DATA606_Lab9_RJM

RJM

2019-12-29

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.

```
download.file("http://www.openintro.org/stat/data/evals.RData", destfile = "evals.RData")

load("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.

variable	description
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

```
head(evals)
```

```
##   score         rank    ethnicity gender language age cls_perc_eval
## 1   4.7 tenure track     minority female  english  36      55.81395
## 2   4.1 tenure track     minority female  english  36      68.80000
## 3   3.9 tenure track     minority female  english  36      60.80000
## 4   4.8 tenure track     minority female  english  36      62.60163
## 5   4.6      tenured not minority   male  english  59      85.00000
## 6   4.3      tenured not minority   male  english  59      87.50000
##   cls_did_eval cls_students cls_level cls_profs  cls_credits bty_f1lower
## 1           24           43     upper    single multi credit           5
## 2           86          125     upper    single multi credit           5
## 3           76          125     upper    single multi credit           5
## 4           77          123     upper    single multi credit           5
## 5           17           20     upper  multiple multi credit           4
## 6           35           40     upper  multiple multi credit           4
##   bty_f1upper bty_f2upper bty_m1lower bty_m1upper bty_m2upper bty_avg
## 1           7           6           2           4           6       5
```

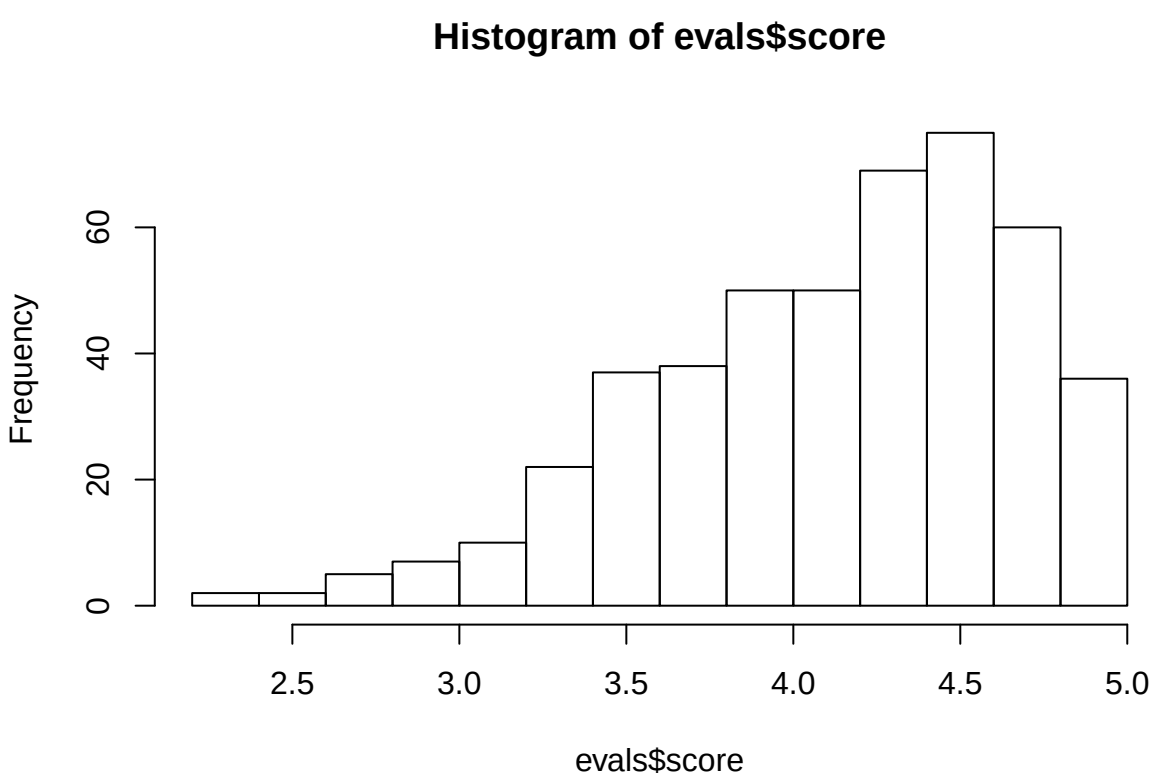
```
## 2           7           6           2           4           6       5
## 3           7           6           2           4           6       5
## 4           7           6           2           4           6       5
## 5           4           2           2           3           3       3
## 6           4           2           2           3           3       3
##   pic_outfit pic_color
## 1 not formal     color
## 2 not formal     color
## 3 not formal     color
## 4 not formal     color
## 5 not formal     color
## 6 not formal     color
```

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 it is not a controlled experiment, a direct relationship between beauty and course evaluations is hard to prove. The question should be rephrased to whether beauty has an impact on the course evaluations.

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?

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



```
summary(evals$score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.300   3.800   4.300   4.175   4.600   5.000
```

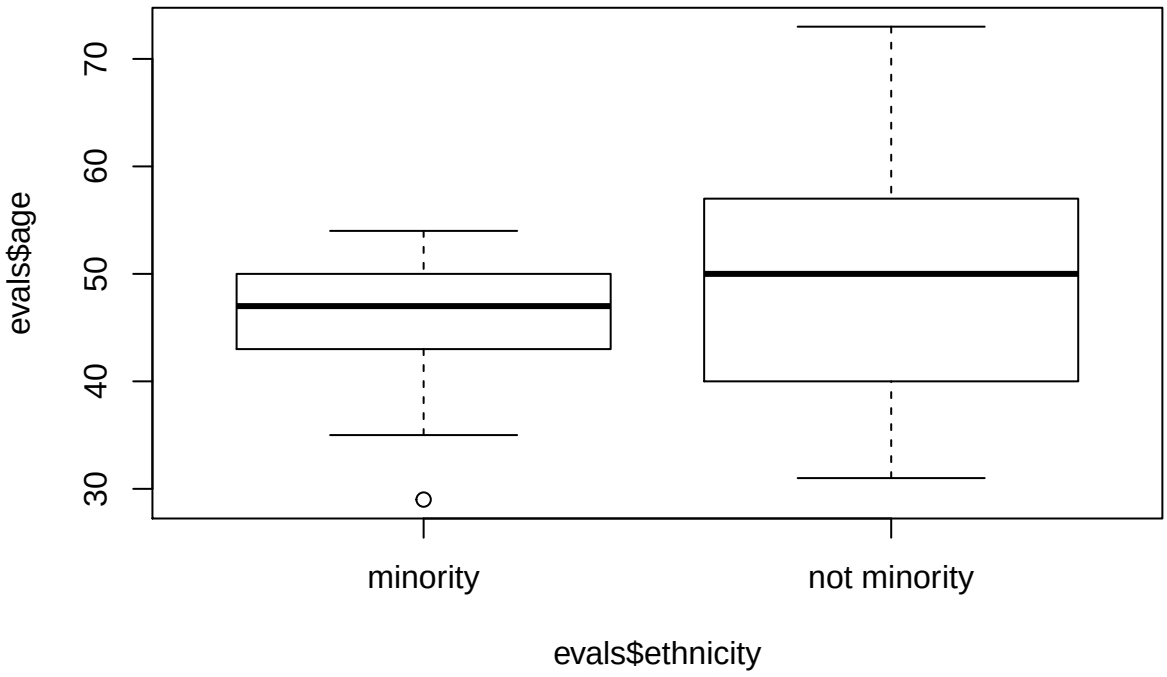
The distribution seems to be normal and left skewed. Most of the students tend to rate the professors higher with 75% of the ratings being over 3.8 and the median is 4.3. This is contrary to what I had thought to be the case as I was expecting a much more lower median and a lot more of the ratings lower than the mean.

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

```
head(evals)
```

```
##   score         rank    ethnicity gender language age cls_perc_eval
## 1   4.7 tenure track     minority female  english  36      55.81395
## 2   4.1 tenure track     minority female  english  36      68.80000
## 3   3.9 tenure track     minority female  english  36      60.80000
## 4   4.8 tenure track     minority female  english  36      62.60163
## 5   4.6      tenured not minority   male  english  59      85.00000
## 6   4.3      tenured not minority   male  english  59      87.50000
##   cls_did_eval cls_students cls_level cls_profs  cls_credits bty_f1lower
## 1           24           43     upper    single multi credit           5
## 2           86          125     upper    single multi credit           5
## 3           76          125     upper    single multi credit           5
## 4           77          123     upper    single multi credit           5
## 5           17           20     upper  multiple multi credit           4
## 6           35           40     upper  multiple multi credit           4
##   bty_f1upper bty_f2upper bty_m1lower bty_m1upper bty_m2upper bty_avg
## 1           7           6           2           4           6       5
## 2           7           6           2           4           6       5
## 3           7           6           2           4           6       5
## 4           7           6           2           4           6       5
## 5           4           2           2           3           3       3
## 6           4           2           2           3           3       3
##   pic_outfit pic_color
## 1 not formal     color
## 2 not formal     color
## 3 not formal     color
## 4 not formal     color
## 5 not formal     color
## 6 not formal     color
```

```
boxplot(evals$age ~ evals$ethnicity)
```



Based on the above plots, it seems that the profs belonging to minorities tend to be younger as compared to profs that are not from minority groups.

Simple linear regression

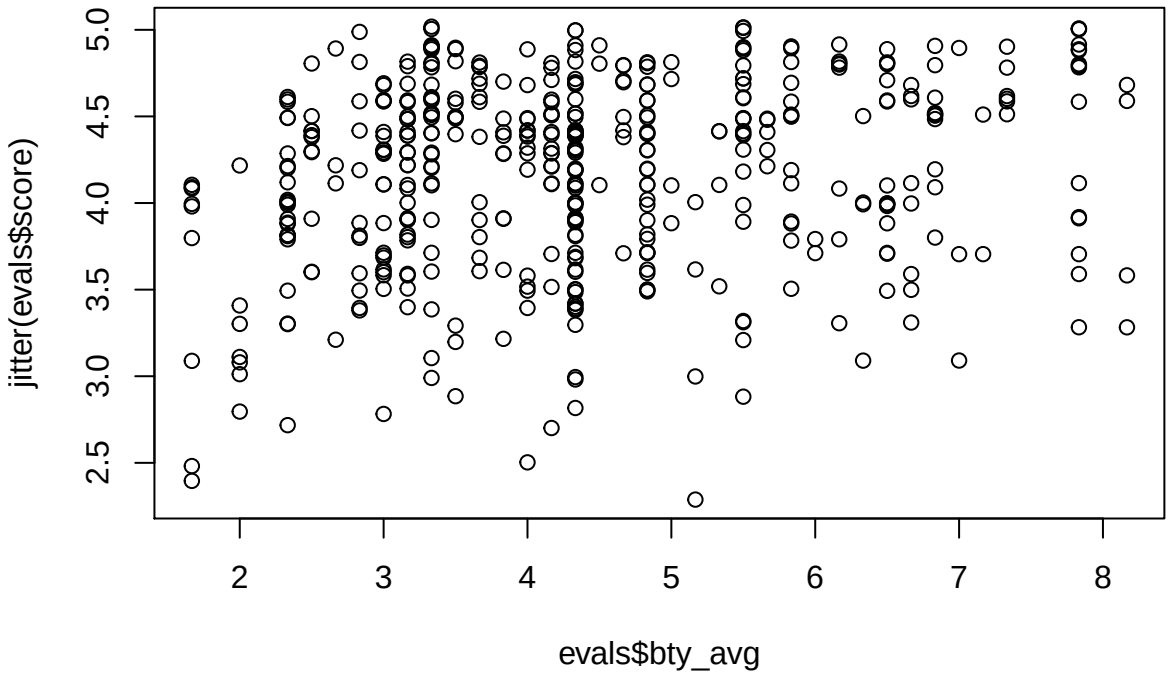
The fundamental phenomenon suggested by the study is that better looking teachers are evaluated more favorably. Let's create a scatterplot to see if this appears to be the case:

```
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?

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



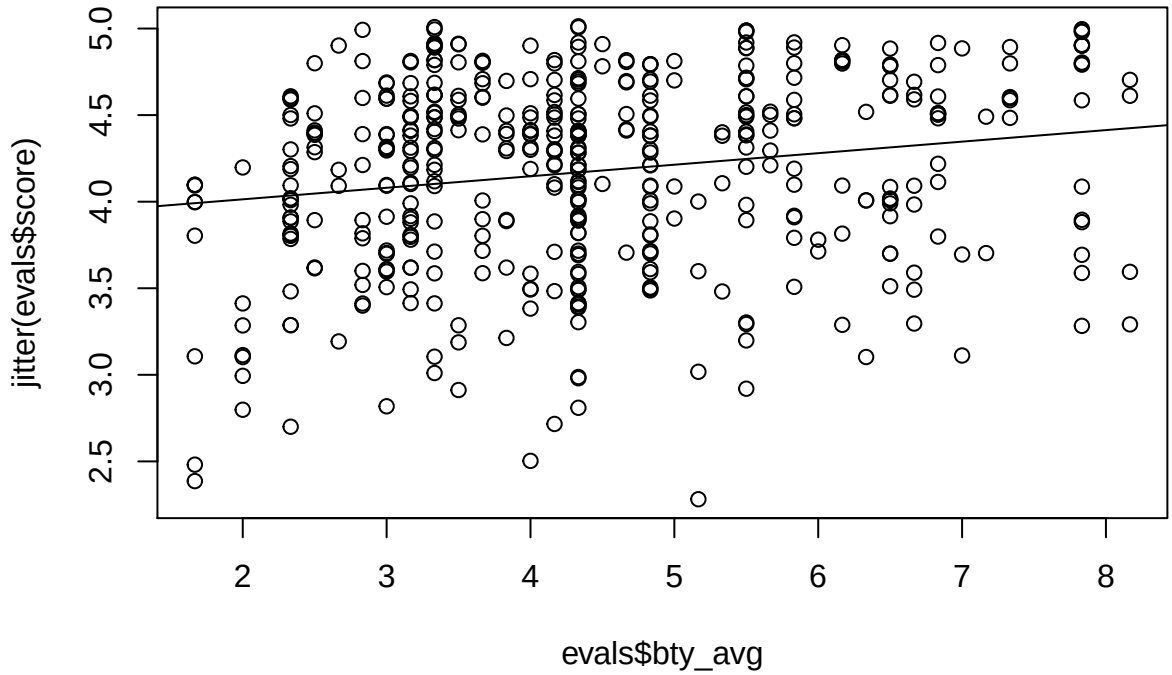
The original plot does not indicate whether there are more than one datapoints represented by a circle. By using the jitter function, we can see that multiple observations with the same value are indicated by the darkening boundry of the circle.

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)
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
```

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

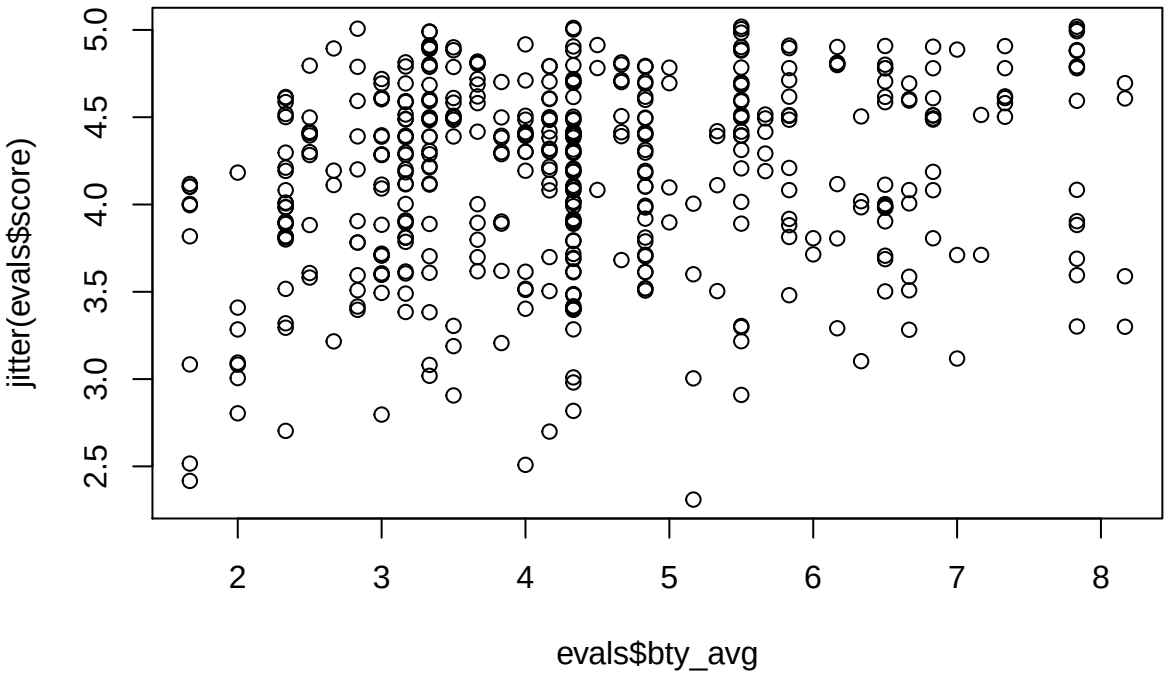


The equation of the linear model is: y_hat = 3.88 + 0.067 * bty_avg

The average beauty is a statistically significant score as p-value is approaching zero. However, it is not a significant predictor as only 3.5% of variation in the score is explained by it.

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(x = evals$bty_avg, y = jitter(evals$score))
```



From the above, it seems that the relationship is stronly linear with a few outliers. The equation of the line is almost same as the linear regression model's line.

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])
```

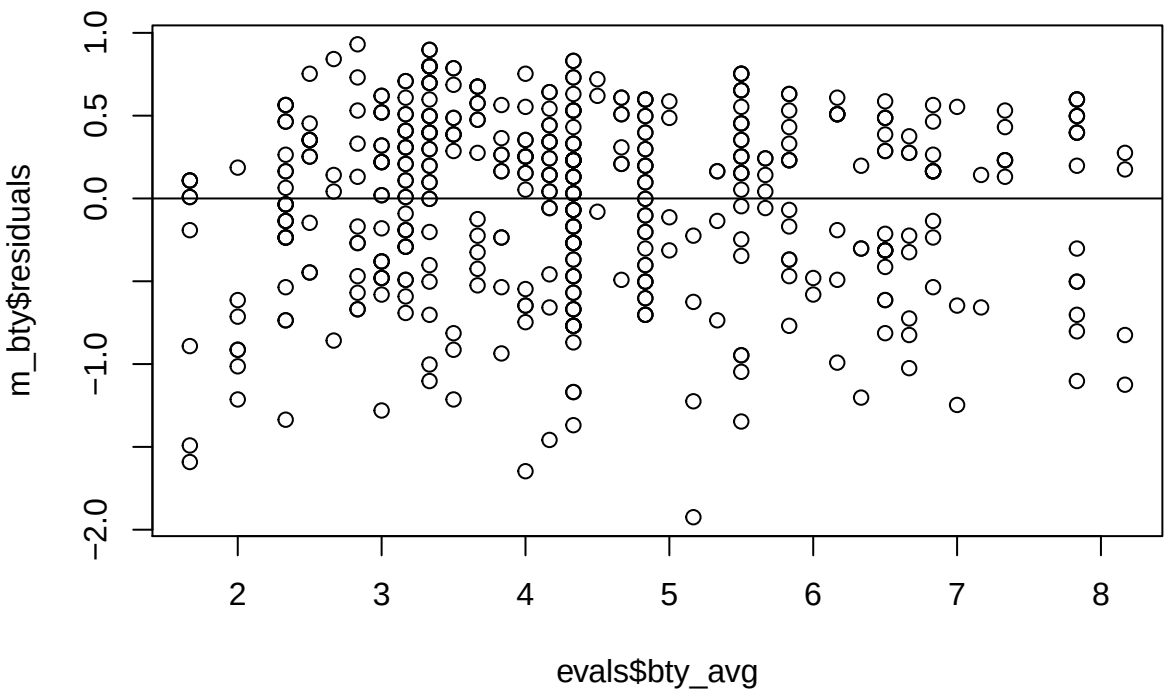
These variables are collinear (correlated), and adding more than one of these variables to the model would not add much value to the model. In this application and with these highly-correlated predictors, it is reasonable to use the average beauty score as the single representative of these variables.

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)
```

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.

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



From the above, we can see that the distribution is nearly normal with a slight left skew. The linearity is already established, variables look independent, and the various of the residuals appears to be constant. The conditions of regression are reasonable.

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 bty_avg still is a significant predictor of score and addition of gender has further reduced the p-value solidifying it is a significant predictor.

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?

The equation of the line is:

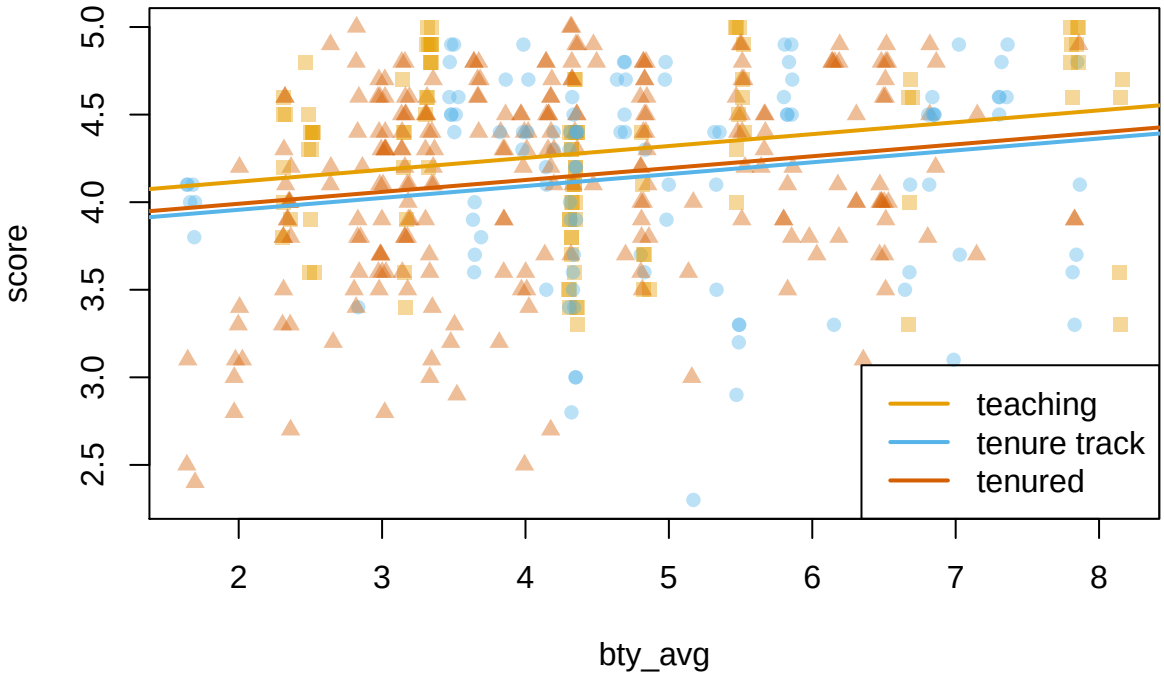
score = 3.74734 + (0.07416 x bty_avg) + (0.17239 x gendermale)

It seems from the equation tht male professors will have higher course evaluation 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)
multiLines(m_bty_rank)
```




```
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
```

R has used two variable with auto-generated names (ranktenure track and ranktenured) to account for the three categories of teaching, tenure track, and tenured. Looking at multilines graph, it seems that teaching classification has the biggest positive impact on the scores.

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. In my opinion, it should be cls_credits as it should not have any impact on the evaluation of a professor as to how many credits the course offers. 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.

The p_value for the cls_creditone credit (dummy variable for cls_credit) is approaching zero which shows that it has a significant association with the evaluation scores. The highest p-value was for cls_profssingle.

13. Interpret the coefficient associated with the ethnicity variable.

It seems that the professors not belonging to a minority have an advantage of 0.1235 to their score compared to their counterparts belonging to minorities.

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_full <- 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)
```

```
##
## 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
```

There are minor changes in the coefficient, p-values, and r-squared values. The adjusted r-square has gone up a little as well.

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

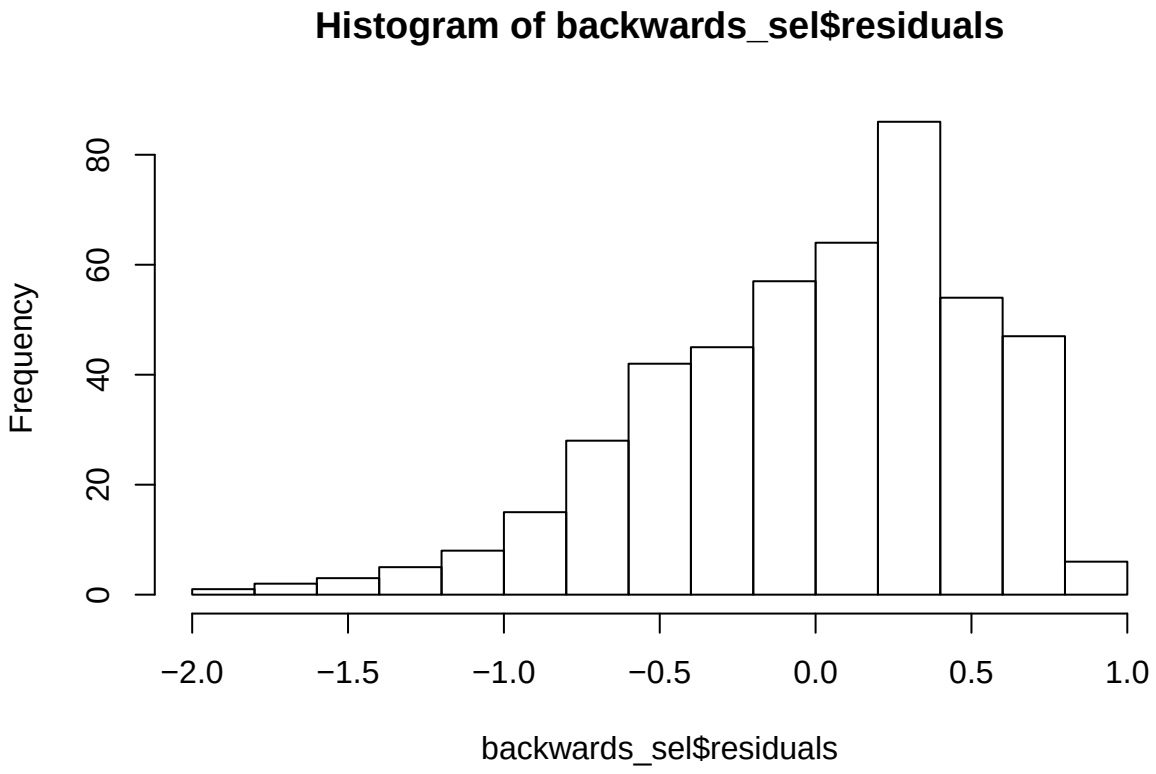
```
##
## Call:
## lm(formula = score ~ cls_credits + cls_perc_eval + gender, data = evals)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.81475 -0.32004  0.07775  0.37782  0.98356
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.636912   0.117171  31.039  < 2e-16 ***
## cls_creditsone credit 0.413725   0.104266   3.968 8.41e-05 ***
## cls_perc_eval         0.005734   0.001464   3.916 0.000104 ***
## gendermale            0.150166   0.049350   3.043 0.002477 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5214 on 459 degrees of freedom
## Multiple R-squared:  0.08684,    Adjusted R-squared:  0.08087
## F-statistic: 14.55 on 3 and 459 DF,  p-value: 4.546e-09
```

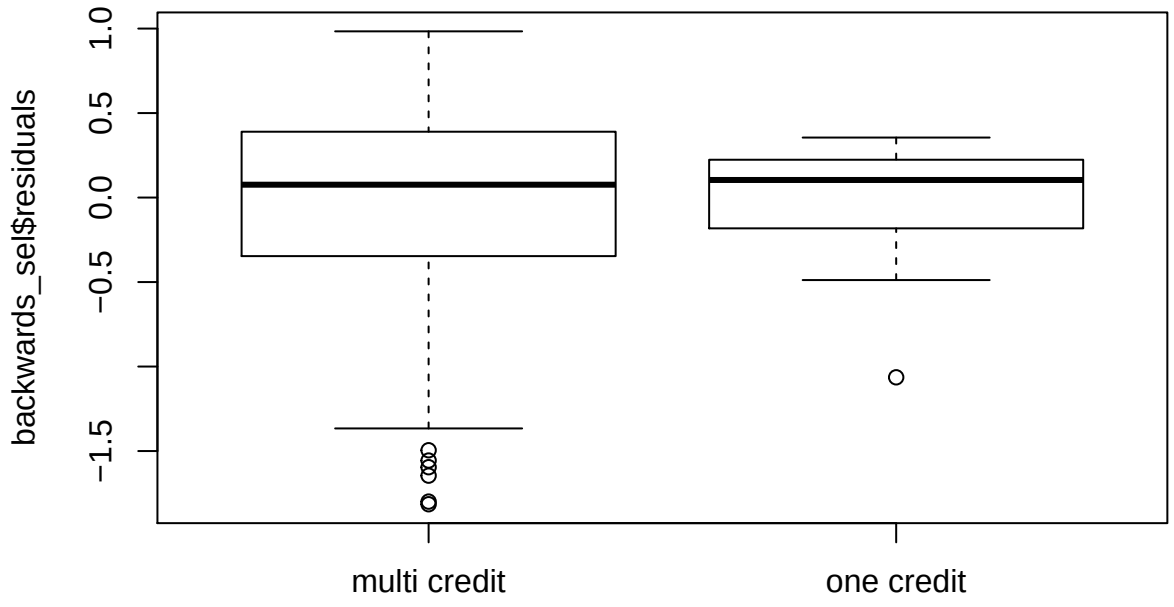
I picked the three variables with the lowest p-values.

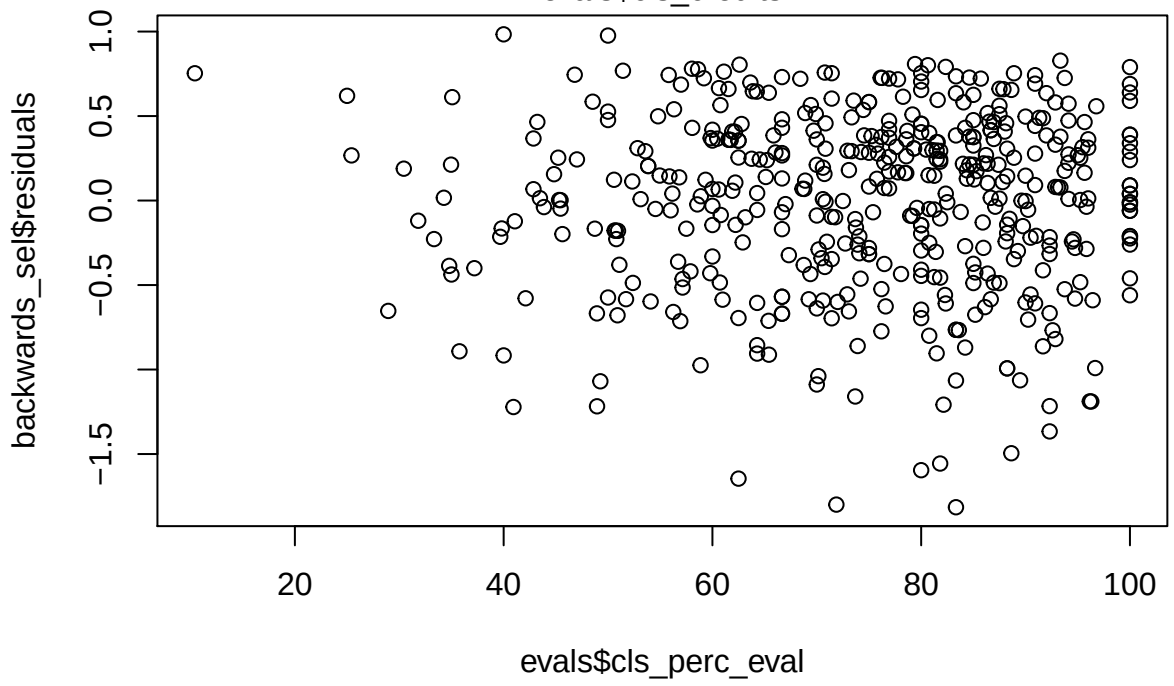
16. Verify that the conditions for this model are reasonable using diagnostic plots.

```
hist(backwards_sel$residuals)
```

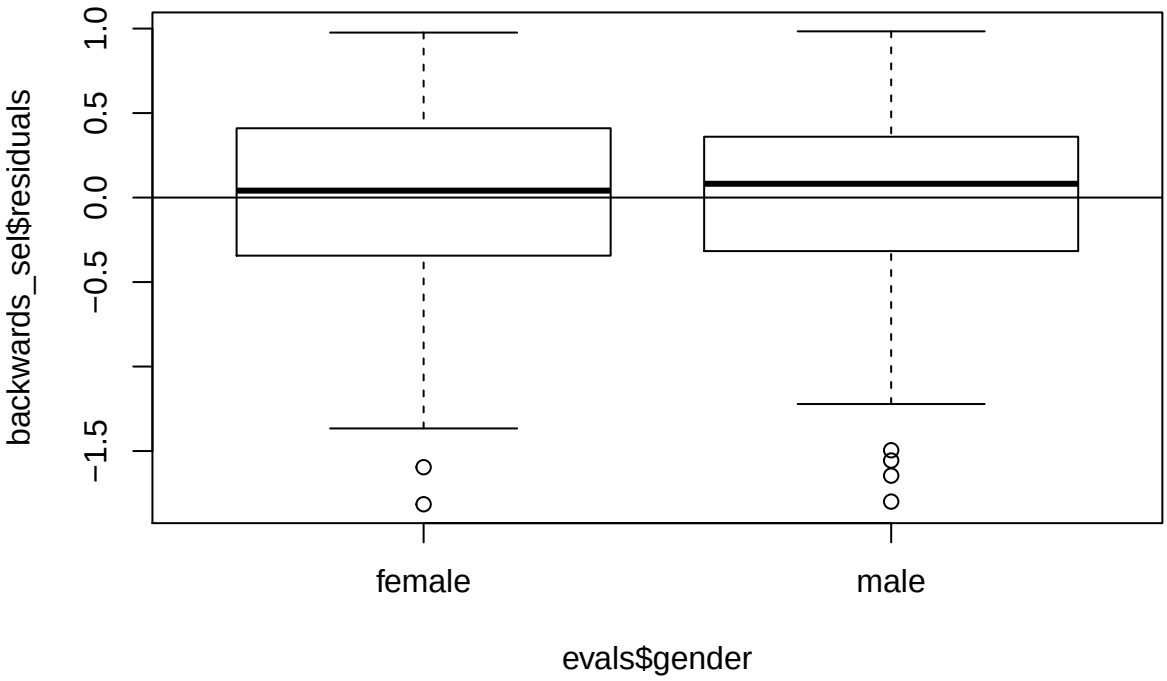



```
plot(backwards_sel$residuals ~ evals$cls_credits + evals$cls_perc_eval + evals$gender)
```


backwards_sel$residuals
1.0
0.5
0.0
−0.5
−1.5
multi credit
one credit
evals$cls_credits


backwards_sel$residuals
1.0
0.5
0.0
−0.5
−1.5
20
40
60
80
100
evals$cls_perc_eval

```
abline(h = 0)
```



The above graph show linear relationship, normal distribution, independence, and constant variance of residuals thus proving that conditions are reasonable.

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?

The condition of independence might be compromised in cases where a student has taken more than one courses with the same professor and would be highly likely to evaluate for both courses in the same manner.

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.

The higher scores will be for male professors who taught one credit courses and the more students from the class evaluated the course.

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

Since it was an obervational study and not a controlled experiment, generalization of the results is not recommended. Perhaps an experimental study, with samples more representative of the population, could be conducted to make it a representative of how beauty impacts the evaluation of professors.