Hao-8

Bruce Hao

12/1/2016

```
library(IS606)
```

```
##
## Welcome to CUNY IS606 Statistics and Probability for Data Analytics
## This package is designed to support this course. The text book used
## is OpenIntro Statistics, 3rd Edition. You can read this by typing
## vignette('os3') or visit www.OpenIntro.org.
##
## The getLabs() function will return a list of the labs available.
##
## The demo(package='IS606') will list the demos that are available.
```

```
library(dplyr)
library(ggplot2)
setwd("~/Google Drive/CUNY/git/DATA606/Lab8")
load("more/evals.RData")
```

Exercise 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.

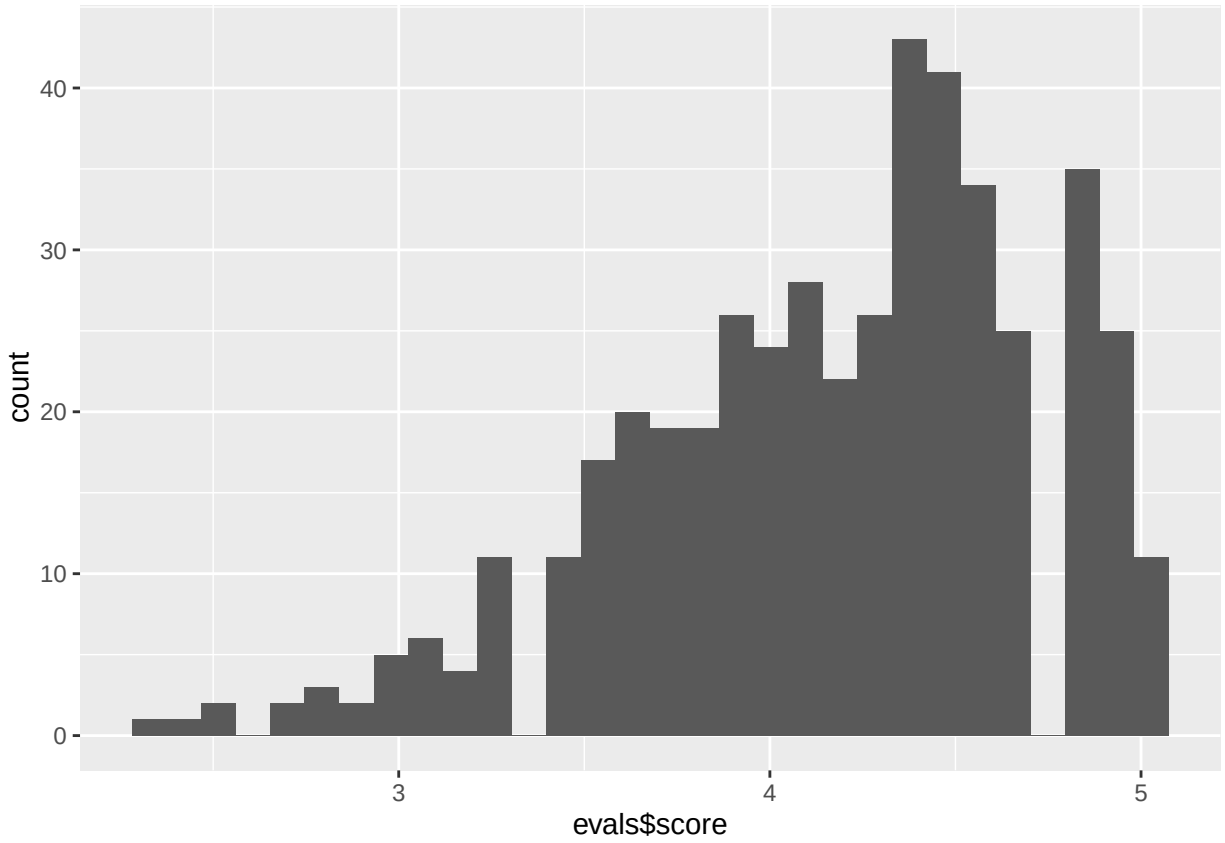
Observational study. Given that the study is an observational one, causal conclusions cannot be drawn. As such, the question might be rephrased as 'Whether beauty is related to differences in course evaluations'.

Exercise 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?

Based on the histogram below, the distribution of score is left skewed, which indicates that most scores are relatively high with a handful of outliers within the lower scores. I'm not sure what I expected to see.

```
qplot(evals$score, geom = 'histogram')
```

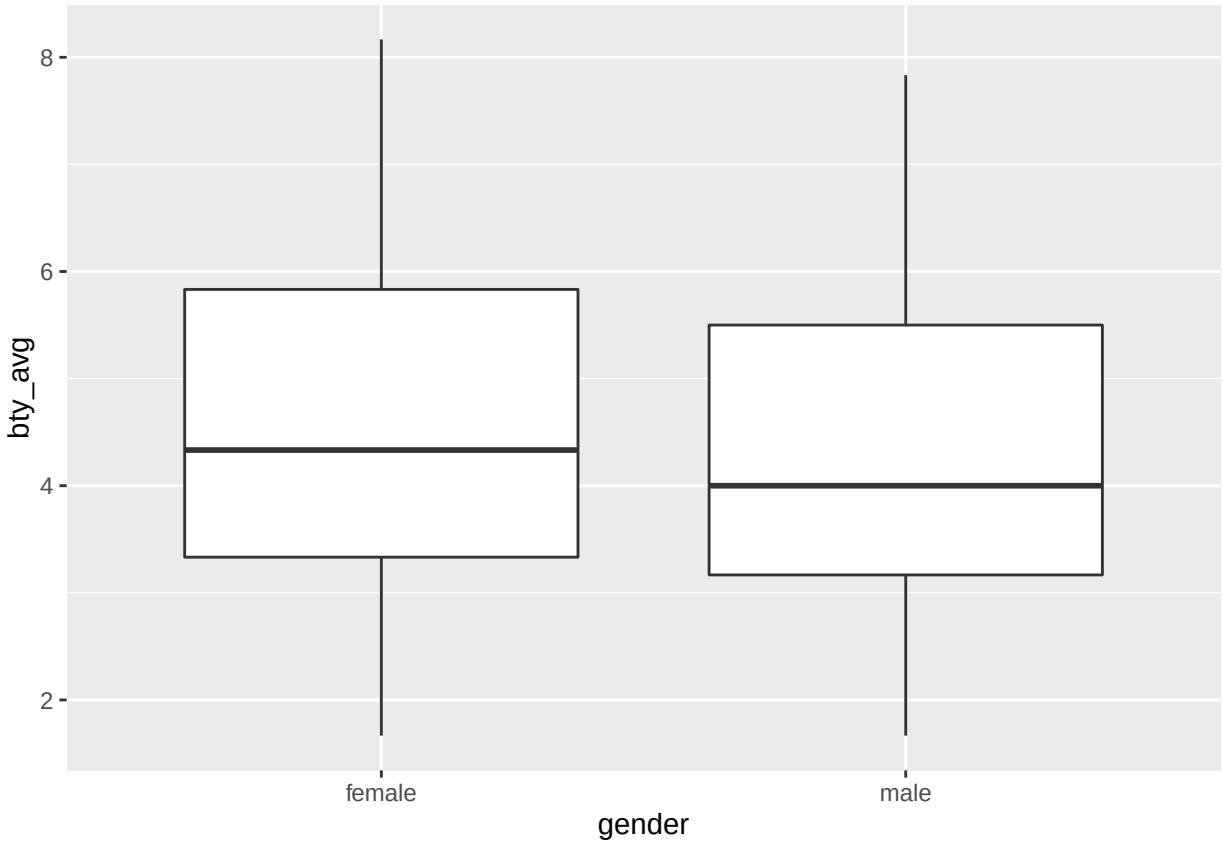
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

According to the box plot below, male professors were rated on average lower than their female counterparts. The distributions do not look materially different between genders.

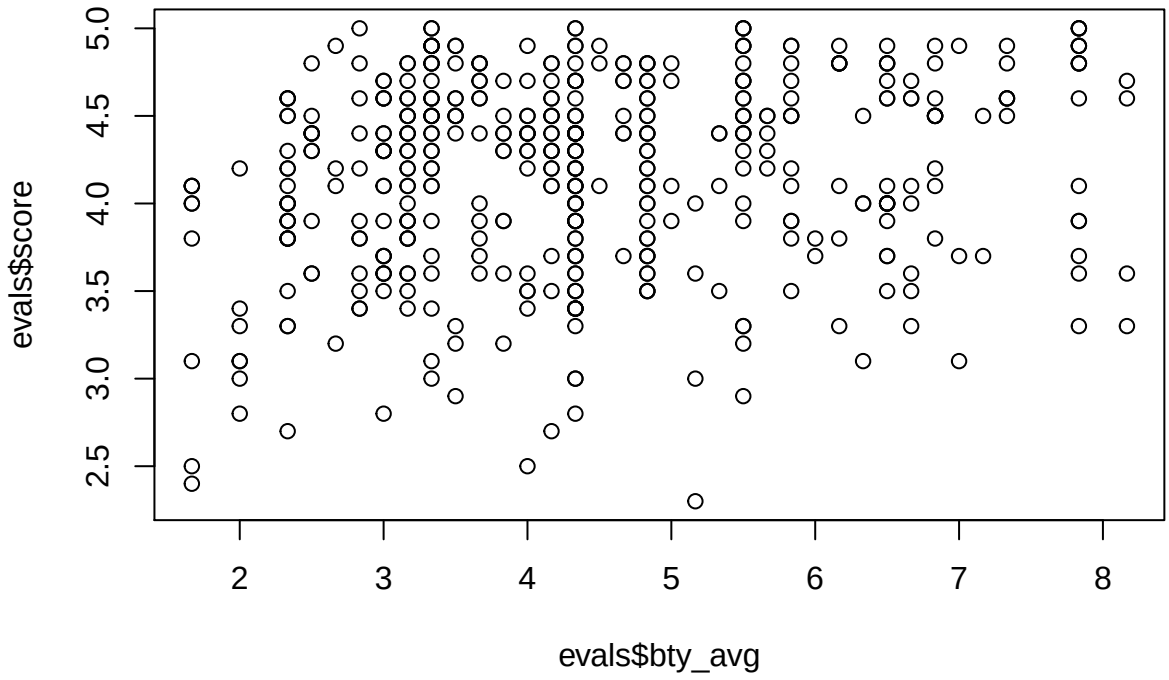
```
evals %>% ggplot(aes(y = bty_avg, x = gender)) + geom_boxplot()
```



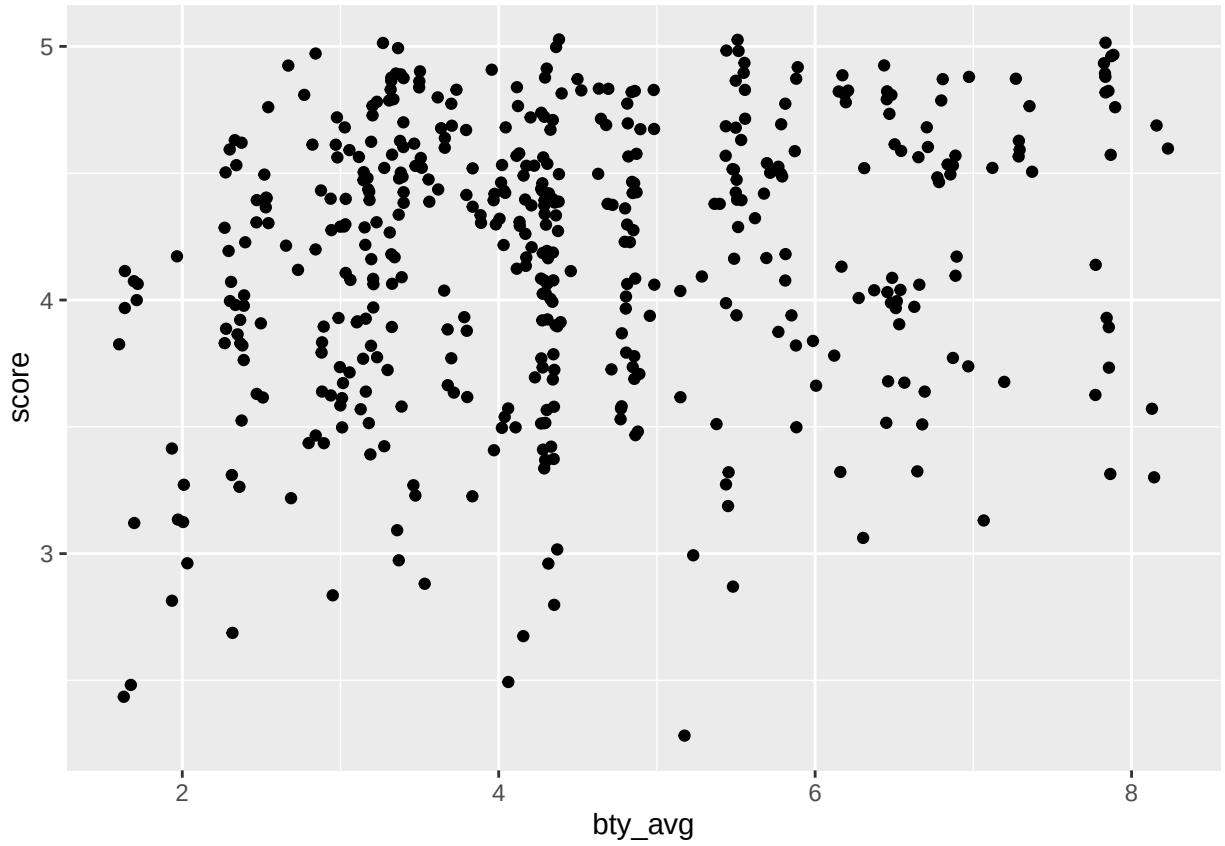
Exercise 4: Replot the scatterplot, but this time use the function jitter() on the yy- or the xx-coordinate. (Use ?jitter to learn more.) What was misleading about the initial scatterplot?

Combinations of beauty and score are repeated, so the points overlap. Using jitter, we can see more of the points.

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




```
evals %>% ggplot(aes(x = bty_avg, y = score)) + geom_point(position = position_jitter())
```



Exercise 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?

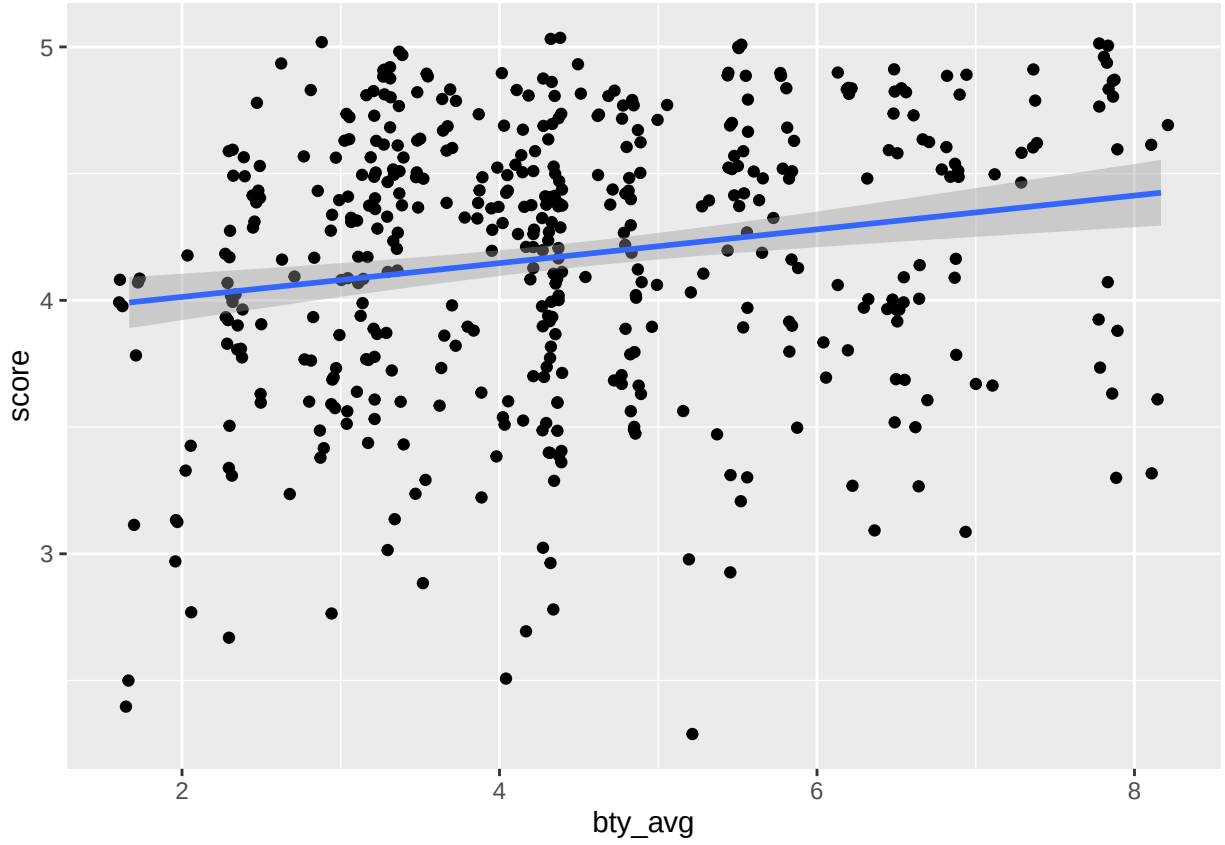
Based on the summary of the linear model, the equation is yˆ = 3.88034 + 0.0664 x bty_avg. The positive slope indicates that for each additional unit in average beauty score, performance score is 0.0664 units greater. Yes, average beauty score is a statistically significant predictor with a t-value > 4. Practically, a maximum beauty score of 10 would only increase performance score by 0.664 points, which is somewhat meaningful on a 1 to 5 scale. The practical significance is arguable.

```
m_bty = lm(evals$score ~ evals$bty_avg)
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
```

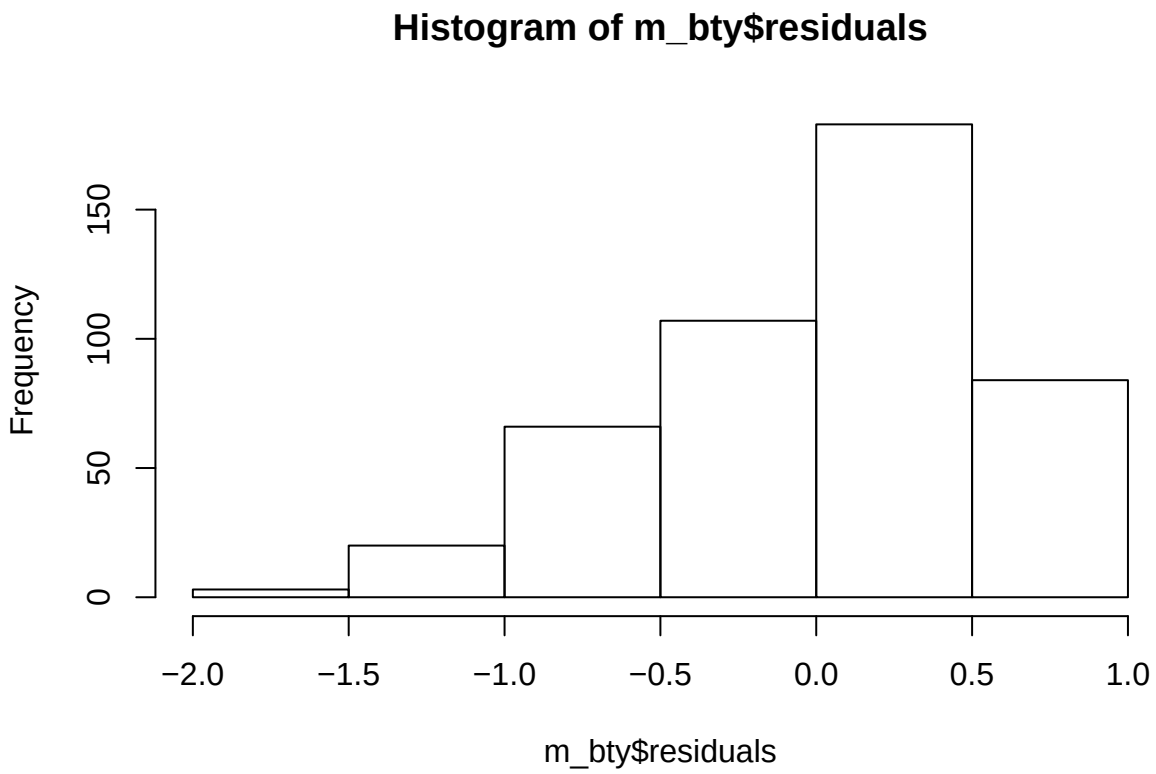
```
evals %>% ggplot(aes(x = bty_avg, y = score)) +
  geom_point(position = position_jitter()) +
  geom_smooth(method = 'lm')
```



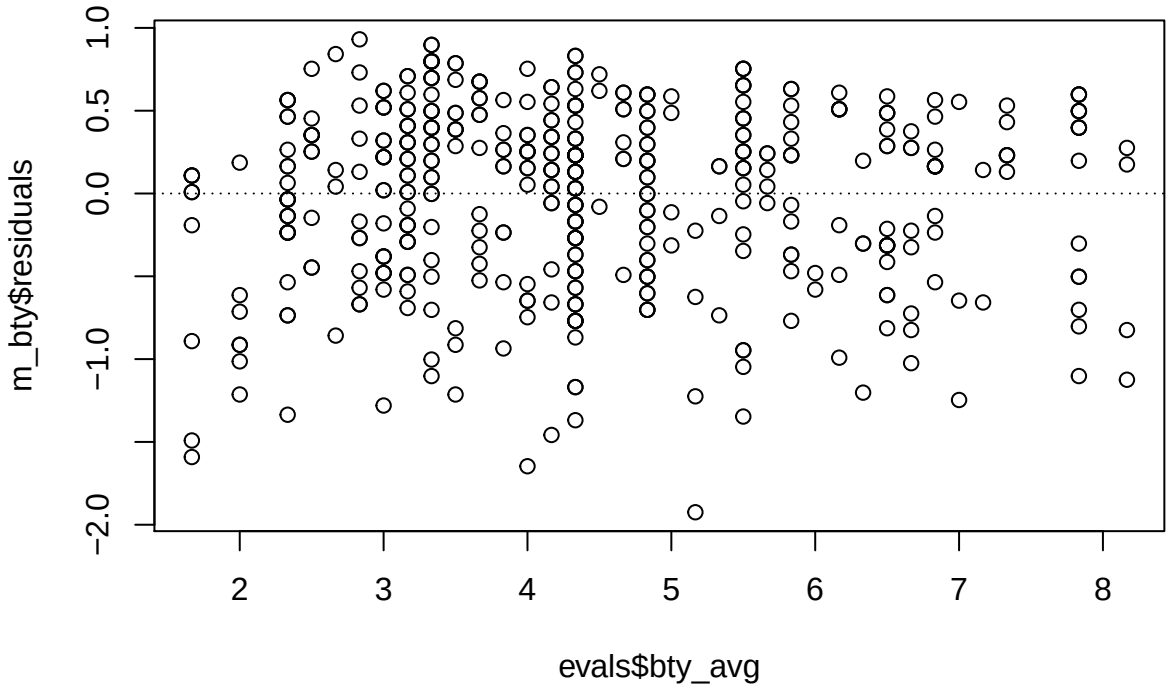
Exercise 6: Use residual plots to evaluate whether the conditions of least squares regression are reasonable. Provide plots and comments for each one (see the Simple Regression Lab for a reminder of how to make these).

The residuals are somewhat left skewed as illustrated by the histogram and qq plots below. It's not clear to me if this is significant enough to negate the use of this linear model.

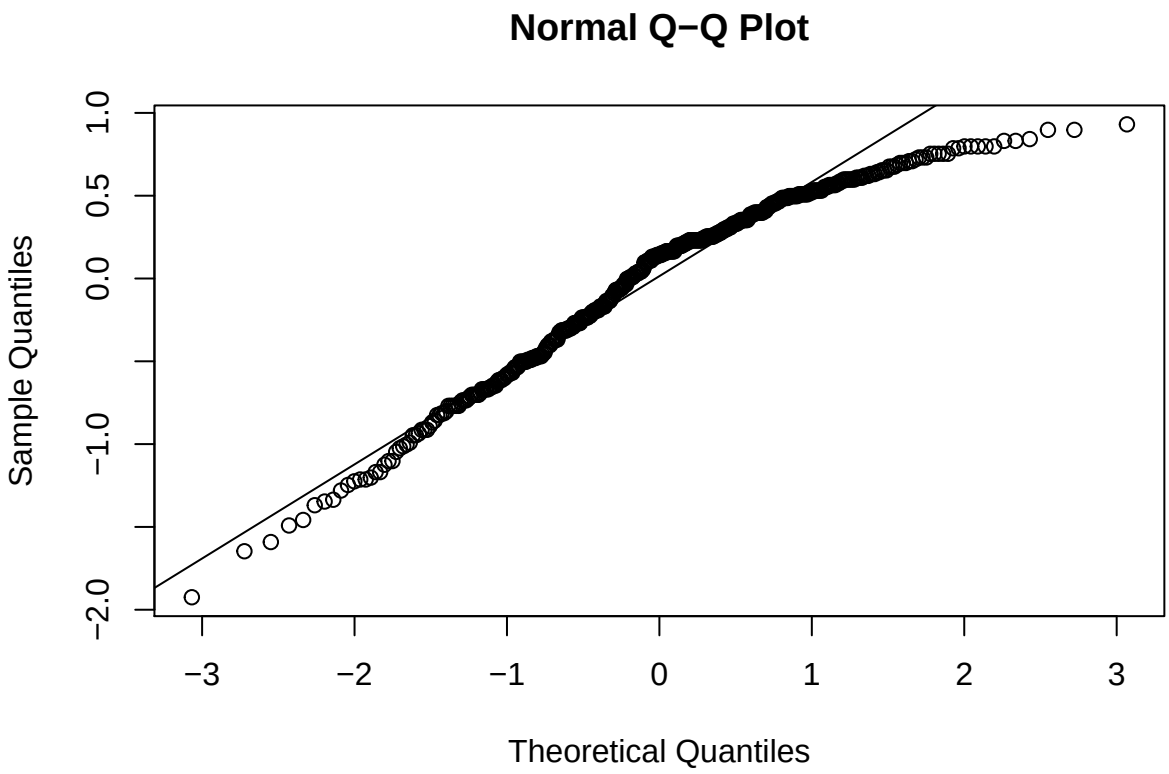
```
hist(m_bty$residuals)
```



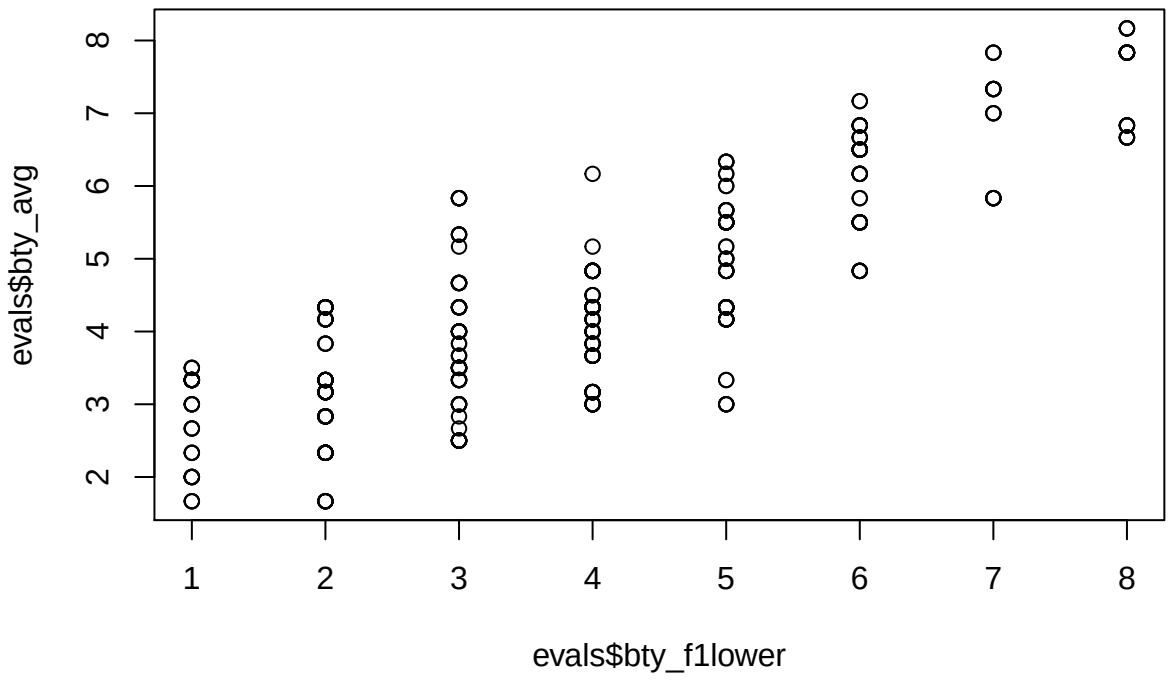

```
plot(m_bty$residuals ~ evals$bty_avg)
abline(h = 0, lty = 3)
```




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



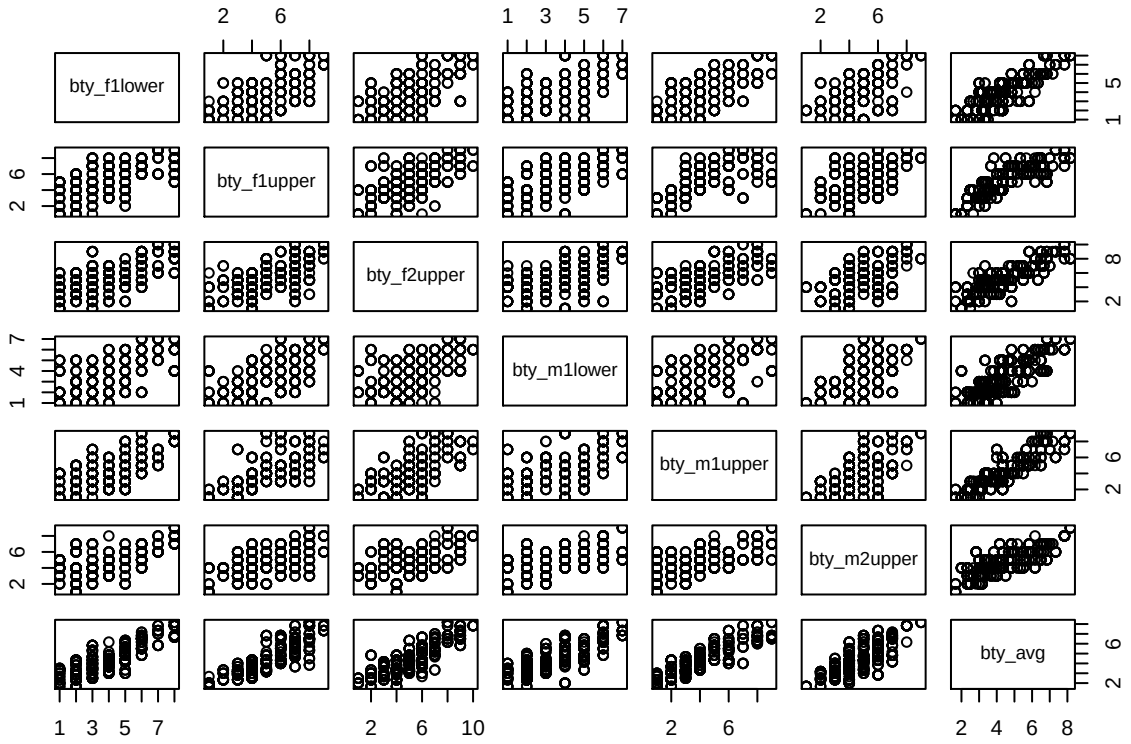

```
plot(evals$bty_avg ~ evals$bty_f1lower)
```




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

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

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



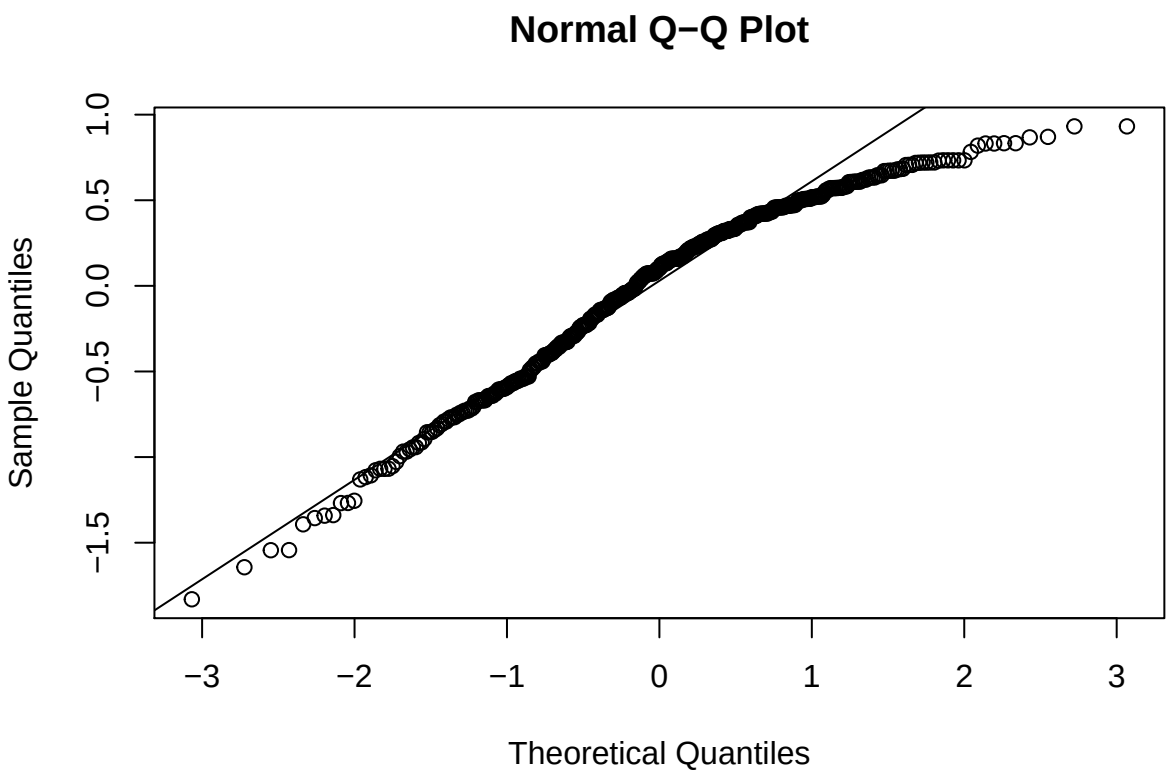

```
m_bty_gen <- lm(score ~ bty_avg + gender, data = evals)
summary(m_bty_gen)
```

```
##
## 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
```

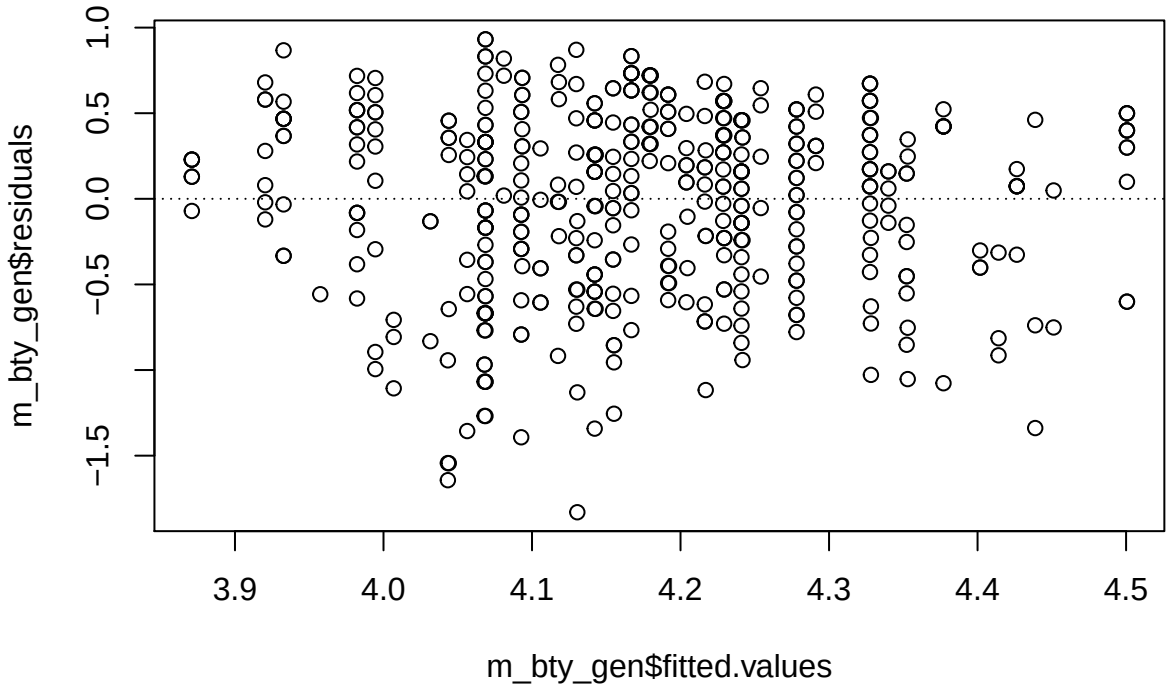
Exercise 7: P-values and parameter estimates should only be trusted if the conditions for the regression are reasonable. Verify that the conditions for this model are reasonable using diagnostic plots.

Based on the diagnostic charts below, while there may be some skew in the residuals, the residuals do not appear to exhibit any identifiable patterns. Based on this, the conditions for regression appear to be reasonably met.

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




```
plot(m_bty_gen$residuals ~ m_bty_gen$fitted.values)
abline(h = 0, lty = 3)
```



Exercise 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?

Yes, bty_avg remains a significant predictor. The slope for the bty_avg predictor actually increased slightly from 0.0664 to 0.07416.

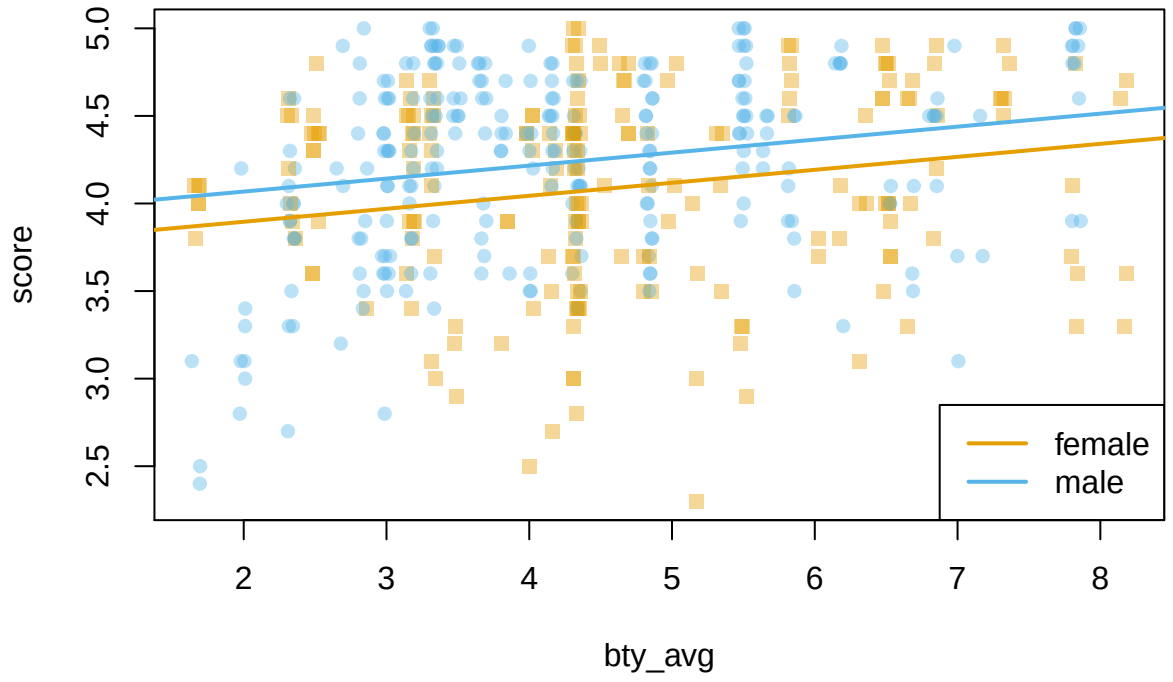
Exercise 9: What is the equation of the line corresponding to males? (Hint: For males, the parameter estimate is multiplied by 1.) For two professors who received the same beauty rating, which gender tends to have the higher course evaluation score?

(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 ***

Males: $\hat{y} = (3.74734 + 0.17239) + 0.07416$ x bty_avg
Females: $\hat{y} = (3.74734) + 0.07416$ x bty_avg

The effect of the dummy variable effectively increases the intercept for male professors. As such, males generally have higher course evaluation scores.

```
multiLines(m_bty_gen)
```



Exercise 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.

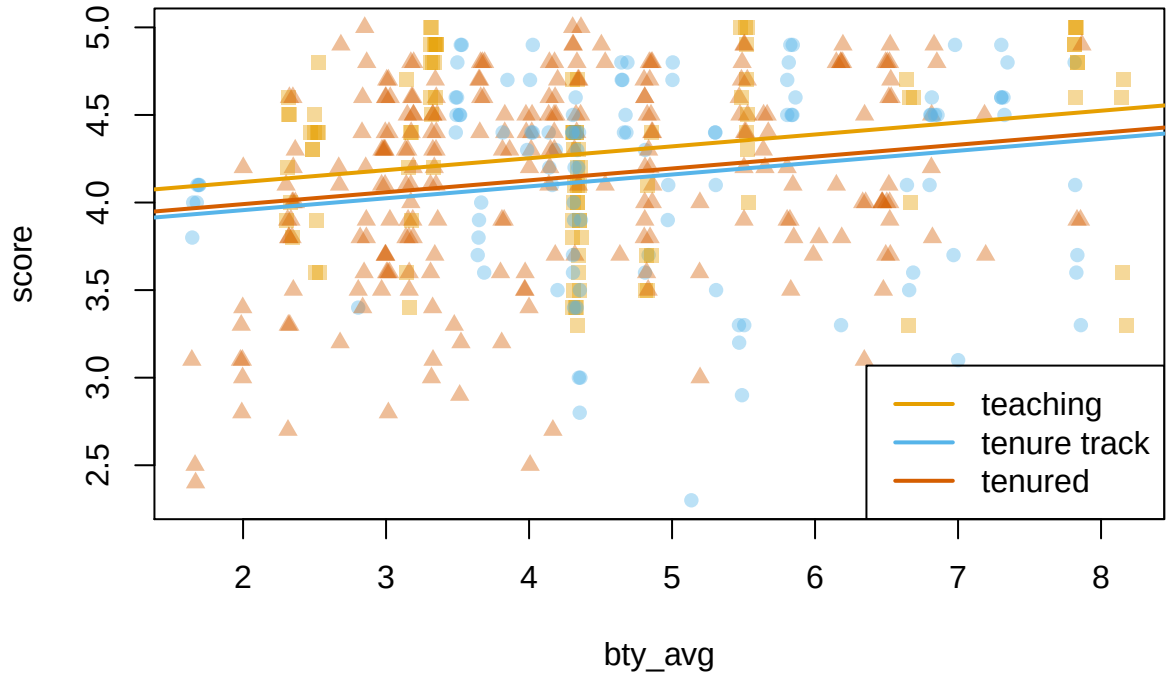
R handles categorical variables with multiple levels by creating dummy variables for all but one of the levels. In this case, the intercept for 'teaching' is simply the intercept parameter 3.98155; the intercept for tenure tracked and tenured are (3.98155 - 0.16070) and (3.98155 - 0.12623), respectively.

```
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
```

```
multiLines(m_bty_rank)
```



Exercise 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.

Variables related to the class, e.g. cls_level, cls_profs, etc., should have high p-values as they describe the class more than the professor.

```
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
```

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

Yes, the p-values for cls_level and cls_profs are among the highest of the predictors.

Exercise 13: Interpret the coefficient associated with the ethnicity variable.

The positive coefficient indicates that professors that are not minorities are rated higher all else being equal.

Exercise 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?

Yes, some of the coefficients changed slightly. But since they didn't change very much, the eliminated predictor (cls_profs) was not highly collinear with the remaining explanatory variables.

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

```
##
## Call:
## lm(formula = score ~ rank + ethnicity + gender + language + age +
##     cls_perc_eval + cls_students + cls_level + cls_credits +
##     bty_avg + pic_outfit + pic_color, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7836 -0.3257  0.0859  0.3513  0.9551
##
## Coefficients:
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.0872523  0.2888562  14.150  < 2e-16 ***
## ranktenure track       -0.1476746  0.0819824  -1.801 0.072327 .
## ranktenured            -0.0973829  0.0662614  -1.470 0.142349
## ethnicitynot minority   0.1274458  0.0772887   1.649 0.099856 .
## gendermale              0.2101231  0.0516873   4.065 5.66e-05 ***
## languagenon-english    -0.2282894  0.1111305  -2.054 0.040530 *
## age                    -0.0089992  0.0031326  -2.873 0.004262 **
## cls_perc_eval           0.0052888  0.0015317   3.453 0.000607 ***
## cls_students            0.0004687  0.0003737   1.254 0.210384
## cls_levelupper          0.0606374  0.0575010   1.055 0.292200
## cls_creditsone credit   0.5061196  0.1149163   4.404 1.33e-05 ***
## bty_avg                 0.0398629  0.0174780   2.281 0.023032 *
## pic_outfitnot formal   -0.1083227  0.0721711  -1.501 0.134080
## pic_colorcolor         -0.2190527  0.0711469  -3.079 0.002205 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4974 on 449 degrees of freedom
## Multiple R-squared:  0.187,  Adjusted R-squared:  0.1634
## F-statistic: 7.943 on 13 and 449 DF,  p-value: 2.336e-14
```

Exercise 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.

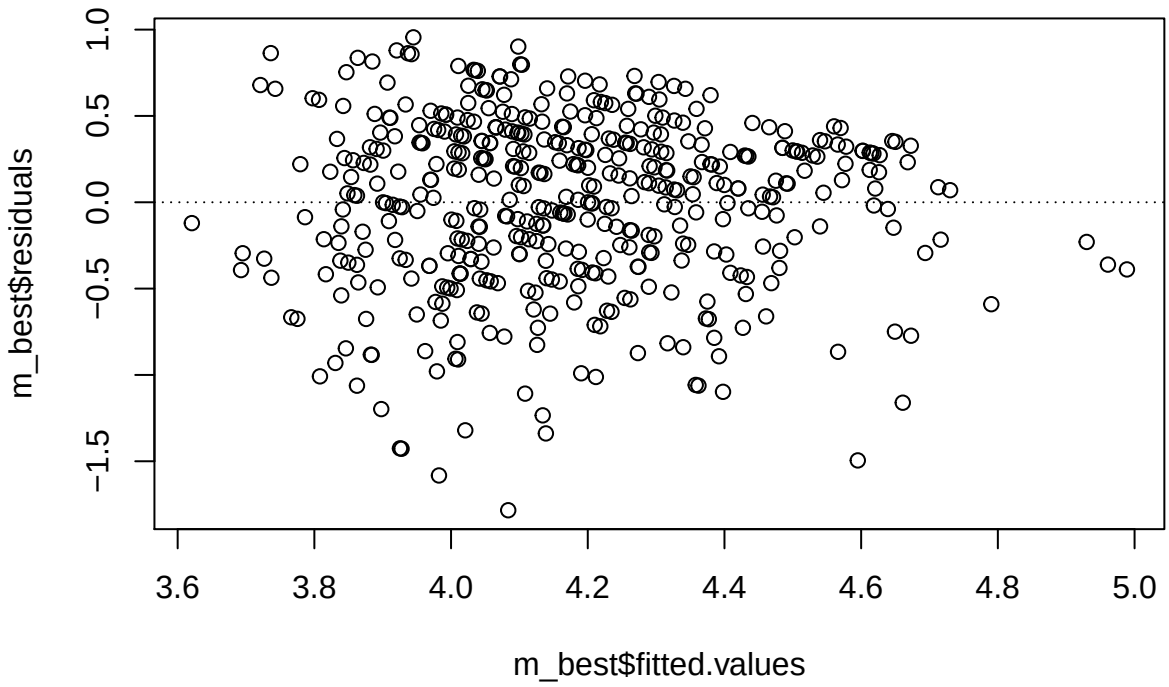
The last model was the best model in terms of adjusted r-squared. The linear model is simply the estimate for each parameter times the respective variable value plus the intercept.

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

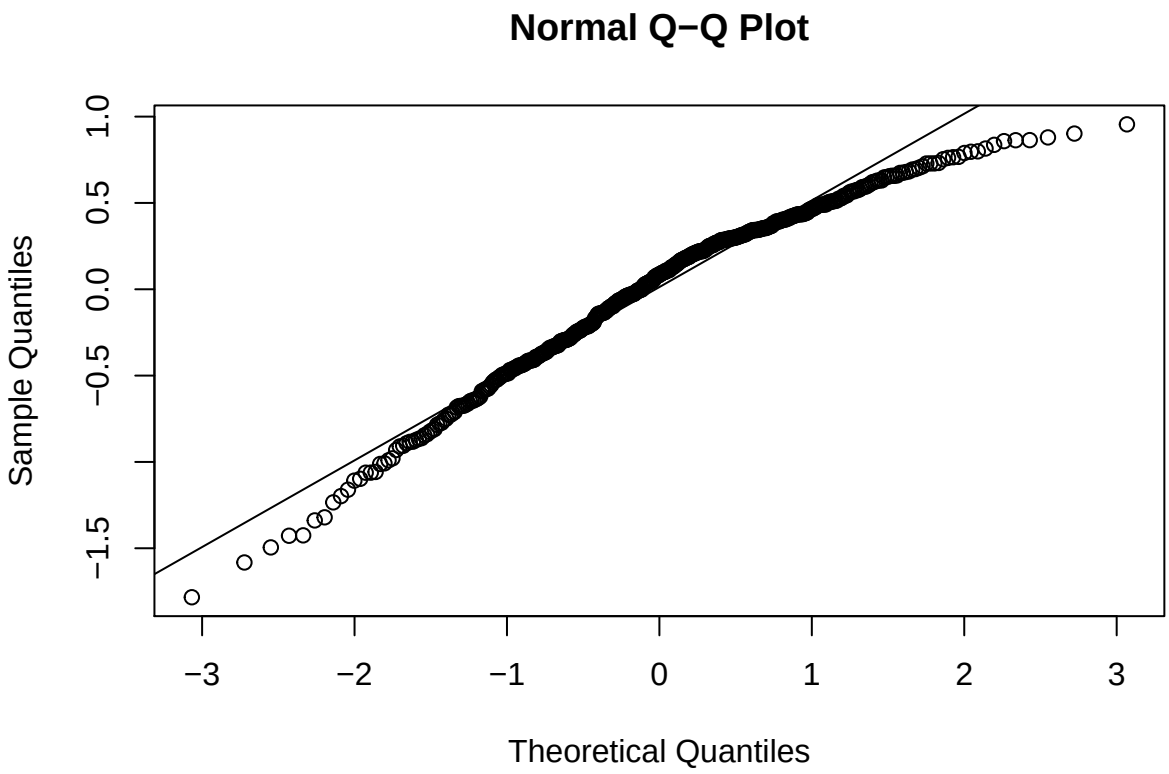
The diagnostic plots below suggest that the conditions are reasonably met.

```
m_best = lm(score ~ rank + ethnicity + gender + language + age + cls_perc_eval
        + cls_students + cls_level + cls_credits + bty_avg
        + pic_outfit + pic_color, data = evals)

plot(m_best$residuals ~ m_best$fitted.values)
abline(h = 0, lty = 3)
```




```
qqnorm(m_best$residuals)
qqline(m_best$residuals)
```



Exercise 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?

Yes, professors who taught multiple courses would appear multiple times as separate rows. Linear regression requires that the observations to be independent, which they cannot be if they are the same person in certain instances.

Exercise 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.

Teaching, not minority, male, english, younger, attractive, with black and white photo.

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

I'm not sure how representative the sample of UT Austin students and professors are of students and professors at other universities; therefore, I am not comfortable generalizing these conclusions.