Homework 6

Anahi Rodriguez

Due 10/12/2022

Homework Instructions

Make sure to add your name to the header of the document. When submitting the assignment on Gradescope, be sure to assign the appropriate pages of your submission to each Exercise.

The point value for each exercise is noted in the exercise title.

For questions that require code, please create or use the code chunk directly below the question and type your code there. Your knitted pdf will then show both the code and the output, so that we can assess your understanding and award any partial credit.

For written questions, please provide your answer after the indicated *Answer* prompt.

You are encouraged to knit your file as you work, to check that your coding and formatting are done so appropriately. This will also help you identify and locate any errors more easily.

Homework Setup

We'll use the following packages for this homework assignment. We'll also read in data from a csv file. To access the data, you'll want to download the dataset from Canvas, and place it in the same folder as this R Markdown document. You'll then be able to use the following code to load in the data.

```
library(ggplot2)
library(MASS)
```

Exercise 1: Revisiting Professor Evaluation Scores [25 points]

Exercises 3 and 4 from Homework 4 involved examining and modeling professor evaluation scores from an average beauty measure as calculated from 6 ratings. We will continue working with the same professor evaluation dataset for Homework 6.

First, we need to load in the data. Make sure that you've downloaded the data from Canvas and that your Homework6.Rmd file is in the same folder as your data. Then, complete the following line of code to load the data as `prof_evals`.

```
# Use this code chunk to load in the data.
setwd("~/Desktop/data")
prof_evals = read.csv("Prof_Evals.csv")
```

part a

We'll fit a linear model that we'll focus on throughout most of this assignment.

Fit a linear model that predicts the evaluation `score` from the following variables:

- `bty_avg`, the average beauty rating given by 6 independent students
- `age`, the age of the professor
- `cls_students`, the size (number of students) in the class
- `cls_perc_eval`, the proportion of the class who completed the evaluations.

Then, write out the fitted linear model. Make sure that the variables are clearly defined for your written model.

```
# Use this code chunk for your answer.
lm1 = lm(data = prof_evals, score ~ bty_avg + age + cls_students + cls_perc_eval)
summary(lm1)
```

```
##
## Call:
## lm(formula = score ~ bty_avg + age + cls_students + cls_perc_eval,
##     data = prof_evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9590 -0.3426  0.1220  0.3851  1.1556
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.5934345  0.2071146  17.350  < 2e-16 ***
## bty_avg        0.0489457  0.0172216   2.842 0.004682 **
## age           -0.0024375  0.0026346  -0.925 0.355364
## cls_students   0.0005651  0.0003545   1.594 0.111606
## cls_perc_eval  0.0060699  0.0016024   3.788 0.000172 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5276 on 458 degrees of freedom
## Multiple R-squared:  0.06708,    Adjusted R-squared:  0.05893
## F-statistic: 8.233 on 4 and 458 DF,  p-value: 2.033e-06
```

Answer: estimated professor evaluation = 3.5934345 + (average beauty rating * 0.0489457) + (age * -0.0024375) + (class size * 0.0005651) + (proportion * 0.0060699)

part b

Write out interpretations for the following coefficients in the model:

- intercept
- slope for beauty average
- slope for age

Answer: we expect a professor to get an average evaluation score of 3.5934345 when his beauty rating, age, class size, and proportion of class who completed the evaluation are equal to 0

for each aditional 1 unit increase in average beauty rating, we expect the estimated professor evaluation score to increase by 0.0489457 on average

for each additional year older the professor is, we expect the professor evaluation score to decrease by 0.0024375 on average

part c

Interpret the R^2 value for this model.

Answer: 6.708% of the variation in profesor evalution score can be explained through its linear relationship with average beauty rating, age, class size, and proportion of the class who completed the survey

---

Exercise 2: Predictions for Professors A & Z [20 points]

We'll continue interpreting the model from Exercise 1.

part a

Calculate the expected evaluation score values for the following two professors with the given features:

- Professor Z, who has an average beauty score of 6.25, an age of 52, a class size of 61, and 83% of the class who completed the evaluations.
- Professor A, who has an average beauty score of 9.5, an age of 34, a class size of 270, and 96% of the class who completed the evaluations.

Print the answers for Professor Z and Professor A, and complete the following statements.

```
# Use this code chunk for your answer.
professor_z = predict(lm1, data.frame('bty_avg' = 6.25, 'age' = 52,
                                      'cls_students' = 61, 'cls_perc_eval' = 83))
professor_z
```

```
##        1
## 4.310864
```

```
professor_a = predict(lm1, data.frame('bty_avg' = 9.5, 'age' = 34,
                                      'cls_students' = 270, 'cls_perc_eval' = 96))
professor_a
```

```
##        1
## 4.710826
```

Answer:

- Professor Z has an expected evaluation score of: 4.310864
- Professor A has an expected evaluation score of: 4.710826

part b

Suppose Professor Z has an evaluation score of 4.6, and Professor A has an evaluation score of 3.7. Calculate and report the residual for each professor.

```
# Use this code chunk as needed for your answer.
4.6 - 4.310864
```

```
## [1] 0.289136
```

```
3.7 - 4.710826
```

```
## [1] -1.010826
```

Answer:

- Professor Z has a residual of: 0.289136
- Professor A has a residual of: -1.010826

part c

Calculate a 85% confidence interval for the mean response of professors with the same characteristics as Professor A.

```
# Use this code chunk for your solution.
predict(lm1, level = 0.85, newdata = data.frame(bty_avg = 9.5, age = 34,
                                                cls_students = 270, cls_perc_eval = 96),
        interval = 'confidence')
```

```
##        fit      lwr      upr
## 1 4.710826 4.543896 4.877756
```

part d

Calculate a 75% prediction interval for an individual response of a new professor with the same characteristics as Professor Z.

```
# Use this code chunk for your solution.
predict(lm1, level = 0.75, newdata = data.frame(bty_avg = 6.25, age = 52,
                                                cls_students = 61, cls_perc_eval = 83),
        interval = 'prediction')
```

```
##        fit     lwr      upr
## 1 4.310864 3.70108 4.920648
```

Exercise 3: Evaluating Professor Coefficients [30 points]

part a

Calculate 80% confidence intervals for the **true intercept**, **true slope for class size**, and **true slope for proportion of the class who complete the evaluation**.

Complete the following statements with your answers.

```
# Use this code chunk for your answer.
confint(lm1, level = 0.80, parm = c('(Intercept)', 'cls_students', 'cls_perc_eval'))
```

```
##                       10 %        90 %
## (Intercept)   3.3276230059 3.859245931
## cls_students  0.0001101363 0.001020059
## cls_perc_eval 0.0040133197 0.008126380
```

Answer:

The confidence intervals for the coefficients are:

- Intercept: (3.3276230059, 3.859245931)
- Slope for Class Size: (0.0001101363, 0.001020059)
- Slope for Class Proportion: (0.0040133197, 0.008126380)

part b

Interpret the confidence interval for the **class size** from part a. Based on your confidence interval, do you believe that the slope for class size is significantly different from 0? Explain. Does the p-value for this coefficient support your claim? Include the p-value in your explanation.

Answer:

Interpretation for **class size**: We are 80% confident that the true class size slope is between 0.0001101363 and 0.001020059

Explanation: Yes, we reject the null that the true slope for class size is equal to 0 at 20% level because 0 is not in our 80% confidence interval. The p-value is 0.111606, supporting the claim at the 20% level becuase $0.111606 < 0.20$

part c

Write the hypotheses being tested for the hypothesis test described in part b.

Answer: H_0: $\beta_{class-size} = 0$ for the model: professor evaluation score = β_0 + β_1 * avg beauty rating + β_2 * age + $\beta_{class-size}$ * class size + β_4 * proportion of class size

H_1: $\beta_{class-size} \neq 0$ for the model: professor evaluation score = β_0 + β_1 * avg beauty rating + β_2 * age + $\beta_{class-size}$ * class size + β_4 * proportion of class size

part d

Calculate an 95% confidence interval for the slope for the average beauty rating. Comment on how your interval compares to the one from Homework 4, Exercise 4 part b. Be sure to discuss the centers and lengths of the two intervals as well as the overlap between the two intervals.

```
# Use this code chunk for your answer.
confint(lm1, level = 0.95, parm = c('bty_avg'))
```

```
##             2.5 %     97.5 %
## bty_avg 0.0151026 0.08278874
```

```
# old model
lm2 = lm(score ~ bty_avg, data = prof_evals)
confint(lm2, level = 0.95, parm = c('bty_avg'))
```

```
##              2.5 %     97.5 %
## bty_avg 0.03462335 0.09865066
```

```
abs((0.0151026  - 0.08278874)/2) - abs((0.0340600000 - 0.09865066)/2) # new center - old center
```

```
## [1] 0.00154774
```

```
abs(0.0151026  - 0.08278874) - abs(0.0340600000 - 0.09865066) # new length -  old length
```

```
## [1] 0.00309548
```

```
 0.08278874 - 0.03462335
```

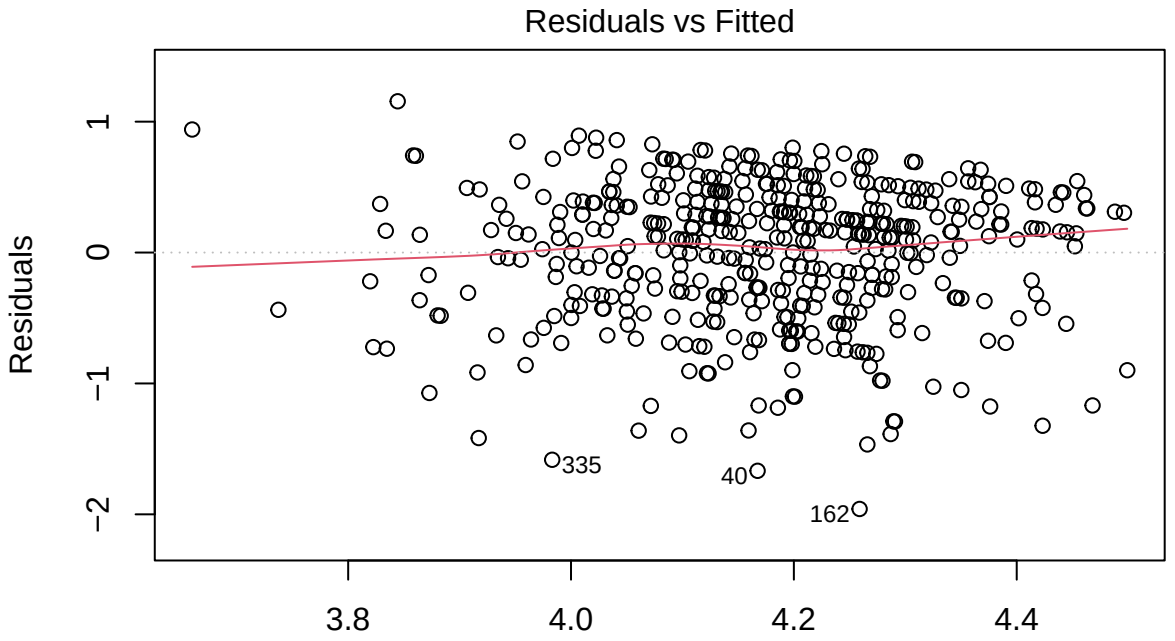
```
## [1] 0.04816539
```

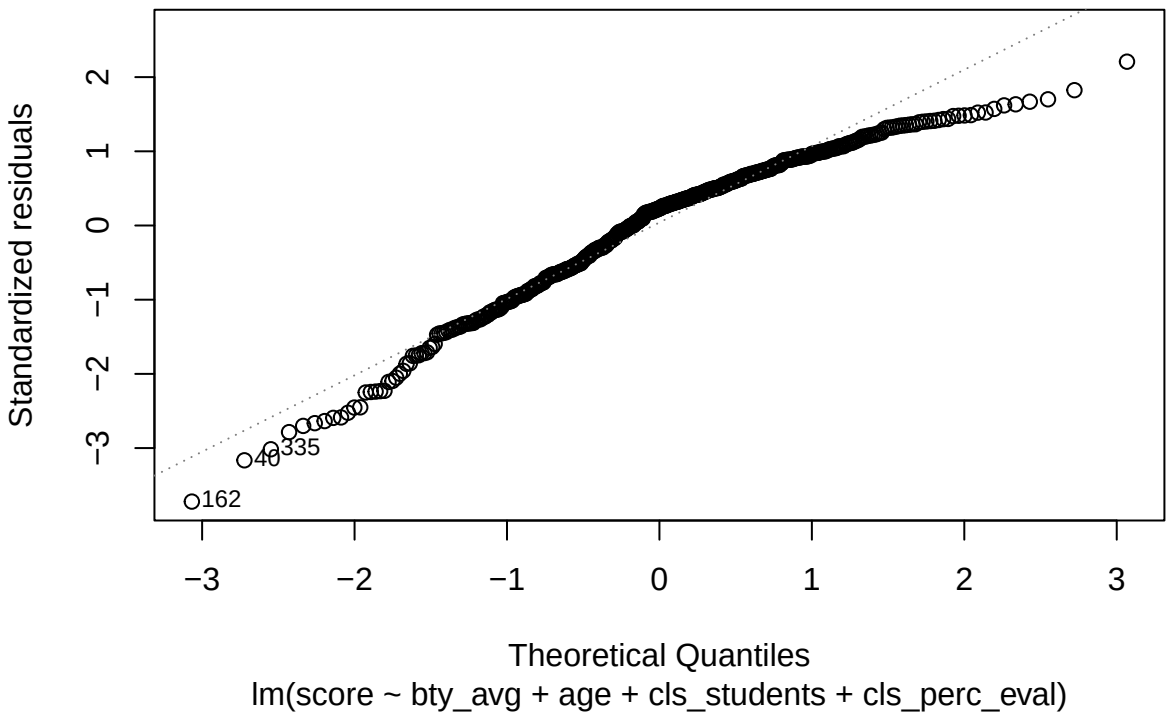
Answer: The new center value is 0.00154774 greater than the old center value previously calculated and similarly, the spread is larger than what was previously calculated by 0.00309548. The old and new confidence intervals have an overlap of 0.04816539 from (0.08278874 - 0.03462335)

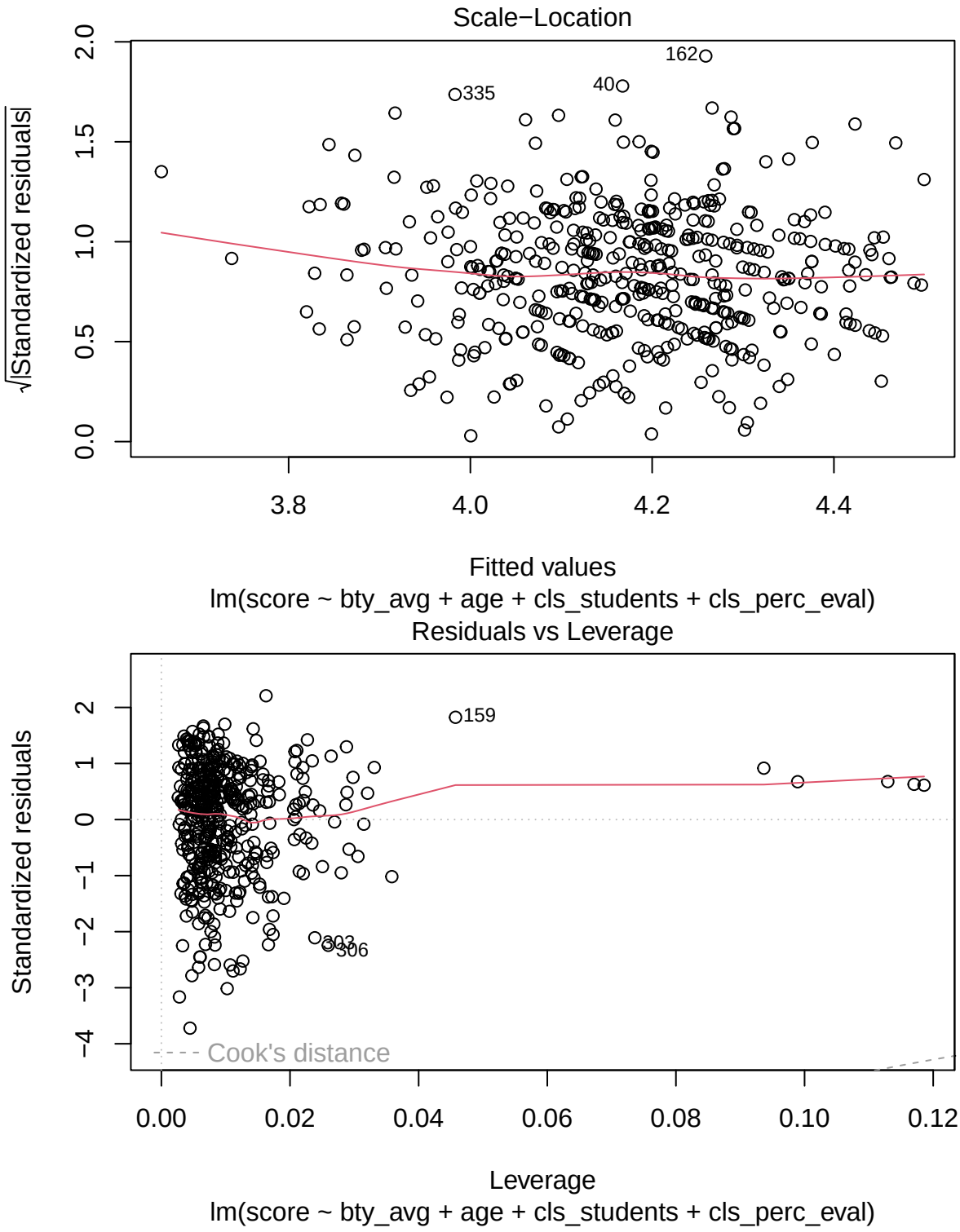
part e

For the inference to be valid, four assumptions need to be met. We can check three of those assumptions using plots. Generate the two plots to check these assumptions (it's ok if four plots are generated). State whether the assumptions seem reasonable from the plots, and explain your answer.

```
# Use this code chunk for your answer.
plot(lm1)
```


Residuals vs Fitted
Residuals
1
0
−1
−2
335
40
162
3.8
4.0
4.2
4.4
Fitted values
lm(score ~ bty_avg + age + cls_students + cls_perc_eval)


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−3
335
40
162
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(score ~ bty_avg + age + cls_students + cls_perc_eval)



Answer: I would say that the constant variance of errors assumption is not reasonably met as the error bars seem to become smaller at larger fitted values, the linear relationship between x and y assumption is not met because the residuals at fitted values are not centered around 0 as represented by the red line. The normality of errors assumption is not met as shown by the observations that fall below the diagonal at high and low theoretical quantile values.

Exercise 4: Comparing Professor Models [20 points]

For this exercise, we will be comparing the models fit in this Homework assignment (HW 6 Exercise 1) and in Homework 4 (HW 4 Exercise 3).

part a

For the two professor models (HW 4 Exercise 3 model and HW 6 Exercise 1 model), which do you expect to have a higher R^2 value (if either)? Explain your answer. Report and compare the actual R^2 values for these two models. No need to recompute these values, although you can if helpful.

```
# Use this code chunk for your answer, if needed.
summary(lm1)
```

```
##
## Call:
## lm(formula = score ~ bty_avg + age + cls_students + cls_perc_eval,
##     data = prof_evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9590 -0.3426  0.1220  0.3851  1.1556
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.5934345  0.2071146  17.350  < 2e-16 ***
## bty_avg        0.0489457  0.0172216   2.842 0.004682 **
## age           -0.0024375  0.0026346  -0.925 0.355364
## cls_students   0.0005651  0.0003545   1.594 0.111606
## cls_perc_eval  0.0060699  0.0016024   3.788 0.000172 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5276 on 458 degrees of freedom
## Multiple R-squared:  0.06708,    Adjusted R-squared:  0.05893
## F-statistic: 8.233 on 4 and 458 DF,  p-value: 2.033e-06
```

```
summary(lm2)
```

```
##
## Call:
## lm(formula = score ~ bty_avg, data = prof_evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9246 -0.3690  0.1420  0.3977  0.9309
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.88033    0.07614   50.96  < 2e-16 ***
## bty_avg      0.06664    0.01629    4.09 5.08e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5348 on 461 degrees of freedom
## Multiple R-squared:  0.03502,    Adjusted R-squared:  0.03293
```

```
## F-statistic: 16.73 on 1 and 461 DF,  p-value: 5.082e-05
```

Answer: I expect the newer model to have a R^2 value at least as high as the older model because there are more variables to explain estimated professor evaluation score through their linear relationship with each other. If none of the new variables have any effect on professor evaluation score, then the R^2 values will be equal to each other. The R^2 value for the old model is 3.5% whereas the new model has a value of 6.7%.

part b

Calculate the SSE and SST values for these two models. Do the observed results match what you would anticipate? Explain.

```
# Use this code chunk for your answer, if needed.
sse = sum(residuals(lm1)^2)
sse
```

```
## [1] 127.4874
```

```
y_bar = mean(prof_evals$score)
sst = sum((prof_evals$score - y_bar)^2)
sst
```

```
## [1] 136.6543
```

```
1 - sse/sst
```

```
## [1] 0.06708106
```

```
# old model
prof_model = lm(score ~ bty_avg, data = prof_evals)
sse_old = sum(residuals(lm2)^2)
sse_old
```

```
## [1] 131.8683
```

```
sst_old = sst
sst_old
```

```
## [1] 136.6543
```

```
1 - sse_old/sst_old
```

```
## [1] 0.03502322
```

```
# higher sse means more error so lower R^2 -- want new model to have a lower error and higher R^2
(sse/sst) <= (sse_old/sst_old)
```

```
## [1] TRUE
```

Answer: These values do match what I would anticipate because the sst is calculated using only y-values, so it will be the same for both models and the sse for the new model should be lower because we have more variables to explain the relationship instead of error to explain the relationship. This new, lower sse value should then lead to a higher R^2 value which we also find to be the case.

part c

For each of these two models, report the dimensions of the X and y matrices that would be used to calculate $\hat{\beta}$. What are the degrees of freedom associated with each of these models?

```
# Use this code chunk for your answer, if needed.
dim(prof_evals)
```

```
## [1] 463  19
```

Answer: For the old model, the dimensions of X are 463 x 2 and for y are 463 x 1. For the new model, the dimensions of X are 463 x 5 and the Y dimensions are still 463 x 1. The degrees of freedom for the old model is 463 -2 = 461 while the degrees of freedom for the new model are 463 - 5 = 458.

part d

Thinking critically about this dataset, do you have any concerns about how it could be used? Any variables in the dataset that you'd like to know more about, or any variables you'd like to have added to the dataset?

You do not need to answer all of these questions, but I am hoping that you will carefully and thoughtfully consider our professor dataset and its applications.

Answer: I understand that the models we have been creating so far are used to primarily predict if evaluation score is affected by beauty, however, I think that focusing on this variable without considering more variables pertaining to the class would be misleading. It should be the case that the content of the class and the professor's teaching style will also change the evaluation score by students, so I would like to see more variables specific to the class content and difficulty level.

Exercise 5: Formatting [5 points]

The last five points of the assignment will be earned for properly formatting your final document. Check that you have:

- included your name on the document
- properly assigned pages to exercises on Gradescope
- selected **page 1 (with your name)** and this page for this exercise (Exercise 5)
- all code is printed and readable for each question
- all output is printed
- generated a pdf file