Multiple linear regression

Grading the professor

Many college courses conclude by giving students the opportunity to evaluate the course and the instructor anonymously. However, the use of these student evaluations as an indicator of course quality and teaching effectiveness is often criticized because these measures may reflect the influence of non-teaching related characteristics, such as the physical appearance of the instructor. The article titled, “Beauty in the classroom: instructors’ pulchritude and putative pedagogical productivity” (Hamermesh and Parker, 2005) found that instructors who are viewed to be better looking receive higher instructional ratings. (Daniel S. Hamermesh, Amy Parker, Beauty in the classroom: instructors pulchritude and putative pedagogical productivity, *Economics of Education Review*, Volume 24, Issue 4, August 2005, Pages 369-376, ISSN 0272-7757, 10.1016/j.econedurev.2004.07.013. http://www.sciencedirect.com/science/article/pii/S0272775704001165.)

In this lab we will analyze the data from this study in order to learn what goes into a positive professor evaluation.

The data

The data were gathered from end of semester student evaluations for a large sample of professors from the University of Texas at Austin. In addition, six students rated the professors’ physical appearance. (This is aslightly modified version of the original data set that was released as part of the replication data for *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Gelman and Hill, 2007).) The result is a data frame where each row contains a different course and columns represent variables about the courses and professors.

```r
load("more/evals.RData")
```

variable	description
`score`	average professor evaluation score: (1) very unsatisfactory - (5) excellent.
`rank`	rank of professor: teaching, tenure track, tenured.
`ethnicity`	ethnicity of professor: not minority, minority.
`gender`	gender of professor: female, male.
`language`	language of school where professor received education: english or non-english.

variable	description
`age`	age of professor.
`cls_perc_eval`	percent of students in class who completed evaluation.
`cls_did_eval`	number of students in class who completed evaluation.
`cls_students`	total number of students in class.
`cls_level`	class level: lower, upper.
`cls_profs`	number of professors teaching sections in course in sample: single, multiple.
`cls_credits`	number of credits of class: one credit (lab, PE, etc.), multi credit.
`bty_f1lower`	beauty rating of professor from lower level female: (1) lowest - (10) highest.
`bty_f1upper`	beauty rating of professor from upper level female: (1) lowest - (10) highest.
`bty_f2upper`	beauty rating of professor from second upper level female: (1) lowest - (10) highest.
`bty_m1lower`	beauty rating of professor from lower level male: (1) lowest - (10) highest.
`bty_m1upper`	beauty rating of professor from upper level male: (1) lowest - (10) highest.

variable	description
`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

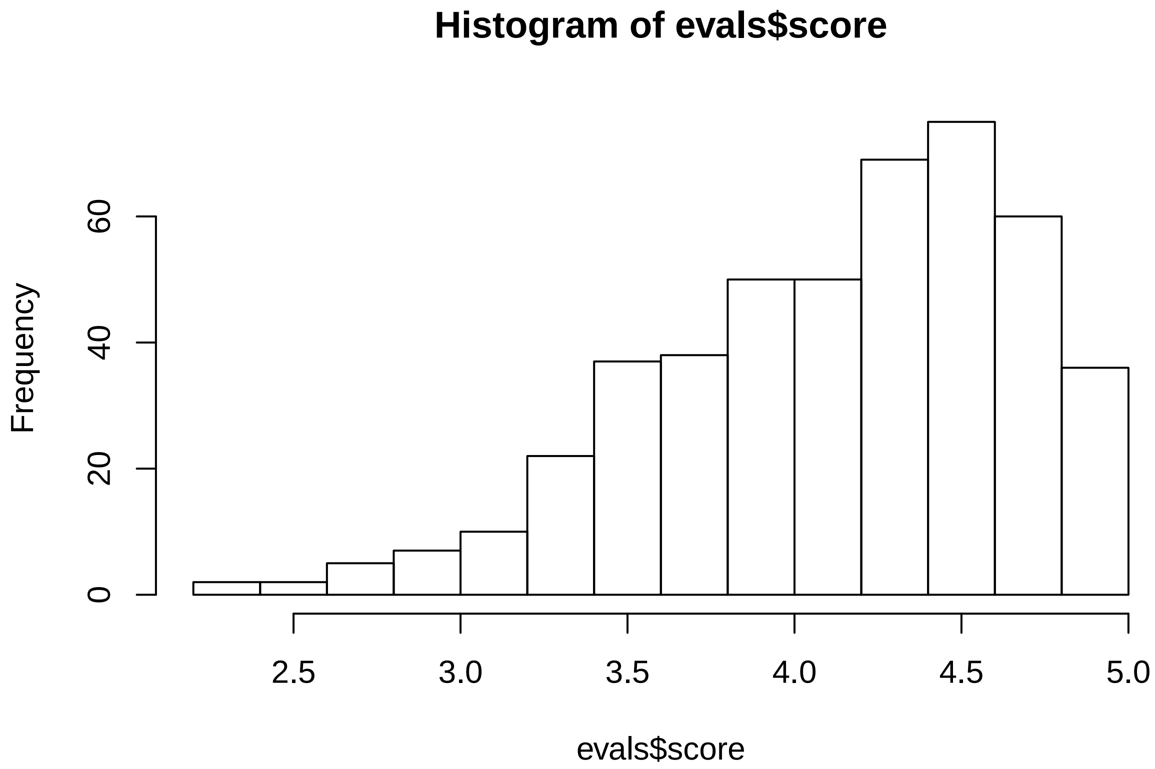
1. Is this an observational study or an experiment? The original research question posed in the paper is whether beauty leads directly to the differences in course evaluations. Given the study design, is it possible to answer this question as it is phrased? If not, rephrase the question.

This is an observational study as the researcher is only collecting data and has no control on the data collected.

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?

As we can see below the distribution looks normal but left or negatively skewed. It does make sense based on my past experience as a student. I never rated a course below the average grade. Usually gave a grade of 80% and above.

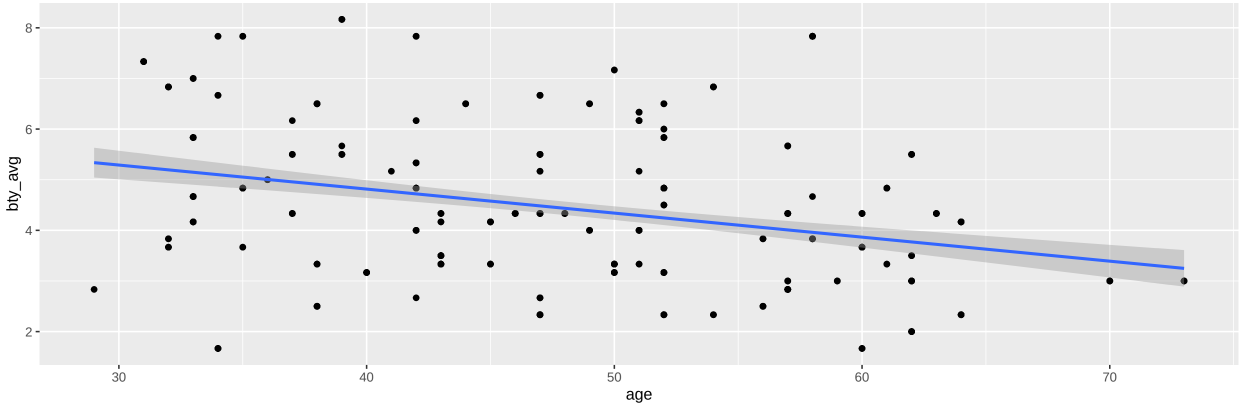
```
hist(evals$score)
```



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

Below we see the relationship between the age of the professor and the average beauty rating of professor given by the students. Using a linear fit line, we see how there is a inverse relationship between the age and the beauty rating given to the professor.

```
library(ggplot2)
ggplot(evals, aes(x=age, y=bty_avg)) + geom_point() + geom_smooth(method='lm')
```



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:

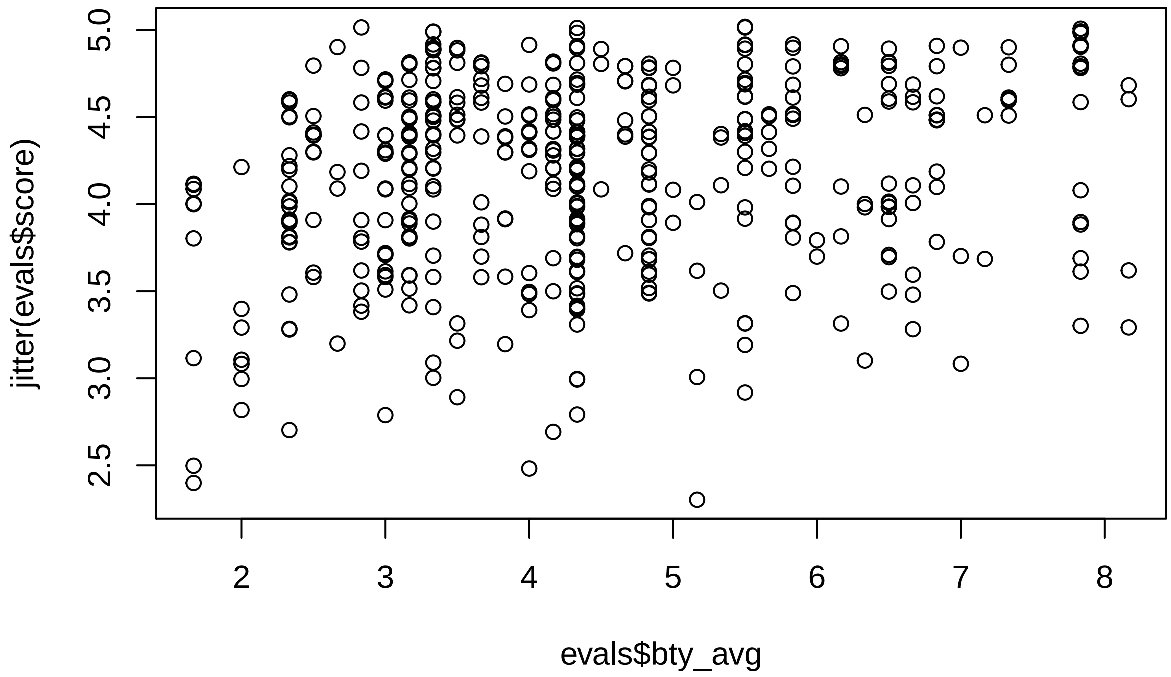
```
plot(evals$score ~ evals$bty_avg)
```

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?

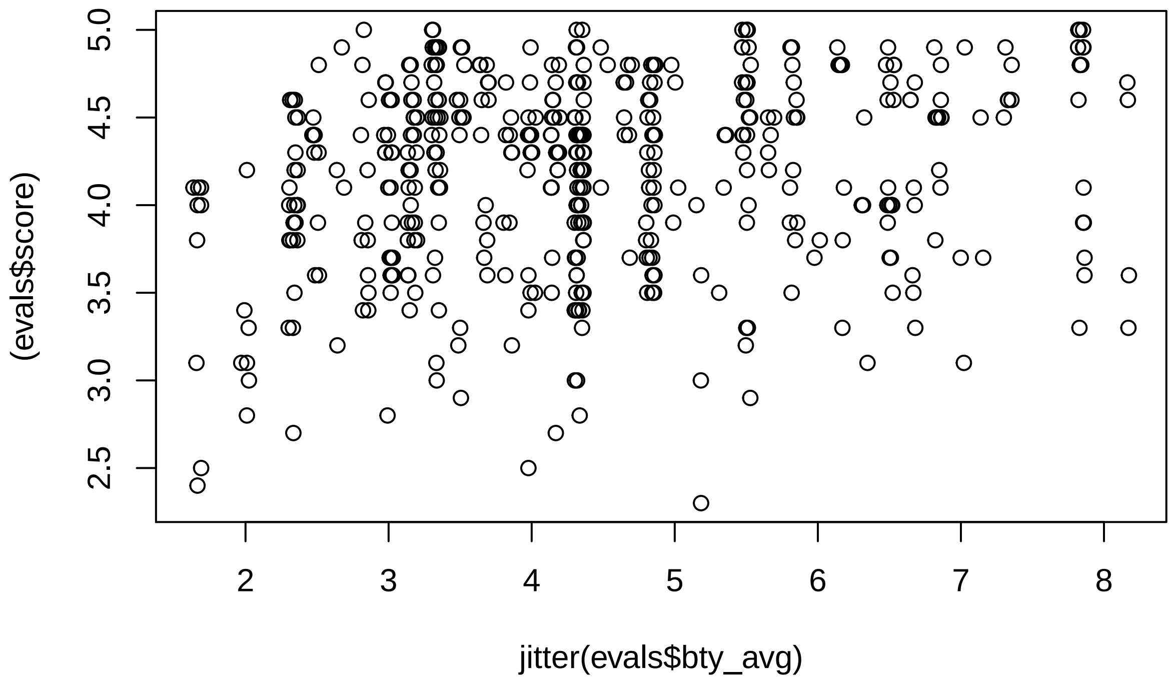
4. Replot the scatterplot, but this time use the function `jitter()` on the y- or the x-coordinate. (Use `?jitter` to learn more.) What was misleading about the initial scatterplot?

We see that there is a lot more datapoints as some of those were superimposed behind each other. Using the jitter() function makes it more apparent by adding some noise behind the integer limited datapoint positions.

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



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



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

The linear fit equation is score = 3.88034 + 0.06664 * bty__avg

The p-value of 5.083e-05 that the average beauty rating is a statistically significant predictor. However, it only explains 3.5% of the variance in the score data. It is statistically significant but not a practical predictor.

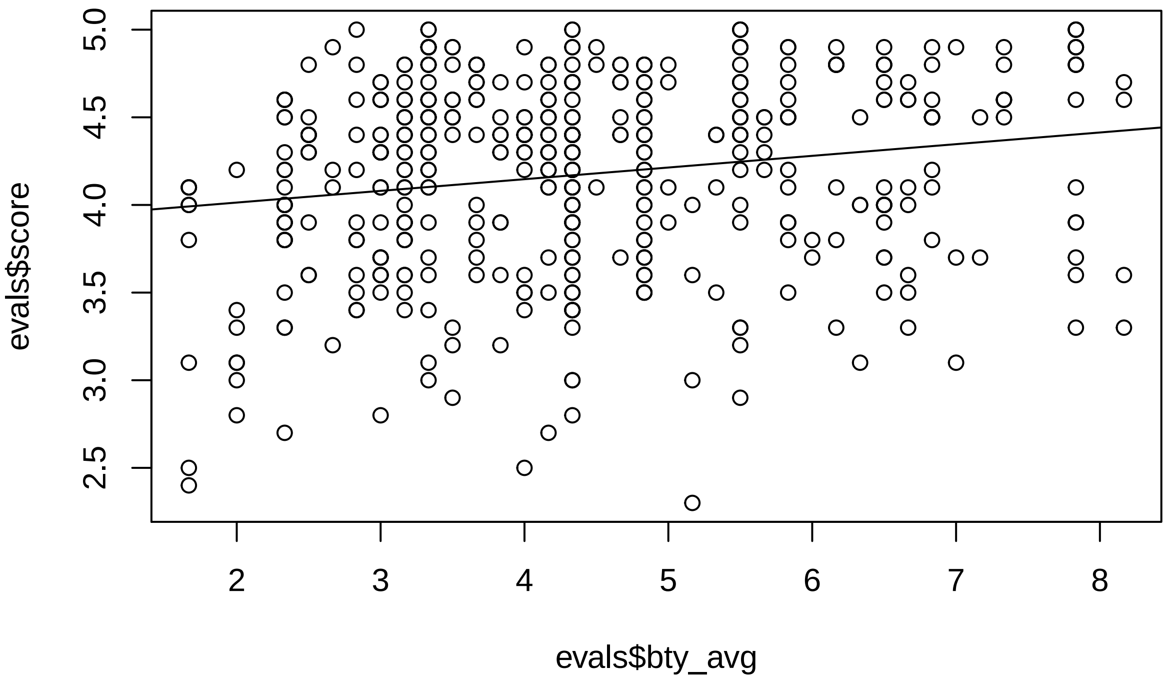
```
attach(evals)
m_bty <- lm(score ~  bty_avg)
m_bty
```

```
##
## Call:
## lm(formula = score ~ bty_avg)
##
## Coefficients:
## (Intercept)      bty_avg
##     3.88034      0.06664
```

```
summary(m_bty)
```

```
##
## Call:
## lm(formula = score ~ 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 ***
## 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(evals$score ~ evals$bty_avg)
abline(m_bty)
```



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).

As see in the 3 plots belows, the conditions for

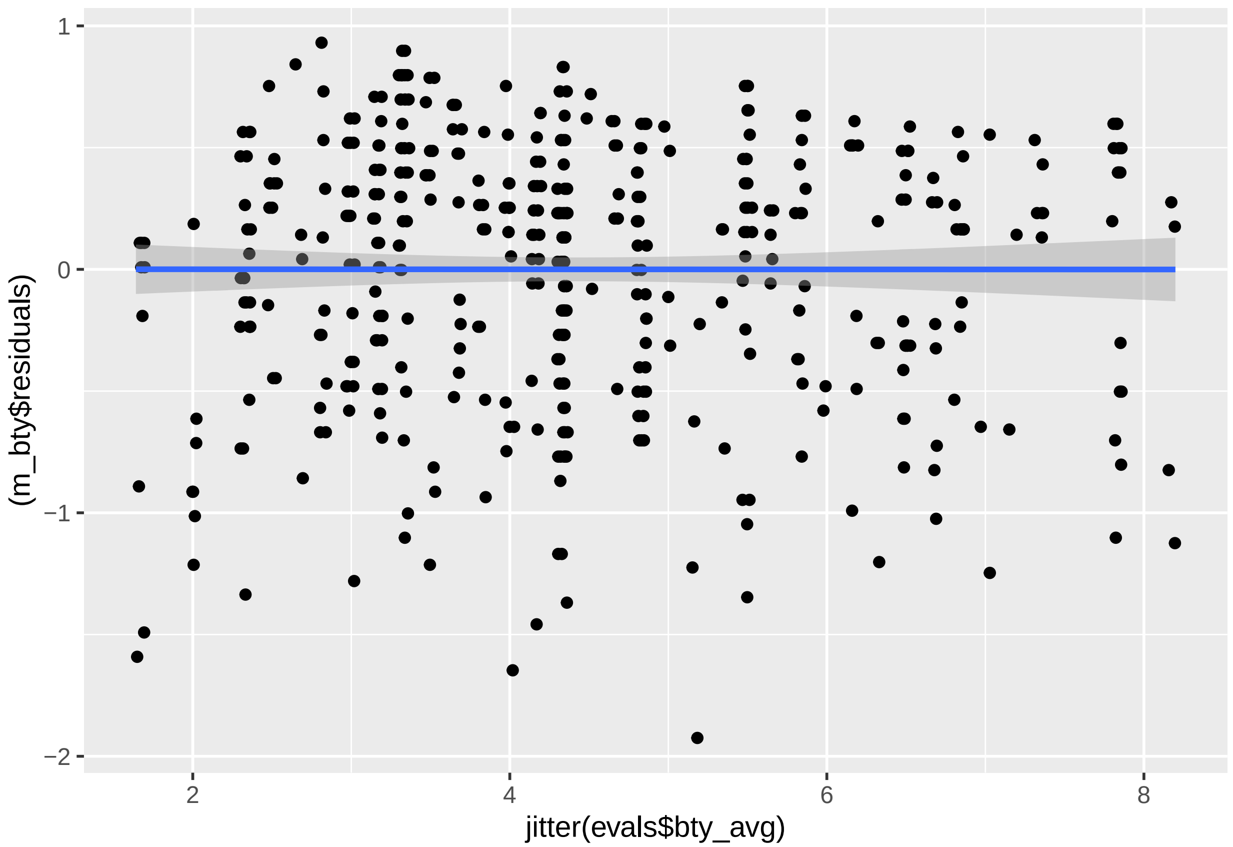
*(1) linearity: Using the residual vs. predictor variable we see that the residuals are not evenly distributed around the zeroth line.

*(2) nearly normal residuals: The residuals look to be left skewed.

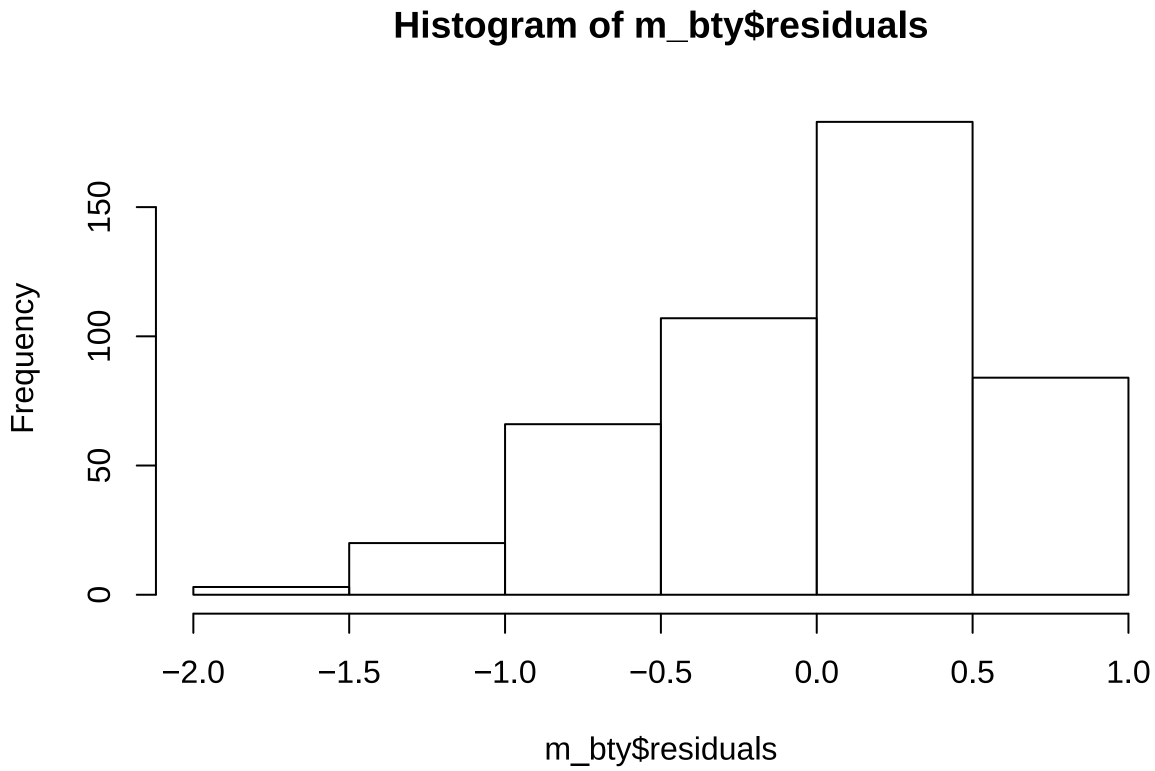
*(3) constant variability: The Q-Q Plot shows that the residuals do not follow a constant fit line.

The conditions least squares regression are not reasonable for this case.

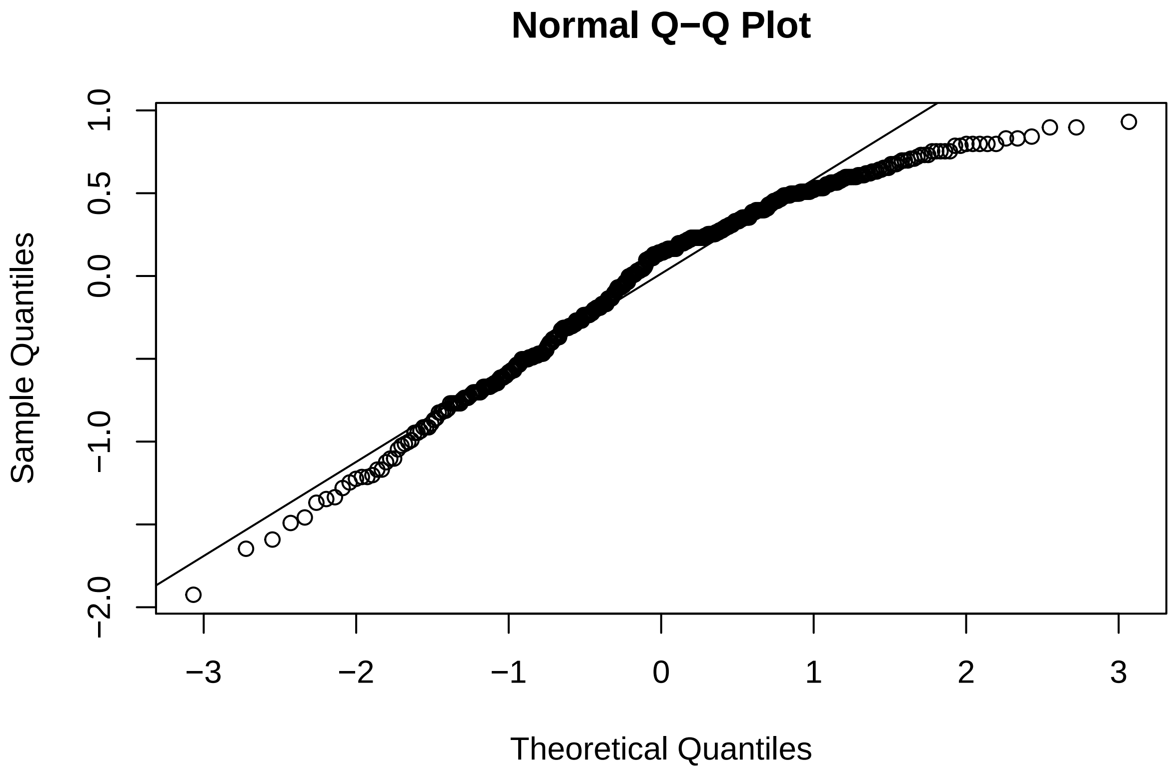
```
ggplot(evals, aes(x=jitter(evals$bty_avg), (m_bty$residuals))) + geom_point() + geom_smooth(method='lm'
```



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



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



Multiple linear regression

The data set contains several variables on the beauty score of the professor: individual ratings from each of the six students who were asked to score the physical appearance of the professors and the average of these six scores. Let's take a look at the relationship between one of these scores and the average beauty score.

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

As expected the relationship is quite strong - after all, the average score is calculated using the individual scores. We can actually take a look at the relationships between all beauty variables (columns 13 through 19) using the following command:

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

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

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

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

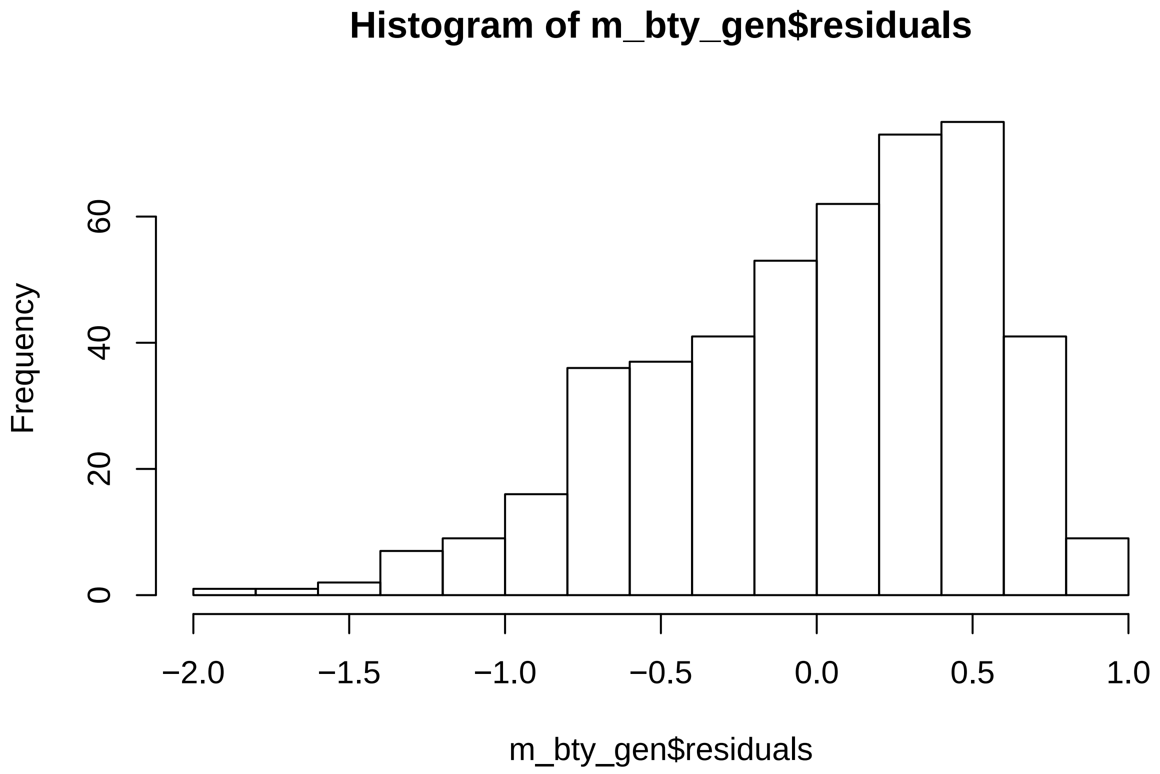
```
##
## Call:
## lm(formula = score ~ bty_avg + gender, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8305 -0.3625  0.1055  0.4213  0.9314
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.74734    0.08466  44.266  < 2e-16 ***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

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

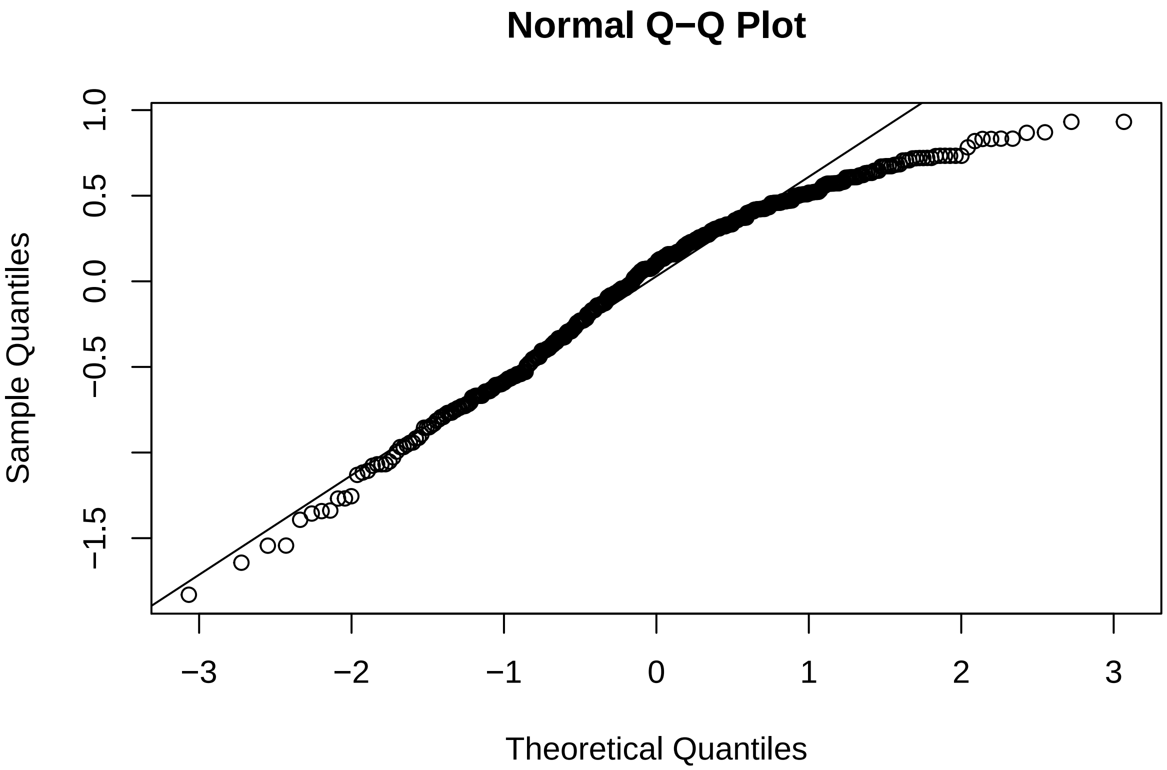
This example of Multiple linear regression has slightly improved the Rˆ2 value from 3.502% to 5.912%. The statistical significance of this fit has also improved with a p-value of 8.177e-07.

The model has improved but the histogram of the residuals still looks heavily left skewed. Also, the Q-Q Plot shows more lineary at the center of the graph but the tails show a negative bias.

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



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



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`?

Before

> bty_avg 0.06664 0.01629 4.09 5.08e-05 ***

After

> bty_avg 0.07416 0.01625 4.563 6.48e-06 ***gendermale 0.17239 0.05022 3.433 0.000652***

The addition of `gender` has increased by an order of magnitude the statistical significance of `bty_avg`

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?

score = 3.74734 + (3.74734 + 0.17239 * 1) * bty__avg

```
m_bty_gen
```

```
##
## Call:
## lm(formula = score ~ bty_avg + gender, data = evals)
##
## Coefficients:
## (Intercept)      bty_avg   gendermale
##     3.74734      0.07416      0.17239
```

The decision to call the indicator variable `gendermale` instead of`genderfemale` has no deeper meaning. R simply codes the category that comes first alphabetically as a 0. (You can change the reference level of a categorical variable, which is the level that is coded as a 0, using the`relevel` function. Use `?relevel` to learn more.)

10. Create a new model called `m_bty_rank` with `gender` removed and `rank` added in. How does R appear to handle categorical variables that have more than two levels? Note that the rank variable has three levels: `teaching`, `tenure track`, `tenured`.

```
m_bty_rank <- lm(score ~ bty_avg + rank, data = evals)
summary(m_bty_rank)
```

Seems that R has created two indicator variable called `ranktenure track` and `ranktenured` to account for the 'gender' descriptor with two factors.

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.

Potentially any of the variables could have an effect on the score. However, I chose the number of the credits as the least relevant predictor variable.

> `cls_credits` | number of credits of class: one credit (lab, PE, etc.), multi credit.

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.

My suspicions were wrong!

> cls_creditsone credit 0.5020432 0.1159388 4.330 1.84e-05 ***

`cls_credits` is one of the 4 predictor variables with a p-score lower than 0.001. It actually has a high statistical significance in explaining the score value.

13. Interpret the coefficient associated with the ethnicity variable.

> ethnicitynot minority 0.1234929 0.0786273 1.571 0.11698

Ethnicity has a positive relationship with the teacher score even if it has low practival statistical significance with a p-score of 0.11698

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?

The coefficients and statistical significance of the rest of the predictor variables did not change at all. There are so many predictor variables in this set, and despite using all of them in our linear fit model, the calculated R^2 value is just at 5%. This indicates that we either have a large number of predictor variables with low practical statistical signficance and that many of these same variables ara also collinear.

```
m_fullV1 <- 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_fullV1)
```

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
m_fullV6 <- lm(score ~ gender + language + cls_perc_eval
              + cls_credits + bty_avg
             + pic_color, data = evals)
m_fullV6
```

```
##
## Call:
## lm(formula = score ~ gender + language + cls_perc_eval + cls_credits +
##     bty_avg + pic_color, data = evals)
##
## Coefficients:
##           (Intercept)             gendermale    languagenon-english
##              3.617542               0.190934              -0.279160
##         cls_perc_eval  cls_creditsone credit                bty_avg
##              0.004468               0.456768               0.060225
##        pic_colorcolor
##             -0.194487
```

```r
summary(m_fullV6)
```

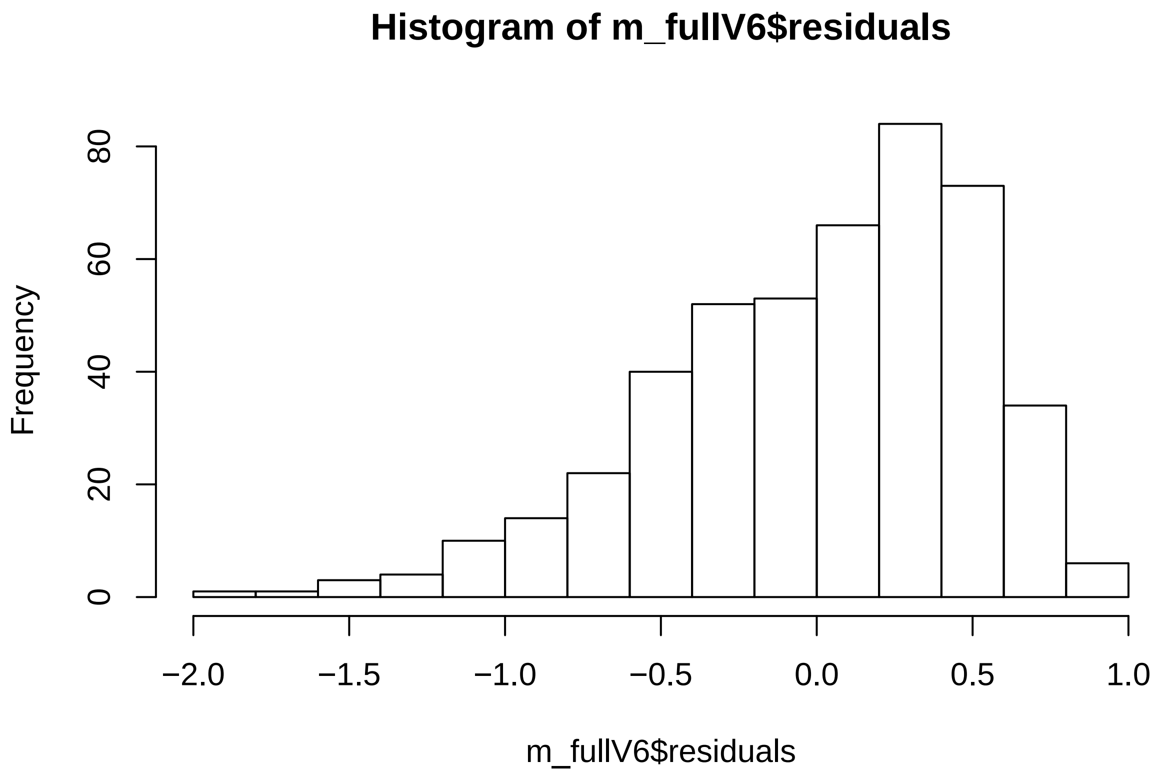
```
##
## Call:
## lm(formula = score ~ gender + language + cls_perc_eval + cls_credits +
##     bty_avg + pic_color, data = evals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.80660 -0.31936  0.09867  0.38572  0.84641
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.617542   0.149744  24.158  < 2e-16 ***
## gendermale             0.190934   0.048255   3.957 8.81e-05 ***
## languagenon-english   -0.279160   0.098766  -2.826 0.004913 **
## cls_perc_eval          0.004468   0.001436   3.111 0.001983 **
## cls_creditsone credit  0.456768   0.101203   4.513 8.13e-06 ***
## bty_avg                0.060225   0.016244   3.708 0.000235 ***
## pic_colorcolor        -0.194487   0.066184  -2.939 0.003464 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5036 on 456 degrees of freedom
## Multiple R-squared:  0.1538, Adjusted R-squared:  0.1427
## F-statistic: 13.82 on 6 and 456 DF,  p-value: 1.879e-14
```

score = 3.617542 + 0.190934 * gendermale - 0.279160 * languagenon-english + 0.004468 * cls_perc_eval + 0.456768 * cls_creditsone credit + 0.060225 *bty_avg - 0.194487 * pic_colorcolor

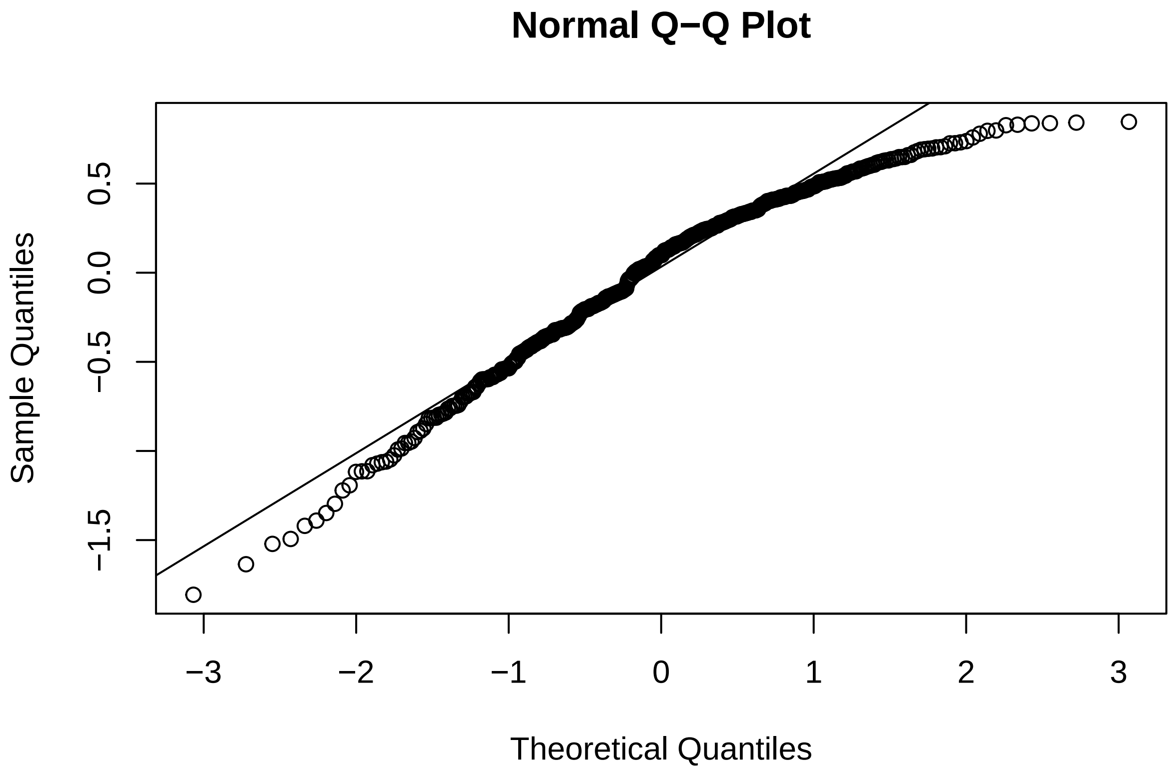
16. Verify that the conditions for this model are reasonable using diagnostic plots.

The model looks normal even if slightly left skewed.

```r
hist(m_fullV6$residuals)
```



```r
qqnorm(m_fullV6$residuals)
qqline(m_fullV6$residuals)  # adds diagonal line to the normal prob plot
```



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 lienar regression model could be bias if professors with ranking higher or lower got more courses assigned than the rest of their peers.

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.

    gendermale 0.190934 0.048255 3.957 8.81e-05 ***languagenon-english -0.279160 0.098766 -2.826 0.004913*** *cls_perc_eval 0.004468 0.001436 3.111 0.001983* ***cls_creditsone credit 0.456768 0.101203 4.513 8.13e-06*** bty_avg 0.060225 0.016244 3.708 0.000235 * **pic_colorcolor -0.194487 0.066184 -2.939 0.003464**

Based on my final model, one credit courses taught by a male professor with a hight beauty average (as ranked by their students) would get a high evaluation score.

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

I would not be comfortable generalizing these results as we don't know if this is a general phenomenon applicable to other universities outside University of Texas at Austin.