Lab8 - Multiple linear regression

Michael Y.

May 5th, 2019

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 a slightly 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.

variable	description
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.

variable	description
`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, not an experiment. As such, it is not possible to infer *causation* from the results of such a study. A better research question would be whether there is an *association* between ratings of instructor attractiveness and student ratings of course evaluations.

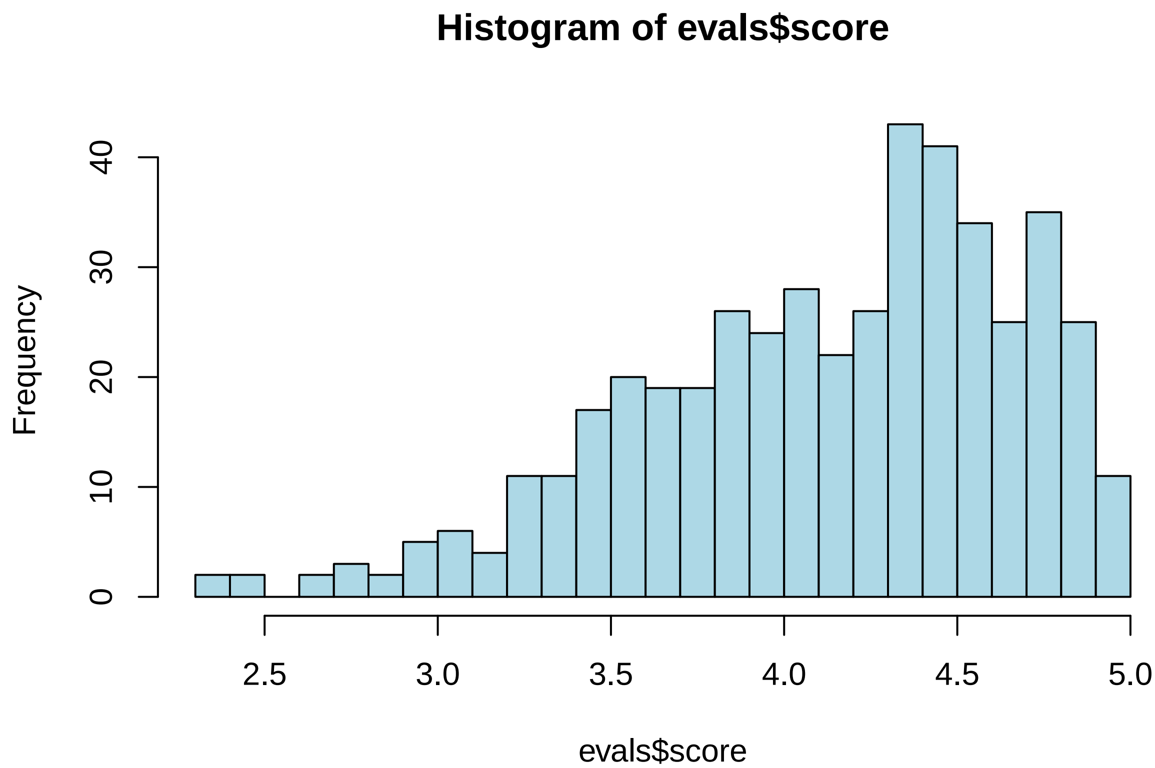
It is also noteworthy that the ratings of instructor attractiveness and the evaluations of courses are not being performed by the same individuals. Rather, students who were enrolled in courses taught by various instructors submitted their evaluations of the course at the end of each term, as is customary (however, varying percentages of such students actually did so.) Subsequently, as part of this study, a panel of six students were shown photographs of those instructors who were selected for analysis in this study and asked to rate the "attractiveness" of each instructor based upon such photos. While there is a high correlation among the ratings assigned by each of the 6 evaluators, this is not necessarily the same result that would have been obtained if the students who were submitting the course evaluations were also asked to rate the attractiveness of their instructors.

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?

```
t(t(table(evals$score)))
```

```
##
##         [,1]
##   2.3      1
##   2.4      1
##   2.5      2
##   2.7      2
##   2.8      3
##   2.9      2
##   3        5
##   3.1      6
##   3.2      4
##   3.3     11
##   3.4     11
##   3.5     17
##   3.6     20
##   3.7     19
##   3.8     19
##   3.9     26
##   4       24
##   4.1     28
##   4.2     22
##   4.3     26
##   4.4     43
##   4.5     41
##   4.6     34
##   4.7     25
##   4.8     35
##   4.9     25
##   5       11
```

```
hist(evals$score, breaks=c(23:50)/10, col="lightblue")
```




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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  2.3000  3.8000  4.3000  4.1747  4.6000  5.0000
```

Describe the distribution of score. Is the distribution skewed?

`score` is a left-skewed distribution, where the mean 4.17473002 is less than the median 4.3 . The distribution is limited on the right by the maximum score of 5. Although it is possible for courses to be rated as low as 1, the minimum actually used is 2.3 .

What does that tell you about how students rate courses?

This indicates that most students rate courses highly, but on rare occasions, a few low ratings are given.

Is this what you expected to see? Why, or why not?

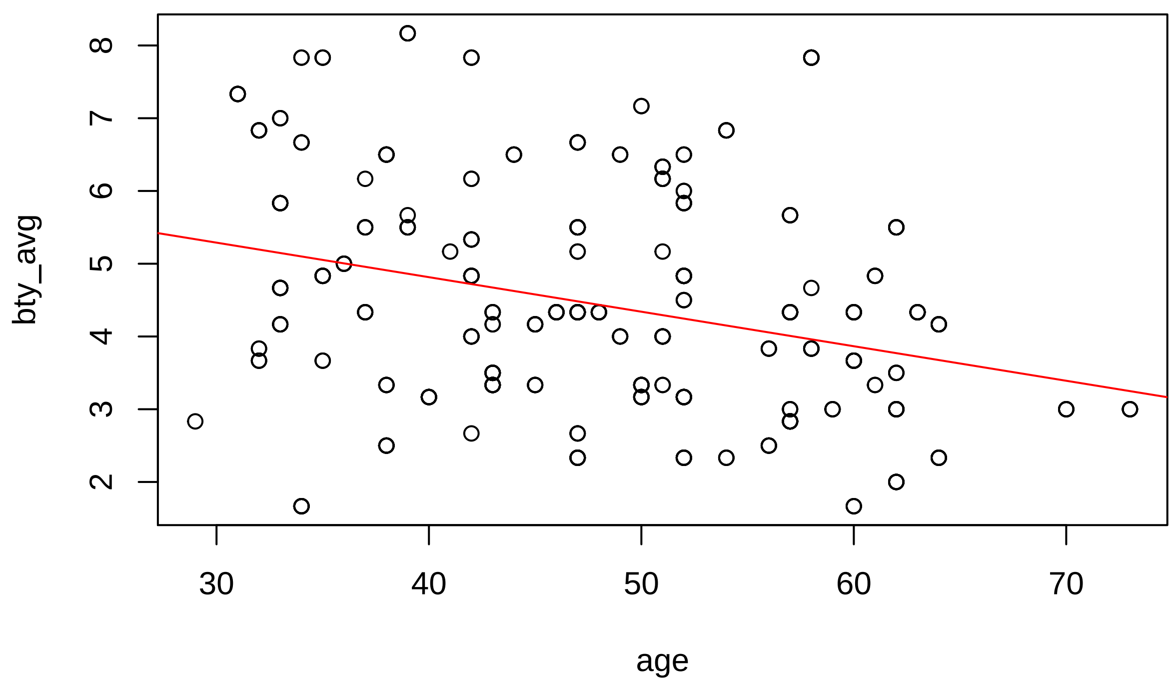
While one might naively 'expect' to see a Normal distribution, the reality is that (akin to 'uber ratings' of taxi drivers) students may feel pressured (or, tempted) to grant high ratings, perhaps in hopes that (despite the anonymity of individual ratings) the instructor would generously grant high grades. (It would be more informative if the course-by-course distribution of ratings were provided, or dispersion information such as within-course standard deviation were provided, but all we have to work with is the point estimate.)

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

Here is a *scatterplot* which displays the relationship between the `age` of the instructor and `bty_avg`, the average beauty rating of the instrutor, where such ratings have been assigned separately by six observers based upon photographs of the instructors, and then averaged.

A linear regression line shows that the average beauty rating *declines* as the age of the instructor increases:

```
plot(bty_avg ~ age, data=evals)
modl <- lm(bty_avg ~ age, data=evals)
abline(modl,col="red")
```



Here is a *boxplot* showing the relative ages of female vs. male instructors.

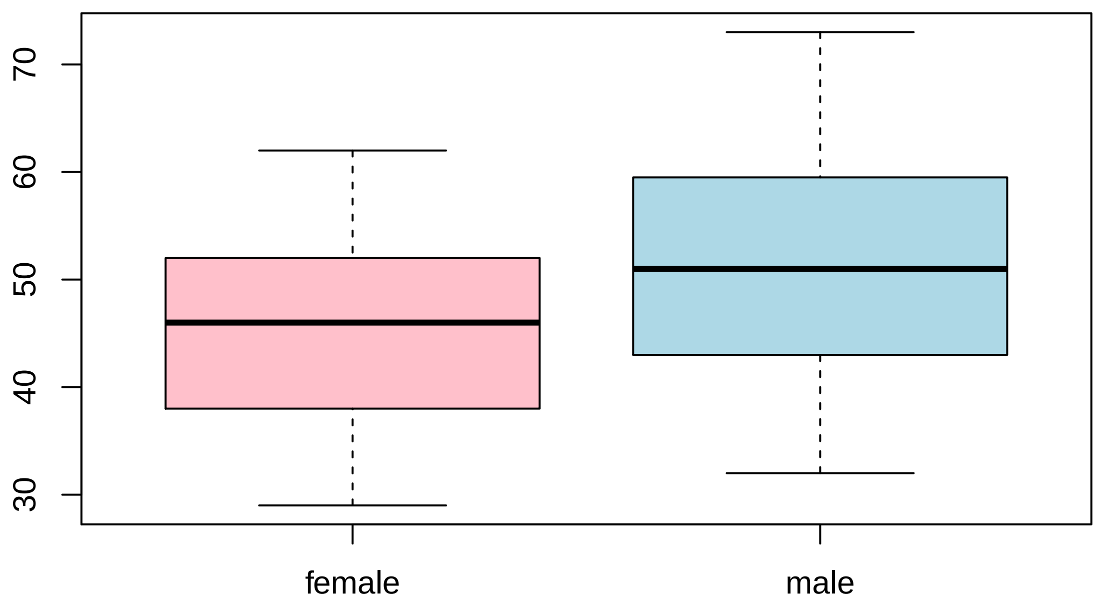
It shows that the male instructors are older than the females.

```
by(data = evals$age, INDICES = evals$gender, FUN = summary)
```

```
## evals$gender: female
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  29.000  38.000  46.000  45.092  52.000  62.000 
## ------------------------------------------------------------------------------------------
## evals$gender: male
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  32.000  43.000  51.000  50.746  59.250  73.000
```

```
boxplot(age~gender, data=evals, col=c("pink","lightblue"))
```



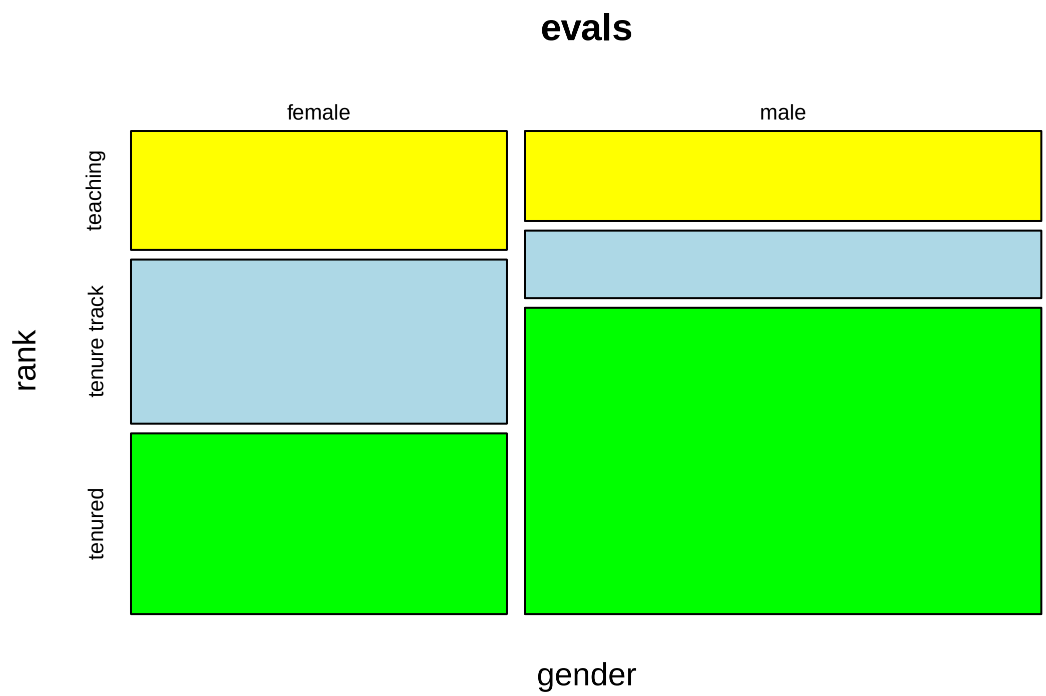
Here is a *mosaicplot* of the instructors divided by `gender` and by faculty `rank` (i.e., teaching, tenure-track, or tenured).

It shows that a much larger number of males than females are tenured, while more females than males are on the tenure-track:

```
cbind(
  rbind(
    table(evals$rank, evals$gender),
    TOTALS=colSums(table(evals$rank, evals$gender))),
  TOTALS=rowSums(rbind(table(evals$rank, evals$gender),
                       totals=colSums(table(evals$rank, evals$gender)))))
```

```
##              female male TOTALS
## teaching         50   52    102
## tenure track     69   39    108
## tenured          76  177    253
## TOTALS          195  268    463
```

```
mosaicplot(formula = gender~rank, data=evals, col=c("yellow","lightblue","green"))
```

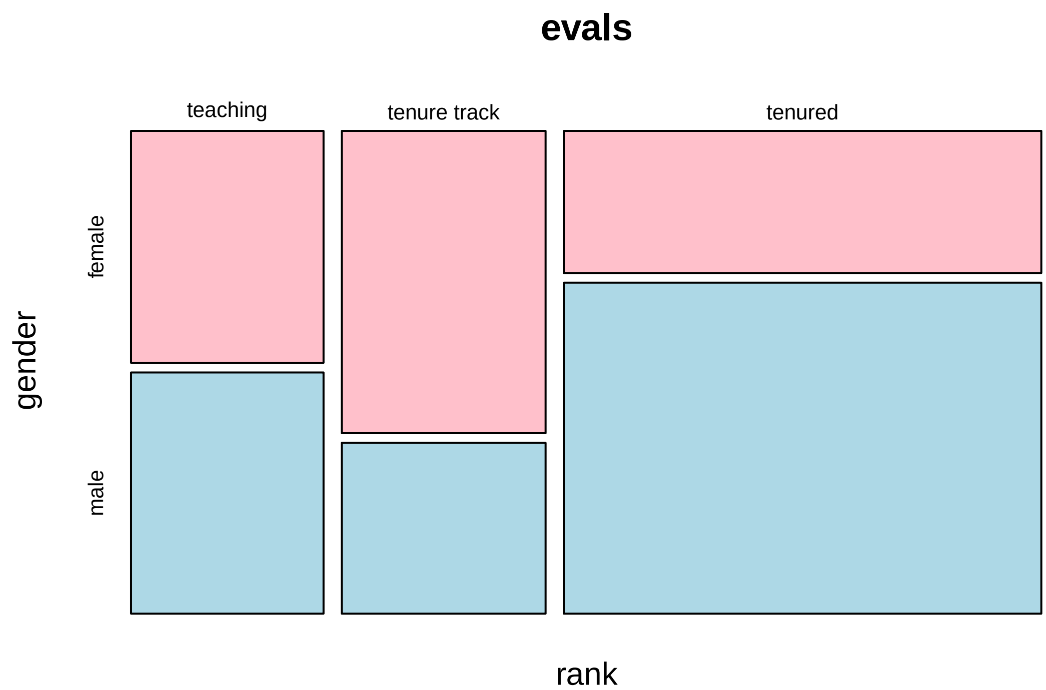


Here is another mosaicplot of the same data, rotated sideways:

```
cbind(
  rbind(
    table(evals$gender, evals$rank),
    TOTALS=colSums(table(evals$gender, evals$rank))),
  TOTALS=rowSums(rbind(table(evals$gender, evals$rank),
                       totals=colSums(table(evals$gender, evals$rank)))))
```

```
##        teaching tenure track tenured TOTALS
## female       50           69      76    195
## male         52           39     177    268
## TOTALS      102          108     253    463
```

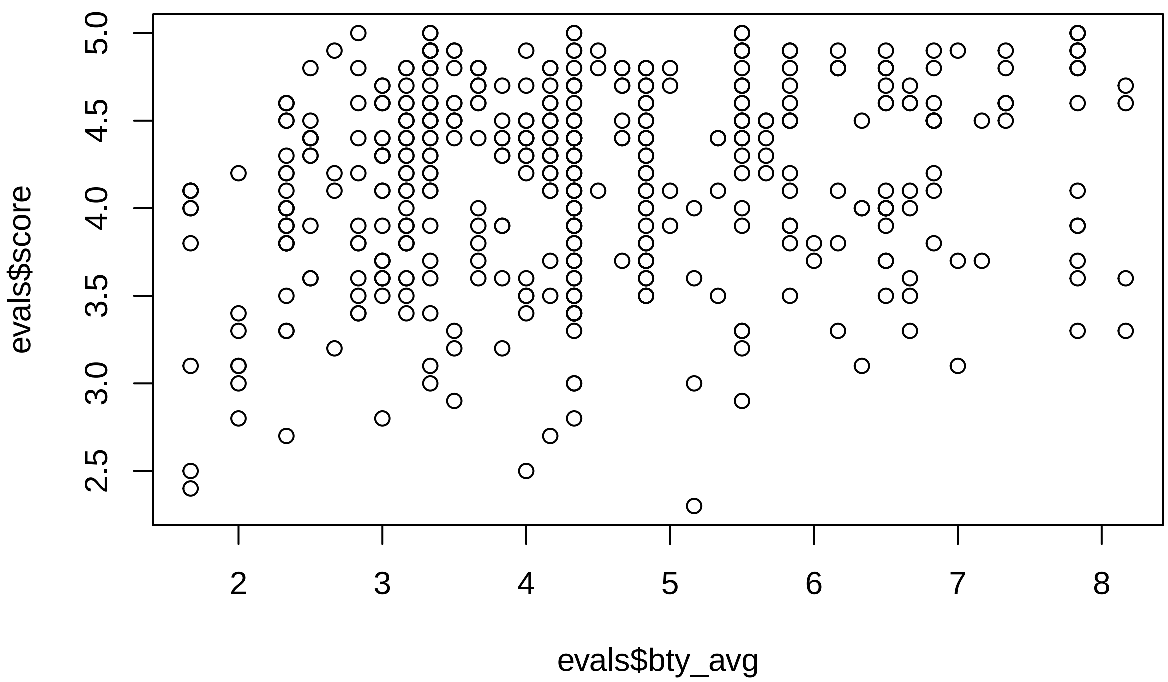
```
mosaicplot(rank ~ gender, data=evals, col=c("pink","lightblue"))
```



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 only 264 distinct points on the scatterplot, while there are 463 total observations. We can easily observe this from the data by combining each pair of (bty_avg, score) into a single number, for example by multiplying one element by a large value and then adding the second element.

```
# create a special column "score_bty_avg" by multiplying score by 10000 and adding bty_avg
evals$score_bty_avg = evals$score*10000+evals$bty_avg

# extract just these columns from the main dataframe
temp1=evals[,c("score_bty_avg","score","bty_avg")]
head(temp1,10)
```

```
##    score_bty_avg score bty_avg
## 1      47005.000   4.7   5.000
## 2      41005.000   4.1   5.000
## 3      39005.000   3.9   5.000
## 4      48005.000   4.8   5.000
## 5      46003.000   4.6   3.000
## 6      43003.000   4.3   3.000
## 7      28003.000   2.8   3.000
## 8      41003.333   4.1   3.333
## 9      34003.333   3.4   3.333
## 10     45003.167   4.5   3.167
```

```r
# sort by this special quantity (score_bty_avg)
temp2=temp1[order(temp1$score_bty_avg),]

# these reflect the items which have the highest score (i.e., 5)
temp2[temp2$score==5,]
```

```
##     score_bty_avg score bty_avg
## 406     50002.833     5   2.833
## 349     50003.333     5   3.333
## 356     50003.333     5   3.333
## 103     50004.333     5   4.333
## 108     50004.333     5   4.333
## 54      50005.500     5   5.500
## 57      50005.500     5   5.500
## 59      50005.500     5   5.500
## 420     50007.833     5   7.833
## 421     50007.833     5   7.833
## 424     50007.833     5   7.833
```

```r
# duplication can be seen among the bty_avg ratings

# tally up a table of distinct occurances of this value
temp3=table(temp2$score_bty_avg)
# This corresponds to the above duplication
tail(t(t(temp3)),5)
```

```
##
##               [,1]
##   50002.833    1
##   50003.333    2
##   50004.333    2
##   50005.5      3
##   50007.833    3
```

```r
# This confirms that all 463 items are still accounted for
sum(temp3)
```

```
## [1] 463
```

```r
# but there are only 264 distinct values
length(temp3)
```

```
## [1] 264
```

```r
# The number of values which do not repeat is 146
sum(temp3==1)
```

```
## [1] 146
```

```r
# The number of values which do repeat is 118
sum(temp3>1)
```

```
## [1] 118
```

```r
#This is the value which occurs most frequently
temp3[temp3==max(temp3)]
```

```
## 44004.333
```

```
##        10
```

```r
# this represents score==4.4 and bty_avg==4.333 ; this pair occurs 10 times

# these are the 10 observations which have the identical "score" and "bty_avg"
evals[evals$score==4.4 & evals$bty_avg==4.333,]
```

```
##     score         rank    ethnicity gender    language age cls_perc_eval cls_did_eval cls_students cl
## 98    4.4     teaching not minority   male     english  48      63.75839           95          149
## 100   4.4     teaching not minority   male     english  48      62.50000           85          136
## 101   4.4     teaching not minority   male     english  48      80.71429          113          140
## 104   4.4      tenured not minority female     english  46      79.31035           23           29
## 164   4.4     teaching not minority   male     english  63      78.57143           11           14
## 166   4.4     teaching not minority   male     english  63      77.77778           14           18
## 180   4.4 tenure track     minority female     english  47     100.00000           16           16
## 182   4.4 tenure track     minority female     english  47      70.00000            7           10
## 451   4.4 tenure track not minority female non-english  60      50.00000           11           22
## 453   4.4 tenure track not minority female non-english  60      88.88889           24           27
##     bty_f1lower bty_f1upper bty_f2upper bty_m1lower bty_m1upper bty_m2upper bty_avg pic_outfit   pic_
## 98            3           5           6           4           4           4   4.333 not formal
## 100           3           5           6           4           4           4   4.333 not formal
## 101           3           5           6           4           4           4   4.333 not formal
## 104           4           4           5           2           6           5   4.333 not formal black&
## 164           5           4           6           4           2           5   4.333 not formal
## 166           5           4           6           4           2           5   4.333 not formal
## 180           2           6           6           3           5           4   4.333 not formal
## 182           2           6           6           3           5           4   4.333 not formal
## 451           4           6           6           2           3           5   4.333     formal black&
## 453           4           6           6           2           3           5   4.333     formal black&
```

There are many cases where points are on top of each other, i.e., multiple observations with the same x and y values.

This is because there are only 146 combinations of (bty_avg,score) which are observed exactly once, while 118 combinations are observed multiple times. In the extreme, there are 10 cases where (bty_avg,score) equals (4.333,4.4). This means that we see only a single point on the above scatterplot in these cases where multiple observations have identical values.

The regular scatter plot doesn't reveal such cases of "overplotting".

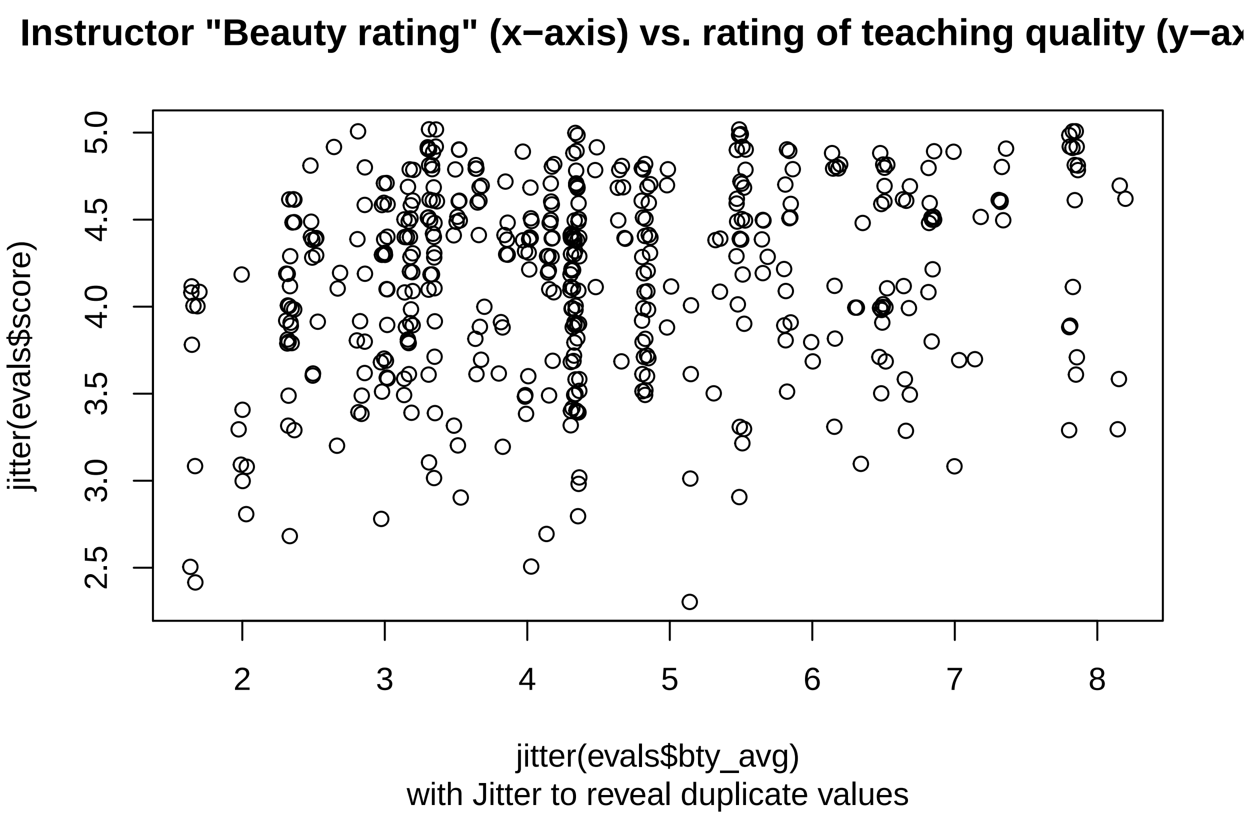
4. Replot the scatterplot, but this time use the function `jitter()` on the y- or the x-coordinate. (Use `?jitter` to learn more.) What was misleading about the initial scatterplot?

There are many cases where points are "on top of" each other, i.e., multiple observations with the same x and y values.

This is because there are only 146 combinations of (bty_avg,score) which are observed exactly once, while 118 combinations are observed multiple times. In the extreme, there are 10 cases where (bty_avg,score) equals (4.333,4.4).

This means that in those cases where multiple observations have identical values, we see only a single point on the initial scatterplot because of such cases of "overplotting".

```
plot(jitter(evals$score) ~ jitter(evals$bty_avg))
title(main = 'Instructor "Beauty rating" (x-axis) vs. rating of teaching quality (y-axis)')
title(sub = "with Jitter to reveal duplicate values")
```



Adding the "jitter" to the individual points makes it easier to observe those values which have multiple observations "on top of" each other.

This is especially evident at the point mentioned above (4.333,4.3) where 10 observations share this value.

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)
m_bty
```

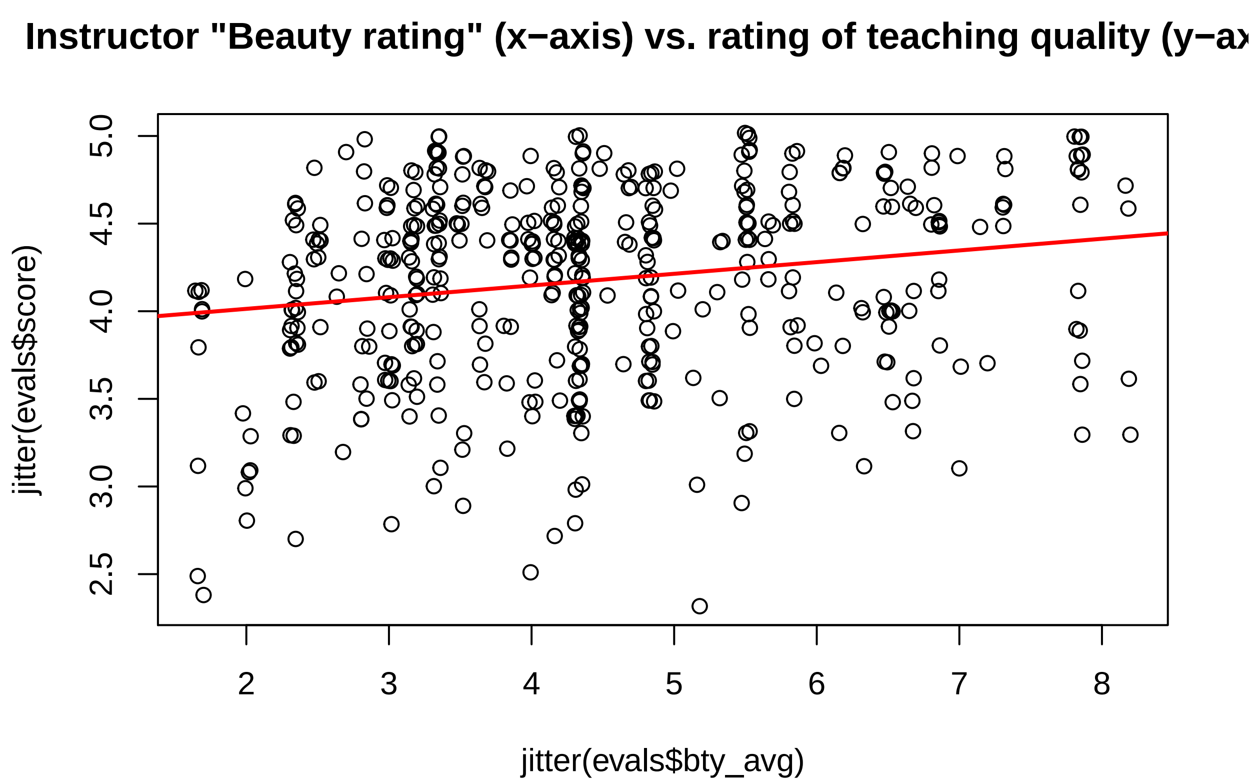
```
##
## Call:
## lm(formula = evals$score ~ evals$bty_avg)
##
## Coefficients:
```

```
##   (Intercept)  evals$bty_avg
##      3.880338       0.066637
```

```r
summary(m_bty)
```

```
##
## Call:
## lm(formula = evals$score ~ evals$bty_avg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.92465 -0.36903  0.14199  0.39769  0.93088
##
## Coefficients:
##                Estimate Std. Error t value              Pr(>|t|)
## (Intercept)    3.880338   0.076143 50.9612 < 0.00000000000000022 ***
## evals$bty_avg 0.066637   0.016291  4.0904            0.00005083 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.53484 on 461 degrees of freedom
## Multiple R-squared:  0.035022,   Adjusted R-squared:  0.032929
## F-statistic: 16.731 on 1 and 461 DF,  p-value: 0.000050827
```

```r
plot(jitter(evals$score) ~ jitter(evals$bty_avg))
abline(m_bty, col='red',lwd=2)
title(main = 'Instructor "Beauty rating" (x-axis) vs. rating of teaching quality (y-axis)')
```



Write out the equation for the linear model and interpret the slope.

$$\widehat{score} = 3.880338 + 0.066637 * bty\_avg$$

The slope value of .066637 indicates that as the average beauty rating increases by 1 point, the average teaching rating increases by .0666 , which is about 1/15 .

Is average beauty score a statistically significant predictor?

It is **statistically** significant, as the regression p-value (0.00005083) is close to zero.

Does it appear to be a practically significant predictor?

It does not appear to be **practically** significant because the slope is very small (0.0666) . Additionally, the R-squared is only 0.035, which indicates that the correlation is only .187 .

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).

To assess whether the linear model is reliable, we need to check for

(a) linearity,

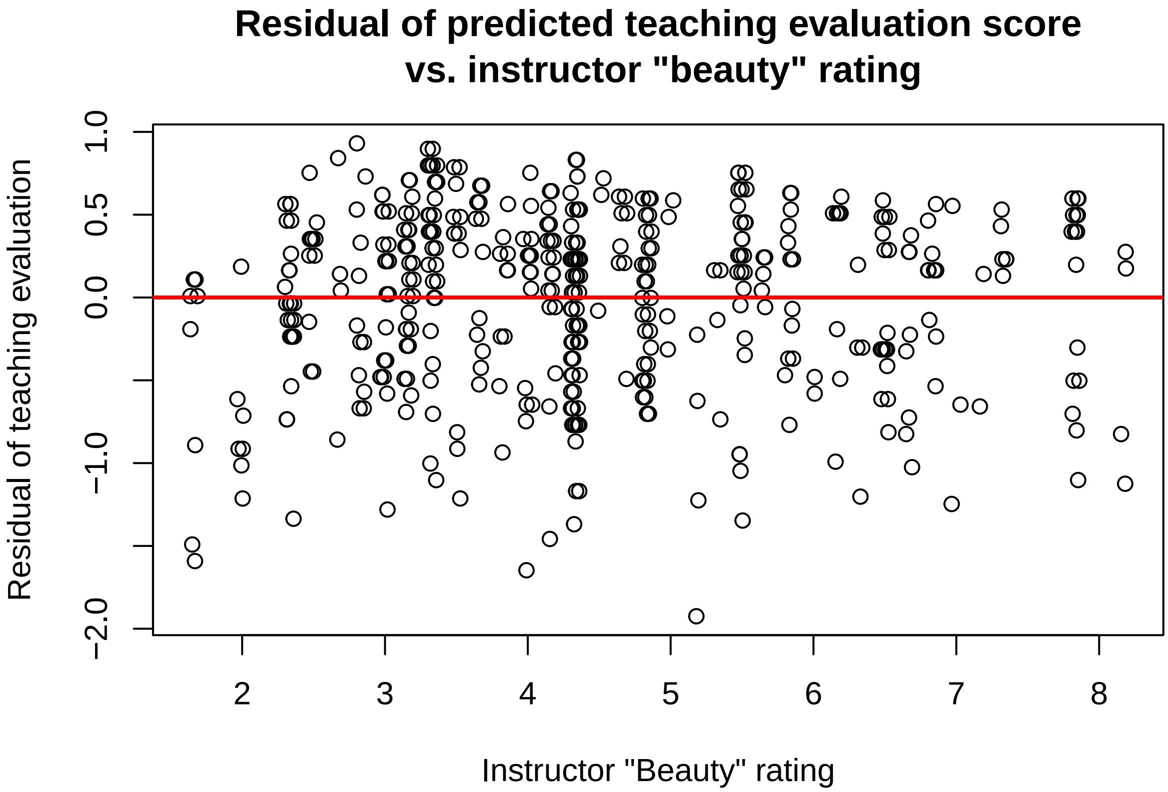
(b) nearly normal residuals, and

(c) constant variability.

***Linearity*:**

We already checked if the relationship between beauty rating and teaching evaluation is linear using a scatterplot. We can also verify this condition with a plot of the residuals of the teaching evaluation vs. the beauty rating:

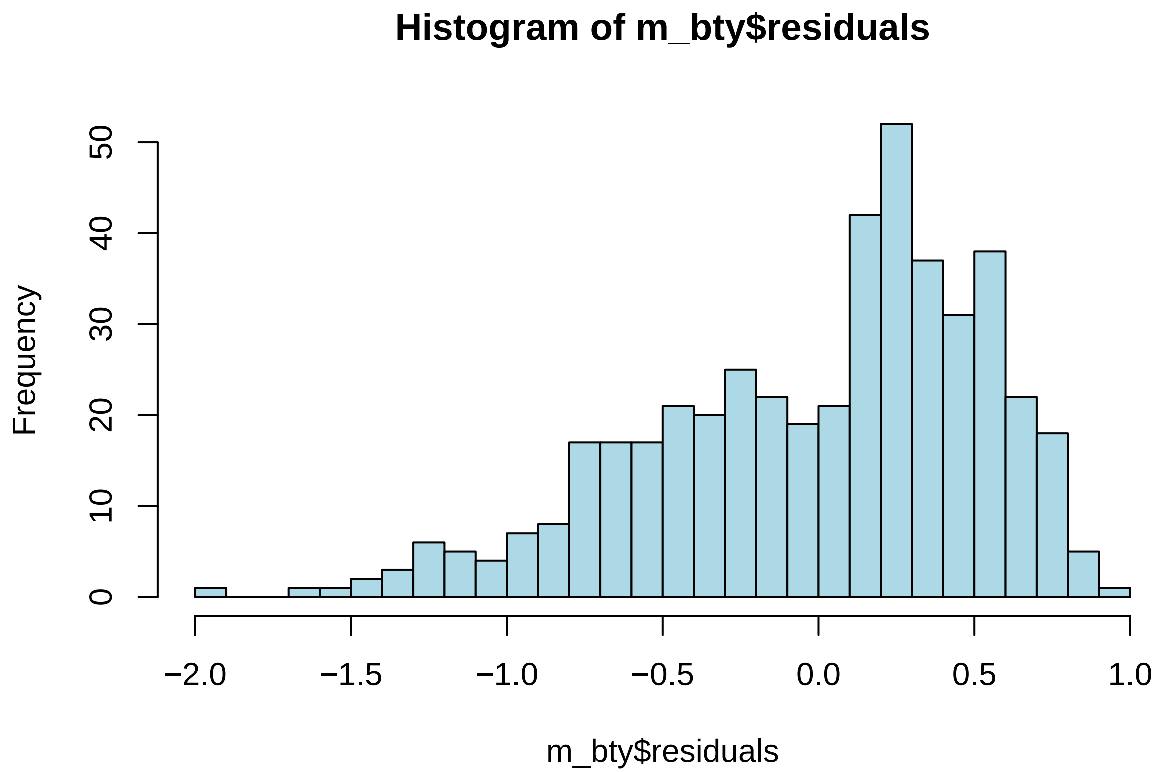
```
plot(m_bty$residuals ~ jitter(evals$bty_avg),xlab="",ylab="")
abline(h = 0, col="red", lwd=2)  # adds a horizontal dashed line at y = 0
title(main='Residual of predicted teaching evaluation score\n vs. instructor "beauty" rating')
title(xlab = 'Instructor "Beauty" rating')
title(ylab = 'Residual of teaching evaluation')
```



Nearly normal residuals:

To check this condition, we can look at a histogram:

```
hist(m_bty$residuals, breaks=30,col="lightblue")
```

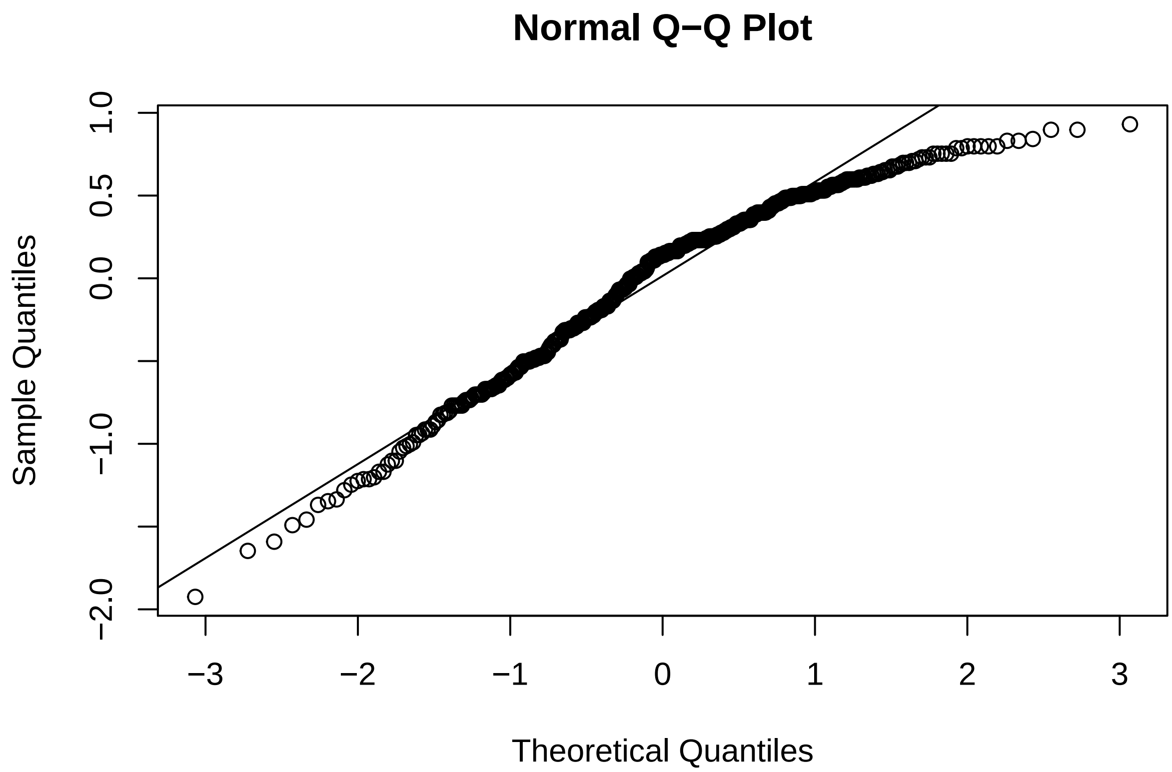



```r
summary(m_bty$residuals)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1.92465 -0.36903  0.14199  0.00000  0.39769  0.93088
```

or a normal probability plot of the residuals:

```r
qqnorm(m_bty$residuals)
qqline(m_bty$residuals)  # adds diagonal line to the normal prob plot
```



The skew reflected in the histogram and in the tails on the QQ plot appear to be so extreme as to be *inconsistent* with normality.

These results would suggest that the "Nearly-normal residuals" condition does not appear to be met.

However, it would be more conclusive to perform actual tests for normality, such as Shapiro-Wilks:

```
shapiro.test(m_bty$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  m_bty$residuals
## W = 0.954907, p-value = 0.00000000010892
```

Because the p-value is small, we *reject* the Null Hypothesis (the residuals *ARE* normal) in favor of the alternative (the residuals are *NOT* normal.)

Another useful test for normality is *Kolmogorov-Smirnov*.

Here we test whether the residuals are consistent with a Normal distribution which has mean=0 and standard deviation matching that of the residuals:

```
ks.test(m_bty$residuals,"pnorm",0,sd(m_bty$residuals))
```

```
## Warning in ks.test(m_bty$residuals, "pnorm", 0, sd(m_bty$residuals)): ties should not be present for
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  m_bty$residuals
## D = 0.11728, p-value = 0.0000058815
## alternative hypothesis: two-sided
```

Here as well, the small p-value indicates that we *reject* the null hypothesis (the residuals are normal) in favor of the alternative (the residuals are *NOT* normal.)

Another useful test of normality is *Anderson-Darling*:

```
require(nortest)
```

```
## Loading required package: nortest
```

```
ad.test(m_bty$residuals)
```

```
##
##  Anderson-Darling normality test
##
## data:  m_bty$residuals
## A = 6.23365, p-value = 0.0000000000000024884
```

Here again, the low p-value indicates that we *reject* the null hypothesis (the residuals are normal) in favor of the alternative (the residuals are *NOT* normal.)

Yet another useful test for normality is *Jarque-Bera*:

```
require(tseries)
```

```
## Loading required package: tseries
```

```
## Warning: package 'tseries' was built under R version 3.5.3
```

```
jarque.bera.test(m_bty$residuals)
```

```
##
##  Jarque Bera Test
##
## data:  m_bty$residuals
## X-squared = 38.7971, df = 2, p-value = 0.0000000037612
```

Here again, the low p-value indicates that we *reject* the null hypothesis (the residuals are normal) in favor of the alternative (the residuals are *NOT* normal.)

The skewed nature of the histogram and the QQ-plot, and the results of these numerical tests of the residuals, cause us to *reject* normality.

Constant variability:

A useful numeric test for constant variance is *Breusch-Pagan*. As it assumes that the data *are* normally distributed, the above tests need to have passed before we can use it.

As the above tests have all *failed*, we should *not* use this test, but will try it for fun:

```r
require(olsrr)
```

```
## Loading required package: olsrr
```

```
## Warning: package 'olsrr' was built under R version 3.5.3
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```r
ols_test_breusch_pagan(m_bty)
```

```
##
##  Breusch Pagan Test for Heteroskedasticity
##  -----------------------------------------
##  Ho: the variance is constant
##  Ha: the variance is not constant
##
##                 Data
##  ---------------------------------------
##  Response : evals$score
##  Variables: fitted values of evals$score
##
##         Test Summary
##  -----------------------------
##  DF            =    1
##  Chi2          =    0.28838718
##  Prob > Chi2   =    0.59125592
```

The high p-value indicates that we *fail to reject H0*, which is that the variance is constant.

However, the conditions to use this test are not met, because the earlier tests for normality failed.

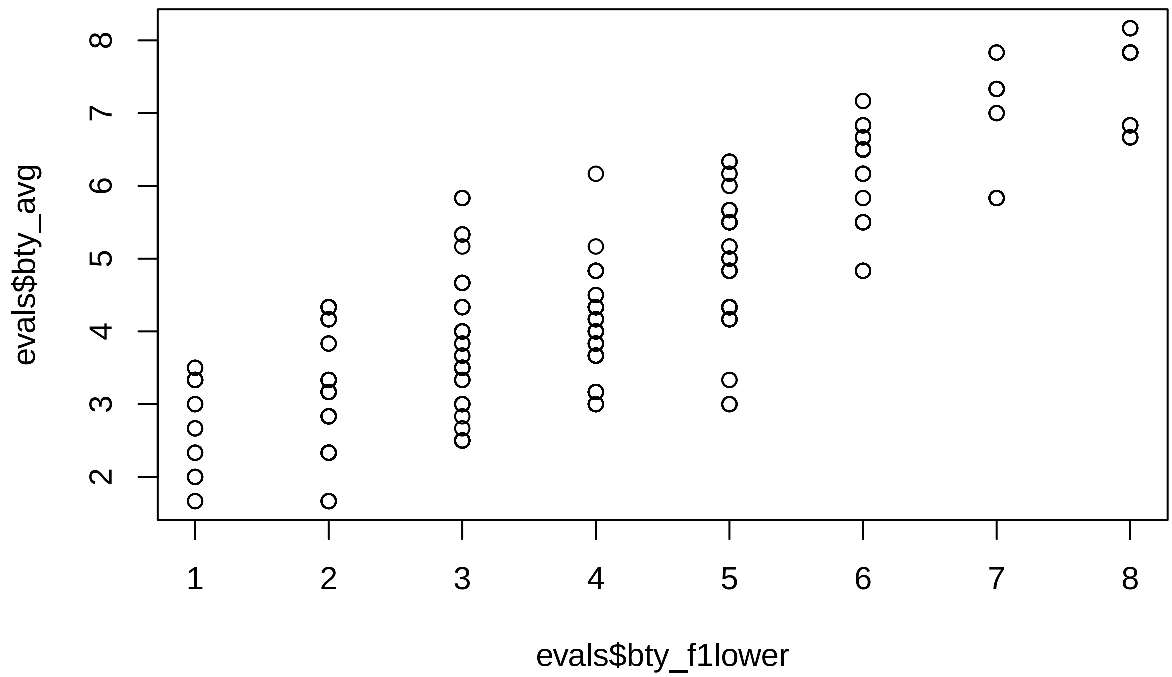
Therefore this test is not valid.

The results indicate that the conditions required for OLS regression (specifically, Normality of residuals) are not met.

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)
```

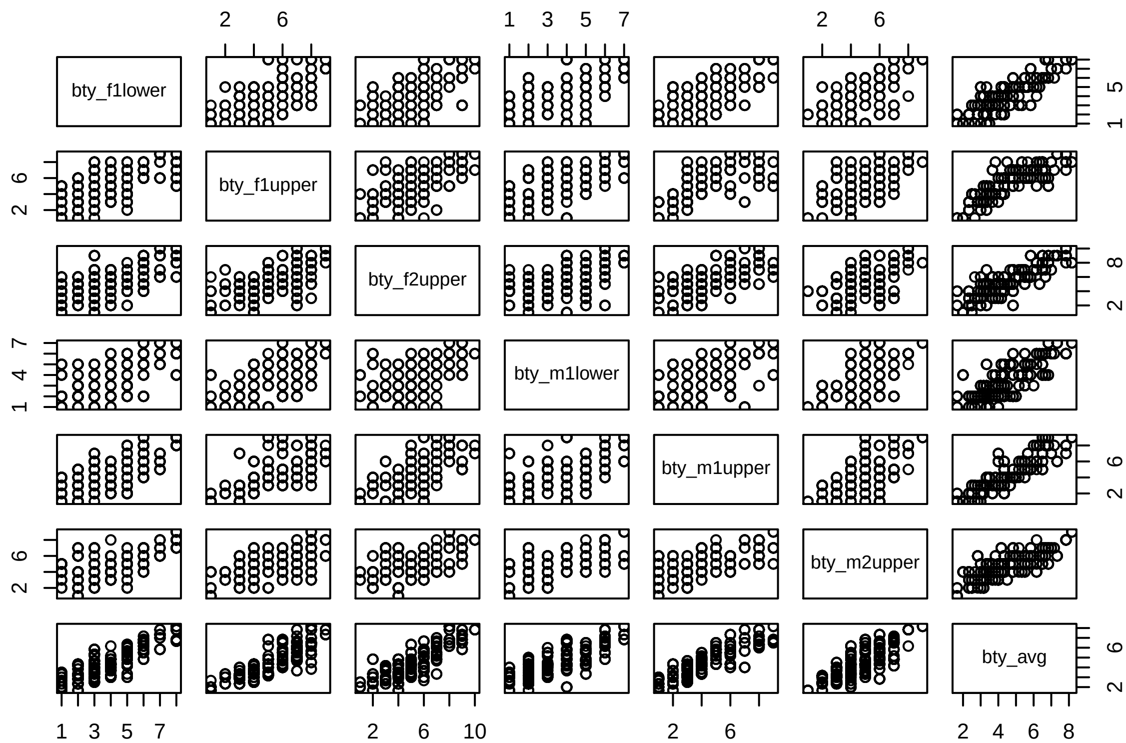



```
cor(evals$bty_avg, evals$bty_f1lower)
```

```
## [1] 0.84391117
```

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
## -1.83050 -0.36250  0.10550  0.42130  0.93135
##
## Coefficients:
##             Estimate Std. Error t value              Pr(>|t|)
## (Intercept) 3.747338   0.084655 44.2660 < 0.00000000000000022 ***
## bty_avg     0.074155   0.016252  4.5628           0.000006484 ***
## gendermale  0.172390   0.050221  3.4326             0.0006518 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.52869 on 460 degrees of freedom
```

```
## Multiple R-squared:  0.059123,   Adjusted R-squared:  0.055032
## F-statistic: 14.453 on 2 and 460 DF,  p-value: 0.00000081767
```

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

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

my_dd_m = data.frame(bty_avg=evals$bty_avg ,
                     score=predict(m_bty_gen,evals),
                     resids=m_bty_gen$residuals,
                     gender=evals$gender)

ggplot(evals) + geom_point(aes(x=(bty_avg),
                               y=(score),
                               colour=gender),
                           position = 'jitter') +
                geom_line(data=my_dd_m,
                          aes(x=bty_avg,
                              y=score,
                              colour=gender),
                          size=2.5,
                          alpha=0.3) +
                ggtitle("Predicted teaching score, by 'beauty' rating and gender")
```

To assess whether the linear model is reliable, we need to check for

(a) linearity,

(b) nearly normal residuals, and

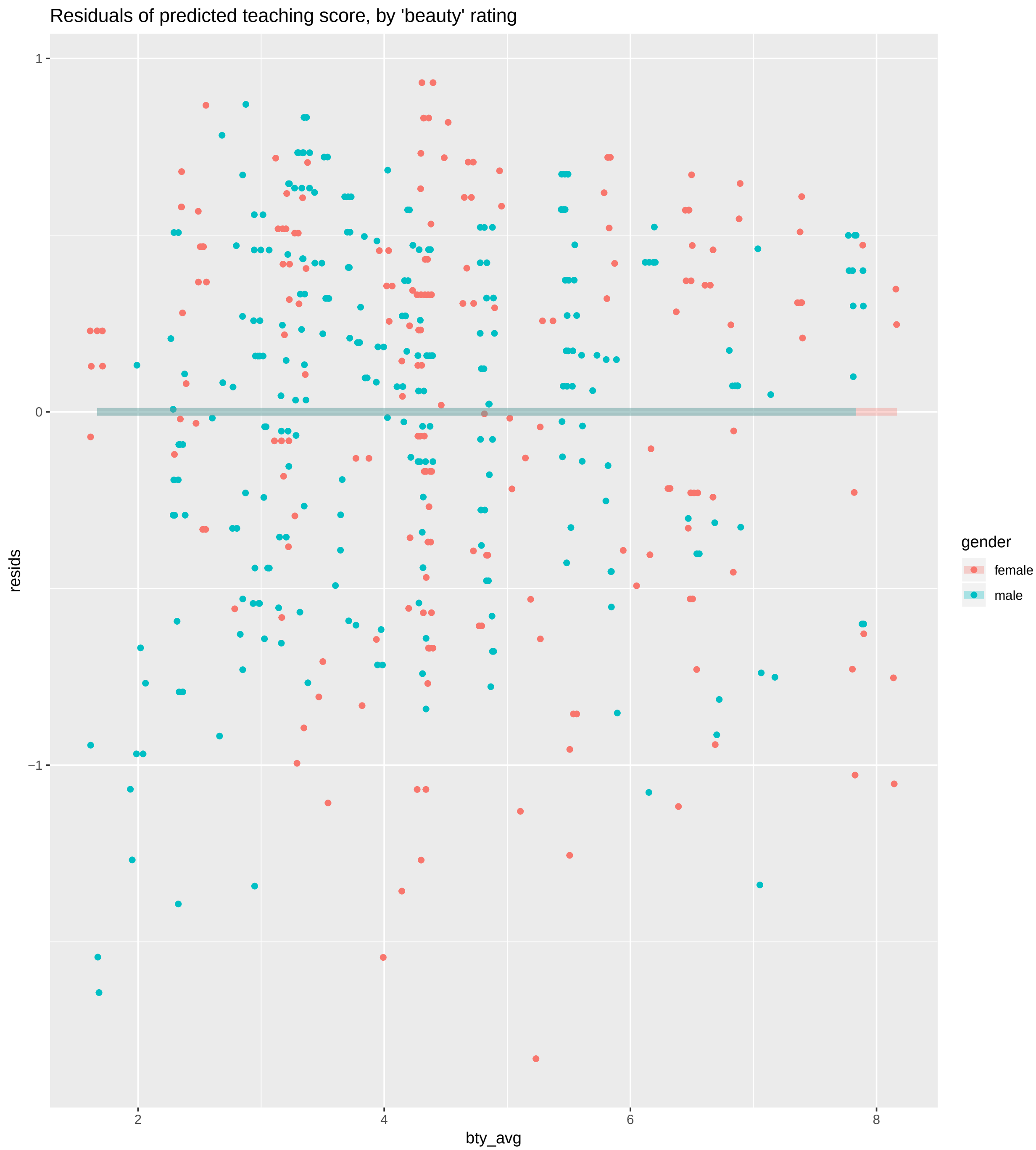
(c) constant variability.

***Linearity*:**

The above scatterplot indicates that the relationship between beauty rating and teaching evaluation, partitioned by gender, appears to be linear, as no other pattern is apparent.

We can also verify this condition with a plot of the residuals of the teaching evaluation vs. the beauty rating + gender:

```
ggplot(my_dd_m) + geom_point(aes(x=bty_avg,
                                  y=resids,
                                  colour=gender),
                              position = 'jitter') +
                  geom_line(data=my_dd_m,
                            aes(x=bty_avg,
                                y=0,
                                colour=gender),
                            size=2.5,
                            alpha=0.3) +
                  ggtitle("Residuals of predicted teaching score, by 'beauty' rating")
```



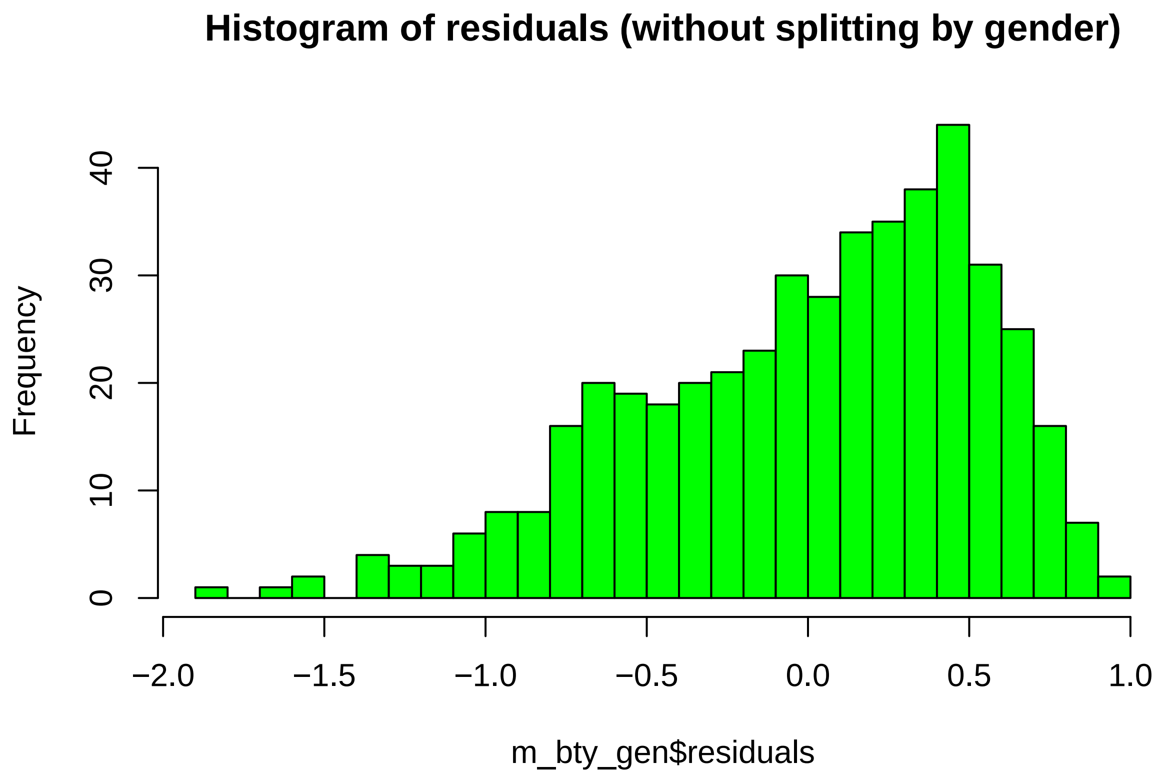

```
#plot(m_bty_gen$residuals ~ jitter(evals$bty_avg),xlab="",ylab="")
#abline(h = 0, col="red", lwd=2)  # adds a horizontal dashed line at y = 0
#title(main='Residual of predicted teaching evaluation score\n vs. instructor "beauty" rating')
#title(xlab = 'Instructor "Beauty" rating')
#title(ylab = 'Residual of teaching evaluation')
```

The plot of the residuals does not appear to reveal any evidence contrary to linearity.

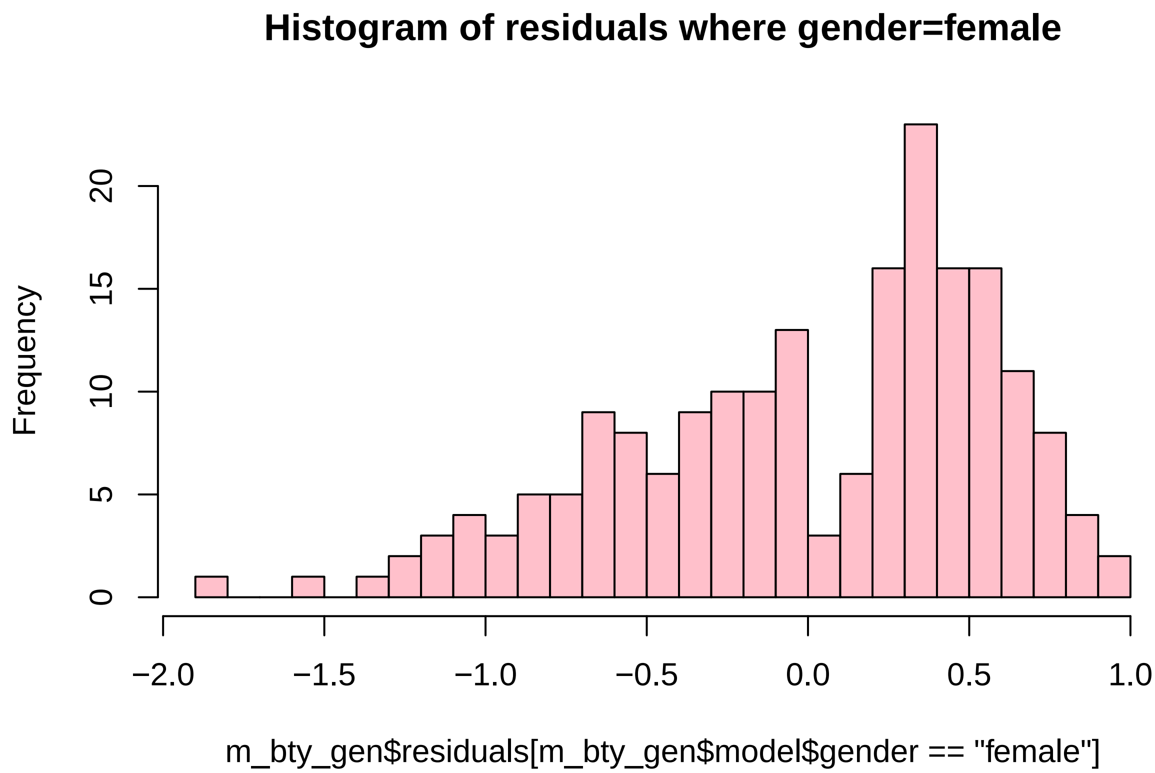
***Nearly normal residuals*:**

To check this condition, we can look at a histogram:

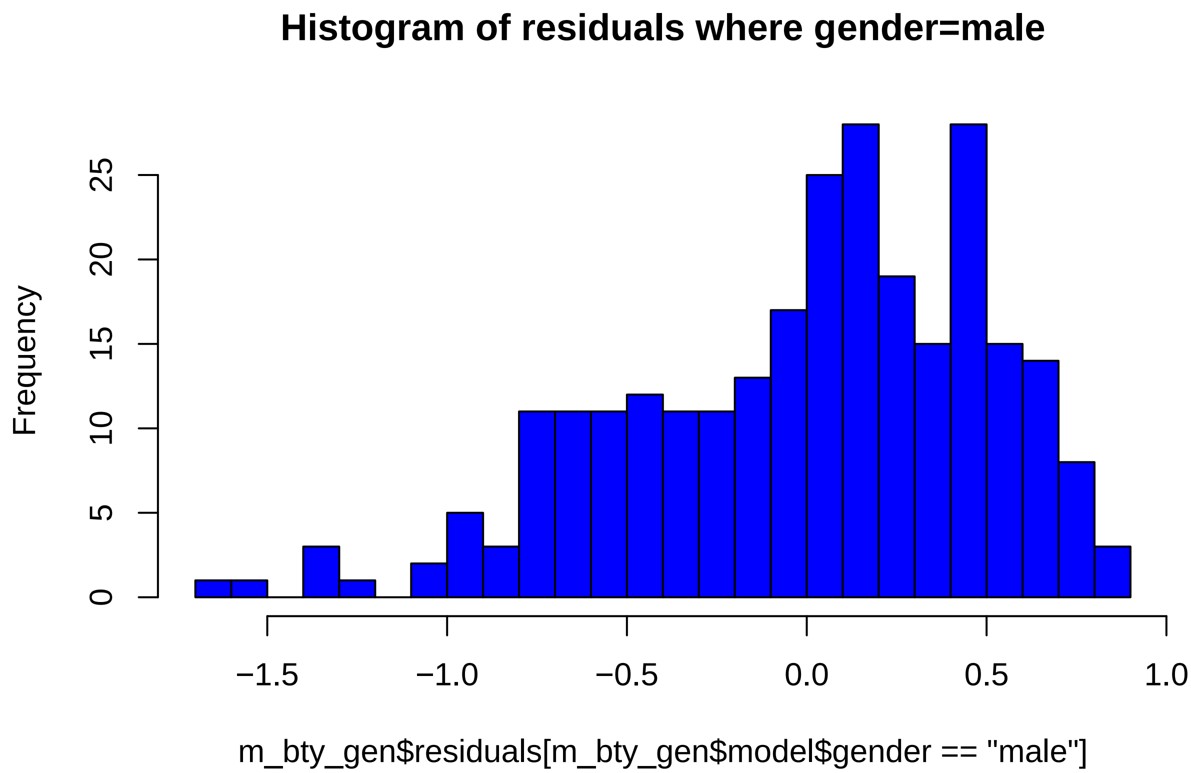
```
hist(m_bty_gen$residuals, breaks=30,col="green", main = "Histogram of residuals (without splitting by ge
```




```
hist(m_bty_gen$residuals[m_bty_gen$model$gender=="female"],breaks=30,col="pink", main = "Histogram of re
```




```
hist(m_bty_gen$residuals[m_bty_gen$model$gender=="male"],breaks=30,col="blue", main = "Histogram of resi
```




```r
summary(m_bty_gen$residuals)
```

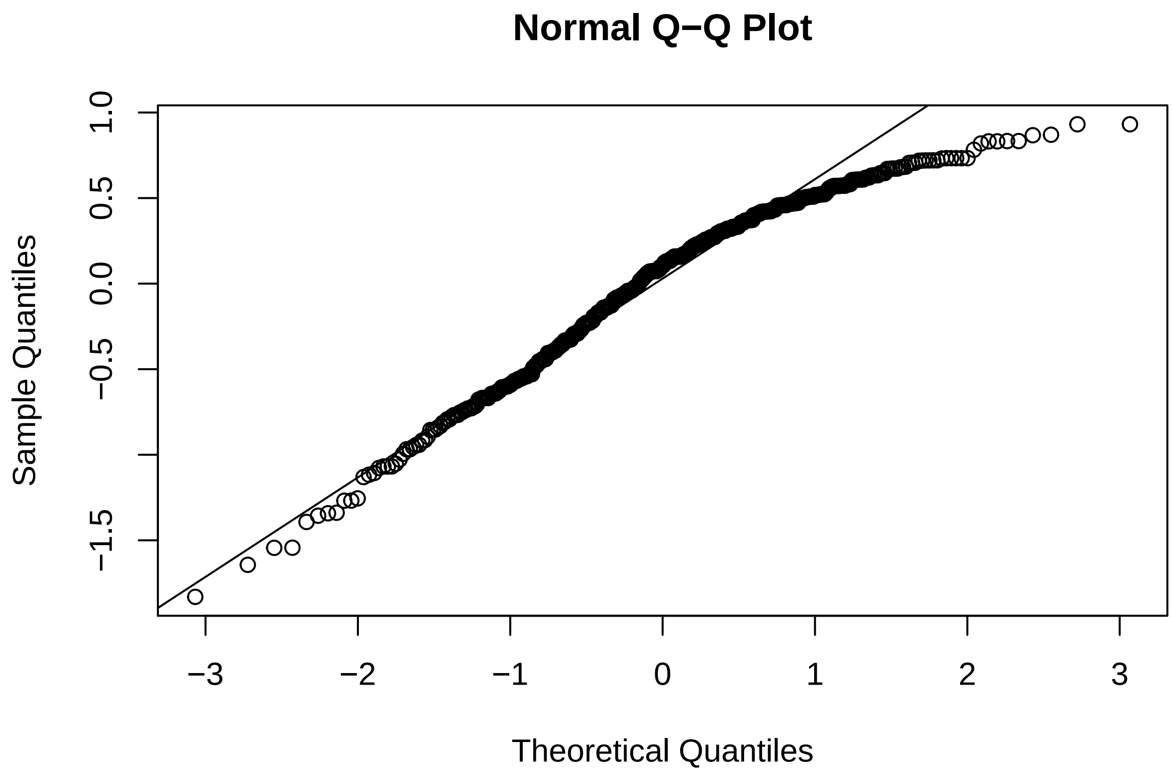
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1.83050 -0.36250  0.10550  0.00000  0.42130  0.93135
```

```r
by(data=m_bty_gen$residuals,INDICES = m_bty_gen$model$gender, FUN = summary )
```

```
## m_bty_gen$model$gender: female
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1.83050 -0.39285  0.13135  0.00000  0.45604  0.93135
## ---------------------------------------------------------------------------------------------
## m_bty_gen$model$gender: male
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -1.643345 -0.329810  0.083075  0.000000  0.408344  0.870190
```

or a normal probability plot of the residuals:

```r
qqnorm(m_bty_gen$residuals)
qqline(m_bty_gen$residuals)  # adds diagonal line to the normal prob plot
```



The histograms and the Q-Q plot suggest that the residuals are *NOT* normally distributed.

***Constant variability*:**

While the above plots suggest that the Normality condition is not satisfied, they do not appear to show heteroscedasticity, suggesting that the constant variability condition is OK.

Because of the apparent failure of the nearly-normal residuals requirement, the conditions for OLS do *not* appear to be satisfied.

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`?

Initially the parameter estimate for `bty_avg` was 0.066637 . Now it is 0.074155 .

The p-value is now 0.000006484 , which continues to indicate significance.

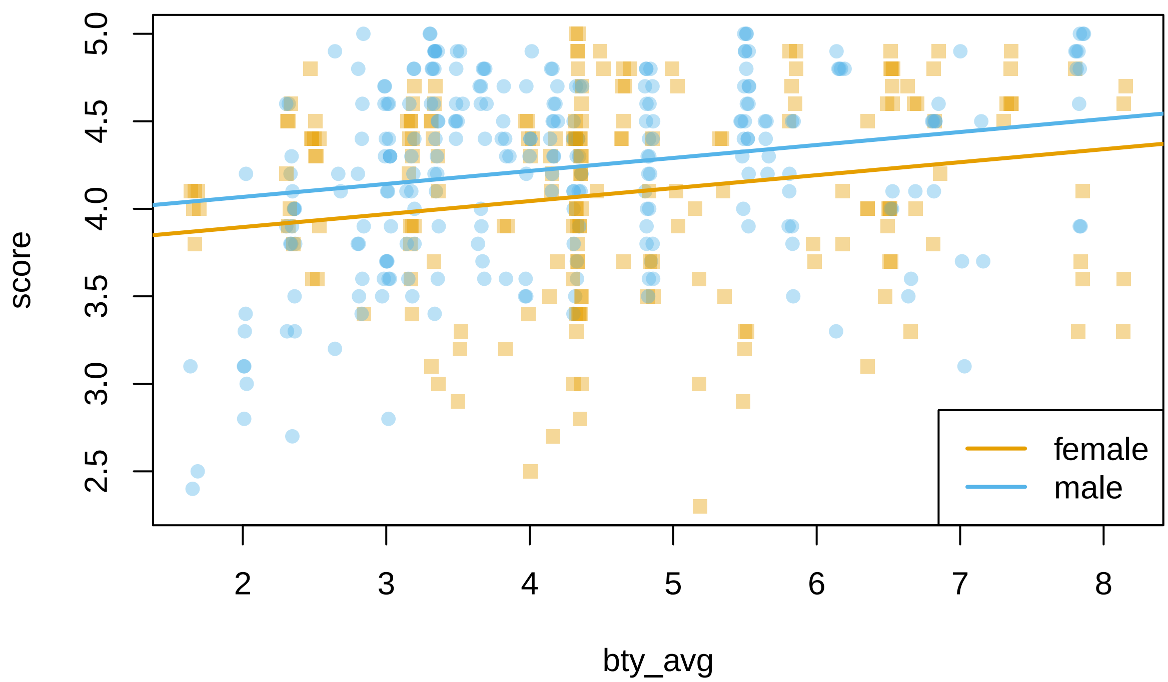
Note that the estimate for `gender` is now called `gendermale`. You'll see this name change whenever you introduce a categorical variable. The reason is that R recodes `gender` from having the values of `female` and `male` to being an indicator variable called `gendermale` that takes a value of 0 for 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?

What is the equation of the line corresponding to males?

$$\begin{aligned}\widehat{score}_{male} &= \hat{\beta}_0 + \hat{\beta}_1 \times bty\_avg + \hat{\beta}_2 \times (1) \\ &= 3.74733824 + 0.07415537 \times bty\_avg + 0.17238955 \times (1) \\ &= 3.91972779 + 0.07415537 \times bty\_avg\end{aligned}$$

Checking the results by switching the default level (male vs. female):

```
m_bty_gen_m <- lm(score ~ bty_avg + (gender=="female"), data = evals)
m_bty_gen_m
```

```
##
## Call:
## lm(formula = score ~ bty_avg + (gender == "female"), data = evals)
##
## Coefficients:
##           (Intercept)                 bty_avg  gender == "female"TRUE
##              3.919728                0.074155               -0.172390
```

```r
summary(m_bty_gen_m)
```

```
##
## Call:
## lm(formula = score ~ bty_avg + (gender == "female"), data = evals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83050 -0.36250  0.10550  0.42130  0.93135
##
## Coefficients:
##                           Estimate Std. Error t value              Pr(>|t|)
## (Intercept)               3.919728   0.076138 51.4822 < 0.00000000000000022 ***
## bty_avg                   0.074155   0.016252  4.5628           0.000006484 ***
## gender == "female"TRUE -0.172390   0.050221 -3.4326             0.0006518 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.52869 on 460 degrees of freedom
## Multiple R-squared:  0.059123,   Adjusted R-squared:  0.055032
## F-statistic: 14.453 on 2 and 460 DF,  p-value: 0.00000081767
```

For two professors who received the same beauty rating, which gender tends to have the higher course evaluation score?

The male instructors, on average, receive a teaching evaluation which is higher by 0.17238955 , given the same "beauty" rating.

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
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
## -1.87130 -0.36418  0.14889  0.41035  0.95253
##
```

```
## Coefficients:
##                   Estimate Std. Error t value              Pr(>|t|)
## (Intercept)       3.981546   0.090779 43.8599 < 0.00000000000000022 ***
## bty_avg           0.067826   0.016550  4.0983            0.00004921 ***
## ranktenure track -0.160702   0.073951 -2.1731               0.03028 *
## ranktenured      -0.126227   0.062662 -2.0144               0.04455 *
## ---
## 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:  0.046519,   Adjusted R-squared:  0.040287
## F-statistic: 7.4647 on 3 and 459 DF,  p-value: 0.000068803
```

How does R appear to handle categorical variables that have more than two levels?

R creates two dummy (indicator) variables for "rank":

"ranktenure track" and "ranktenured",

with rank="teaching" as the base level (which is represented by setting both of the above dummies to zero).

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.

`cls_profs`: number of professors teaching sections in course in sample: single, multiple.

This variable indicates whether a course has one instructor or multiple instructors. As the evaluation of the teaching score rests with each individual instructor, this variable should not have any association with the teaching score.

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

```
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
## -1.773971 -0.324325  0.090673  0.351834  0.950357
##
## Coefficients:
##                          Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)            4.09521408  0.29052766 14.0958 < 0.0000000000000002 ***
## ranktenure track      -0.14759325  0.08206709 -1.7984            0.0727793 .
## ranktenured           -0.09733776  0.06632958 -1.4675            0.1429455
## ethnicitynot minority  0.12349292  0.07862732  1.5706            0.1169791
## gendermale             0.21094813  0.05182296  4.0706           0.00005544 ***
## languagenon-english   -0.22981119  0.11137542 -2.0634            0.0396509 *
## age                   -0.00900719  0.00313591 -2.8723            0.0042688 **
## cls_perc_eval          0.00532724  0.00153932  3.4608            0.0005903 ***
## cls_students           0.00045463  0.00037739  1.2047            0.2289607
## cls_levelupper         0.06051396  0.05756166  1.0513            0.2936925
## cls_profssingle       -0.01466192  0.05198850 -0.2820            0.7780566
## cls_creditsone credit  0.50204318  0.11593877  4.3302           0.00001839 ***
## bty_avg                0.04003330  0.01750642  2.2868            0.0226744 *
## pic_outfitnot formal  -0.11268169  0.07388004 -1.5252            0.1279153
## pic_colorcolor        -0.21726300  0.07150214 -3.0386            0.0025162 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.49795 on 448 degrees of freedom
## Multiple R-squared:  0.18711,    Adjusted R-squared:  0.16171
## F-statistic: 7.3657 on 14 and 448 DF,  p-value: 0.000000000000065525
```

The cls_profs(single) has a p-value of 0.7780566, which is indeed the highest value across all the p-values.

13. Interpret the coefficient associated with the ethnicity variable.

The coefficient for `ethnicitynot minority` is 0.12349292 . This means that *Non-minority* instructors are expected to receive an evaluation 0.1235 points higher than an equivalent minority instructor, where all other variables are unchanged.

However, this variable has a high p-value of 0.1169791 , which indicates that it is not statistically significant under the present model (incorporating all the above variables.)

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?

Drop the variable with the highest p-value and re-fit the model.

```r
m_2 <- 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_2)
```

```
##
## 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
## -1.783645 -0.325748  0.085899  0.351316  0.955121
##
## Coefficients:
##                         Estimate  Std. Error t value              Pr(>|t|)
## (Intercept)           4.08725232  0.28885625 14.1498 < 0.0000000000000022 ***
## ranktenure track     -0.14767458  0.08198242 -1.8013             0.0723271 .
## ranktenured          -0.09738288  0.06626137 -1.4697             0.1423493
## ethnicitynot minority  0.12744576  0.07728865  1.6490             0.0998556 .
## gendermale            0.21012314  0.05168727  4.0653            0.00005665 ***
## languagenon-english  -0.22828945  0.11113055 -2.0542             0.0405303 *
## age                  -0.00899919  0.00313257 -2.8728             0.0042616 **
## cls_perc_eval         0.00528876  0.00153169  3.4529             0.0006072 ***
## cls_students          0.00046872  0.00037369  1.2543             0.2103843
## cls_levelupper        0.06063743  0.05750097  1.0545             0.2922000
## cls_creditsone credit  0.50611955  0.11491627  4.4042            0.00001329 ***
## bty_avg               0.03986289  0.01747804  2.2807             0.0230315 *
## pic_outfitnot formal -0.10832274  0.07217113 -1.5009             0.1340803
## pic_colorcolor       -0.21905269  0.07114694 -3.0789             0.0022052 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.49744 on 449 degrees of freedom
## Multiple R-squared:  0.18697,    Adjusted R-squared:  0.16343
## F-statistic: 7.9425 on 13 and 449 DF,  p-value: 0.000000000000023359
```

Did the coefficients and significance of the other explanatory variables change?

There are very slight adjustments to the coefficients when `cls_profs` is dropped from the model.

As for significance, the only noteworthy change is the following:

Full Model:

ethnicitynot minority 0.12349292 0.07862732 1.5706 0.1169791

Model without cls_profs:

ethnicitynot minority 0.12744576 0.07728865 1.6490 0.0998556 .

When cls_profs is dropped from the model, the p-value for the ethnicity variable is reduced to a level where it (narrowly!) becomes significant at the 0.10 level.

This is the only variable which demonstrates such a change in significance.

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.

To automate this task, I'll use the "ols_step_backward_p" function from the "olsrr" package:

```r
require(olsrr)
# perform stepwise backward selection, eliminating all variables with p-values greater than 0.10
ols_step_backward_p(model = m_full, details=T, prem = .10)
```

```
## Backward Elimination Method
## ---------------------------
##
## Candidate Terms:
##
## 1 . rank
## 2 . ethnicity
## 3 . gender
## 4 . language
## 5 . age
## 6 . cls_perc_eval
## 7 . cls_students
## 8 . cls_level
## 9 . cls_profs
## 10 . cls_credits
## 11 . bty_avg
## 12 . pic_outfit
## 13 . pic_color
##
## We are eliminating variables based on p value...
##
## - cls_profs
##
## Backward Elimination: Step 1
##
##  Variable cls_profs Removed
##
##                         Model Summary
## --------------------------------------------------------------
## R                       0.432       RMSE               0.497
## R-Squared               0.187       Coef. Var         11.916
## Adj. R-Squared          0.163       MSE                0.247
## Pred R-Squared          0.139       MAE                0.397
## --------------------------------------------------------------
```

```
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
##                 Sum of
##                Squares         DF    Mean Square      F        Sig.
## --------------------------------------------------------------------
## Regression      25.550         13          1.965    7.943    0.0000
## Residual       111.105        449          0.247
## Total          136.654        462
## --------------------------------------------------------------------
##
##                                         Parameter Estimates
## ----------------------------------------------------------------------------------------------------
##                   model      Beta    Std. Error    Std. Beta       t        Sig      lower     upper
## ----------------------------------------------------------------------------------------------------
##             (Intercept)     4.087         0.289                 14.150    0.000      3.520     4.655
##      ranktenure track    -0.148         0.082        -0.115    -1.801    0.072     -0.309     0.013
##           ranktenured    -0.097         0.066        -0.089    -1.470    0.142     -0.228     0.033
## ethnicitynot minority     0.127         0.077         0.081     1.649    0.100     -0.024     0.279
##            gendermale     0.210         0.052         0.191     4.065    0.000      0.109     0.312
##   languagenon-english    -0.228         0.111        -0.100    -2.054    0.041     -0.447    -0.010
##                    age    -0.009         0.003        -0.162    -2.873    0.004     -0.015    -0.003
##         cls_perc_eval     0.005         0.002         0.163     3.453    0.001      0.002     0.008
##          cls_students     0.000         0.000         0.065     1.254    0.210      0.000     0.001
##        cls_levelupper     0.061         0.058         0.053     1.055    0.292     -0.052     0.174
## cls_creditsone credit     0.506         0.115         0.218     4.404    0.000      0.280     0.732
##               bty_avg     0.040         0.017         0.112     2.281    0.023      0.006     0.074
##  pic_outfitnot formal    -0.108         0.072        -0.074    -1.501    0.134     -0.250     0.034
##        pic_colorcolor    -0.219         0.071        -0.151    -3.079    0.002     -0.359    -0.079
## ----------------------------------------------------------------------------------------------------
##
##
## - cls_level
##
## Backward Elimination: Step 2
##
##  Variable cls_level Removed
##
##                         Model Summary
## --------------------------------------------------------------
## R                       0.430       RMSE                0.498
## R-Squared               0.185       Coef. Var          11.917
## Adj. R-Squared          0.163       MSE                 0.248
## Pred R-Squared          0.140       MAE                 0.397
## --------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## --------------------------------------------------------------------
```

```
##               Sum of
##              Squares         DF    Mean Square      F        Sig.
## --------------------------------------------------------------------
## Regression     25.275         12          2.106     8.51    0.0000
## Residual      111.380        450          0.248
## Total         136.654        462
## --------------------------------------------------------------------
##
##                                             Parameter Estimates
## ----------------------------------------------------------------------------------------------------
##                    model      Beta    Std. Error    Std. Beta        t        Sig      lower     upper
## ----------------------------------------------------------------------------------------------------
##              (Intercept)     4.086         0.289                 14.143     0.000      3.518     4.653
##       ranktenure track    -0.142         0.082       -0.111     -1.736     0.083     -0.303     0.019
##             ranktenured    -0.090         0.066       -0.082     -1.360     0.174     -0.219     0.040
## ethnicitynot minority     0.142         0.076        0.090      1.875     0.061     -0.007     0.292
##              gendermale     0.204         0.051        0.185      3.969     0.000      0.103     0.305
##    languagenon-english    -0.209         0.110       -0.092     -1.908     0.057     -0.425     0.006
##                     age    -0.009         0.003       -0.157     -2.795     0.005     -0.015    -0.003
##          cls_perc_eval     0.005         0.002        0.165      3.498     0.001      0.002     0.008
##            cls_students     0.000         0.000        0.049      0.997     0.319      0.000     0.001
## cls_creditsone credit     0.473         0.111        0.204      4.278     0.000      0.256     0.691
##                 bty_avg     0.041         0.017        0.115      2.352     0.019      0.007     0.075
##   pic_outfitnot formal    -0.117         0.072       -0.080     -1.635     0.103     -0.258     0.024
##         pic_colorcolor    -0.197         0.068       -0.136     -2.897     0.004     -0.331    -0.063
## ----------------------------------------------------------------------------------------------------
##
##
## - cls_students
##
## Backward Elimination: Step 3
##
##  Variable cls_students Removed
##
##                         Model Summary
## --------------------------------------------------------------
## R                       0.428       RMSE                0.498
## R-Squared               0.183       Coef. Var          11.917
## Adj. R-Squared          0.163       MSE                 0.248
## Pred R-Squared          0.142       MAE                 0.398
## --------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                               ANOVA
## --------------------------------------------------------------------
##               Sum of
##              Squares         DF    Mean Square      F        Sig.
## --------------------------------------------------------------------
## Regression     25.029         11          2.275    9.193    0.0000
## Residual      111.626        451          0.248
## Total         136.654        462
## --------------------------------------------------------------------
```

```
##
##                                       Parameter Estimates
## ---------------------------------------------------------------------------------------------------
##                model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
## ---------------------------------------------------------------------------------------------------
##          (Intercept)     4.153         0.281                 14.785    0.000     3.601     4.705
##     ranktenure track    -0.142         0.082       -0.111    -1.738    0.083    -0.303     0.019
##          ranktenured    -0.083         0.066       -0.076    -1.268    0.205    -0.212     0.046
## ethnicitynot minority     0.144         0.076        0.091     1.889    0.060    -0.006     0.293
##           gendermale     0.208         0.051        0.189     4.067    0.000     0.108     0.309
##   languagenon-english    -0.223         0.109       -0.098    -2.044    0.042    -0.436    -0.009
##                  age    -0.009         0.003       -0.164    -2.924    0.004    -0.015    -0.003
##        cls_perc_eval     0.005         0.001        0.149     3.359    0.001     0.002     0.008
## cls_creditsone credit     0.473         0.111        0.204     4.272    0.000     0.255     0.690
##              bty_avg     0.044         0.017        0.122     2.525    0.012     0.010     0.077
##  pic_outfitnot formal    -0.137         0.069       -0.094    -1.980    0.048    -0.272    -0.001
##       pic_colorcolor    -0.190         0.068       -0.131    -2.805    0.005    -0.323    -0.057
## ---------------------------------------------------------------------------------------------------
##
##
## - rank
##
## Backward Elimination: Step 4
##
##  Variable rank Removed
##
##                        Model Summary
## --------------------------------------------------------------
## R                       0.421       RMSE               0.498
## R-Squared               0.177       Coef. Var         11.933
## Adj. R-Squared          0.161       MSE                0.248
## Pred R-Squared          0.143       MAE                0.398
## --------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                               ANOVA
## --------------------------------------------------------------------
##              Sum of
##             Squares         DF    Mean Square      F         Sig.
## --------------------------------------------------------------------
## Regression    24.239          9          2.693    10.853    0.0000
## Residual     112.415        453          0.248
## Total        136.654        462
## --------------------------------------------------------------------
##
##                                       Parameter Estimates
## ---------------------------------------------------------------------------------------------------
##                model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
## ---------------------------------------------------------------------------------------------------
##          (Intercept)     3.907         0.245                 15.954    0.000     3.426     4.388
## ethnicitynot minority     0.164         0.075        0.104     2.180    0.030     0.016     0.312
##           gendermale     0.203         0.050        0.184     4.044    0.000     0.104     0.301
```

```
##   languagenon-english    -0.247         0.106       -0.108    -2.324    0.021    -0.455    -0.038
##                   age    -0.007         0.003       -0.125    -2.606    0.009    -0.012    -0.002
##         cls_perc_eval     0.005         0.001        0.152     3.427    0.001     0.002     0.008
## cls_creditsone credit     0.517         0.104        0.223     4.966    0.000     0.313     0.722
##               bty_avg     0.047         0.017        0.131     2.734    0.006     0.013     0.080
##  pic_outfitnot formal    -0.114         0.067       -0.078    -1.696    0.091    -0.246     0.018
##        pic_colorcolor    -0.181         0.067       -0.125    -2.681    0.008    -0.313    -0.048
## ----------------------------------------------------------------------------------------------------
##
##
##
## No more variables satisfy the condition of p value = 0.1
##
##
## Variables Removed:
##
## - cls_profs
## - cls_level
## - cls_students
## - rank
##
##
## Final Model Output
## ------------------
##
##                         Model Summary
## ---------------------------------------------------------------
## R                       0.421       RMSE                0.498
## R-Squared               0.177       Coef. Var          11.933
## Adj. R-Squared          0.161       MSE                 0.248
## Pred R-Squared          0.143       MAE                 0.398
## ---------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                                ANOVA
## ---------------------------------------------------------------------
##               Sum of
##              Squares        DF    Mean Square      F         Sig.
## ---------------------------------------------------------------------
## Regression    24.239         9          2.693    10.853    0.0000
## Residual     112.415       453          0.248
## Total        136.654       462
## ---------------------------------------------------------------------
##
##                                        Parameter Estimates
## ----------------------------------------------------------------------------------------------------
##                 model      Beta    Std. Error    Std. Beta        t        Sig      lower     upper
## ----------------------------------------------------------------------------------------------------
##           (Intercept)     3.907         0.245                 15.954    0.000     3.426     4.388
## ethnicitynot minority     0.164         0.075        0.104     2.180    0.030     0.016     0.312
##            gendermale     0.203         0.050        0.184     4.044    0.000     0.104     0.301
##   languagenon-english    -0.247         0.106       -0.108    -2.324    0.021    -0.455    -0.038
```

```
##                   age    -0.007          0.003       -0.125     -2.606    0.009     -0.012    -0.002
##         cls_perc_eval     0.005          0.001        0.152      3.427    0.001      0.002     0.008
## cls_creditsone credit     0.517          0.104        0.223      4.966    0.000      0.313     0.722
##               bty_avg     0.047          0.017        0.131      2.734    0.006      0.013     0.080
##  pic_outfitnot formal    -0.114          0.067       -0.078     -1.696    0.091     -0.246     0.018
##        pic_colorcolor    -0.181          0.067       -0.125     -2.681    0.008     -0.313    -0.048
## ----------------------------------------------------------------------------------------------------
##
##
##                              Elimination Summary
## -----------------------------------------------------------------------------
##         Variable                      Adj.
## Step      Removed       R-Square    R-Square     C(p)        AIC        RMSE
## -----------------------------------------------------------------------------
##    1    cls_profs          0.187      0.1634    11.0795    683.1181    0.4974
##    2    cls_level          0.185      0.1632    10.1893    682.2634    0.4975
##    3    cls_students      0.1832      0.1632     9.1809    681.2844    0.4975
##    4    rank              0.1774       0.161    10.3638    680.5463    0.4982
## -----------------------------------------------------------------------------
```

The result of the above is that 4 variables have been dropped: `cls_profs`, `cls_level`, `cls_students`, and `rank`.

The "final model" is the following:

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

```
##
## Call:
## lm(formula = score ~ ethnicity + gender + language + age + cls_perc_eval +
##     cls_credits + bty_avg + pic_outfit + pic_color, data = evals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.84547 -0.32206  0.10128  0.37448  0.90511
##
## Coefficients:
##                        Estimate Std. Error t value             Pr(>|t|)
## (Intercept)           3.9070305  0.2448894 15.9543 < 0.0000000000000022 ***
## ethnicitynot minority  0.1638182  0.0751583  2.1796            0.0297983 *
## gendermale             0.2025970  0.0501022  4.0437         0.0000618401 ***
## languagenon-english   -0.2466834  0.1061463 -2.3240            0.0205673 *
## age                   -0.0069246  0.0026577 -2.6055            0.0094749 **
## cls_perc_eval          0.0049425  0.0014421  3.4272            0.0006655 ***
## cls_creditsone credit  0.5172051  0.1041413  4.9664         0.0000009681 ***
## bty_avg                0.0467322  0.0170914  2.7343            0.0064972 **
## pic_outfitnot formal  -0.1139392  0.0671680 -1.6963            0.0905102 .
## pic_colorcolor        -0.1808705  0.0674557 -2.6813            0.0076009 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.49815 on 453 degrees of freedom
## Multiple R-squared:  0.17738,    Adjusted R-squared:  0.16103
## F-statistic: 10.853 on 9 and 453 DF,  p-value: 0.0000000000000024411
```

All of the remaining variables are significant at the 0.10 level. (If this were decreased, say to 0.05, then the next variable to be dropped would be `pic_outfit(not formal)` .)

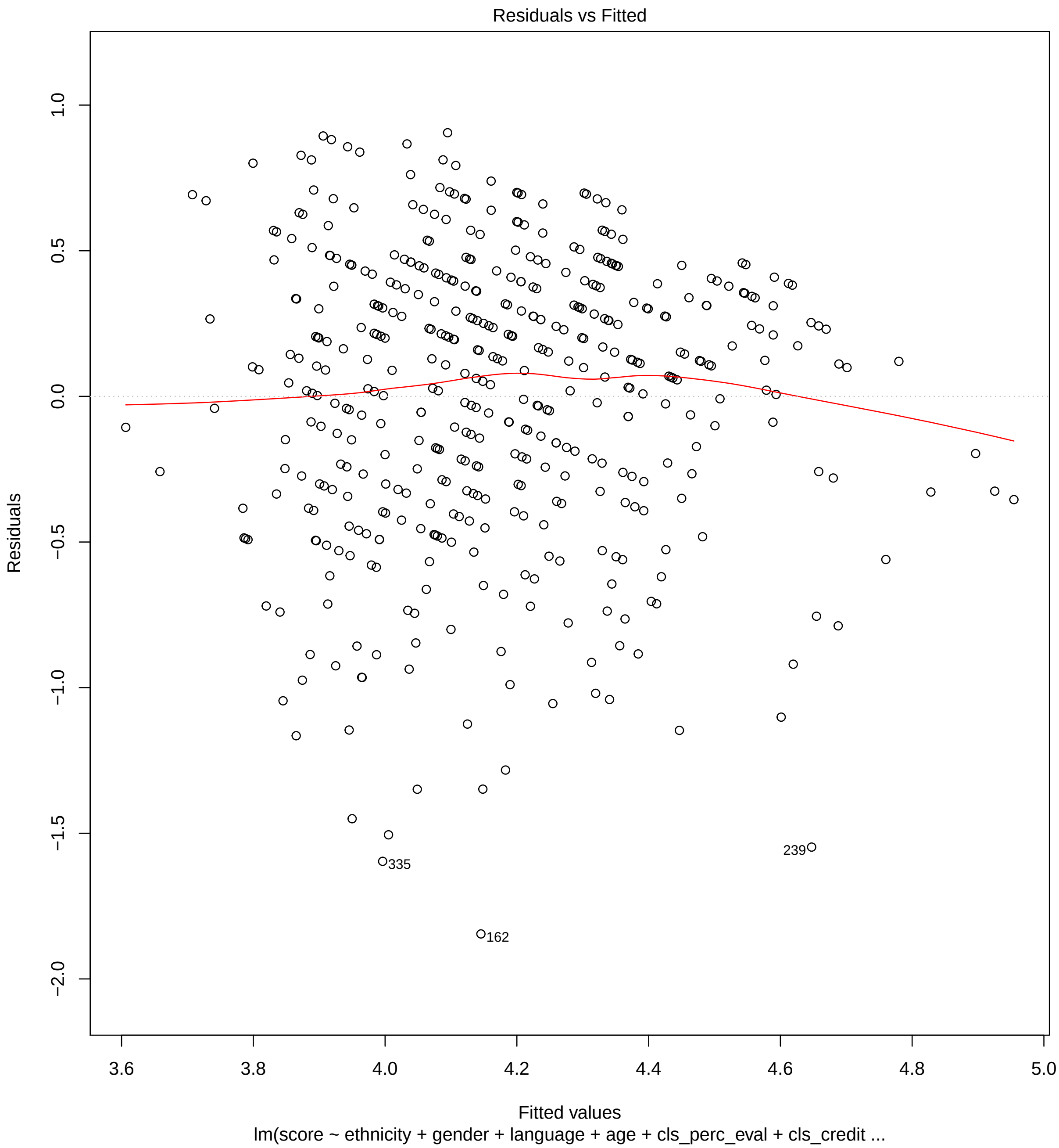
write out the linear model for predicting score based on the final model you settle on

The model is:

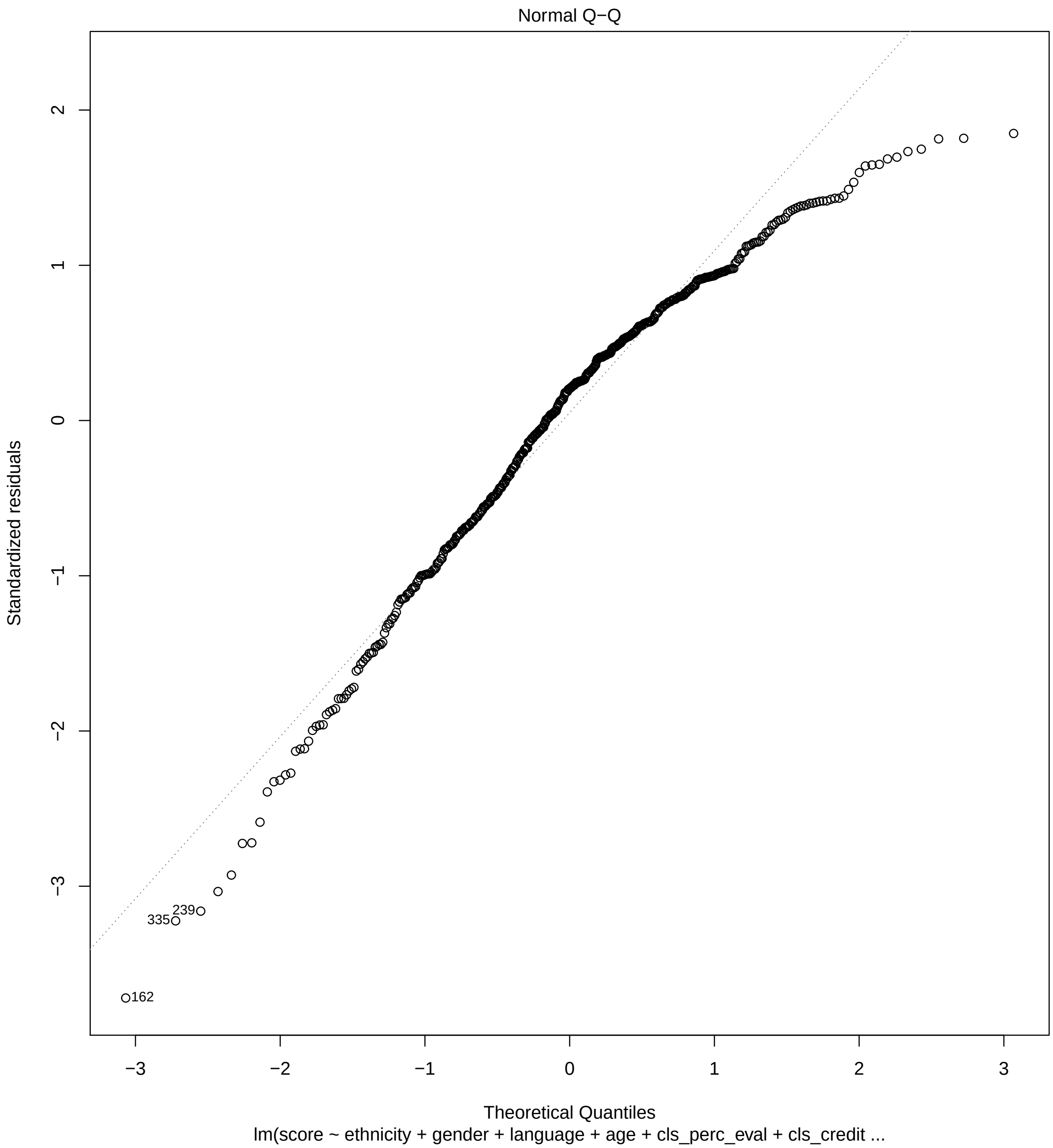
$$\begin{aligned}
\widehat{score} = 3.9070305 &+ 0.1638182 \times ethnicity_{(notMinority)} \\
&+ 0.2025970 \times gender_{(male)} \\
&- 0.2466834 \times language_{(nonEnglish)} \\
&- 0.0069246 \times age \\
&+ 0.0049425 \times cls_perc_eval \\
&+ 0.5172051 \times cls_credits_{(oneCredit)} \\
&+ 0.0467322 \times bty_avg \\
&- 0.1139392 \times pic_outfit_{(notFormal)} \\
&- 0.1808705 \times pic_color_{(color)}
\end{aligned}$$

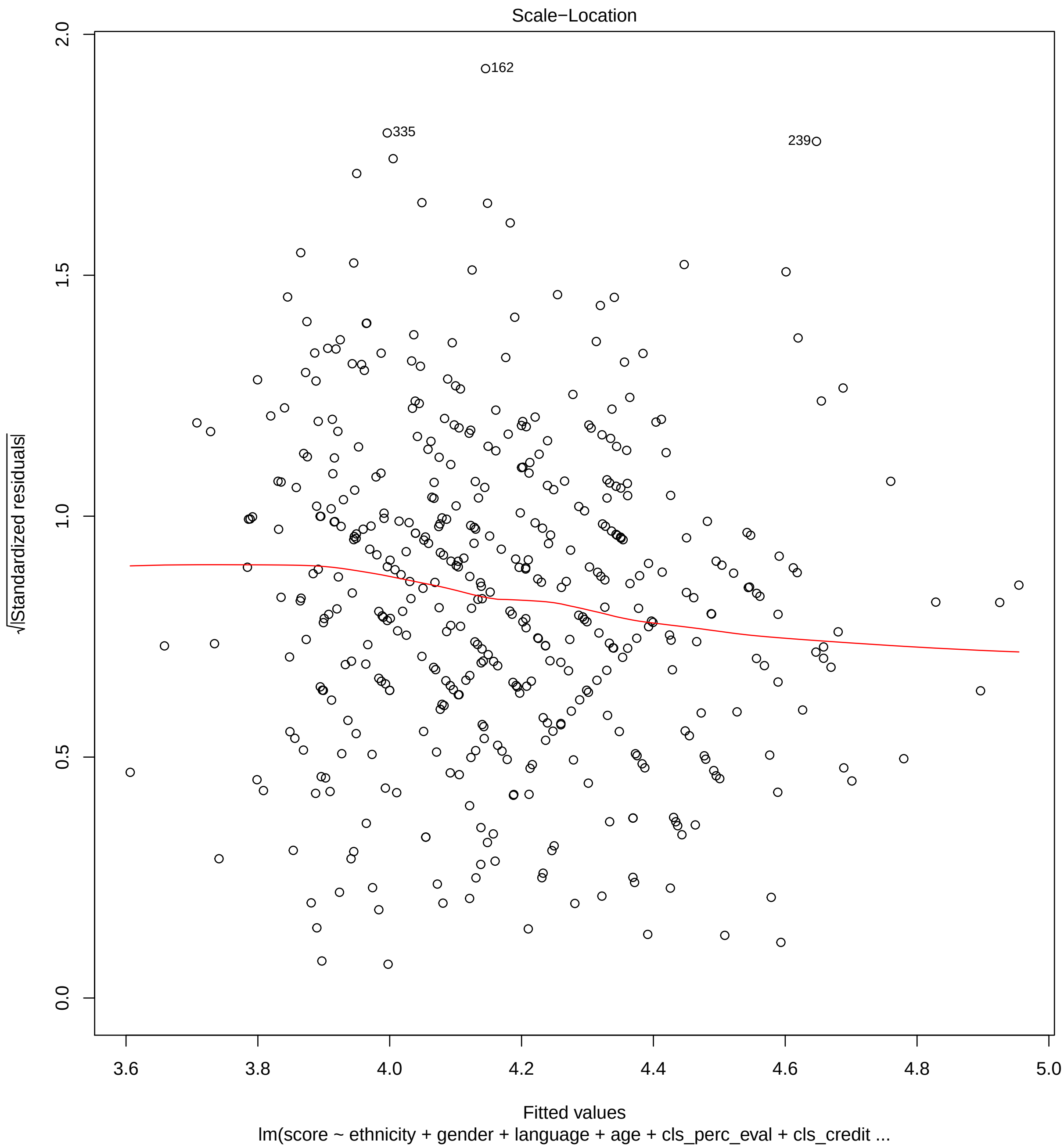
16. Verify that the conditions for this model are reasonable using diagnostic plots.

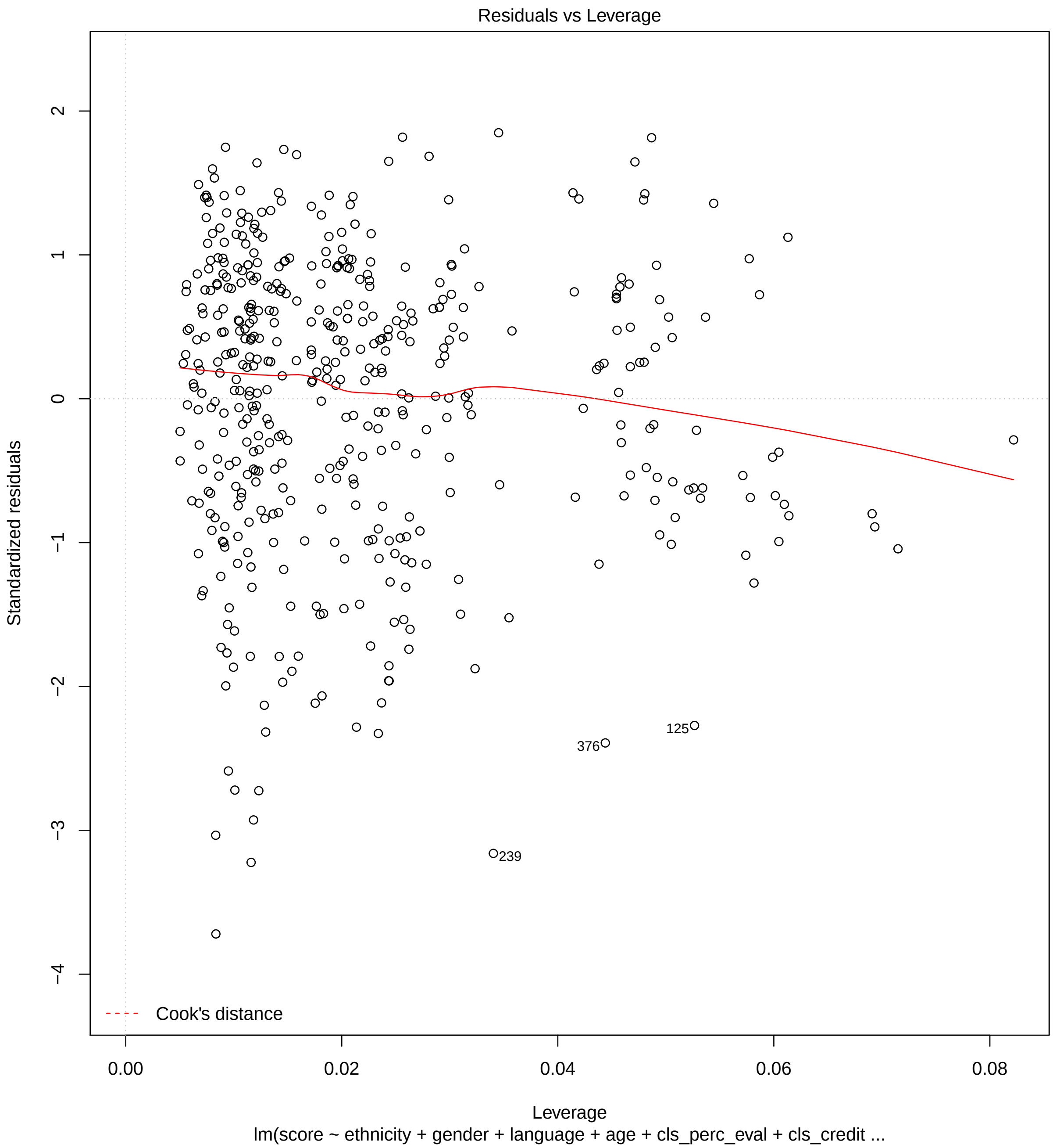
```r
plot(m_final)
```



Residuals vs Fitted
Residuals
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
3.6
3.8
4.0
4.2
4.4
4.6
4.8
5.0
335
239
162
Fitted values
lm(score ~ ethnicity + gender + language + age + cls_perc_eval + cls_credit ...


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(score ~ ethnicity + gender + language + age + cls_perc_eval + cls_credit ...
162
335
239


Scale−Location
√|Standardized residuals|
Fitted values
lm(score ~ ethnicity + gender + language + age + cls_perc_eval + cls_credit ...
162
335
239
0.0
0.5
1.0
1.5
2.0
3.6
3.8
4.0
4.2
4.4
4.6
4.8
5.0



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 conditions assume independence of observations. While the "beauty" assessment is assigned to the instructor not by the students in each course who are also assessing the instructor's teaching, but by a separate half-dozen observers who are evaluating the looks of the instructors based upon photographs, such assessments of physical appearance would be the same for each instructor across all of his/her courses taught. Indeed, the data show that the

entries for the following variables match for a numerous clusters of courses, all of which would appear to map to a single instructor:

`rank`, `ethnicity`, `gender`, `language`, `age`, The 6 raw "beauty" variables, and their average, `bty_avg`, `pic_outfit`, and `pic_color`

Indeed, tabling the data based on identical matches on the above attributes suggests that there are only 94 different instructors, a fact which is confirmed by the original paper.

Thus, these variables are not independent of each other with respect to each of the line items in the data set. This means that OLS may not be the most appropriate statistical method for such data. Rather, techniques such as instrumental variables, two-stage least squares, fixed-effects, and structural equation modeling should be considered.

In the paper, the authors indicate that they utilize a weighted-least-squares technique because of the differing percentage of students in each course who respond to the end-of-term evaluation surveys.

Indeed, the authors explain "*As weights we use the number of students completing the evaluation forms in each class, because the error variances in the average teaching ratings are larger the fewer students completing the instructional evaluations.*"

Additionally, the authors note that "We present robust standard errors that account for the clustering of the observations (because we observe multiple classes for the overwhelming majority of instructors) for each of the parameter estimates."

Thus, the authors of the paper have recognized that OLS is not appropriate given the clustering of the data; they have taken steps to address this in their modeling.

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 model coefficients, the following characteristics of a professor would be associated with a high evaluation score:

`Ethnicity` **not minority**

`Gender` **== male**

`Language` **(of instructor's undergraduate institution) is English**

`Age` **is young (this is a numeric variable)**

Posed for "Formal" photograph (e.g., wearing a necktie)

Such photograph is not in color.

For the courses, the sole remaining characteristics which correlate to a high evaluation incude:

`cls_pct` **– a large percentage of students in the course did respond to the survey of instrutor's teaching, and**

`cls_credits` **– the course is a one-credit course (which applies to less than 6 percent of the line items - there are only 27 such entries out of 463; the original paper explains that each of these are laboratory sections)**

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

The authors explain that to construct their superset of instructors whose courses were considered for inclusion in the study, they could only consider those instructors who included photographs of themselves on their departmental websites. The set of instructors who choose to post their photo in this way may be biased, as the authors indicate, because perhaps only "better-looking" instructors would agree to post their photos, while more modest-looking instructors may have been inelgible for consideration for the study if they chose not to post their photos.

Other universities may have different policies, perhaps automatically including the photos of ALL instructors on their websites. This could impact results at such other schools, if true.

Additionally, the perception of beauty may vary from region to region, which could impact ratings.

Therefore I would not be comfortable generalizing my conclusions to apply to professors generally.