Multiple linear regression

Student Name : Sachid Deshmukh

Date : 12/14/2018

- GitHub Location for rmd file
- GitHub Location for pdf file
- RPubs location of published file

Grading the professor

Many college courses conclude by giving students the opportunity to evaluate the course and the instructor anonymously. However, the use of these student evaluations as an indicator of course quality and teaching effectiveness is often criticized because these measures may reflect the influence of non-teaching related characteristics, such as the physical appearance of the instructor. The article titled, "Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity" (Hamermesh and Parker, 2005) found that instructors who are viewed to be better looking receive higher instructional ratings. (Daniel S. Hamermesh, Amy Parker, Beauty in the classroom: instructors pulchritude and putative pedagogical productivity, *Economics of Education Review*, Volume 24, Issue 4, August 2005, Pages 369-376, ISSN 0272-7757, 10.1016/j.econedurev.2004.07.013. http://www.sciencedirect.com/science/article/pii/S0272775704001165.)

In this lab we will analyze the data from this study in order to learn what goes into a positive professor evaluation.

The data

The data were gathered from end of semester student evaluations for a large sample of professors from the University of Texas at Austin. In addition, six students rated the professors' physical appearance. (This is aslightly modified version of the original data set that was released as part of the replication data for *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Gelman and Hill, 2007).) The result is a data frame where each row contains a different course and columns represent variables about the courses and professors.

```
load("more/evals.RData")
```

variable	description
`score`	average professor evaluation score: (1) very unsatisfactory - (5) excellent.
`rank`	rank of professor: teaching, tenure track, tenured.

variable	description
ethnicity	ethnicity of professor: not minority, minority.
gender	gender of professor: female, male.
language	language of school where professor received education: english or non-english.
age	age of professor.
cls_perc_eval	percent of students in class who completed evaluation.
cls_did_eval	number of students in class who completed evaluation.
cls_students	total number of students in class.
cls_level	class level: lower, upper.
cls_profs	number of professors teaching sections in course in sample: single, multiple.
cls_credits	number of credits of class: one credit (lab, PE, etc.), multi credit.
bty_f1lower	beauty rating of professor from lower level female: (1) lowest - (10) highest.
bty_f1upper	beauty rating of professor from upper level female: (1) lowest - (10) highest.

variable	description
`bty_f2upper`	beauty rating of professor from second upper level female: (1) lowest - (10) highest.
`bty_m1lower`	beauty rating of professor from lower level male: (1) lowest - (10) highest.
`bty_m1upper`	beauty rating of professor from upper level male: (1) lowest - (10) highest.
`bty_m2upper`	beauty rating of professor from second upper level male: (1) lowest - (10) highest.
`bty_avg`	average beauty rating of professor.
`pic_outfit`	outfit of professor in picture: not formal, formal.
`pic_color`	color of professor's picture: color, black & white.

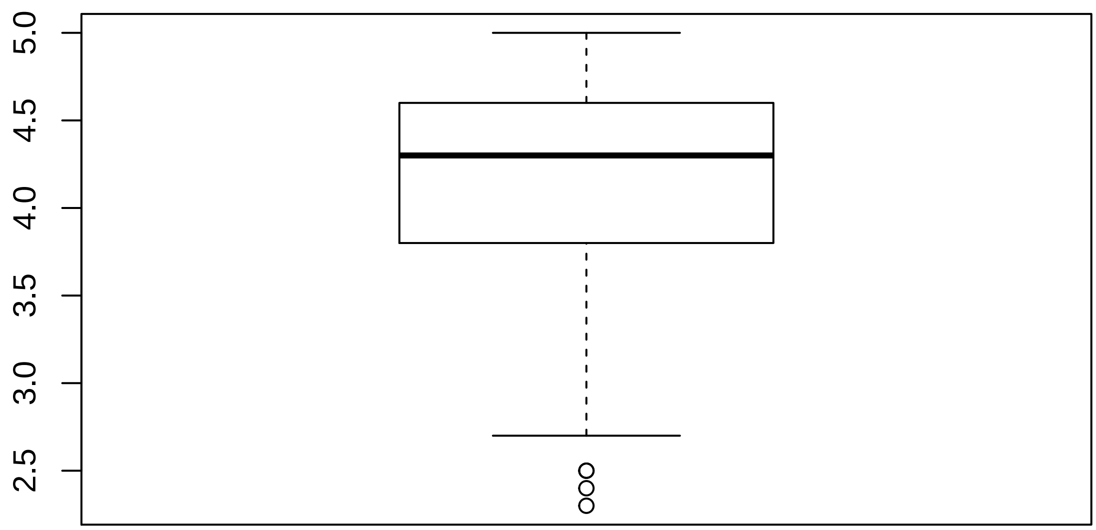
Exploring the data

1. Is this an observational study or an experiment? The original research question posed in the paper is whether beauty leads directly to the differences in course evaluations. Given the study design, is it possible to answer this question as it is phrased? If not, rephrase the question.

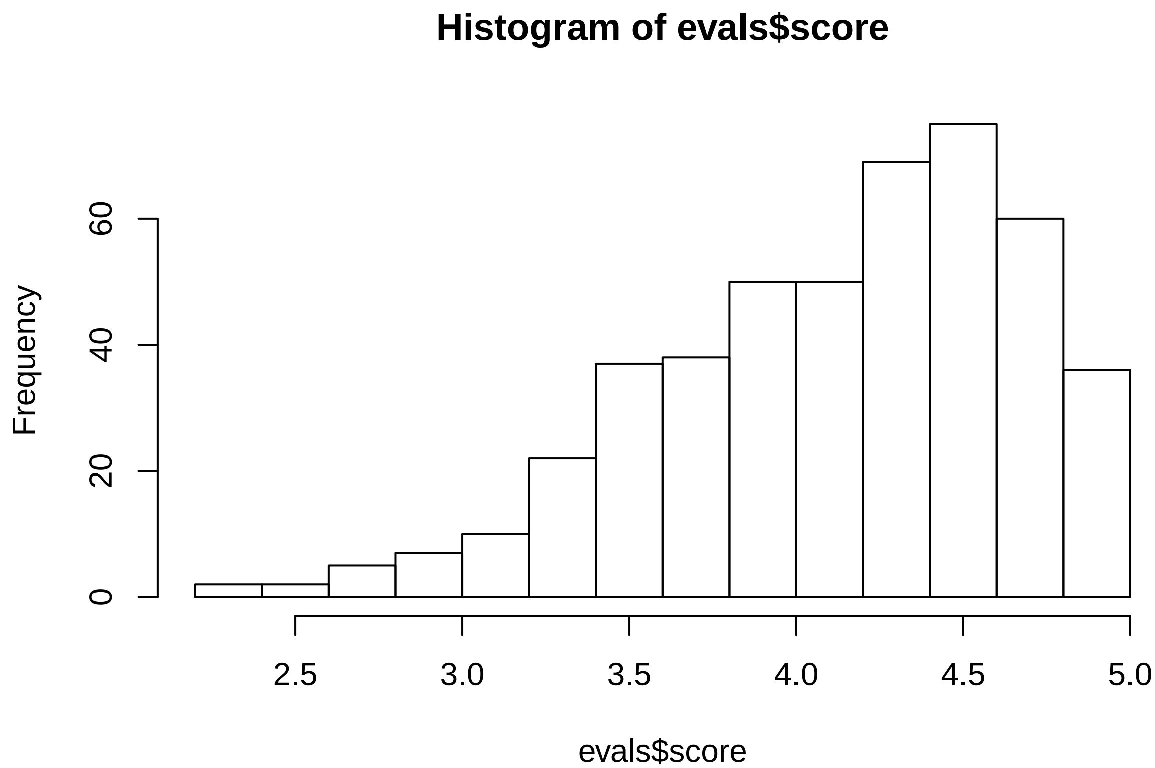
Answer: This is an observational study. This observational study is conducted with multiple independent variables and cumulative impact of those variables on course evaluation (output variable), Due to this nature of the study it is not feasible to answer a very specific question whether beauty of the teacher alone leads to differences in the course evaluations. We can re-phrase the question as Which variables amogst those collected in observational study leads to differences in the course evaluations and by what magnitude

2. Describe the distribution of `score`. Is the distribution skewed? What does that tell you about how students rate courses? Is this what you expected to see? Why, or why not?

```
boxplot(evals$score)
```



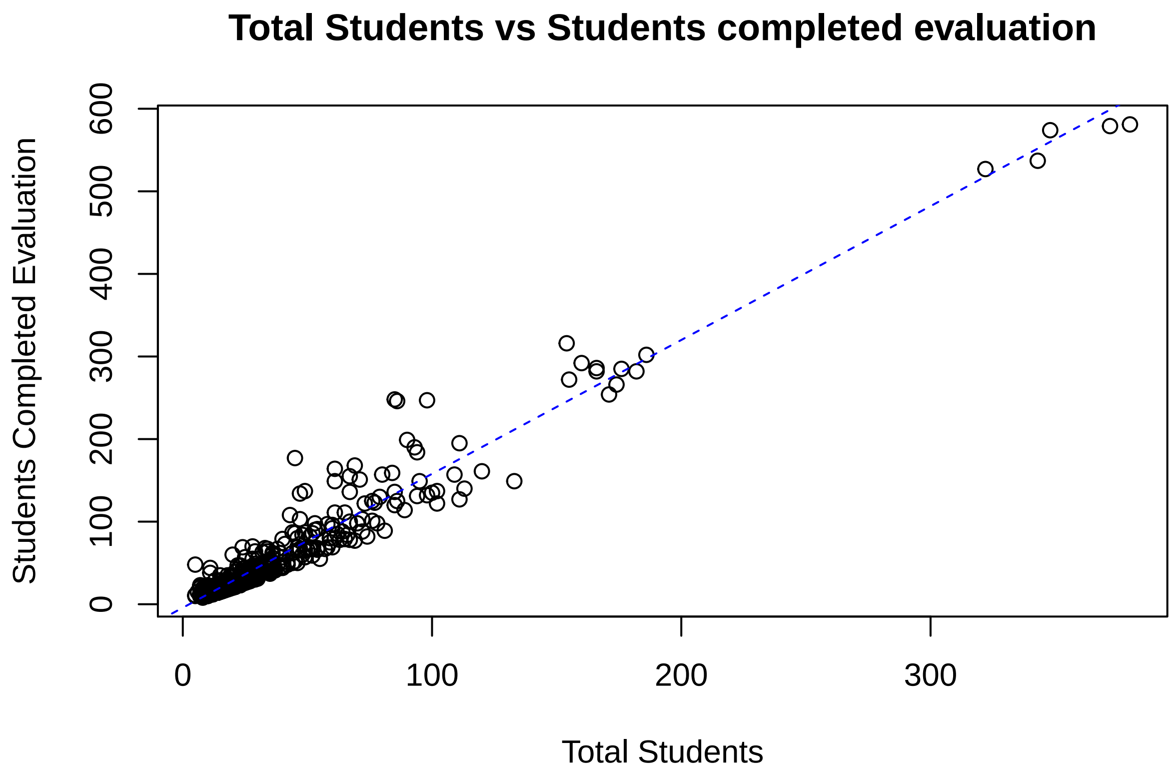

```
hist(evals$score)
```



Answer: We can see from the box plot and histogram above that students ratings are left skewed. 50% students rate teachers from 2.7 to 4.2 and another 50% students rate them from 4.2 to 5. This indicates that more students agree on higher ratings and student opinions differ a lot on lower ratings. The first 25% quartile is vary tall, indicating more diverse opinion amongst students when it comes to lower ratings. Students tend to agree more with each other as rating goes higher. This is expected because most teachers are well qualified and are very experienced and they acquire the specific teaching style which makes them popular amognst students. However there are some new teachers who are still trying to figure out optimum teaching technique and may receive less ratings from students. Since portion of new teachers is comparably lower than experenced teachers there are few observations with lower satisfaction ratings from students

3. Excluding `score`, select two other variables and describe their relationship using an appropriate visualization (scatterplot, side-by-side boxplots, or mosaic plot).

```
fit = lm(evals$cls_students~evals$cls_did_eval)
plot(evals$cls_students~evals$cls_did_eval, xlab="Total Students", ylab="Students Completed Evaluation"
abline(coefficients(fit), lty=2, col="blue")
```

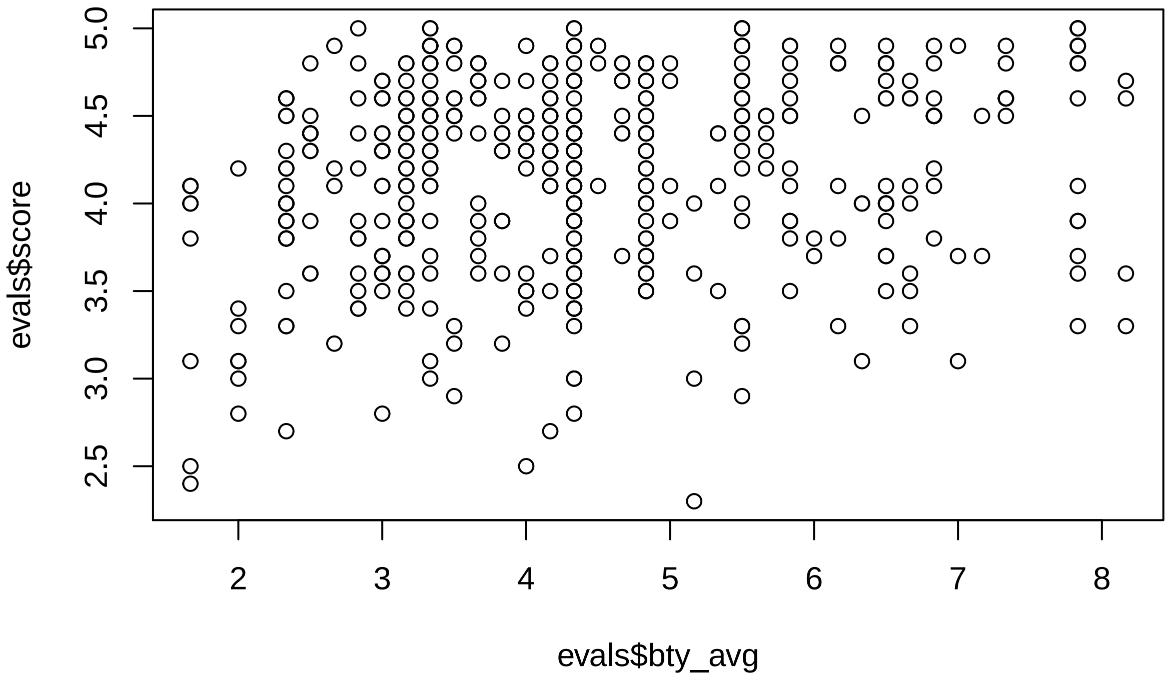


Answer: The above scatter plot shows relationship between Total Students and Students who completed evaluation. From the scatter plot it can be seen that higher students particiapate in teachers evaluation as class size increased. This shows that there is a positive correlation between Total Students and Students particiapted in evaluation

Simple linear regression

The fundamental phenomenon suggested by the study is that better looking teachers are evaluated more favorably. Let's create a scatterplot to see if this appears to be the case:

```
plot(evals$score ~ evals$bty_avg)
```



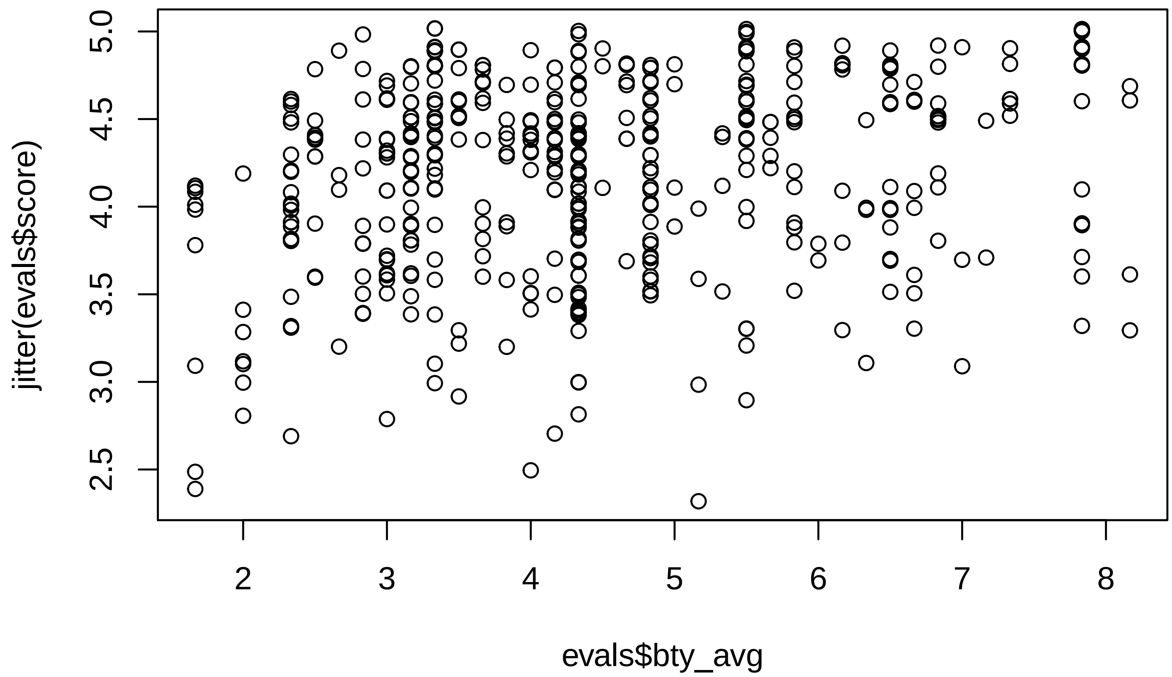
Answer: The above scatter plot is homoscadastic. This means there is no appearent relationship between avg beauty level and teachers grading. From the above scatter plot we can conclude that avg beauty level of a teacher have no impact on evaluation score

Before we draw conclusions about the trend, compare the number of observations in the data frame with the approximate number of points on the scatterplot. Is anything awry?

Answer: The points on scatter plots are way lesser than actual points in data frame. This suggests that points are overlapped

4. Replot the scatterplot, but this time use the function `jitter()` on the y- or the x-coordinate. (Use `?jitter` to learn more.) What was misleading about the initial scatterplot?

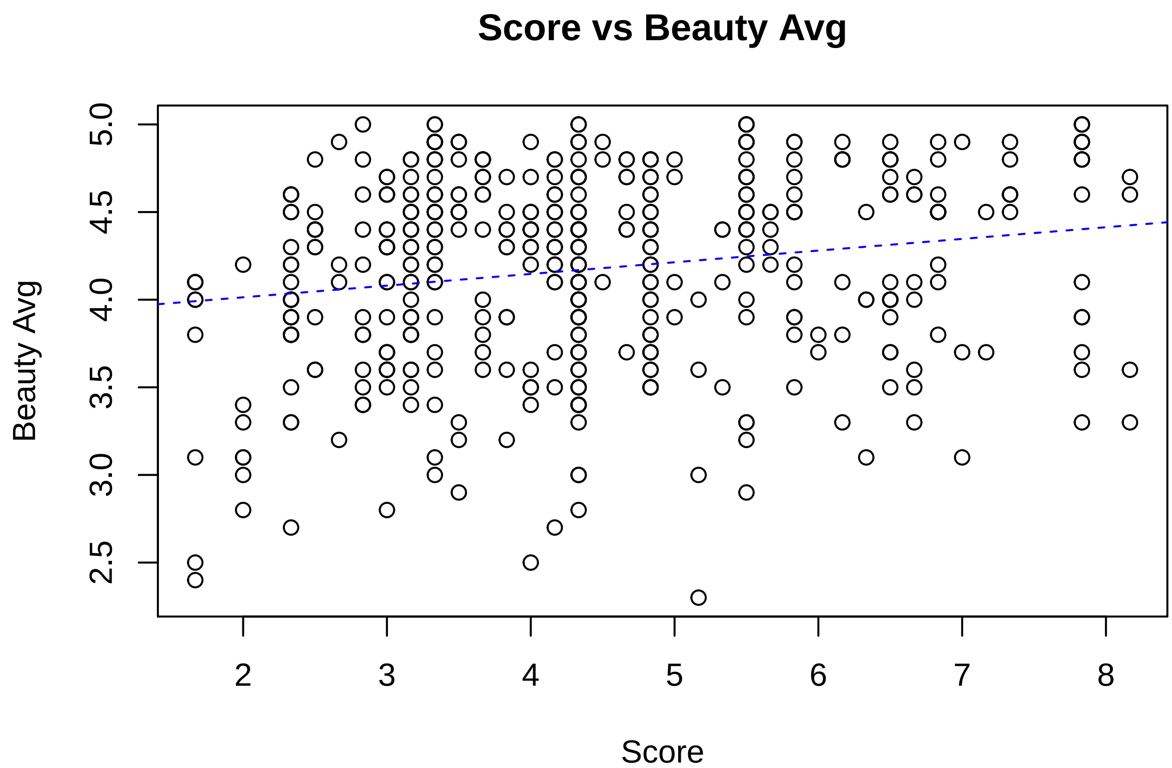
```
plot(jitter(evals$score) ~ evals$bty_avg)
```



Answer : Initial scatter plot was not able to differentiate overlapping of observations. That is misleading since each point have equal weight. We can't see that more variables are concentrated around certain beauty avg level which changes our conclusion

5. Let's see if the apparent trend in the plot is something more than natural variation. Fit a linear model called `m_bty` to predict average professor score by average beauty rating and add the line to your plot using `abline(m_bty)`. Write out the equation for the linear model and interpret the slope. Is average beauty score a statistically significant predictor? Does it appear to be a practically significant predictor?

```
m_bty = lm(evals$score~evals$bty_avg)
plot(evals$score~evals$bty_avg, xlab="Score", ylab="Beauty Avg", main="Score vs Beauty Avg")
abline(coefficients(m_bty), lty=2, col="blue")
```




```r
summary(m_bty)
```

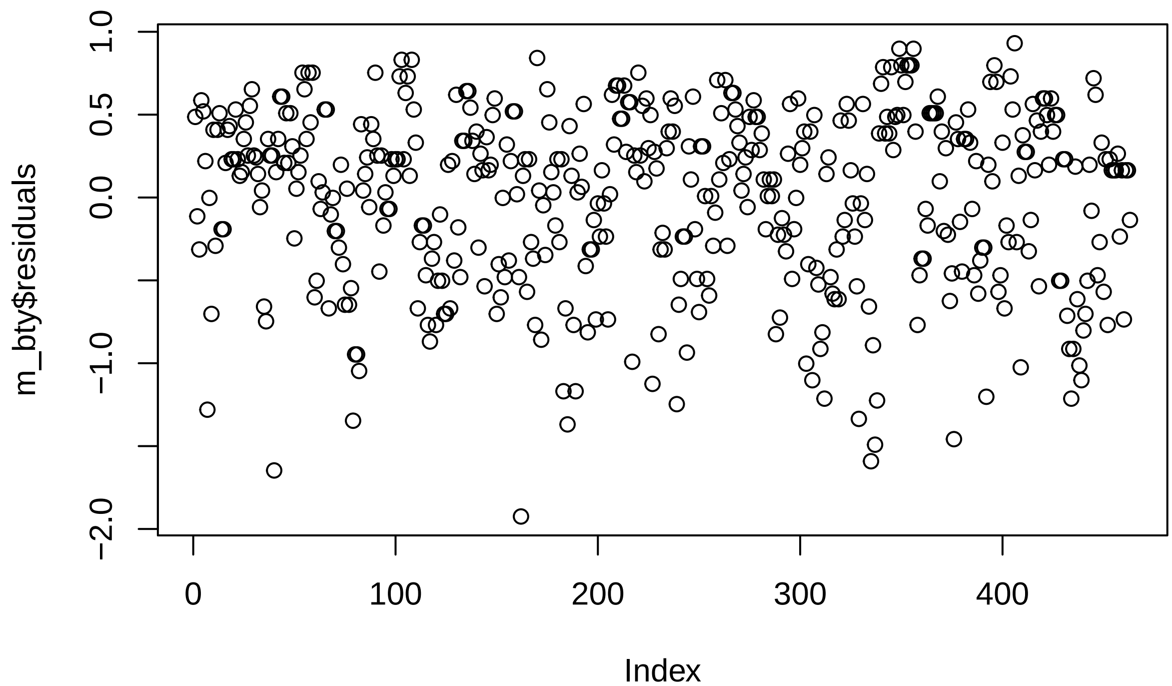
```
##
## Call:
## lm(formula = evals$score ~ evals$bty_avg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9246 -0.3690  0.1420  0.3977  0.9309
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.88034    0.07614   50.96  < 2e-16 ***
## evals$bty_avg  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: 5.083e-05
```

Answer: From the summary of the fitted model we can see that coefficient of beauty average is positive (0.06664) and p value is close to 0 making it statistically significant. We can conclude that beauty average impacts the teachers score positively and each one unit increase in beauty average increases overal teacher rating by 0.06. The equation of the linear model can be written as score = 3.88 + 0.06664 * Avg Beauty

Even though avg beauty is a statistically significant predictor for practical purposes it just improves the teacher score by 0.06 for each one point increase in avg beauty level. This is not a lot and for practical purpose we can ignore this impact

6. Use residual plots to evaluate whether the conditions of least squares regression are reasonable. Provide plots and comments for each one (see the Simple Regression Lab for a reminder of how to make these).

```
plot(m_bty$residuals)
```

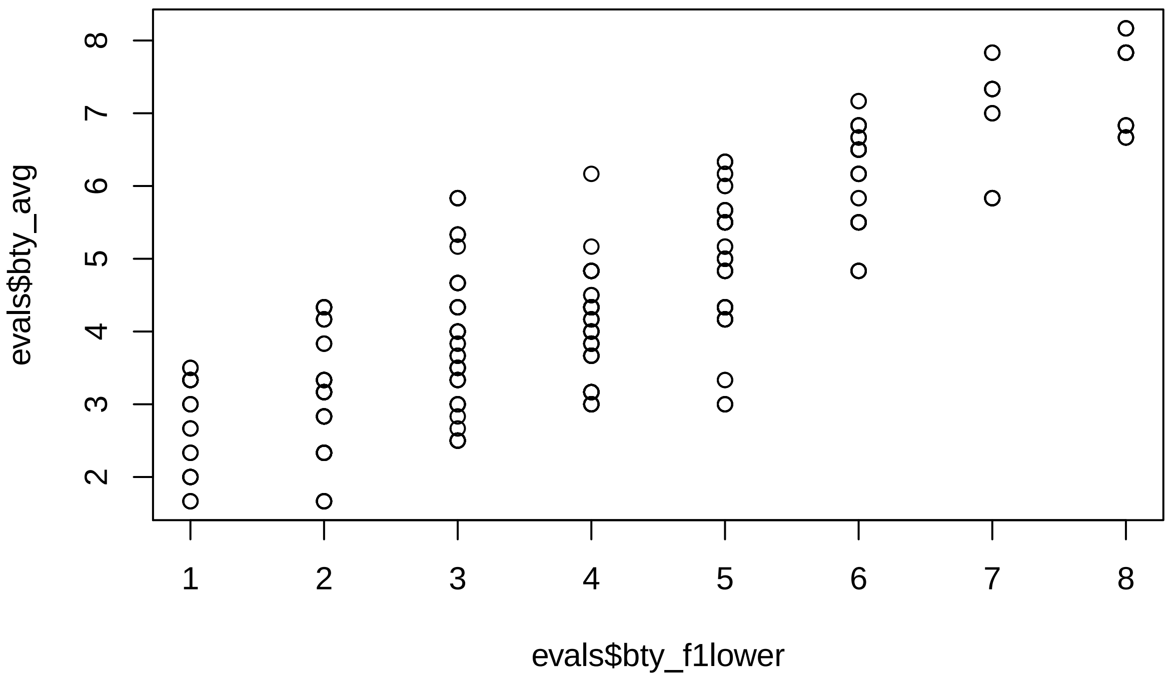


Answer : For linear regression we assume that residuals are nearly normally distributed and those are homoscadastic. Homoscadastic means there is no appearent pattern in the residual plot. From the above plot we can see that plot is y axis imbalance and errors are increasing as beauty avg value increases. Y axis imbalance tells us that we are predicting consistently lower value comapared to actual value. Residual for most of the points are negative. There is increasing pattern for high errors as avg beauty value increases. This means that we need transformation on the feature variables to make it more normally distributed

Multiple linear regression

The data set contains several variables on the beauty score of the professor: individual ratings from each of the six students who were asked to score the physical appearance of the professors and the average of these six scores. Let's take a look at the relationship between one of these scores and the average beauty score.

```
plot(evals$bty_avg ~ evals$bty_f1lower)
```

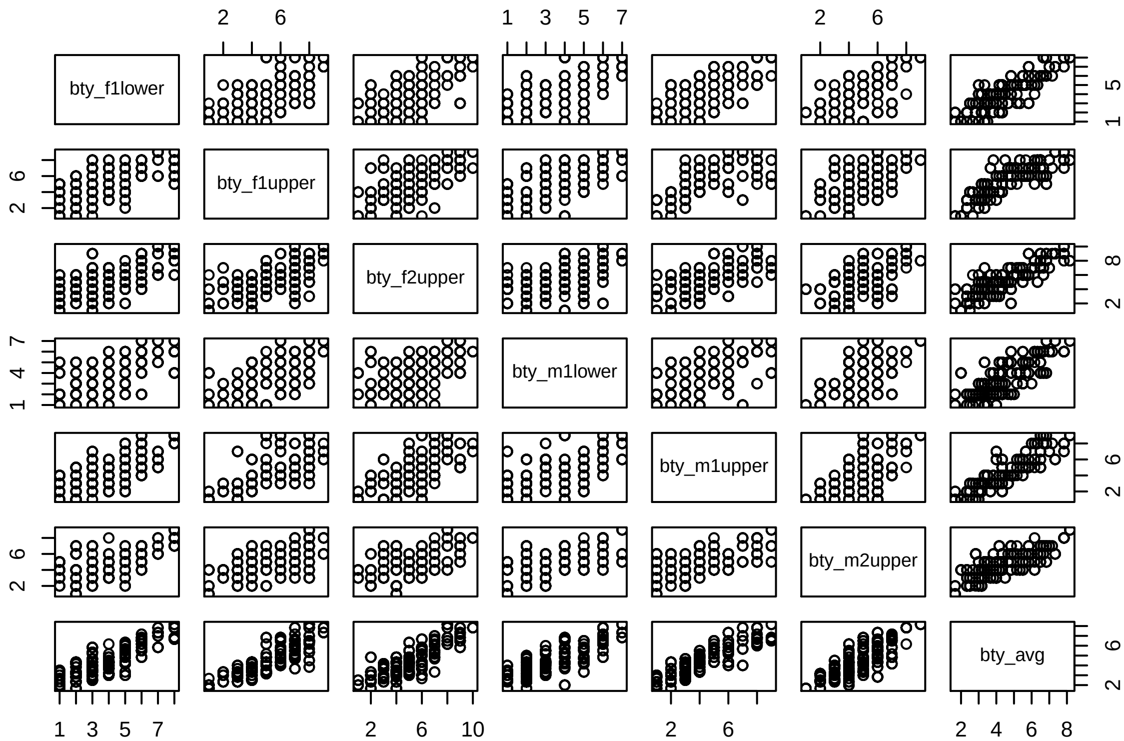



```
cor(evals$bty_avg, evals$bty_f1lower)
```

```
## [1] 0.8439112
```

As expected the relationship is quite strong - after all, the average score is calculated using the individual scores. We can actually take a look at the relationships between all beauty variables (columns 13 through 19) using the following command:

```
plot(evals[,13:19])
```



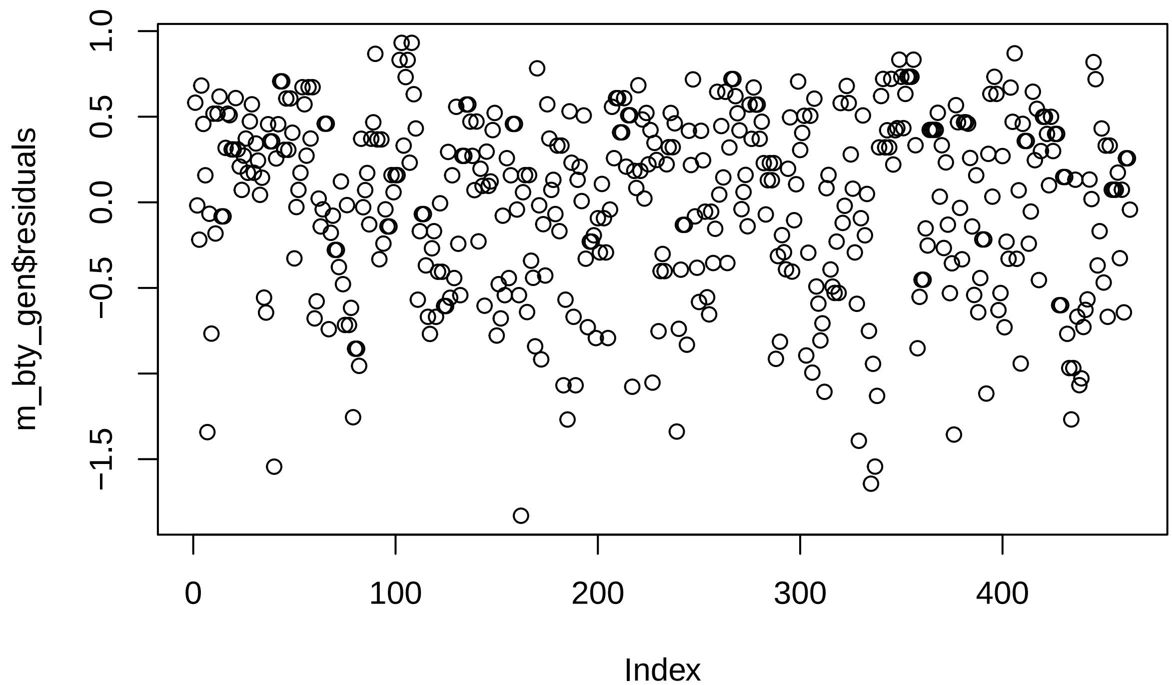
These variables are collinear (correlated), and adding more than one of these variables to the model would not add much value to the model. In this application and with these highly-correlated predictors, it is reasonable to use the average beauty score as the single representative of these variables.

In order to see if beauty is still a significant predictor of professor score after we've accounted for the gender of the professor, we can add the gender term into the model.

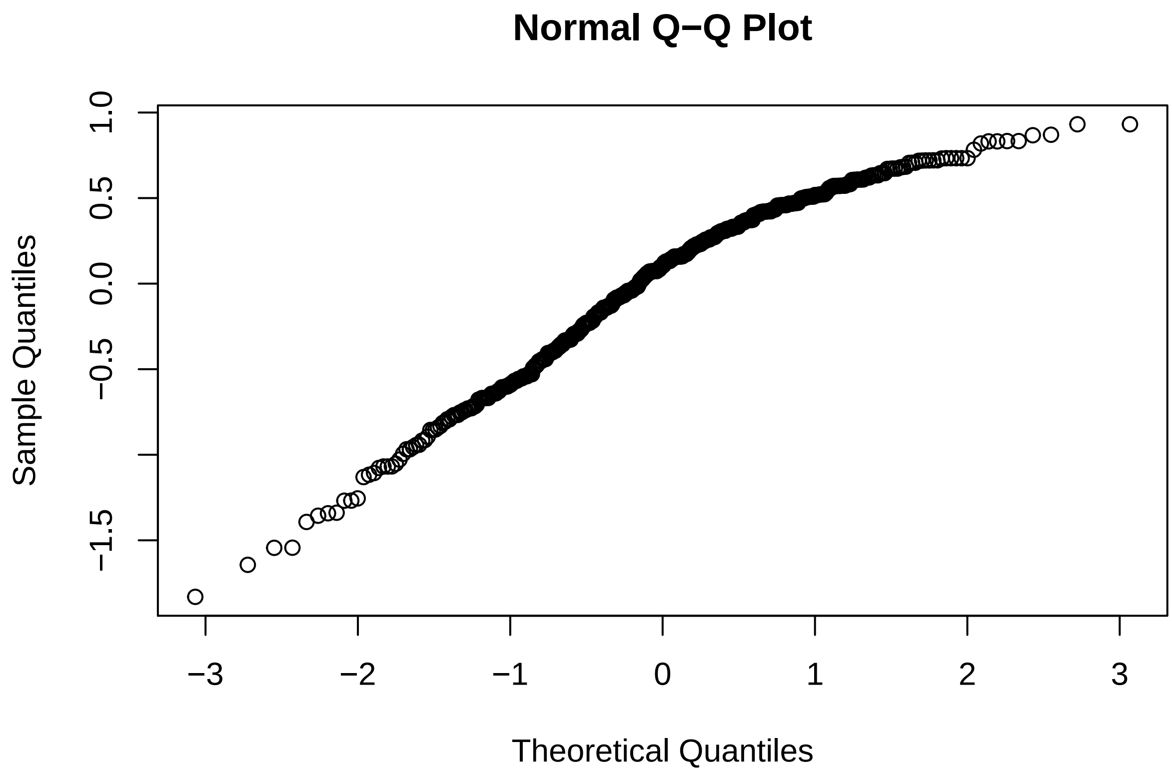
```
m_bty_gen <- lm(score ~ bty_avg + gender, data = evals)
summary(m_bty_gen)
```

7. P-values and parameter estimates should only be trusted if the conditions for the regression are reasonable. Verify that the conditions for this model are reasonable using diagnostic plots.

```
m_bty_gen <- lm(score ~ bty_avg + gender, data = evals)
plot(m_bty_gen$residuals)
```




```r
qqnorm(m_bty_gen$residuals)
```



Answer: From the diagnostic plot we can see that two important conditions for linear regression are not satsfied. The residuals are not homoscadastic. There is pattern where residuals increases on x axis. From the probability density plot we can see that residuals are not nearly normally distributed. We can't trust the p values here since conditions for normal distribution are not met

8. Is `bty_avg` still a significant predictor of `score`? Has the addition of `gender` to the model changed the parameter estimate for `bty_avg`?

Answer: From the fitted model summary we can say the avg beauty is still a significant predictor of the score (P value is statistically significant). Addition of the gender to the model has changed the coefficient value of beauty avg parameter

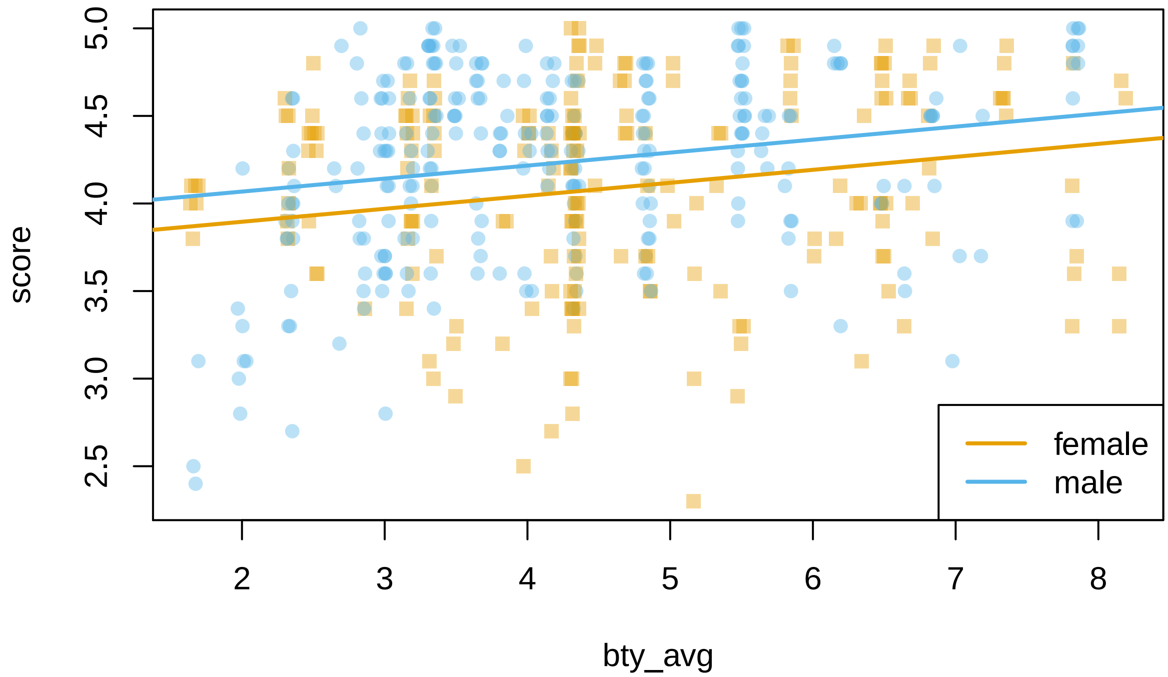
Note that the estimate for `gender` is now called `gendermale`. You'll see this name change whenever you introduce a categorical variable. The reason is that R recodes `gender` from having the values of `female` and `male` to being an indicator variable called `gendermale` that takes a value of 0 for females and a value of 1 for males. (Such variables are often referred to as "dummy" variables.)

As a result, for females, the parameter estimate is multiplied by zero, leaving the intercept and slope form familiar from simple regression.

$$\begin{aligned}\widehat{score} &= \hat{\beta}_0 + \hat{\beta}_1 \times bty\_avg + \hat{\beta}_2 \times (0) \\ &= \hat{\beta}_0 + \hat{\beta}_1 \times bty\_avg\end{aligned}$$

We can plot this line and the line corresponding to males with the following custom function.

```
multiLines(m_bty_gen)
```



9. What is the equation of the line corresponding to males? (*Hint:* For males, the parameter estimate is multiplied by 1.) For two professors who received the same beauty rating, which gender tends to have the higher course evaluation score?

Answer: The linear model equation for males can be written as score = 3.74 + (0.07 * beauty avg) + (0.17 * 1) From this equation we can see that for the same avg beauty, male gender receives 0.17 more avg score compared to female gender

The decision to call the indicator variable `gendermale` instead of`genderfemale` has no deeper meaning. R simply codes the category that comes first alphabetically as a 0. (You can change the reference level of a categorical variable, which is the level that is coded as a 0, using the`relevel` function. Use `?relevel` to learn more.)

10. Create a new model called `m_bty_rank` with `gender` removed and `rank` added in. How does R appear to handle categorical variables that have more than two levels? Note that the rank variable has three levels: `teaching`, `tenure track`, `tenured`.

```
m_bty_rank <- lm(score ~ bty_avg + rank, data = evals)
summary(m_bty_rank)
```

```
##
## Call:
## lm(formula = score ~ bty_avg + rank, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8713 -0.3642  0.1489  0.4103  0.9525
##
```

```
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.98155    0.09078  43.860  < 2e-16 ***
## bty_avg           0.06783    0.01655   4.098 4.92e-05 ***
## ranktenure track -0.16070    0.07395  -2.173   0.0303 *
## ranktenured      -0.12623    0.06266  -2.014   0.0445 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5328 on 459 degrees of freedom
## Multiple R-squared:  0.04652,    Adjusted R-squared:  0.04029
## F-statistic: 7.465 on 3 and 459 DF,  p-value: 6.88e-05
```

Answer: R uses one hot encoding for handling categorical variable. From the regression coefficient we can see that R has automatically added rank tenuretrak and rank tenured variables to the regression equation. rank teaching is excluded since it can be represented as 0 coefficient. We can interpret the above result as follows

** 1] If teacher is ranked teaching then each increase in avg beauty of the teacher results in 0.06783 increase in avg score**

** 2] If teacher is ranked tenure track then each increase in avg beauty of the teacher results in 0.06783 - 0.160 = -0.0921 increase in avg score **

** 3] If teacher is ranked tenure then each increase in avg beauty of the teacher results in 0.06783 - 0.12623 = -0.0584 increase in avg score **

We can conclude that teaching teacher are awarded more for their beauty average followed by tenured teachers and then followed by tenure track teachers

The interpretation of the coefficients in multiple regression is slightly different from that of simple regression. The estimate for `bty_avg` reflects how much higher a group of professors is expected to score if they have a beauty rating that is one point higher *while holding all other variables constant*. In this case, that translates into considering only professors of the same rank with `bty_avg` scores that are one point apart.

The search for the best model

We will start with a full model that predicts professor score based on rank, ethnicity, gender, language of the university where they got their degree, age, proportion of students that filled out evaluations, class size, course level, number of professors, number of credits, average beauty rating, outfit, and picture color.

11. Which variable would you expect to have the highest p-value in this model? Why? *Hint:* Think about which variable would you expect to not have any association with the professor score.

Answer: I expect following variables to hace highest p value in this model : - ethnicity, gender, language, age, students proportions, class size ,average beauty rating, outfit and picture color. These I think teachers score are the true reflection of teachers effectiveness and variables mentioned above don't influence teacher's effectiveness

Let's run the model…

```
m_full <- lm(score ~ rank + ethnicity + gender + language + age + cls_perc_eval
             + cls_students + cls_level + cls_profs + cls_credits + bty_avg
             + pic_outfit + pic_color, data = evals)
summary(m_full)
```

```
##
## Call:
```

```
## lm(formula = score ~ rank + ethnicity + gender + language + age +
##     cls_perc_eval + cls_students + cls_level + cls_profs + cls_credits +
##     bty_avg + pic_outfit + pic_color, data = evals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77397 -0.32432  0.09067  0.35183  0.95036
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4.0952141  0.2905277  14.096  < 2e-16 ***
## ranktenure track      -0.1475932  0.0820671  -1.798  0.07278 .
## ranktenured           -0.0973378  0.0663296  -1.467  0.14295
## ethnicitynot minority  0.1234929  0.0786273   1.571  0.11698
## gendermale             0.2109481  0.0518230   4.071 5.54e-05 ***
## languagenon-english   -0.2298112  0.1113754  -2.063  0.03965 *
## age                   -0.0090072  0.0031359  -2.872  0.00427 **
## cls_perc_eval          0.0053272  0.0015393   3.461  0.00059 ***
## cls_students           0.0004546  0.0003774   1.205  0.22896
## cls_levelupper         0.0605140  0.0575617   1.051  0.29369
## cls_profssingle       -0.0146619  0.0519885  -0.282  0.77806
## cls_creditsone credit  0.5020432  0.1159388   4.330 1.84e-05 ***
## bty_avg                0.0400333  0.0175064   2.287  0.02267 *
## pic_outfitnot formal  -0.1126817  0.0738800  -1.525  0.12792
## pic_colorcolor        -0.2172630  0.0715021  -3.039  0.00252 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.498 on 448 degrees of freedom
## Multiple R-squared:  0.1871, Adjusted R-squared:  0.1617
## F-statistic: 7.366 on 14 and 448 DF,  p-value: 6.552e-14
```

12. Check your suspicions from the previous exercise. Include the model output in your response.

Answer : From the actual model output the variables which are not significant are rank tenure track, rank tenured, ethnicity, class student, class level upper, class prof single, picture outfit not formal. Most of the variables are aligned to my conclusion in quaestion-11 above. I am surprised to see that color picture variable is significant and male gender makes a difference on teachers ratings

13. Interpret the coefficient associated with the ethnicity variable.

Answer: The ethnicity variable has p value of 0.11698 which is greater than critical value 0.05 making it statistically insignificant. From this we can conclude that ethnicity is not statistically important for predicting avg teachers rating and can be removed from the model

14. Drop the variable with the highest p-value and re-fit the model. Did the coefficients and significance of the other explanatory variables change? (One of the things that makes multiple regression interesting is that coefficient estimates depend on the other variables that are included in the model.) If not, what does this say about whether or not the dropped variable was collinear with the other explanatory variables?

```r
m_imp <- lm(score ~ gender + language + age + cls_perc_eval
             + cls_credits + bty_avg
             +  pic_color, data = evals)
summary(m_imp)
```

```
##
## Call:
## lm(formula = score ~ gender + language + age + cls_perc_eval +
##     cls_credits + bty_avg + pic_color, data = evals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.81919 -0.32035  0.09272  0.38526  0.88213
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.967255   0.215824  18.382  < 2e-16 ***
## gendermale             0.221457   0.049937   4.435 1.16e-05 ***
## languagenon-english   -0.281933   0.098341  -2.867  0.00434 **
## age                   -0.005877   0.002622  -2.241  0.02551 *
## cls_perc_eval          0.004295   0.001432   2.999  0.00286 **
## cls_creditsone credit  0.444392   0.100910   4.404 1.33e-05 ***
## bty_avg                0.048679   0.016974   2.868  0.00432 **
## pic_colorcolor        -0.216556   0.066625  -3.250  0.00124 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5014 on 455 degrees of freedom
## Multiple R-squared:  0.1631, Adjusted R-squared:  0.1502
## F-statistic: 12.67 on 7 and 455 DF,  p-value: 6.996e-15
```

Answer: We can see that once we fit the model with only keeping important variables there is a change in both significance level and value of coefficient of the input feature variables. Since removing un-important variables results in change of both significance level and value of coefficients we can conclude that removed variables were correlated to one or more important variables

15. Using backward-selection and p-value as the selection criterion, determine the best model. You do not need to show all steps in your answer, just the output for the final model. Also, write out the linear model for predicting score based on the final model you settle on.

```r
m_bktest <- lm(score ~ gender + cls_perc_eval + cls_credits +  pic_color, data = evals)
summary(m_bktest)
```

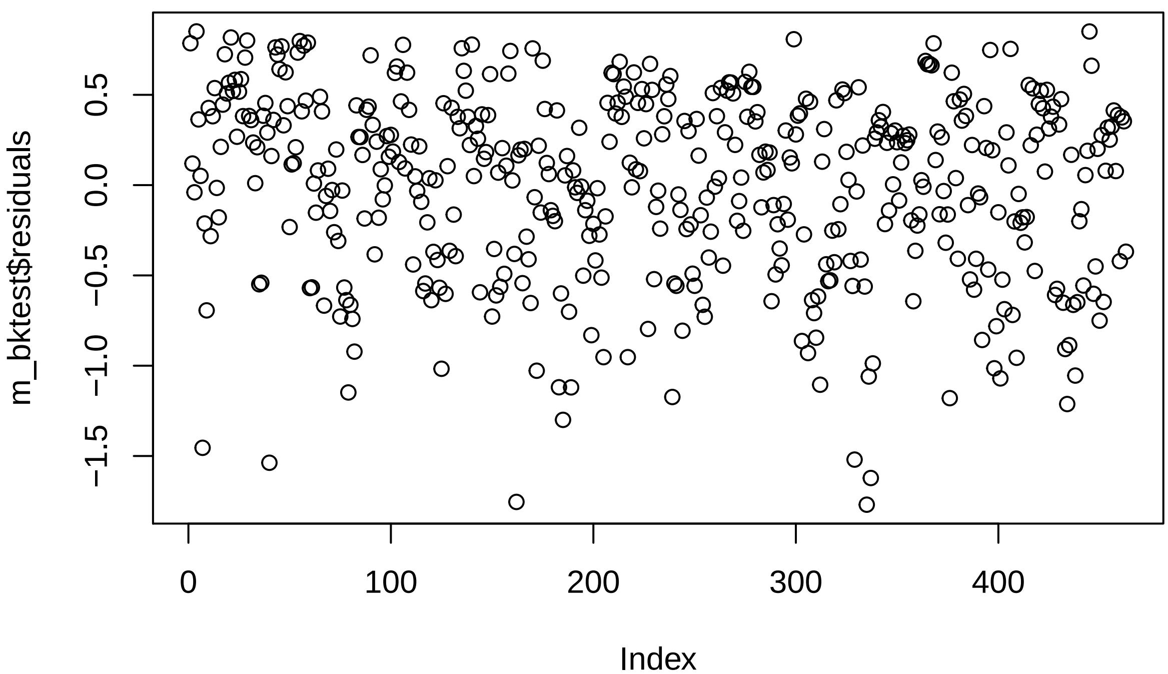
```
##
## Call:
## lm(formula = score ~ gender + cls_perc_eval + cls_credits + pic_color,
##     data = evals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.76952 -0.36391  0.08773  0.38610  0.85105
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.869080   0.131869  29.340  < 2e-16 ***
## gendermale             0.173455   0.049110   3.532 0.000454 ***
## cls_perc_eval          0.005081   0.001456   3.490 0.000529 ***
## cls_creditsone credit  0.431468   0.102999   4.189 3.36e-05 ***
## pic_colorcolor        -0.238193   0.065063  -3.661 0.000281 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5145 on 458 degrees of freedom
## Multiple R-squared:  0.1128, Adjusted R-squared:  0.1051
## F-statistic: 14.56 on 4 and 458 DF,  p-value: 3.349e-11
```

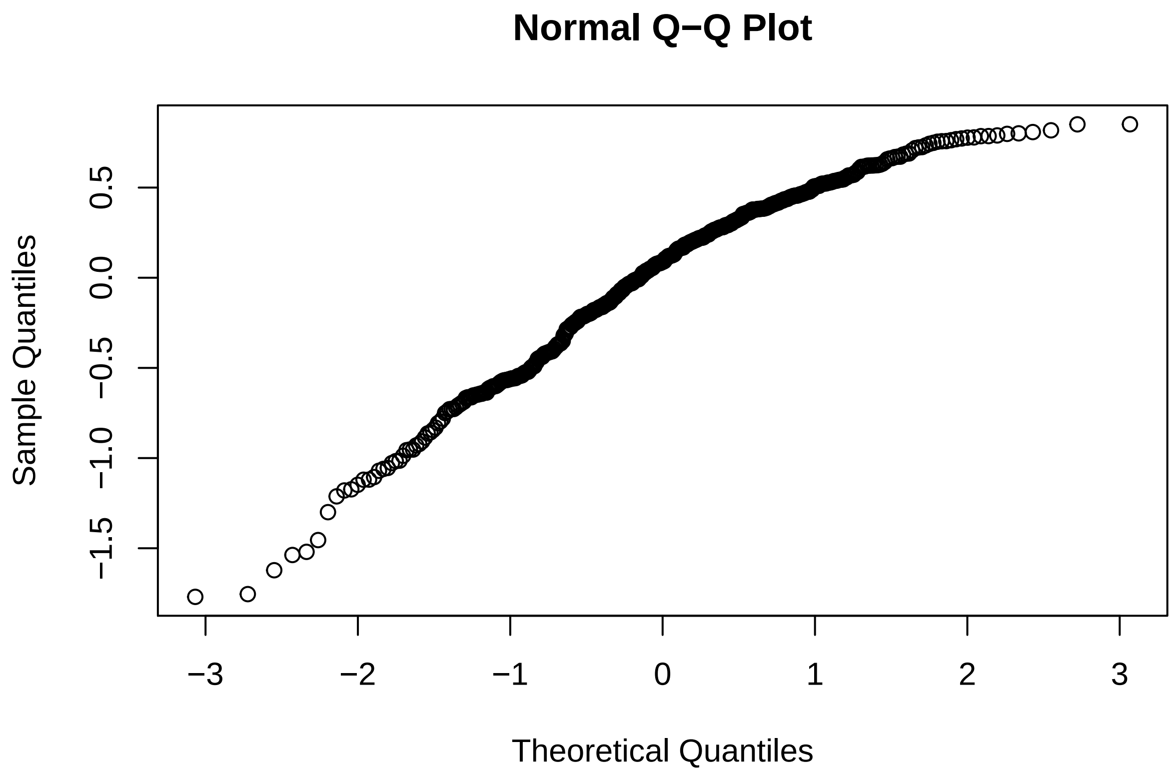
Answer: Based on final model we can write below equation for linear model Avg Rating = 3.869 + (0.173 * Gender Male) + (0.0050 * cls_perc_eval) + (0.431 * cls_creditsone credit) - (0.238 * pic_colorcolor)

16. Verify that the conditions for this model are reasonable using diagnostic plots.

```r
m_bktest <- lm(score ~ gender + cls_perc_eval + cls_credits +  pic_color, data = evals)
plot(m_bktest$residuals)
```




```r
qqnorm(m_bktest$residuals)
```



Answer : From the residual plot we can dee that the conditions for Linear model like residuals being homoscadastic and nearly normally distributed are not truely satisfied here

17. The original paper describes how these data were gathered by taking a sample of professors from the University of Texas at Austin and including all courses that they have taught. Considering that each row represents a course, could this new information have an impact on any of the conditions of linear regression?

Answer : Since the sample represents only one location, it is same as adding a constant value to the model input. This will not be very useful information for the model to predict the avg teachers score and will not make any difference

18. Based on your final model, describe the characteristics of a professor and course at University of Texas at Austin that would be associated with a high evaluation score.

Answer : Based on the final model we can see that variables which are statistically important are gender, class percentage filling evaluation, class credit and color picture. Based on these parameters we can say the a male prefessor teaching for a larger class size with one credit course (lab, PE, etc.) and having black and white picture on the profile would be associated with high avg score

19. Would you be comfortable generalizing your conclusions to apply to professors generally (at any university)? Why or why not?

Based on this study, I wouldn't be comfortable to ganearalize these results to any university. Sample just represents very specific university (University of Texas in Austin). Sample is not randomized and doesn't represent overall population of the students and teachers from nationwide universities. This can be very biased sample and could reveal a trend of a very specific university targetted in this observational study. However to generalize the results we

need to make our sampling strategy more comprehensive and unbiased. Observation collected should be randomized and should be true representation of the overall population for us to generalize the results. From the model output we can see that the conditions for linear model are not fully satisfied in the result. We don't have a very strong conclusion whether identified important variables are truely doing good job in predicting the outcome variable. That is also one of the factor contributing towards not generalizing the result of this observational study