463 course at UT Ausitn

evals\_disc11 %>%   
 sample\_n(size = 10)

## # A tibble: 10 x 4  
## ID score bty\_avg age  
## <int> <dbl> <dbl> <int>  
## 1 135 4.8 4.17 64  
## 2 441 3.7 7.83 35  
## 3 355 4.9 3.33 50  
## 4 60 3.6 4.83 35  
## 5 449 4.5 4.33 60  
## 6 107 4.3 4.33 46  
## 7 261 4.6 3.17 52  
## 8 153 4.2 4.83 52  
## 9 215 4.7 3.67 60  
## 10 397 4.8 3.33 45

### Statistics

evals\_disc11 %>%   
 select(score, bty\_avg) %>% skim()

Data summary

|  |  |
| --- | --- |
| Name | Piped data |
| Number of rows | 463 |
| Number of columns | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| numeric | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| score | 0 | 1 | 4.17 | 0.54 | 2.30 | 3.80 | 4.30 | 4.6 | 5.00 | ▁▁▅▇▇ |
| bty\_avg | 0 | 1 | 4.42 | 1.53 | 1.67 | 3.17 | 4.33 | 5.5 | 8.17 | ▃▇▇▃▂ |

### Correlation Coefficient

evals\_disc11 %>%   
 get\_correlation(formula = score ~ bty\_avg)

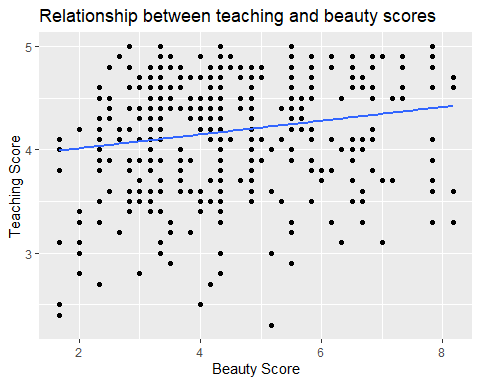
## # A tibble: 1 x 1  
## cor  
## <dbl>  
## 1 0.187

**correlation coefficient 0.187 indicates that the relationship between teaching evaluation and is**

**unjittered scatterplot**

ggplot(data = evals\_disc11, aes(x = bty\_avg, y = score)) +   
 geom\_point() +  
 labs(x = "Beauty Score", y = "Teaching Score",  
 title = "Relationship between teaching and beauty scores") +   
 geom\_smooth(method = "lm", se = F)

## `geom\_smooth()` using formula 'y ~ x'

 Regression line is consistent with the earlier result of correlation coefficient 0.187

**as instructors have higher beauty scores they receive higher teacher evaluations**

### Fit Model

score\_model <- lm(score ~ bty\_avg, data = evals\_disc11)  
  
get\_regression\_table(score\_model)

## # A tibble: 2 x 7  
## term estimate std\_error statistic p\_value lower\_ci upper\_ci  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 intercept 3.88 0.076 51.0 0 3.73 4.03   
## 2 bty\_avg 0.067 0.016 4.09 0 0.035 0.099

The intercept / average teaching score =

Relationship between teaching and beauty =

**Note**

* Sign is positive
* positive relationship
* **teachers with higher beauty tend to have higher teaching scores**
* Correlation Coefficient 0.187 is positive
* slope interpretation
* \*for every increase of 1 unit in bty\_avg  
  there is an associated increase of,  
  on average, 0.0067 units of score\*

### Residuals

regression\_points <- get\_regression\_points(score\_model)  
regression\_points

## # A tibble: 463 x 5  
## ID score bty\_avg score\_hat residual  
## <int> <dbl> <dbl> <dbl> <dbl>  
## 1 1 4.7 5 4.21 0.486  
## 2 2 4.1 5 4.21 -0.114  
## 3 3 3.9 5 4.21 -0.314  
## 4 4 4.8 5 4.21 0.586  
## 5 5 4.6 3 4.08 0.52   
## 6 6 4.3 3 4.08 0.22   
## 7 7 2.8 3 4.08 -1.28   
## 8 8 4.1 3.33 4.10 -0.002  
## 9 9 3.4 3.33 4.10 -0.702  
## 10 10 4.5 3.17 4.09 0.409  
## # ... with 453 more rows

**Note**

* score = y
* bty\_avg =
* score\_hat =
* residual =

## Conclusion

**sum of squared residuals**

if we compute the residuals for all 463 courses’ instructors and compute the sum of squared residuals we would obtain the “lack of a fit in a model”

# compute the square of residuals  
  
regression\_points %>%   
 mutate(squared\_residuals = residual^2) %>%   
 summarise(sum\_of\_squared\_residuals = sum(squared\_residuals))

## # A tibble: 1 x 1  
## sum\_of\_squared\_residuals  
## <dbl>  
## 1 132.

**if the regression line fits all the points perfectly, then the sum of the squared residuals is 0. In this case as we can see this linear model was not the appropriate choice for this regression.**