Multiple linear regression

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.
`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.

variable	description
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.
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

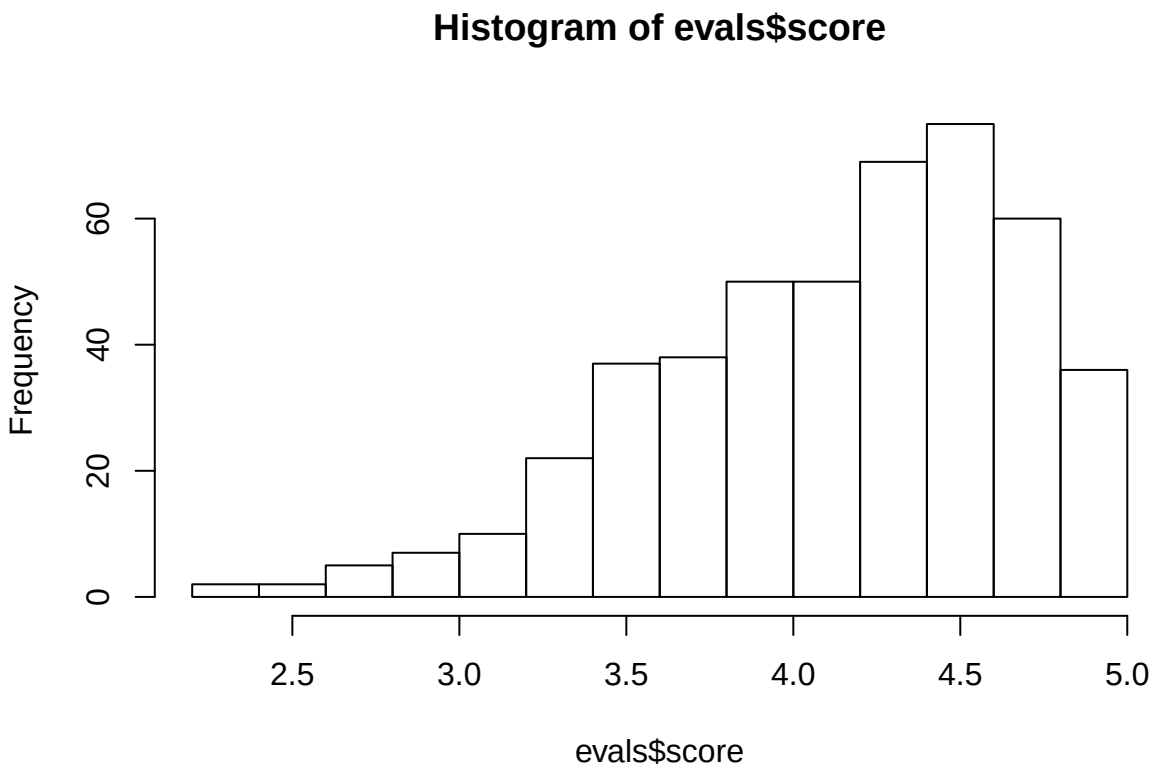
1. Is this an observational study or an experiment? The original research question posed in the paper is whether beauty leads directly to the differences in course evaluations. Given the study design, is it possible to answer this question as it is phrased? If not, rephrase the question.

This is an obsevational study and cannot be generalized to the population but can show associations from this sample set. Thus, it is not a question that can be answered directly from this study alone. A rephrasing might be, Is there an association between evaluation and preceived beauty in some university students?

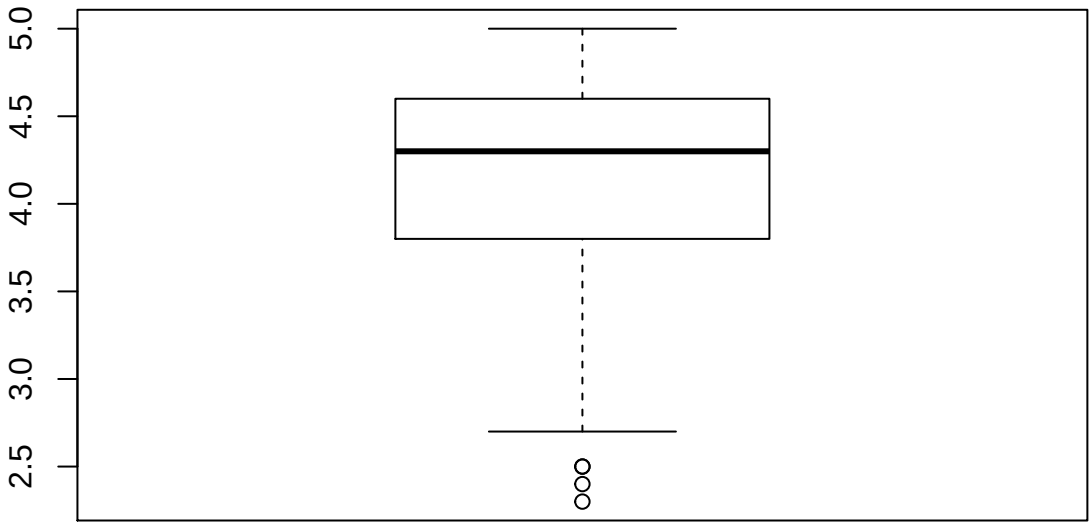
2. Describe the distribution of `score`. Is the distribution skewed? What does that tell you about how students rate courses? Is this what you expected to see? Why, or why not?

This definitely left-skewed data as the bulk of the scores are centered around 4.25. I would expect this as the de facto grading system in all school in the west 3-5 rather than 1-5. You really have to dislike a teacher to given them a 1.

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

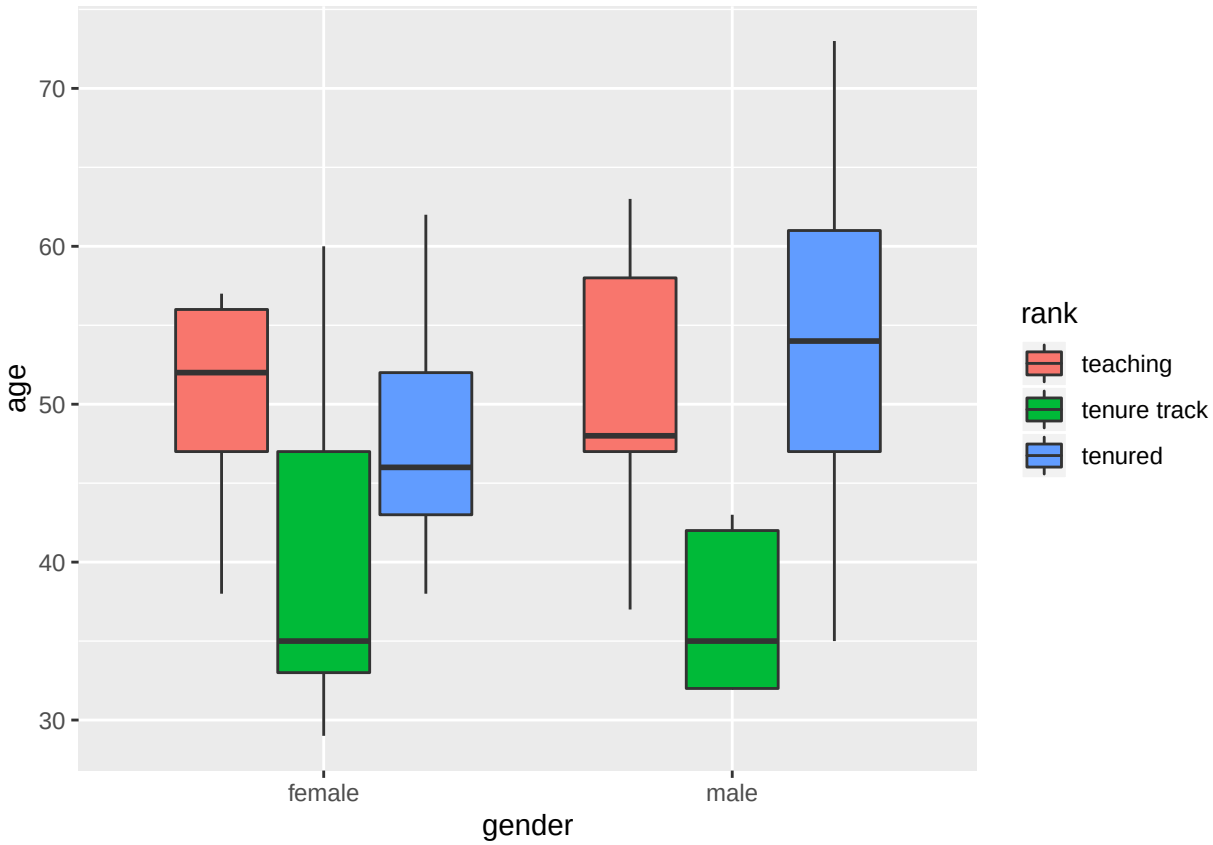



```
boxplot(evals$score)
```



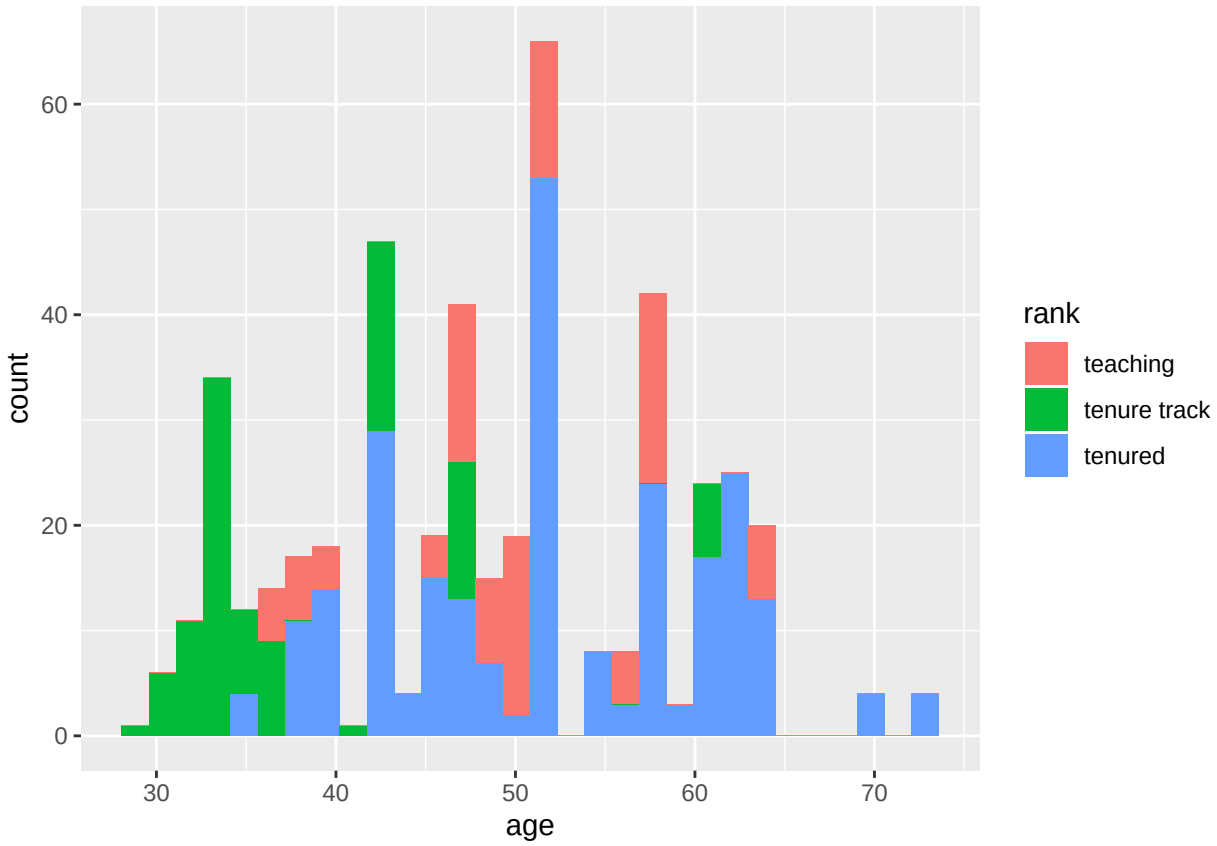
3. Excluding `score`, select two other variables and describe their relationship using an appropriate visualization (scatterplot, side-by-side boxplots, or mosaic plot).

```
library(ggplot2)
ggplot(data = evals) +
  geom_boxplot(aes(x=gender,y=age, fill=rank))
```




```
ggplot(data = evals) +
  geom_histogram(aes(x=age, fill=rank))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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)
```

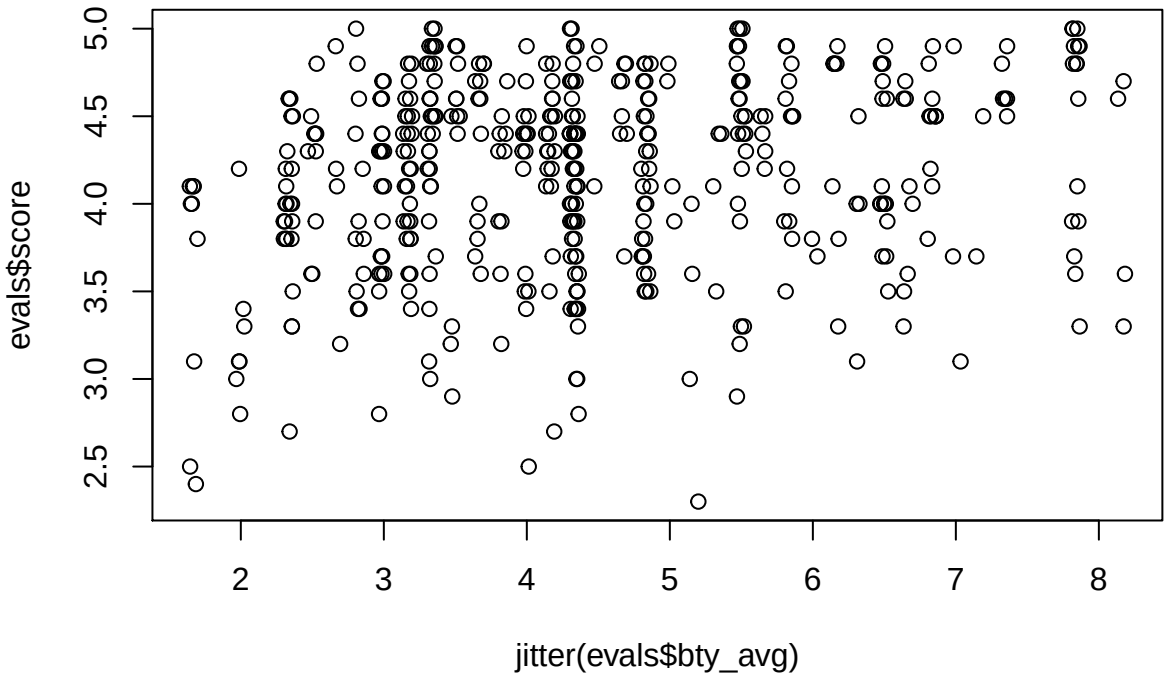
Before we draw conclusions about the trend, compare the number of observations in the data frame with the approximate number of points on the scatterplot. Is anything awry?

There are many overlapping points.

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?

All the overlapping points were shown as one singular point.

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



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(data = evals, score ~ bty_avg)
summary(m_bty)
```

```
##
## Call:
## lm(formula = score ~ bty_avg, data = 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)  3.88034    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: 5.083e-05
```

```
paste("equation of the line: score =",m_bty$coefficients[2],"*(average beauty) +",m_bty$coefficients[1])
```

```
## [1] "equation of the line: score = 0.0666370370198143 *(average beauty) + 3.88033795460773"
```

The p-values for the slope being different from 0 are low but the R^2 values are terrible. Doesn't look linear to me

```
library(tidyverse)
```

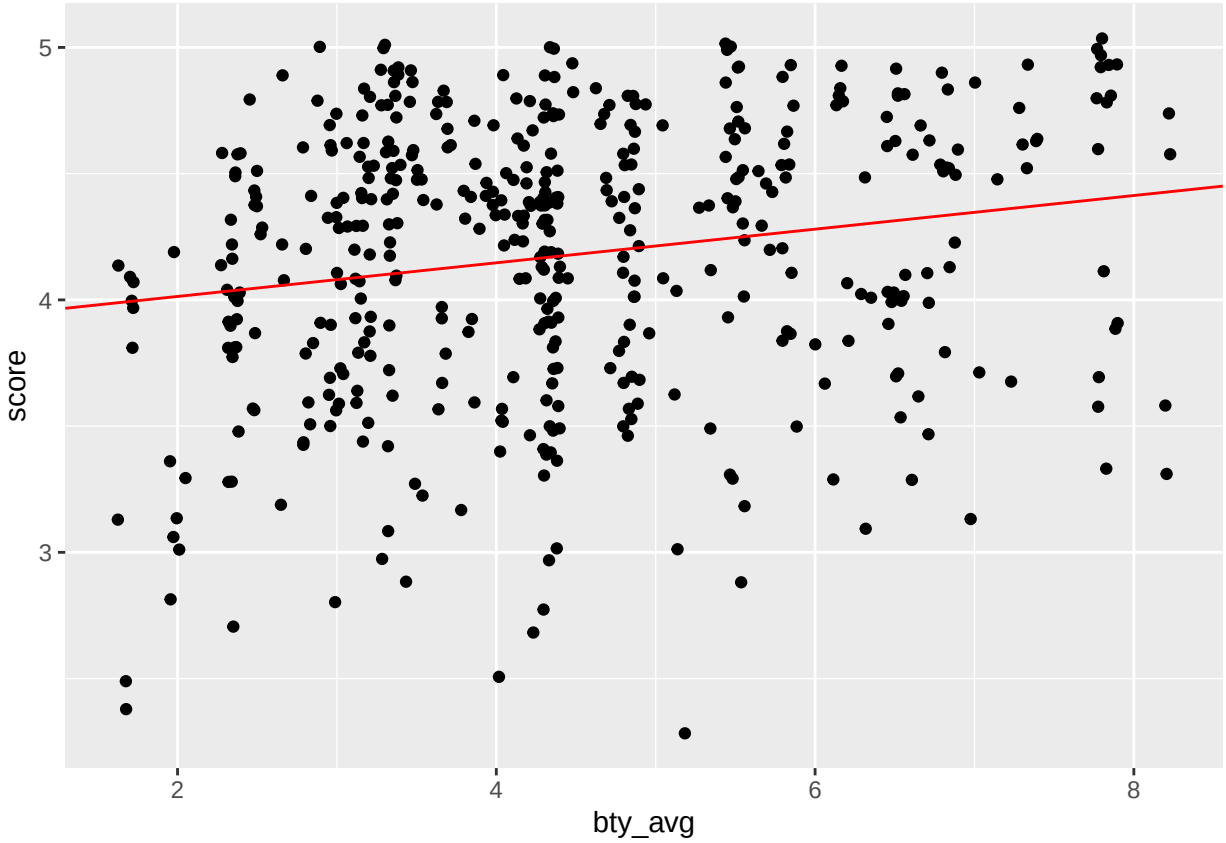
```
## Registered S3 method overwritten by 'rvest':
##   method            from
##   read_xml.response xml2
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v tibble  2.1.3     v purrr   0.3.3
## v tidyr   1.0.2     v dplyr   0.8.3
## v readr   1.3.1     v stringr 1.4.0
## v tibble  2.1.3     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

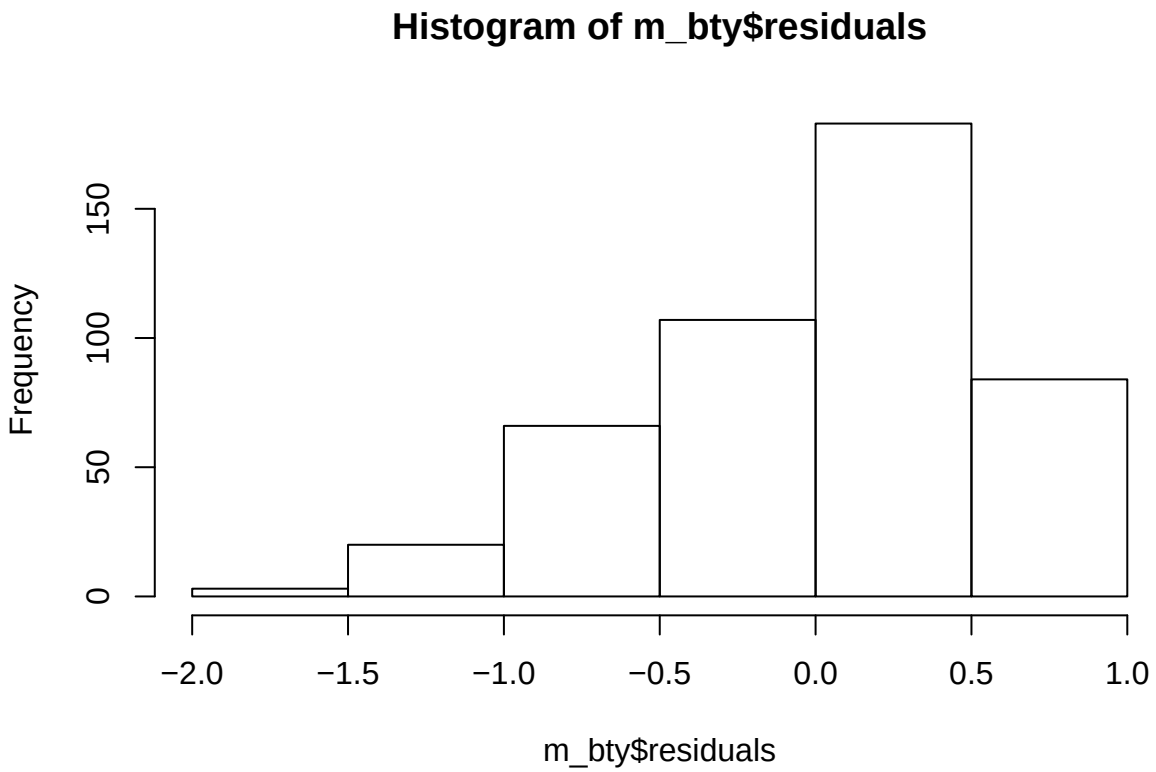
```
m_bty %>%
ggplot(aes(x=bty_avg, y=score)) +
  geom_jitter(color="black") +
  geom_abline(slope = m_bty$coefficients[2],
              intercept = m_bty$coefficients[1], color="red")
```



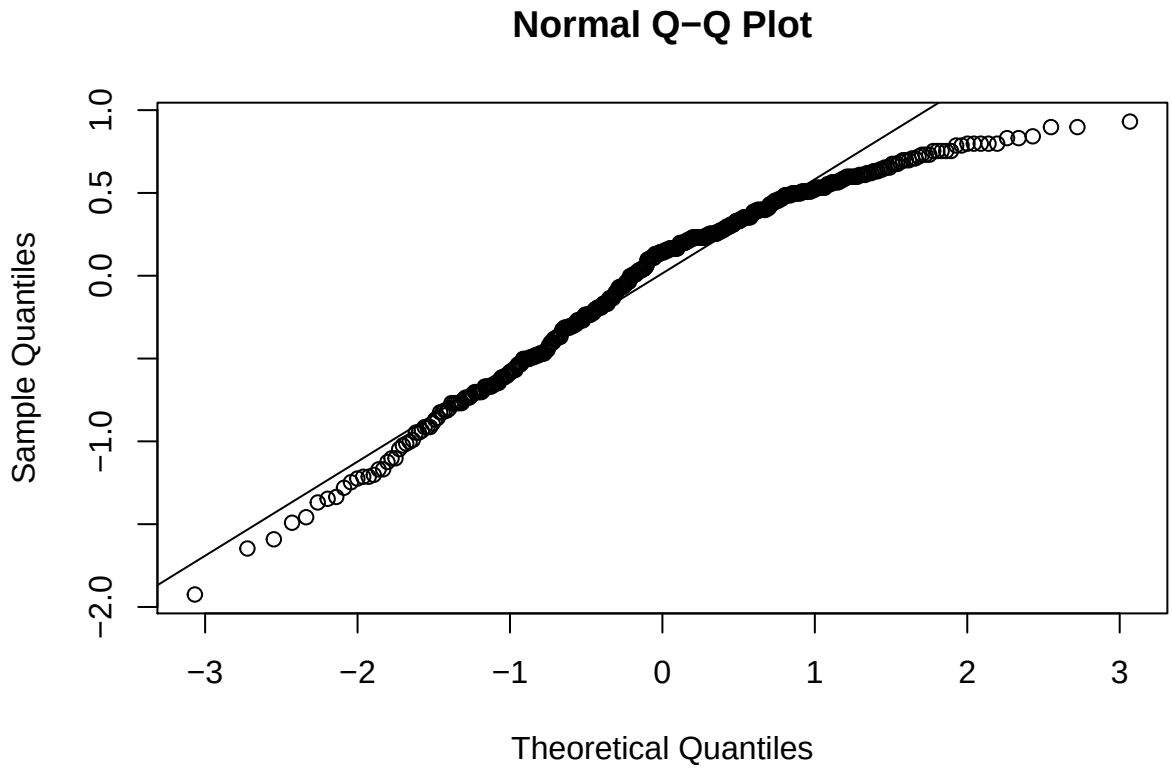
6. Use residual plots to evaluate whether the conditions of least squares regression are reasonable. Provide plots and comments for each one (see the Simple Regression Lab for a reminder of how to make these).

The residual plot is left-skewed and the qqplot has heavy tails; the upper tail especially so. The residual may be on the fence regarding nearly normal distribution.

```r
hist(m_bty$residuals)
```




```r
qqnorm(m_bty$residuals)
qqline(m_bty$residuals)
```



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)
```

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.

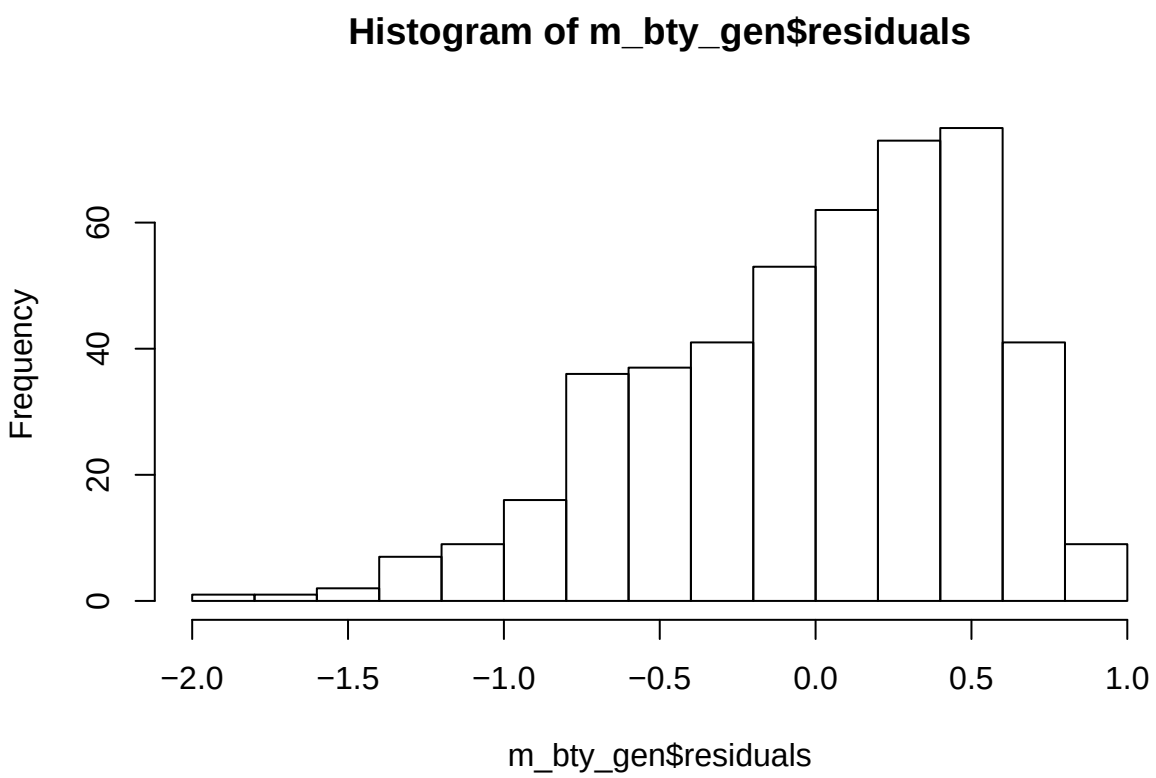
In order to see if beauty is still a significant predictor of professor score after we've accounted for the gender of the professor, we can add the gender term into the model.

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

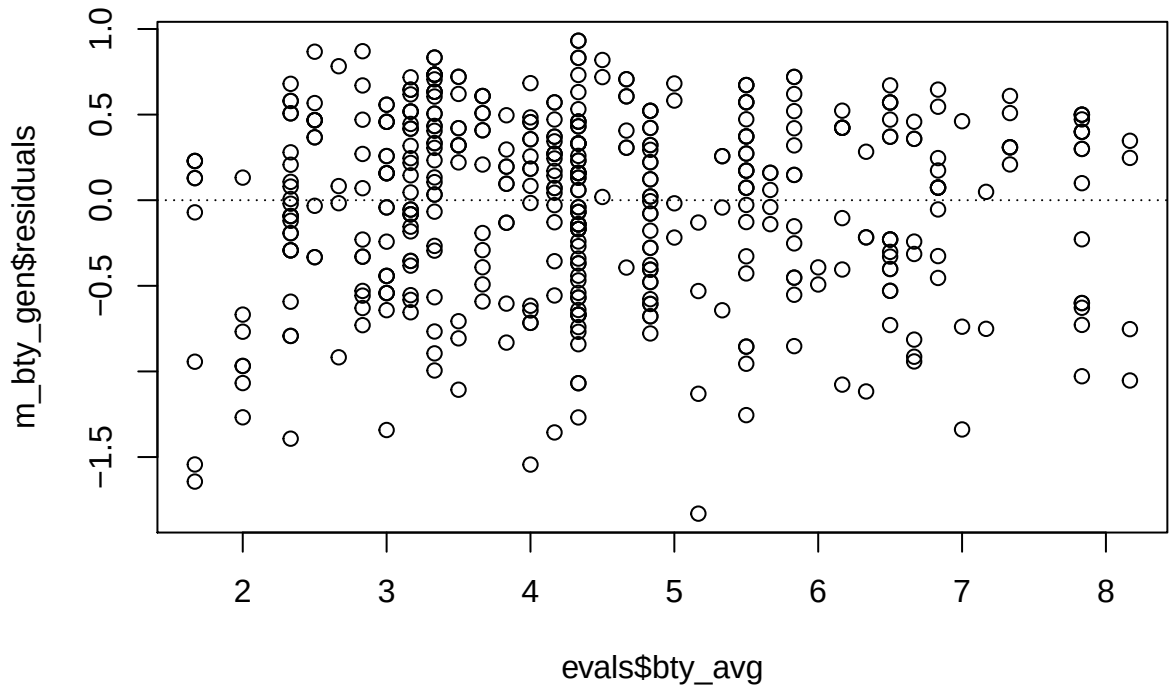
```
##
## Call:
## lm(formula = score ~ bty_avg + gender, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8305 -0.3625  0.1055  0.4213  0.9314
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.74734    0.08466  44.266  < 2e-16 ***
## bty_avg      0.07416    0.01625   4.563 6.48e-06 ***
## gendermale   0.17239    0.05022   3.433 0.000652 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5287 on 460 degrees of freedom
## Multiple R-squared:  0.05912,    Adjusted R-squared:  0.05503
## F-statistic: 14.45 on 2 and 460 DF,  p-value: 8.177e-07
```

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.

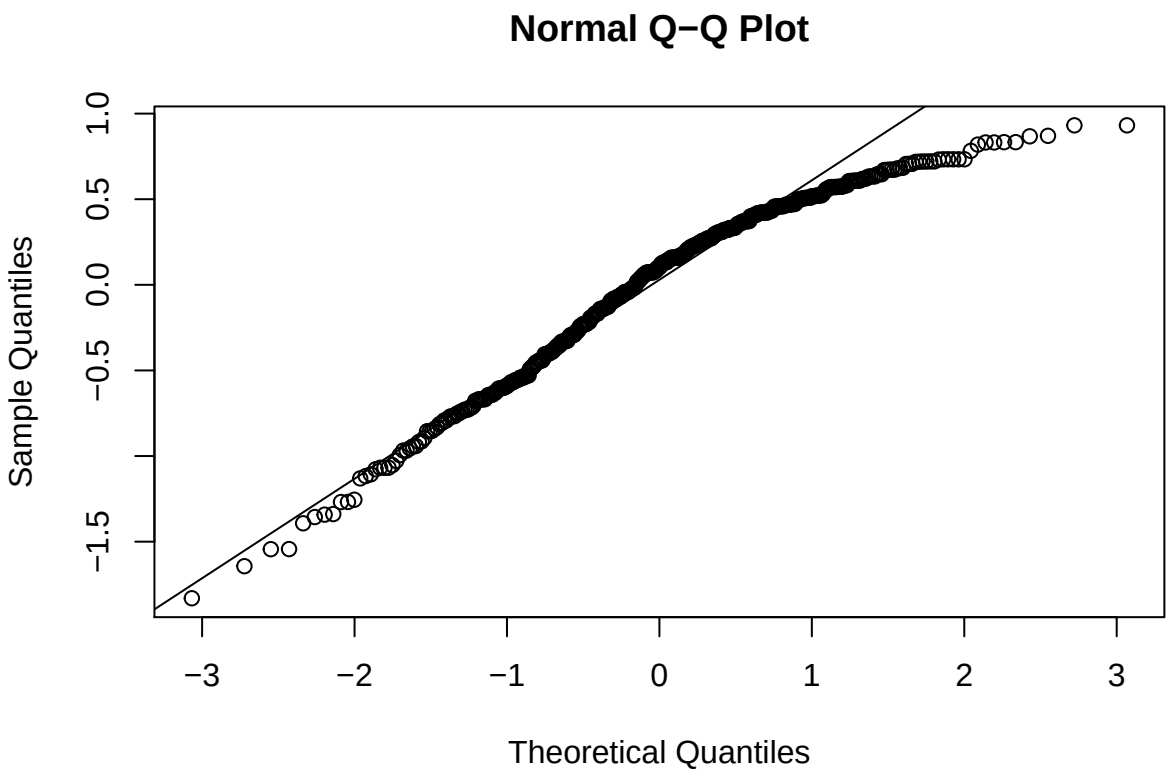
```
hist(m_bty_gen$residuals)
```




```
plot(m_bty_gen$residuals ~ evals$bty_avg)
abline(h = 0, lty = 3)
```




```
qqnorm(m_bty_gen$residuals)
qqline(m_bty_gen$residuals)
```



Seems a bit better and reasonably normal. The qqplot is a bit more smooth and the lower tail is better aligned.

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`?

The R square and adjusted R square are better than previous and the residual indicate an improvement in residual distribution.

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?

```
paste("equation score =",m_bty_gen$coefficients[2],"*(beauty score) +",m_bty_gen$coefficients[3]+m_bty_
```

```
## [1] "equation score = 0.0741553729841086 *(beauty score) + 3.91972778541286"
```

Men would have a higher tendency to have a higher score since they have a non-negative gender coefficient, whereas women have no coefficient for 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`.

R creates a dummy variable that encodes the tracks with n-1 variables since the last category will just be zero.

```
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
```

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.

If I had to guess, I would say something like course credit would the least strong predictor.

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)
```

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

Hey, I guessed right! Course credit had the highest p-value at 0.778.

13. Interpret the coefficient associated with the ethnicity variable.

There appears to be a positive association between score and not being a minority. The p-value puts it in the middle of predictor strength.

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?

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

```
##
## Call:
## lm(formula = score ~ rank + ethnicity + gender + language + age +
##     cls_perc_eval + cls_students + cls_level + cls_profs + bty_avg +
##     pic_outfit + pic_color, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7498 -0.3200  0.1056  0.3679  0.9200
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              4.3098194  0.2918733  14.766  < 2e-16 ***
## ranktenure track        -0.1957586  0.0829015  -2.361 0.018635 *
## ranktenured             -0.1809000  0.0647027  -2.796 0.005398 **
## ethnicitynot minority    0.0429967  0.0778938   0.552 0.581229
## gendermale               0.2366593  0.0524895   4.509 8.33e-06 ***
## languagenon-english     -0.2589399  0.1133484  -2.284 0.022810 *
## age                     -0.0090463  0.0031973  -2.829 0.004873 **
## cls_perc_eval            0.0059006  0.0015636   3.774 0.000182 ***
## cls_students             0.0002954  0.0003829   0.771 0.440863
## cls_levelupper          -0.0065495  0.0565243  -0.116 0.907807
## cls_profssingle         -0.0427280  0.0525927  -0.812 0.416974
## bty_avg                  0.0315543  0.0177371   1.779 0.075917 .
## pic_outfitnot formal    -0.1362125  0.0751223  -1.813 0.070467 .
## pic_colorcolor          -0.2091633  0.0728769  -2.870 0.004297 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5077 on 449 degrees of freedom
## Multiple R-squared:  0.1531, Adjusted R-squared:  0.1286
## F-statistic: 6.243 on 13 and 449 DF,  p-value: 7.671e-11
```

All predictors moved a little bit both in estimate and p-value, as expected.

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.

After taking out the obviously poor predictors one-by-one a choice had to be made regarding bty_avg as it's p-value was higher than the typical critical value of 0.05. I kept it due the fact that having bty_avg produced a better adjusted R squared score than not.

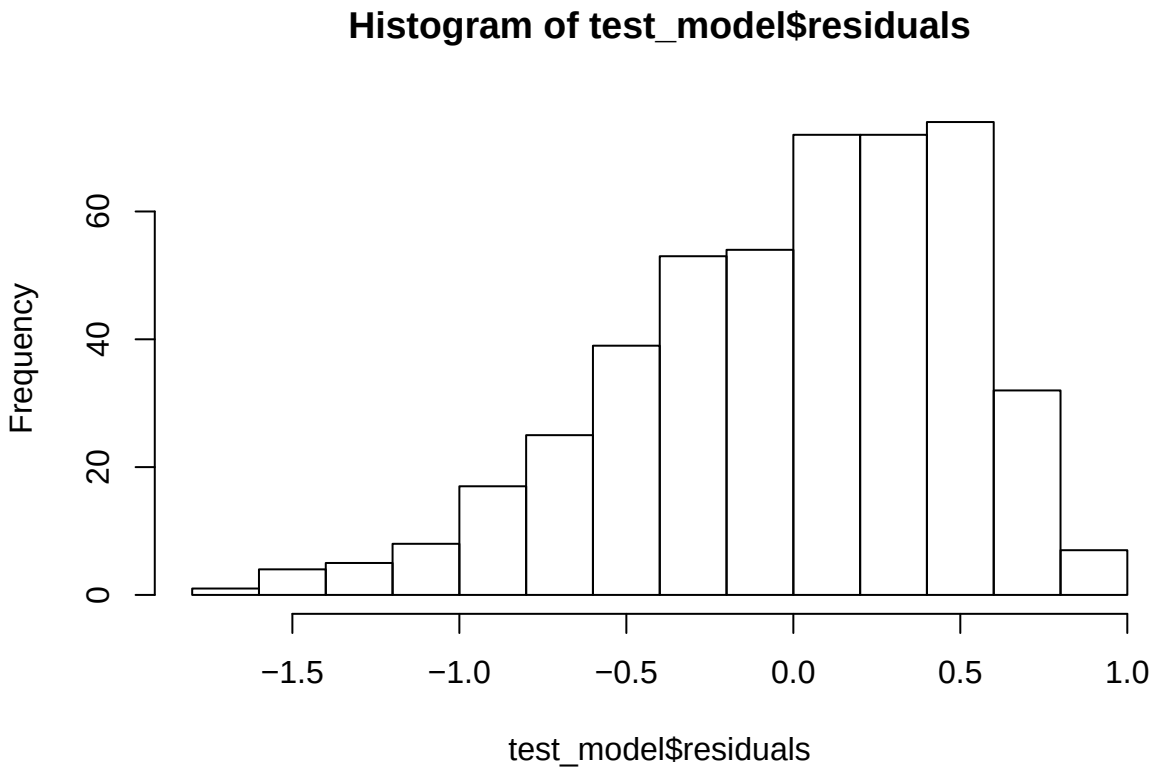
```
test_model <- lm(score ~ rank + gender + language + age + cls_perc_eval + bty_avg
             + pic_outfit + pic_color, data = evals)
summary(test_model)
```

```
##
## Call:
## lm(formula = score ~ rank + gender + language + age + cls_perc_eval +
##     bty_avg + pic_outfit + pic_color, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7725 -0.3435  0.1013  0.3869  0.9344
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.428380   0.263039  16.835  < 2e-16 ***
## ranktenure track     -0.203938   0.081849  -2.492 0.013072 *
## ranktenured          -0.180170   0.062692  -2.874 0.004245 **
## gendermale            0.243721   0.051264   4.754 2.68e-06 ***
## languagenon-english  -0.292573   0.105651  -2.769 0.005849 **
## age                  -0.009440   0.003158  -2.989 0.002952 **
## cls_perc_eval         0.005155   0.001449   3.557 0.000415 ***
## bty_avg               0.032528   0.017375   1.872 0.061838 .
## pic_outfitnot formal -0.144593   0.070119  -2.062 0.039767 *
## pic_colorcolor       -0.217967   0.067794  -3.215 0.001397 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5066 on 453 degrees of freedom
## Multiple R-squared:  0.1492, Adjusted R-squared:  0.1323
## F-statistic: 8.827 on 9 and 453 DF,  p-value: 2.78e-12
```

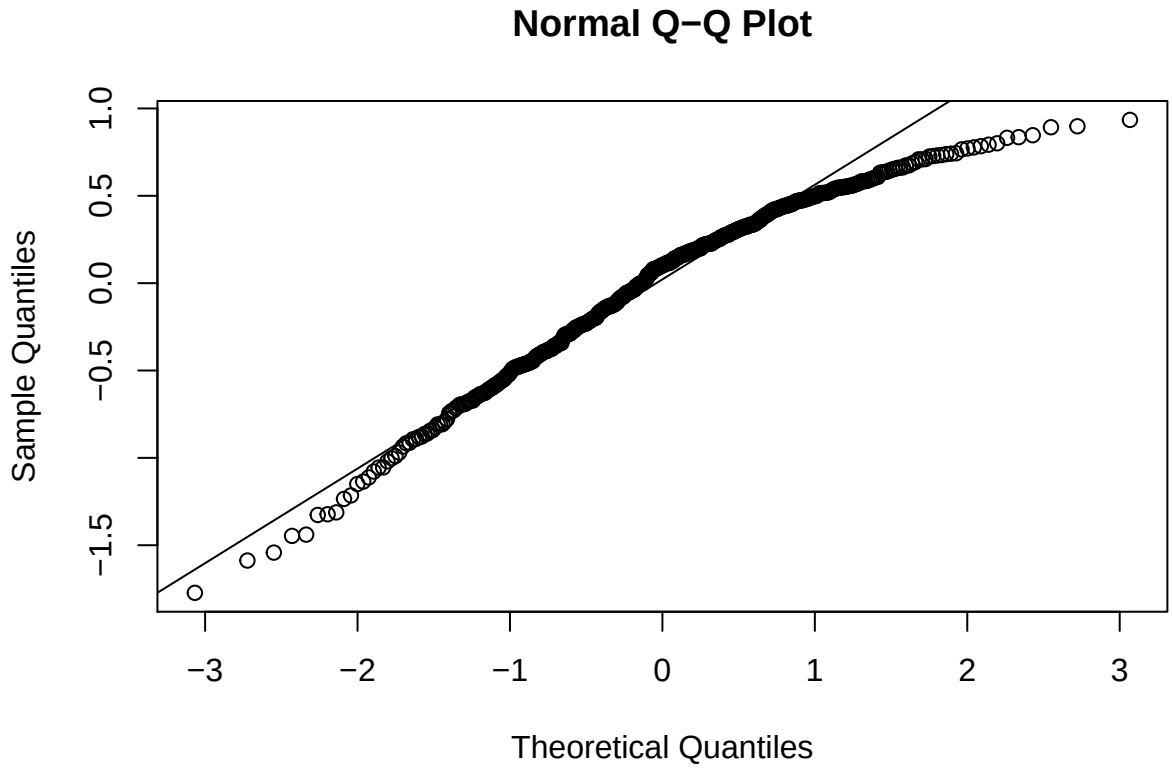
16. Verify that the conditions for this model are reasonable using diagnostic plots.

These seem reasonable considering the data. Still left-skewed, still a bit heavy tailed.

```
hist(test_model$residuals)
```




```
qqnorm(test_model$residuals)
qqline(test_model$residuals)
```



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?

There does not seem to be any predictors that are course specific other than the percent of students who completed the evaluation. As such, I doubt think this would impact the regression.

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.

```
test_model$coefficients
```

```
##          (Intercept)     ranktenure track          ranktenured
##          4.428380053         -0.203937662         -0.180169941
##           gendermale  languagenon-english                  age
##          0.243720547         -0.292572836         -0.009440274
##        cls_perc_eval              bty_avg pic_outfitnot formal
##          0.005154830          0.032527545         -0.144593342
##       pic_colorcolor
##         -0.217967046
```

The highest scoring professor would be male, young, native english speaker that is "teaching" professor with a high percent of completed evals. They would dress formally, and have a color picture somewhere.

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

No way. This doesn't meet the requirments for generalization and there could be hidden factors in how the evals are completed and how beauty scores are assigned. More data is needed from a variety of sources at a minimum first.