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.
`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

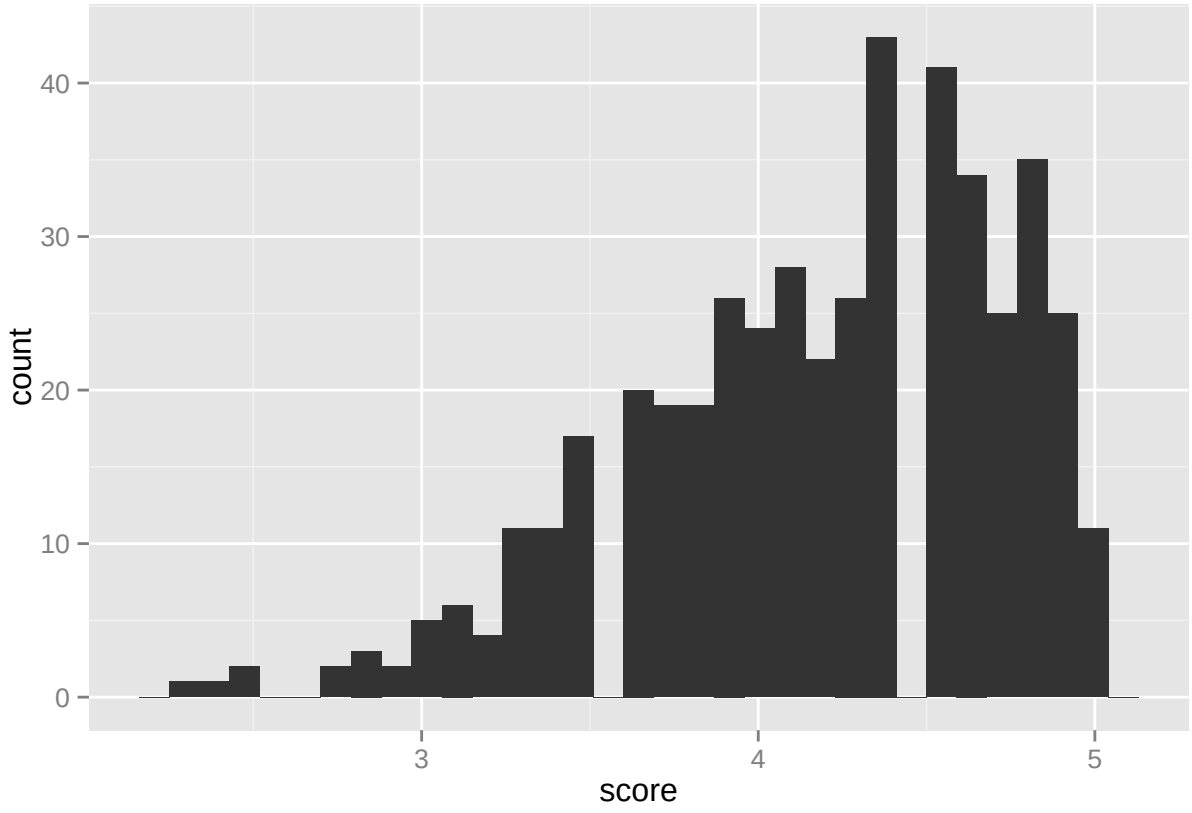
1. Is this an observational study or an experiment? The original research question posed in the paper is whether beauty leads directly to the differences in course evaluations. Given the study design, is it possible to answer this question as it is phrased? If not, rephrase the question.

   This is an observational study since we're observing whether or not a professor's looks correlates with ratings.

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?

```
library(ggplot2)

ggplot(evals,aes(x=score)) + geom_histogram()
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```



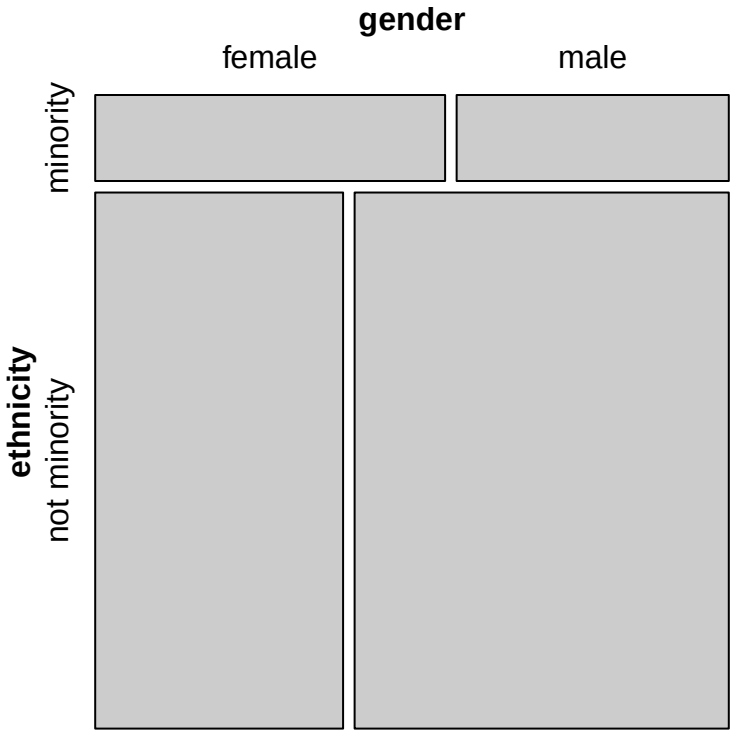

```
It looks like the distribution of scores is negatively skewed, with more
students giving higher ratings than lower ones. I actually expected to see
a positive skew for scores, with students who were upset with a course more
likely to give a (more negative) grade.
```

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

```
library(vcd)
```

```
## Loading required package: grid
```

```
mosaic(~ ethnicity + gender, data=evals)
```

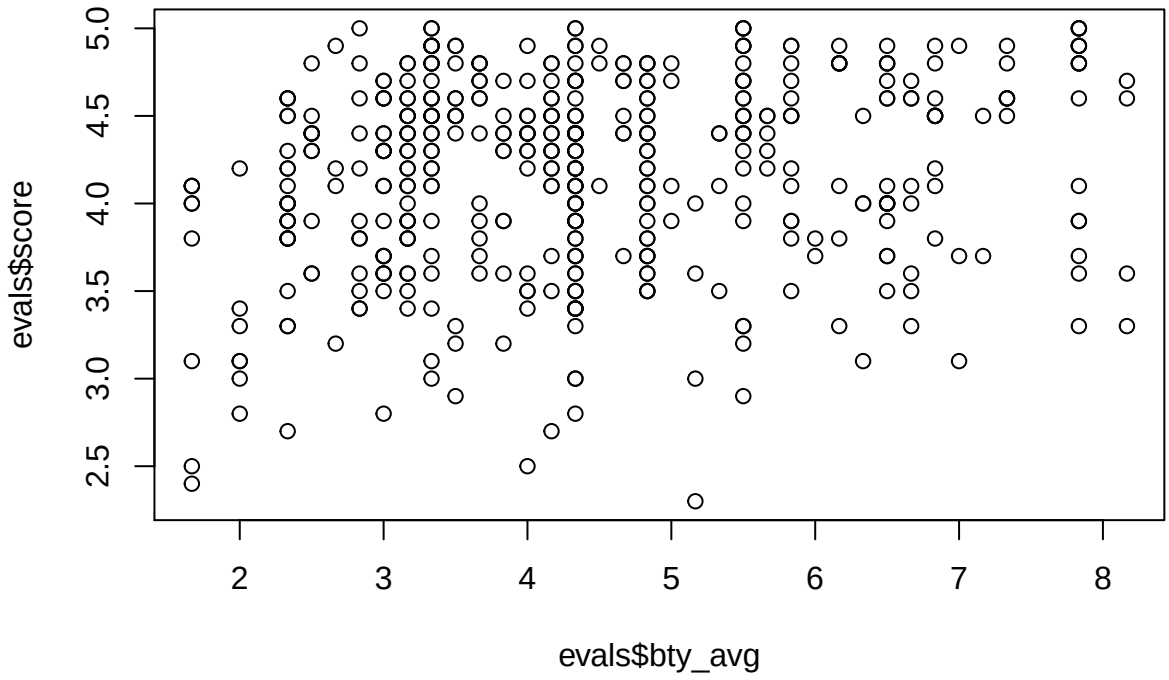


Using a mosaic plot to compare professors by ethnicity and gender, it seems like it is more likely for minority professors to be women than non-minority professors.

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:

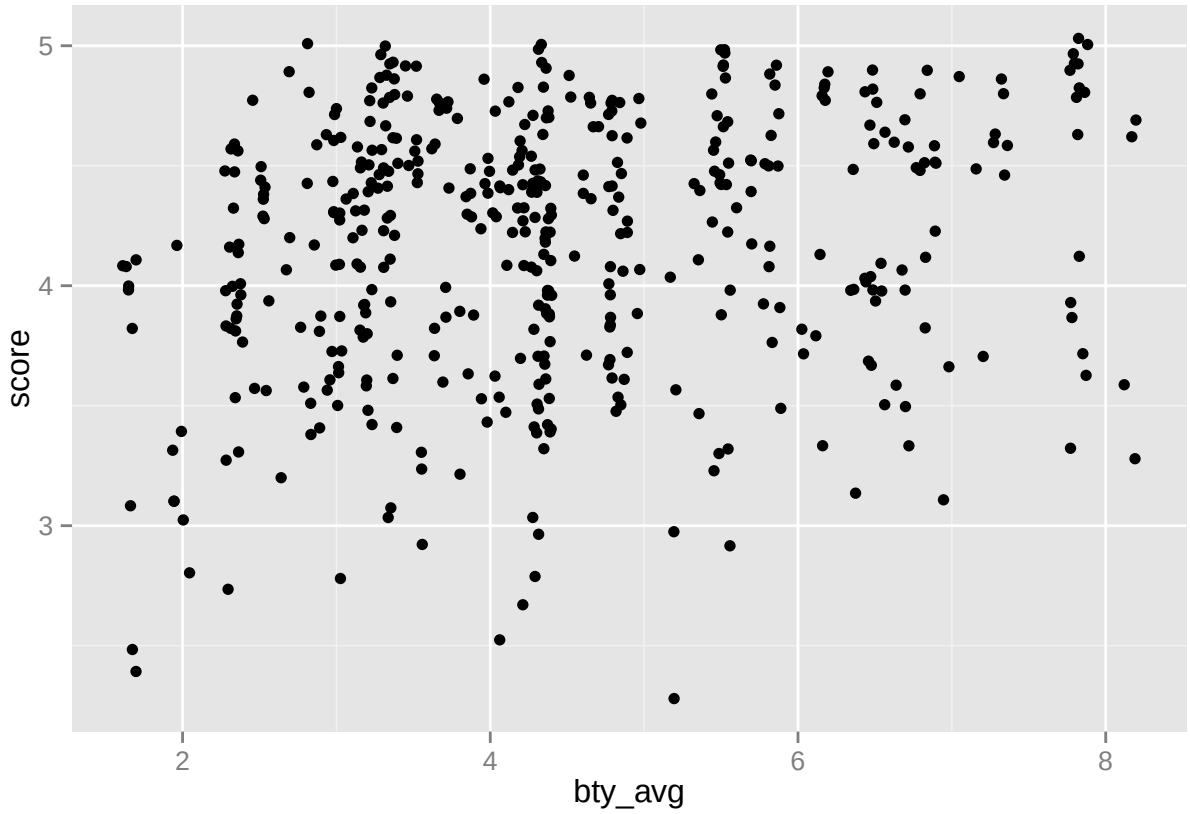
```
plot(evals$score ~ evals$bty_avg)
```



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?

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?

```
ggplot(evals,aes(x=bty_avg, y=score)) + geom_point(position = "jitter")
```



It looks like there were points that had the same value. When viewed in a scatter plot without jittering, the points would appear to be exactly the same. By jittering, we can see clusters of points that were otherwise hidden in this way.

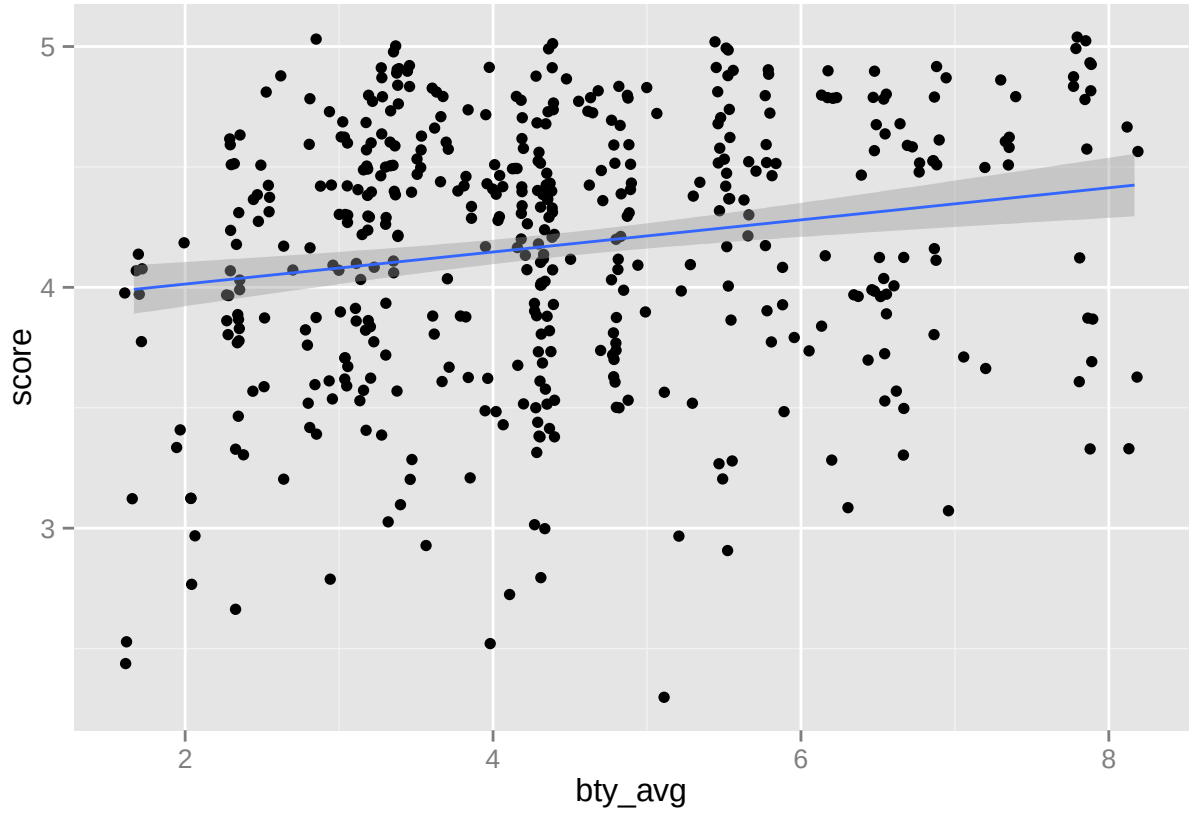
5. Let's see if the apparent trend in the plot is something more than natural variation. Fit a linear model called `m_bty` to predict average professor score by average beauty rating and add the line to your plot using `abline(m_bty)`. Write out the equation for the linear model and interpret the slope.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

```
m_bty <- lm(score ~ bty_avg, data=evals)

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
```

```
ggplot(evals,aes(x=bty_avg, y=score)) + geom_point(position = "jitter") +
  stat_smooth(method="lm")
```

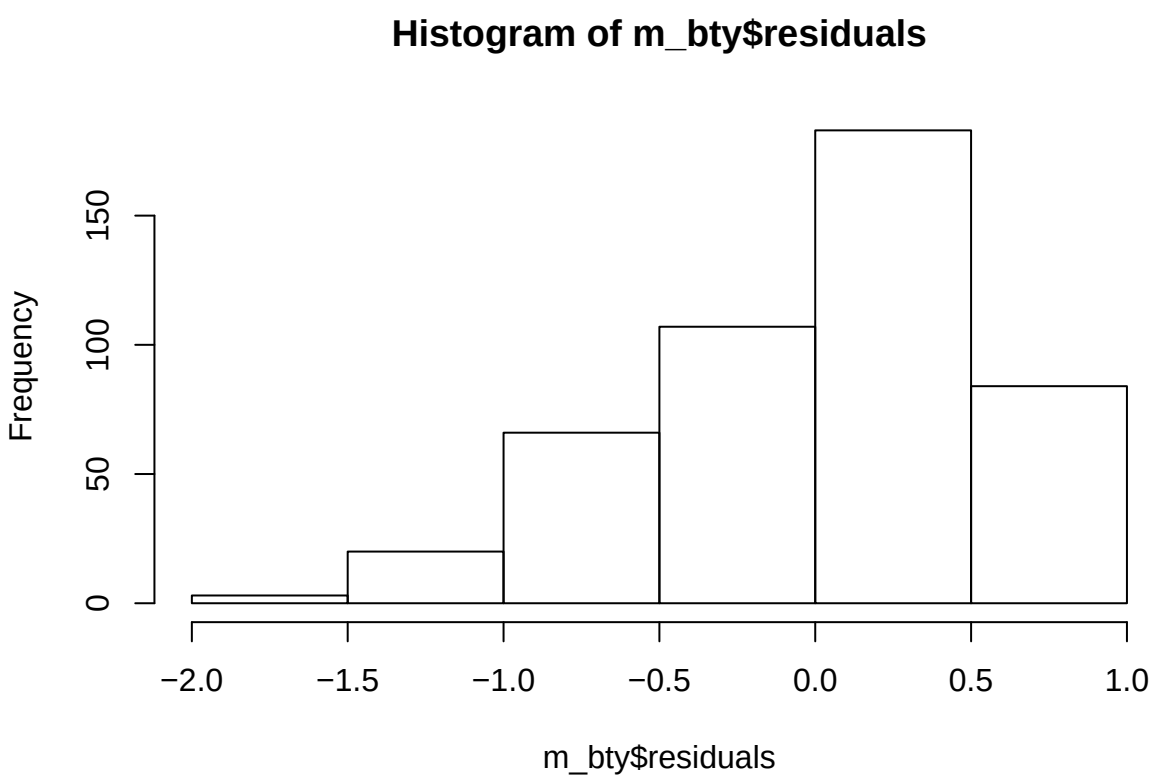


The equation for this line is: score = 3.88034 + 0.06664 $\times$ bty_avg

It appears to be somewhat predictive, but there also appear to be lots of residuals. Our R-squared tells us the same thing: at only 0.03502, this model isn't very strong.

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).

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

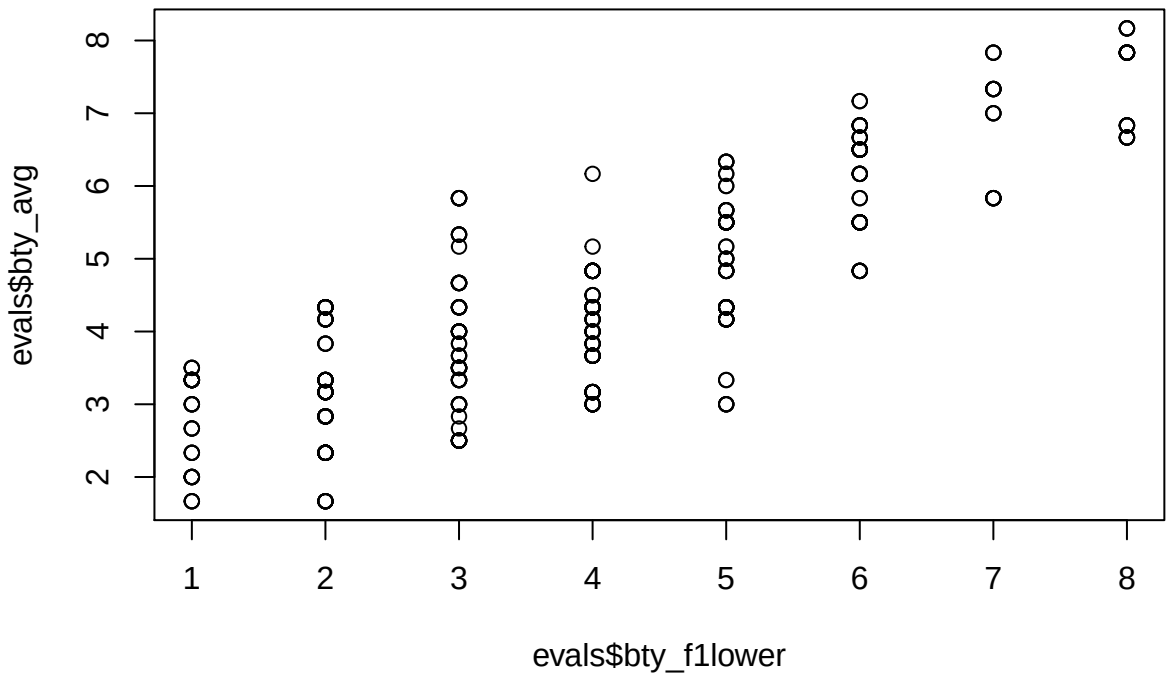



```
Linearity: This condition appears to be partially met, there is a wide
variance, but there is a certain degree of linearity.

Nearly normal residuals: This condition seems to be somewhat met... The
data iss a bit negatively skewed however.

Constant variability: This condition, according to the scatter plot, might
not be met. First of all, there is a chunk of scores between 4 and 5 that
might be skewing this condition. Higher beauty scores are more sparse, and
the score cap at 5 seems to be affecting the variability as the beauty score
gets higher.
```

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.

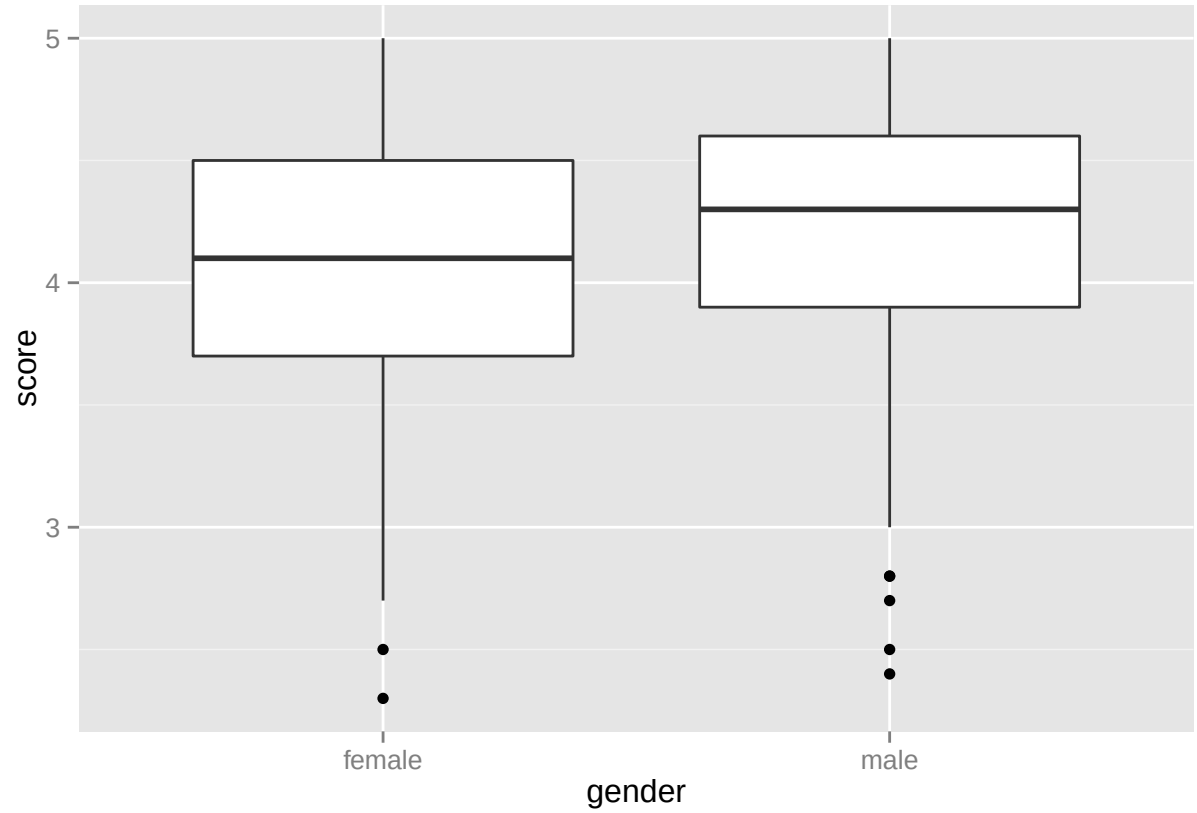
```
m_bty_gen <- lm(score ~ bty_avg + gender, data = evals)
summary(m_bty_gen)
```

```
##
## 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
```

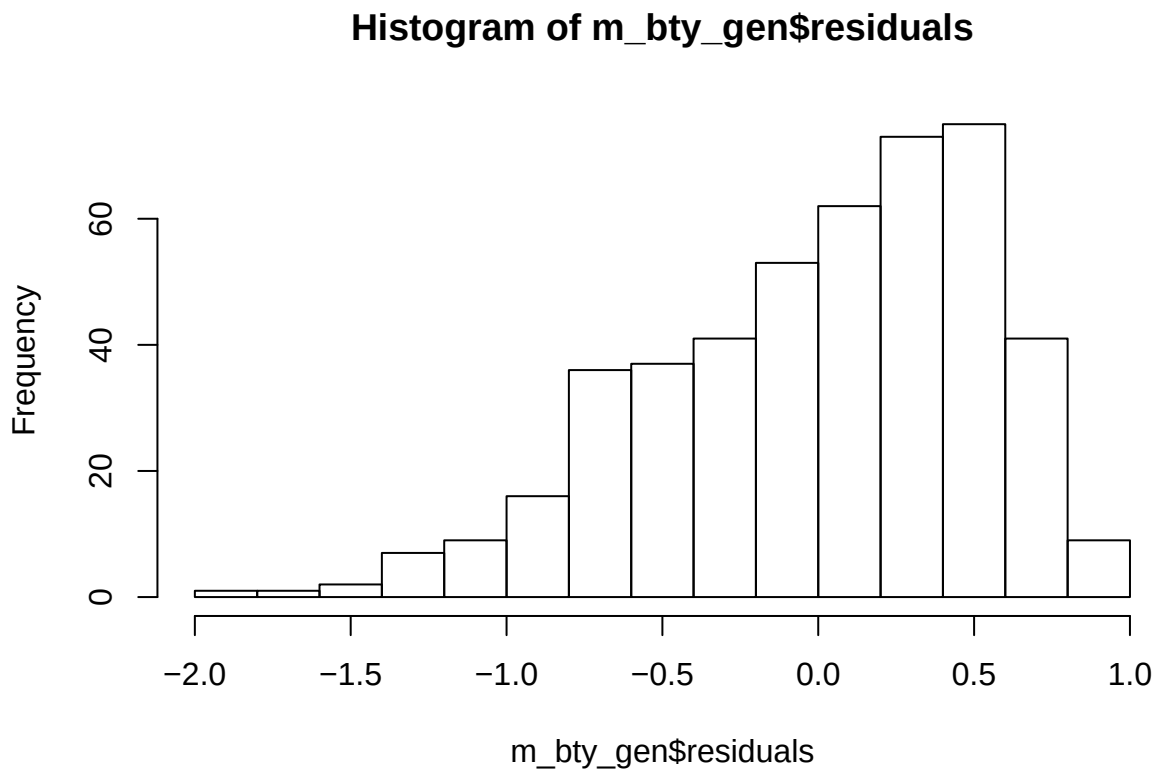
7. P-values and parameter estimates should only be trusted if the conditions for the regression are reasonable. Verify that the conditions for this model are reasonable using diagnostic plots.

```
ggplot(evals,aes(x=gender,y=score)) + geom_boxplot()
```




```
Here we're dealing with a categorical variable, so it made more sense to
view these as boxplots. It does seem like there is a difference between
male and female scores.
```

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



The residuals seem to be fairly normal as well, if a bit skewed once again.

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`?

```
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
```

```
It does still seem like average beauty is a significant predictor of
score, and the gender of the student has only improved our $R^2$ a bit
from 0.03 to 0.05 (still not a great $R^2$)
```

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?

   For males, the equation of the line is: score = 3.91973 + 0.07416 × bty_avg

   This is just the genderemale parameter plus the intercept, since if we're only looking at males this will always be 1. For two professors who received the same beauty rating, males would be predicted to have higher scores.

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
```

```
levels(evals$rank)
```

```
## [1] "teaching"     "tenure track" "tenured"
```

```
If there are n levels, R adds n-1 variables. These represent n-1 categories,
with the nth category represented by all the variables being 0.
```

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.

    This is actually a toss up for me... All of these variables could conceptually affect student rankings. If I had to guess, I would think that ethnicity might play the lowest role.

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.

    Ethnicty played a low role in accordance to what I guessed. According to the p-values, other variables that didn't predict too well were stats about the class, and tenure (which was actually my second guess.)

13. Interpret the coefficient associated with the ethnicity variable.

As mentioned above, ethnicity does not appear to be significant in predicting the scores of professors. The p-value here is 0.12, which is above our 5% threshold.

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?

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

```
##
## 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
```

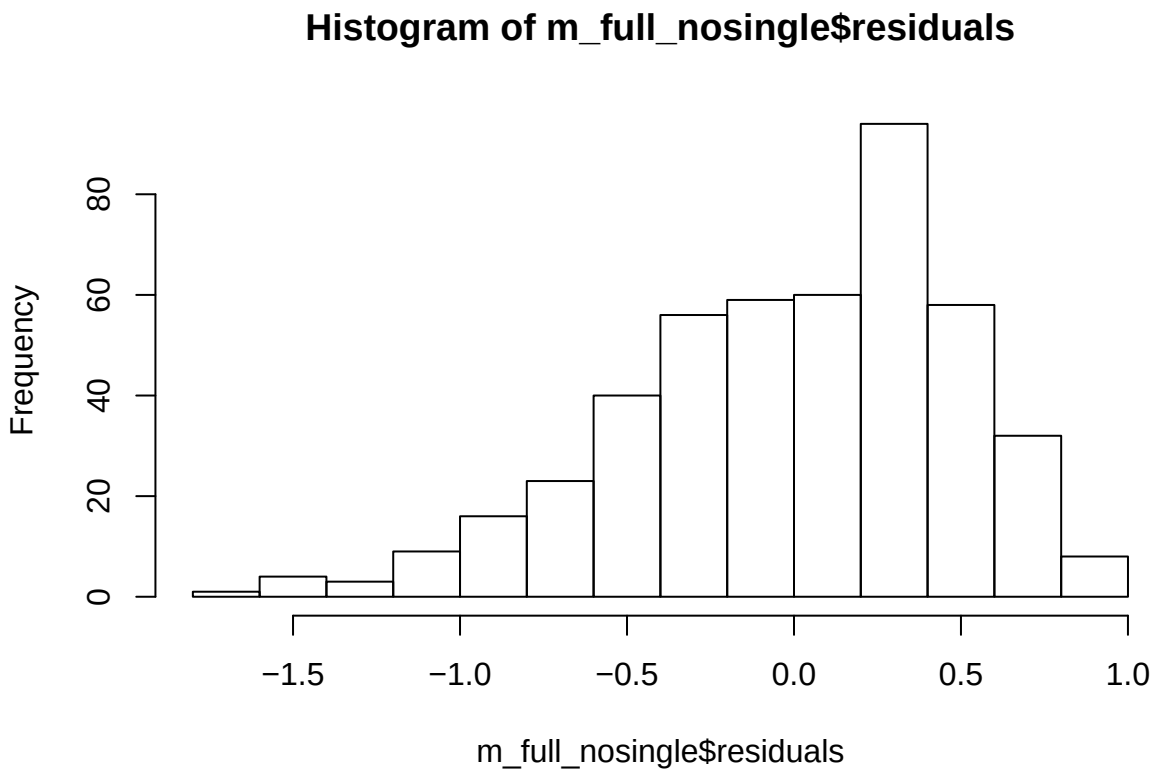
```
Our R squared adjusted changed slightly, but more importantly, our
p-values seemed to have changed. It seems like now we have no p-value
above 0.3, while before we had some up to 0.7.
```

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 model created above actually seems to be best. I tried to remove the next highest p-value, cls_level, but the R-squared fell slightly, and a few of the p-values were higher than in the previous model.

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

```
hist(m_full_nosingle$residuals)
```



It seems like the residuals are as normal as in the models we were studying before. If we wanted to truly check all diagnostic plots, we'd want to check the scatter plots to make sure our data looks fairly linear

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?

    I'd be interested to see if they double count professors in classes. In order to properly model this, they should link the scores given to professors in different classes, and average them (making sure to account for the increased sample)

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.

    Based on the final model, the qualities being associated with a high evaluation score are being of teaching rank, non-minority, male, speaking english, younger, and teaching one credit classes.

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

    It seems like this might be fairly generalizeable. Personal qualities of the professor, such as age, language, gender, and ethnicity might be