Data 606 - Lab 9

Avery Davidowitz

2022-12-09

Multiple linear regression

Load packages

```
library(tidyverse)
library(openintro)
library(GGally)
```

Data

```
glimpse(evals)
```

```
## Rows: 463
## Columns: 23
## $ course_id     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1~
## $ prof_id       <int> 1, 1, 1, 1, 2, 2, 2, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5,~
## $ score         <dbl> 4.7, 4.1, 3.9, 4.8, 4.6, 4.3, 2.8, 4.1, 3.4, 4.5, 3.8, 4~
## $ rank          <fct> tenure track, tenure track, tenure track, tenure track, ~
## $ ethnicity     <fct> minority, minority, minority, minority, not minority, no~
## $ gender        <fct> female, female, female, female, male, male, male, male, ~
## $ language      <fct> english, english, english, english, english, english, en~
## $ age           <int> 36, 36, 36, 36, 59, 59, 59, 51, 51, 40, 40, 40, 40, 40, ~
## $ cls_perc_eval <dbl> 55.81395, 68.80000, 60.80000, 62.60163, 85.00000, 87.500~
## $ cls_did_eval  <int> 24, 86, 76, 77, 17, 35, 39, 55, 111, 40, 24, 24, 17, 14,~
## $ cls_students  <int> 43, 125, 125, 123, 20, 40, 44, 55, 195, 46, 27, 25, 20, ~
## $ cls_level     <fct> upper, upper, upper, upper, upper, upper, upper, upper, ~
## $ cls_profs     <fct> single, single, single, single, multiple, multiple, mult~
## $ cls_credits   <fct> multi credit, multi credit, multi credit, multi credit, ~
## $ bty_f1lower   <int> 5, 5, 5, 5, 4, 4, 4, 5, 5, 2, 2, 2, 2, 2, 2, 2, 2, 7, 7,~
## $ bty_f1upper   <int> 7, 7, 7, 7, 4, 4, 4, 2, 2, 5, 5, 5, 5, 5, 5, 5, 5, 9, 9,~
## $ bty_f2upper   <int> 6, 6, 6, 6, 2, 2, 2, 5, 5, 4, 4, 4, 4, 4, 4, 4, 4, 9, 9,~
## $ bty_m1lower   <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 7, 7,~
## $ bty_m1upper   <int> 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 6, 6,~
## $ bty_m2upper   <int> 6, 6, 6, 6, 3, 3, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 6, 6,~
## $ bty_avg       <dbl> 5.000, 5.000, 5.000, 5.000, 3.000, 3.000, 3.000, 3.333, ~
## $ pic_outfit    <fct> not formal, not formal, not formal, not formal, not form~
## $ pic_color     <fct> color, color, color, color, color, color, color, color, ~
```

Exercise 1

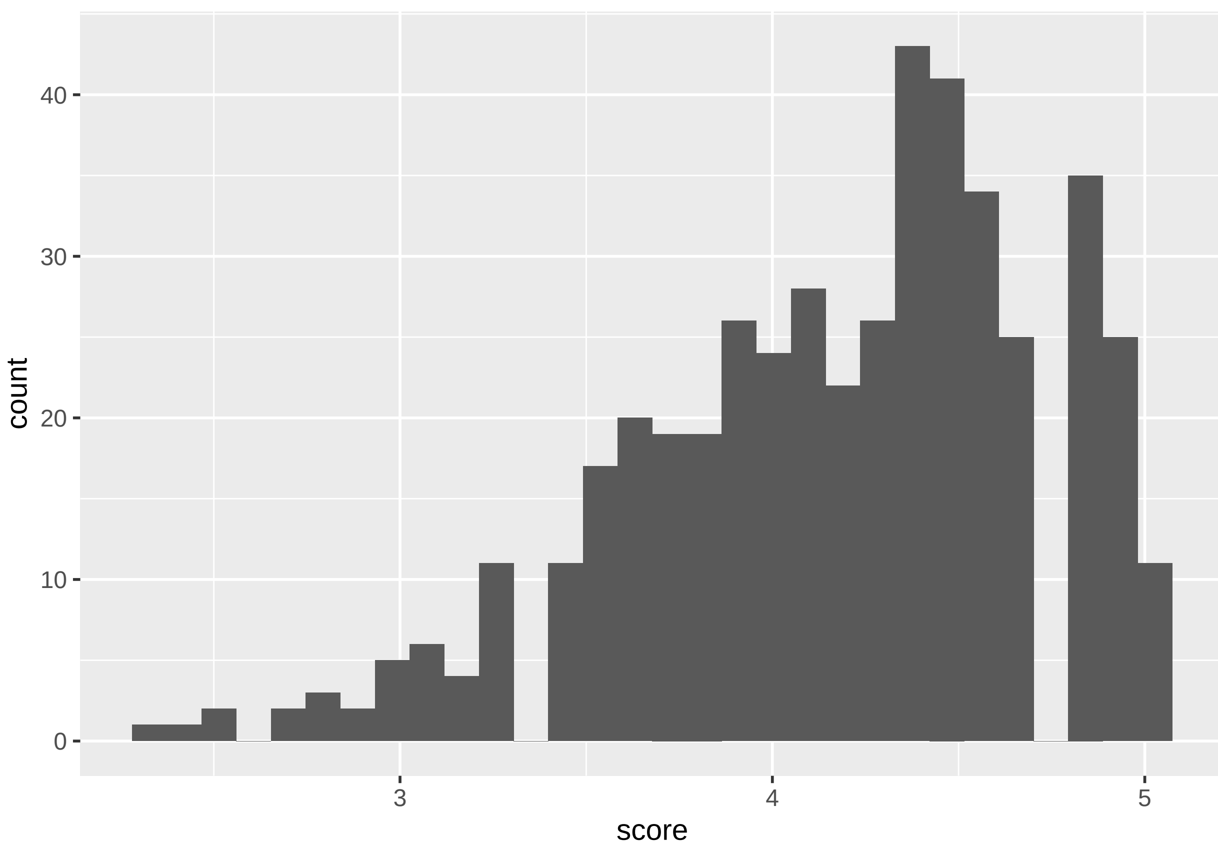
Is this an observational study or an experiment? The original research question posed in the paper is whether beauty leads directly to the differences in course evaluations. Given the study design, is it possible to answer this question as it is phrased? If not, rephrase the question.

This is an observational study due to the lack of random sampling and assignment. The original research question is somewhat problematic because it implies direct causality. It would be better to rephrase it if beauty is significantly correlated with differences in course evaluations.

Exercise 2

Describe the distribution of score. Is the distribution skewed? What does that tell you about how students rate courses? Is this what you expected to see? Why, or why not?

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

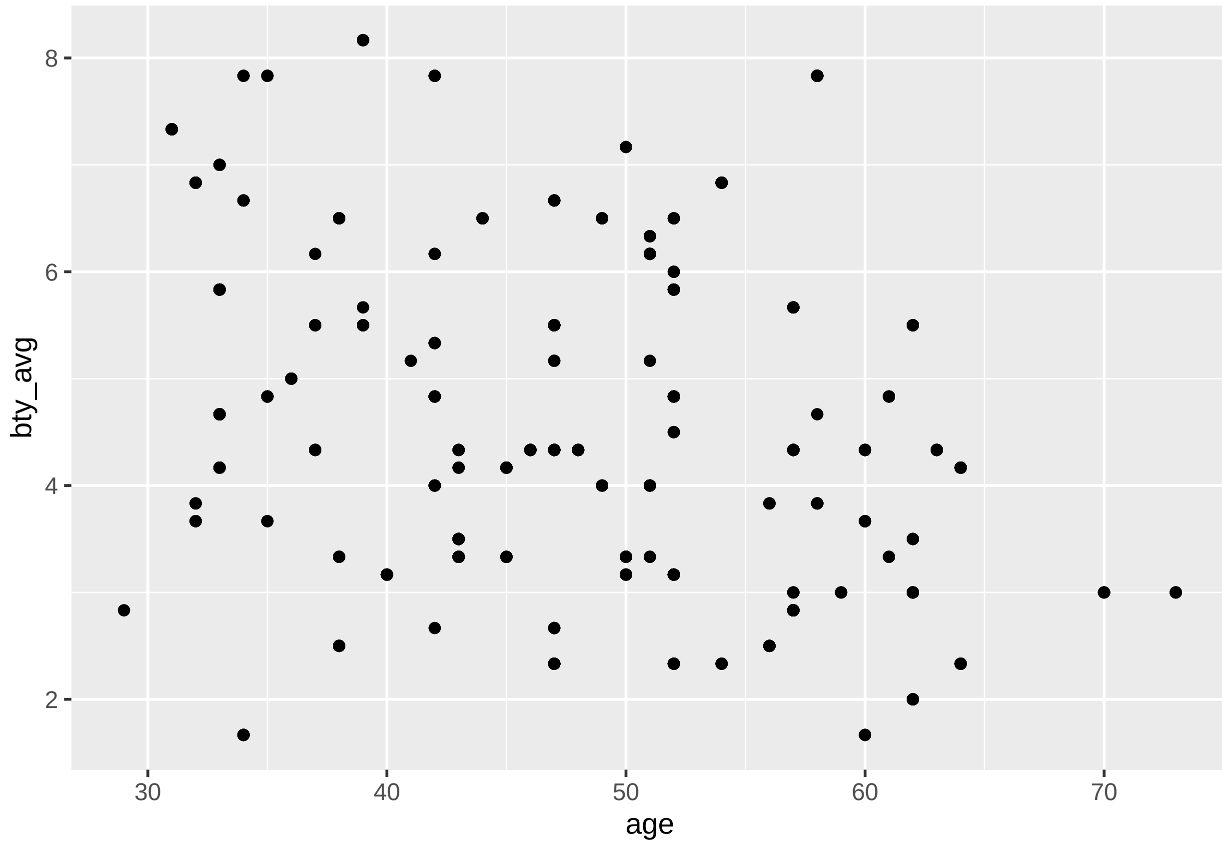


The distribution somewhat resembles a skewed left nearly normal distribution centered near 4.5. This seems pretty reasonable with a well regarded institution.

Exercise 3

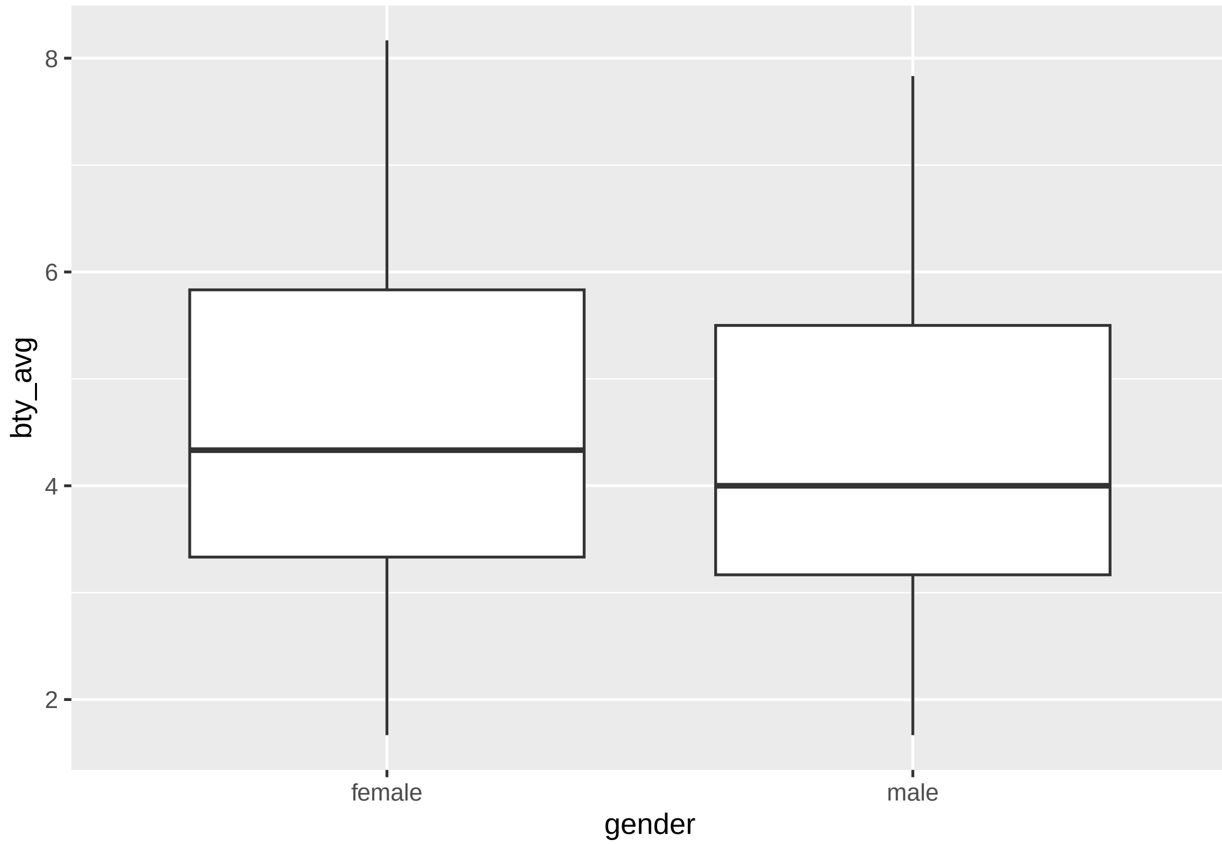
Excluding score, select two other variables and describe their relationship with each other using an appropriate visualization.

```
ggplot(data = evals, aes(x=age, y=bty_avg)) + geom_point()
```



The relationship between age and average beauty score seems to be mostly noise.

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

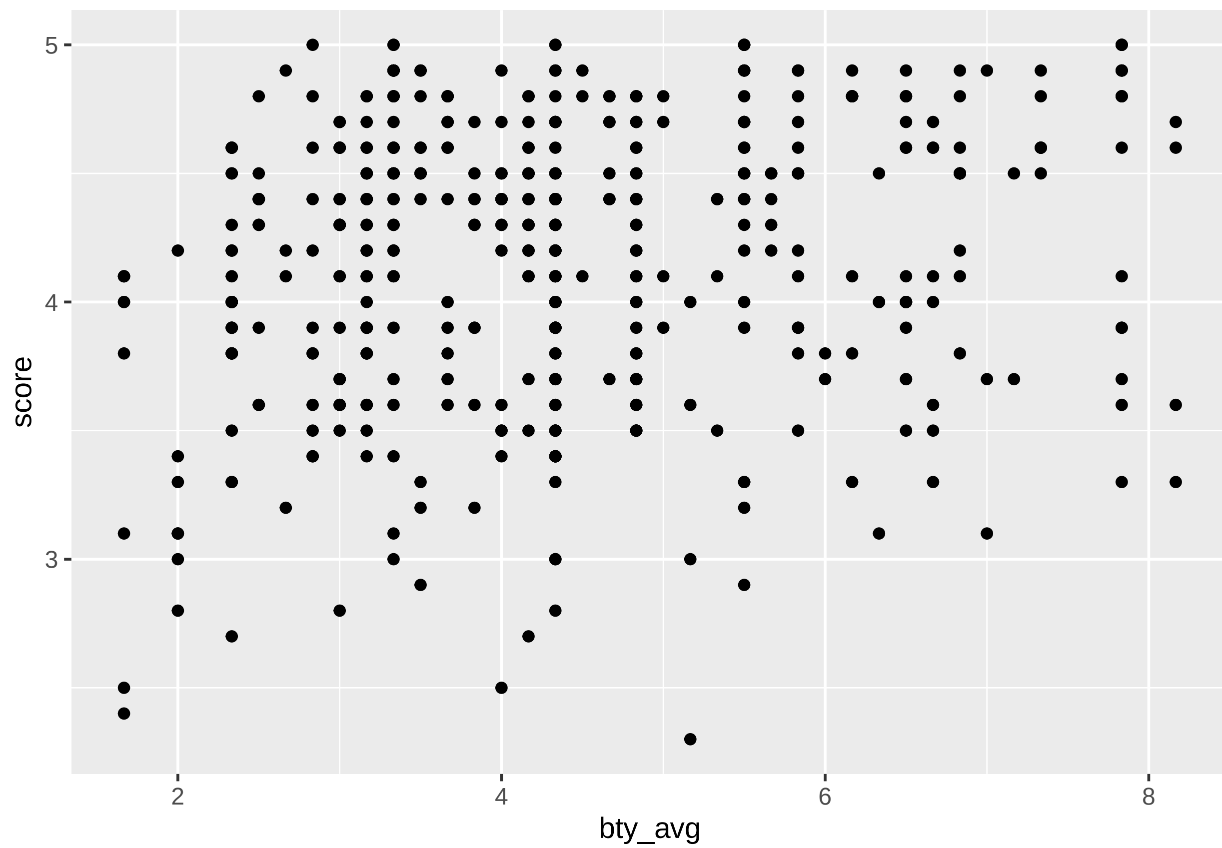


It looks like men and women were rated very simliarly on average beauty score with similar means and quartiles. The women rated slightly higher across most measures but by visualizing alone it does not appear to be significant.

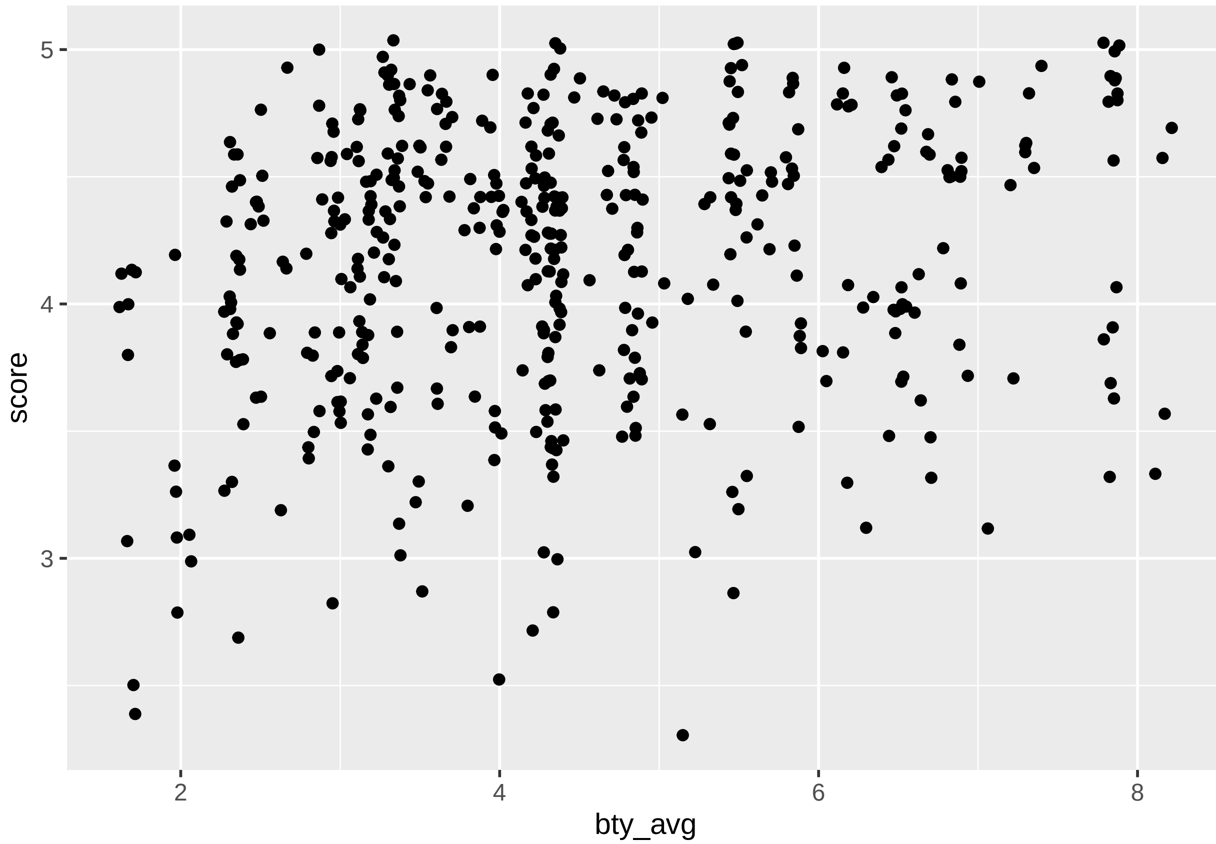
Exercise 4

Replot the scatterplot, but this time use geom_jitter as your layer. What was misleading about the initial scatterplot?

```
ggplot(data = evals, aes(x = bty_avg, y = score)) +
  geom_point()
```




```
ggplot(data = evals, aes(x = bty_avg, y = score)) +
  geom_jitter()
```



From the ggplot2 documentation Jitter "adds a small amount of random variation to the location of each point, and is a useful way of handling overplotting caused by discreteness in smaller datasets." So it may be misleading to use a regular scatterplot when there are many overlapping points which could happen with a dataset like this where the variable isn't continuous (rating being 1-5). You could ignore the strenth of the trend because some data points would contribute multiple times to the overall correlation.

Exercise 5

Let's see if the apparent trend in the plot is something more than natural variation. Fit a linear model called m_bty to predict average professor score by average beauty rating. 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?

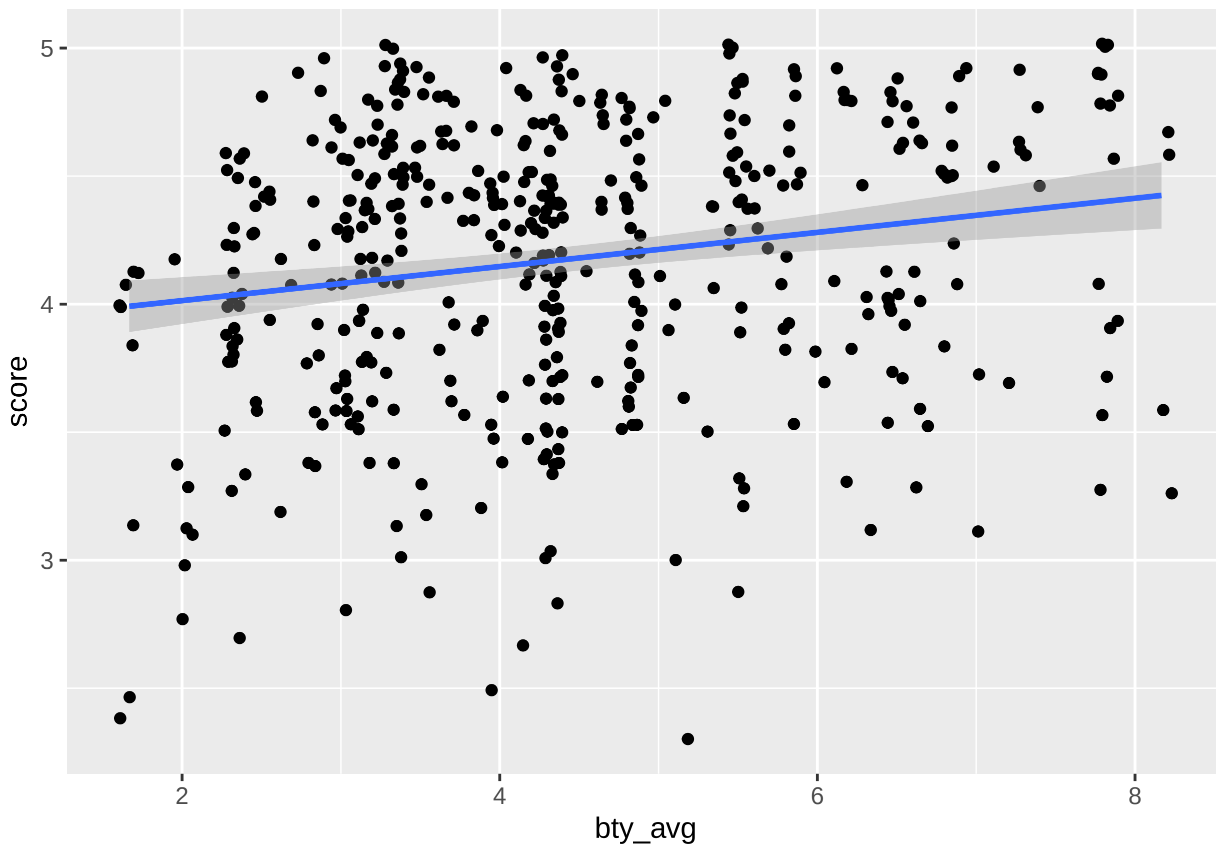
```
score_bty_lm <- lm(score ~ bty_avg, data=evals)

summary(score_bty_lm)
```

```
##
## Call:
## lm(formula = score ~ bty_avg, data = evals)
##
## Residuals:
##      Min       1Q  Median       3Q      Max
## -1.9246 -0.3690  0.1420  0.3977  0.9309
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.88034    0.07614   50.96  < 2e-16 ***
## bty_avg      0.06664    0.01629    4.09 5.08e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5348 on 461 degrees of freedom
## Multiple R-squared:  0.03502,    Adjusted R-squared:  0.03293
## F-statistic: 16.73 on 1 and 461 DF,  p-value: 5.083e-05
```

```
ggplot(data = evals, aes(x = bty_avg, y = score)) +
  geom_jitter() +
  geom_smooth(method = "lm")
```

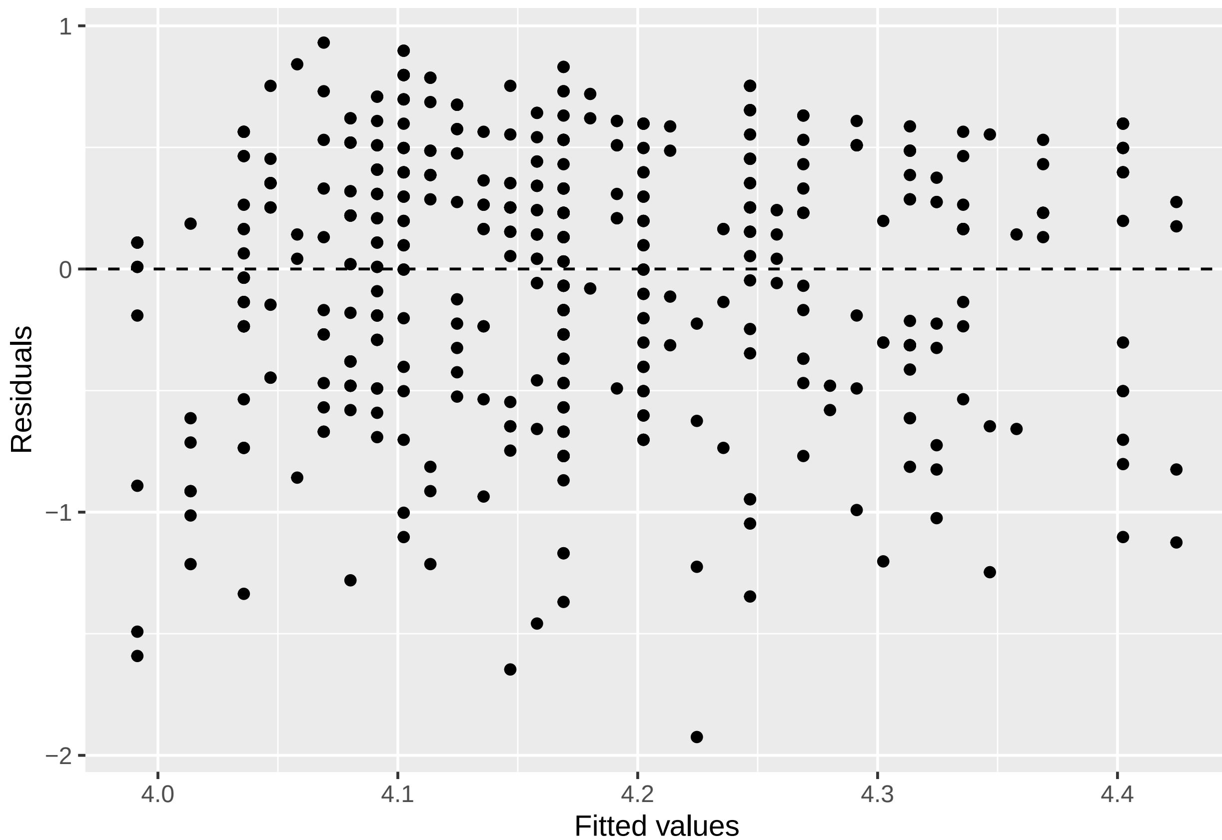


The formula for the regression line would be score = .06664 bty_avg + 3.88034. The p value shows a significant relationship. However, the R squared indicates that the beauty average accounts for a very small amount of variability in the independent variable. Such a small slope would mean that the difference in expected score of a 1 and 10 beauty would be only account for a difference of ~ .6 in a scale between 1-5.

Exercise 6

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

```r
ggplot(data = score_bty_lm, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```



The conditions are indeed met. The residuals appear to be normally distributed and the variance seems similar along the fitted values so homoscedasticity is met.

Exercise 7

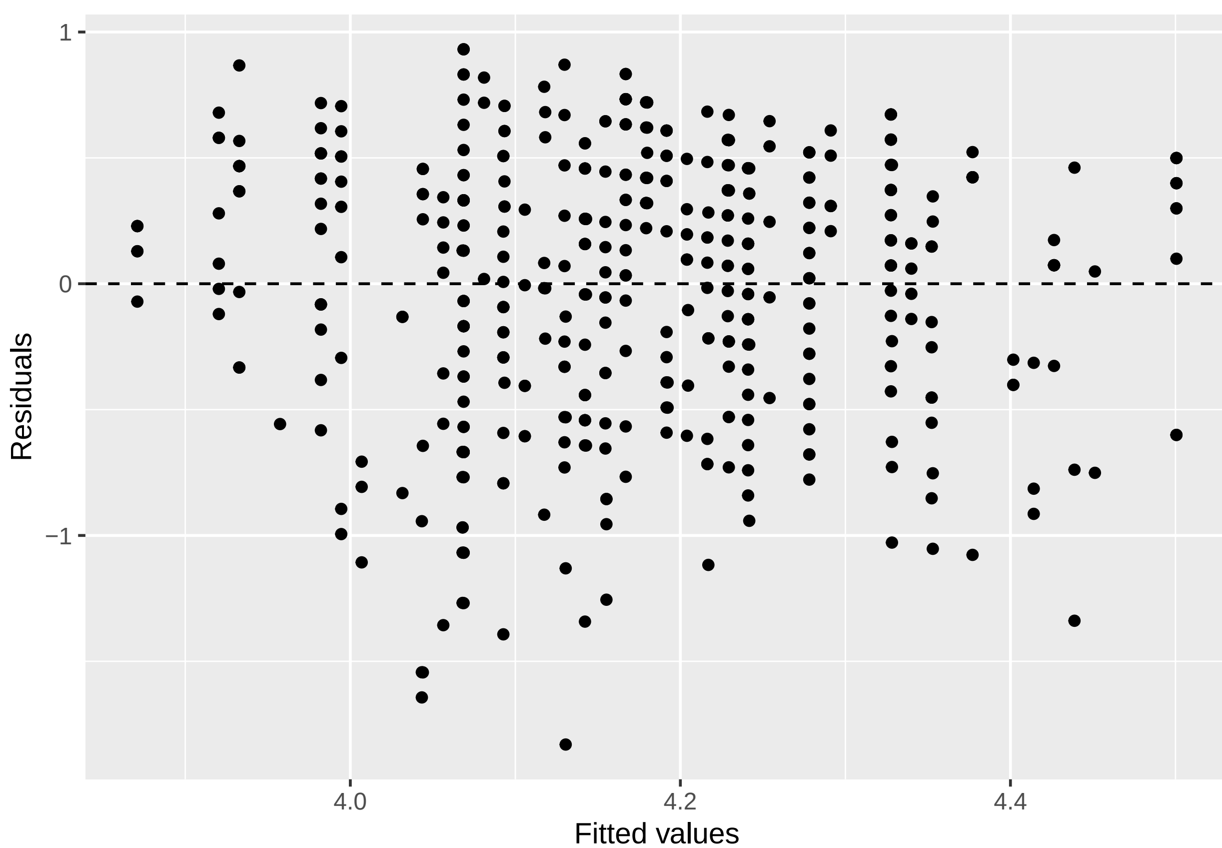
P-values and parameter estimates should only be trusted if the conditions for the regression are reasonable. Verify that the conditions for this model are reasonable using diagnostic plots.

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

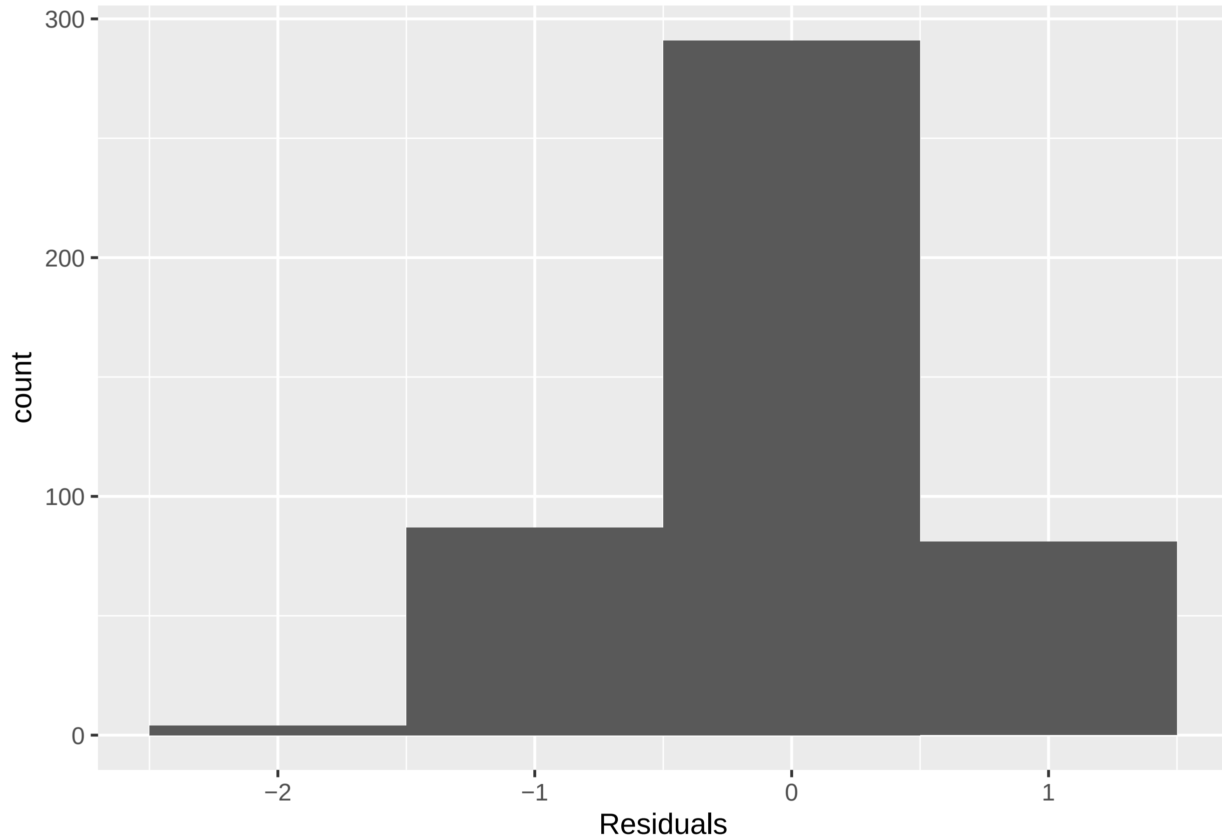
```
##
## Call:
## lm(formula = score ~ bty_avg + gender, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8305 -0.3625  0.1055  0.4213  0.9314
```

```
##
## 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
```

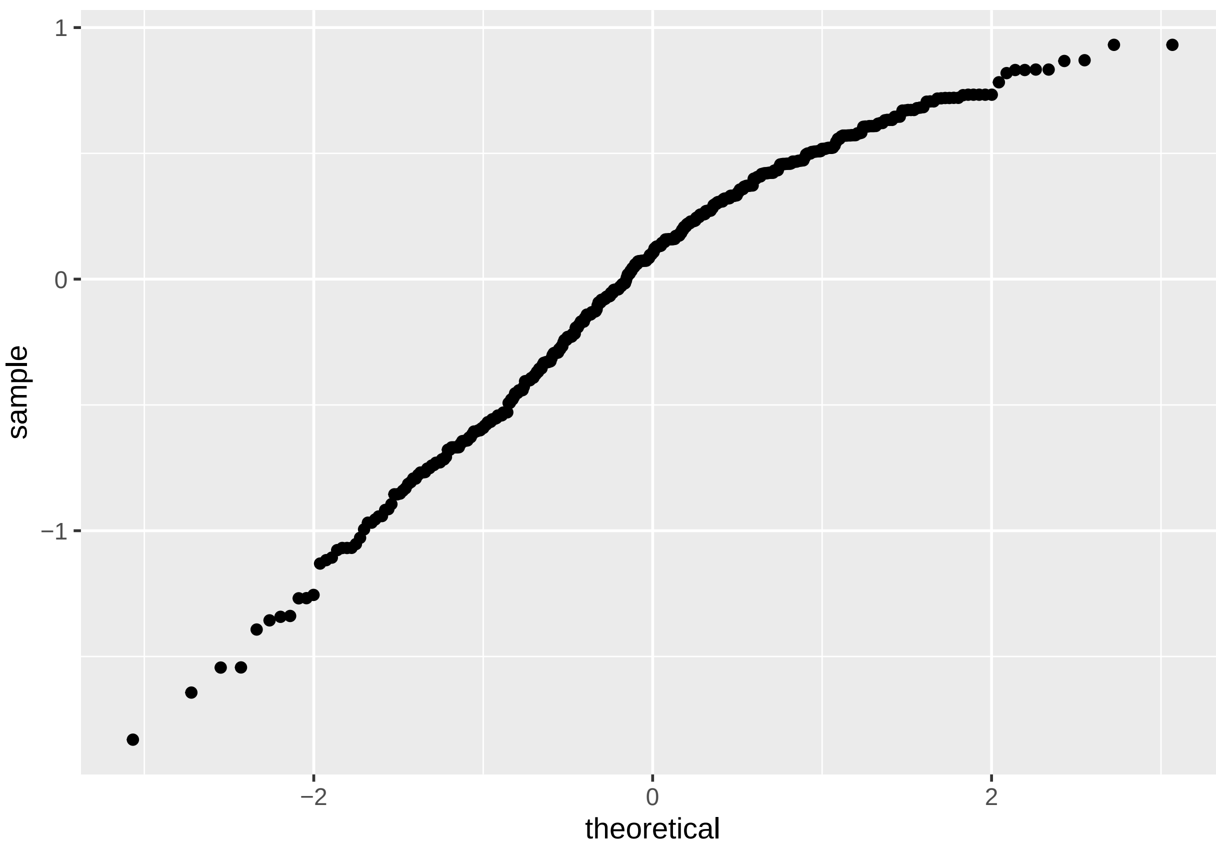
```
ggplot(data = m_bty_gen, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```




```
ggplot(data = m_bty_gen, aes(x = .resid)) +
  geom_histogram(binwidth = 1) +
  xlab("Residuals")
```




```
ggplot(data = m_bty_gen, aes(sample = .resid)) +
  stat_qq()
```



The conditions are also met for the multiple linear model with features of gender and beauty score. There is a slight tail but overall normality and homoscedasticity are satisfied.

Exercise 8

Is bty_avg still a significant predictor of score? Has the addition of gender to the model changed the parameter estimate for bty_avg?

Yes, bty_avg is still significant. Both independent variables have very small p values and the resulting model is an improvement with a higher adjust R squared score as well.

Exercise 9

What is the equation of the line corresponding to those with color pictures? (Hint: For those with color pictures, the parameter estimate is multiplied by 1.) For two professors who received the same beauty rating, which color picture tends to have the higher course evaluation score?

```r
score_bty_color_lm <- lm(score ~ bty_avg + pic_color, data = evals)
summary(score_bty_color_lm)
```

```
##
## Call:
## lm(formula = score ~ bty_avg + pic_color, data = evals)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -1.8892 -0.3690  0.1293  0.4023  0.9125
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.06318    0.10908  37.249  < 2e-16 ***
## bty_avg         0.05548    0.01691   3.282  0.00111 **
## pic_colorcolor -0.16059    0.06892  -2.330  0.02022 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5323 on 460 degrees of freedom
## Multiple R-squared:  0.04628,    Adjusted R-squared:  0.04213
## F-statistic: 11.16 on 2 and 460 DF,  p-value: 1.848e-05
```

The equation for the score of professors with color pictures is score = 4.06318 + .05548 * bty_avg - .16059 * 1(for pic_color= color). Because the estimate is negative for color pictures, professors with black and white pictures will have a slightly higher score for all other considerations being equal.

Exercise 10

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

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

```
##
## Call:
## lm(formula = score ~ bty_avg + rank, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8713 -0.3642  0.1489  0.4103  0.9525
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.98155    0.09078  43.860  < 2e-16 ***
## bty_avg           0.06783    0.01655   4.098 4.92e-05 ***
## ranktenure track -0.16070    0.07395  -2.173   0.0303 *
## ranktenured      -0.12623    0.06266  -2.014   0.0445 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5328 on 459 degrees of freedom
## Multiple R-squared:  0.04652,    Adjusted R-squared:  0.04029
## F-statistic: 7.465 on 3 and 459 DF,  p-value: 6.88e-05
```

R handles multiple levels by creating an additional dummy variable. You will have a total of n-1 variables for n levels of a categorical variable. So the resulting equation is score = 3.98155 + .06783 * bty_avg - 0.16070 * 1 (if rank=tenure track) - .12623 * 1 (if rank=tenured).

Exercise 11

Which variable would you expect to have the highest p-value in this model? Why? Hint: Think about which variable would you expect to not have any association with the professor score.

A number of variables seem to me to not be influential in a professors level of atraction. I don't see how the number of credits a course is for would impact their atraction(cls_credits).

Exercise 12

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

```
m_full <- lm(score ~ rank + gender + ethnicity + 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 + gender + ethnicity + language + age +
##     cls_perc_eval + cls_students + cls_level + cls_profs + cls_credits +
##     bty_avg + pic_outfit + pic_color, data = evals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77397 -0.32432  0.09067  0.35183  0.95036
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4.0952141  0.2905277  14.096  < 2e-16 ***
## ranktenure track      -0.1475932  0.0820671  -1.798  0.07278 .
## ranktenured           -0.0973378  0.0663296  -1.467  0.14295
## gendermale             0.2109481  0.0518230   4.071 5.54e-05 ***
## ethnicitynot minority  0.1234929  0.0786273   1.571  0.11698
## languagenon-english   -0.2298112  0.1113754  -2.063  0.03965 *
## age                   -0.0090072  0.0031359  -2.872  0.00427 **
## cls_perc_eval          0.0053272  0.0015393   3.461  0.00059 ***
## cls_students           0.0004546  0.0003774   1.205  0.22896
## cls_levelupper         0.0605140  0.0575617   1.051  0.29369
## cls_profssingle       -0.0146619  0.0519885  -0.282  0.77806
## cls_creditsone credit  0.5020432  0.1159388   4.330 1.84e-05 ***
## bty_avg                0.0400333  0.0175064   2.287  0.02267 *
## pic_outfitnot formal  -0.1126817  0.0738800  -1.525  0.12792
## pic_colorcolor        -0.2172630  0.0715021  -3.039  0.00252 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.498 on 448 degrees of freedom
## Multiple R-squared:  0.1871, Adjusted R-squared:  0.1617
## F-statistic: 7.366 on 14 and 448 DF,  p-value: 6.552e-14
```

I was incorrect in my assumption. The p value for credits is significant. The number of professors teaching a course has the least significance.

Exercise 13

Interpret the coefficient associated with the ethnicity variable.

Non-minority professors have a slightly higher score in general by .1234.

Exercise 14

Drop the variable with the highest p-value and re-fit the model. Did the coefficients and significance of the other explanatory variables change? (One of the things that makes multiple regression interesting is that coefficient estimates depend on the other variables that are included in the model.) If not, what does this say about whether or not the dropped variable was collinear with the other explanatory variables?

```r
score_lm1 <- lm(score ~ rank + gender + ethnicity + language + age + cls_perc_eval
             + cls_students + cls_level + cls_credits + bty_avg
             + pic_outfit + pic_color, data = evals)
summary(score_lm1)
```

```
##
## Call:
## lm(formula = score ~ rank + gender + ethnicity + language + age +
##     cls_perc_eval + cls_students + cls_level + cls_credits +
##     bty_avg + pic_outfit + pic_color, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7836 -0.3257  0.0859  0.3513  0.9551
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4.0872523  0.2888562  14.150  < 2e-16 ***
## ranktenure track      -0.1476746  0.0819824  -1.801 0.072327 .
## ranktenured           -0.0973829  0.0662614  -1.470 0.142349
## gendermale             0.2101231  0.0516873   4.065 5.66e-05 ***
## ethnicitynot minority  0.1274458  0.0772887   1.649 0.099856 .
## languagenon-english   -0.2282894  0.1111305  -2.054 0.040530 *
## age                   -0.0089992  0.0031326  -2.873 0.004262 **
## cls_perc_eval          0.0052888  0.0015317   3.453 0.000607 ***
## cls_students           0.0004687  0.0003737   1.254 0.210384
## cls_levelupper         0.0606374  0.0575010   1.055 0.292200
## cls_creditsone credit  0.5061196  0.1149163   4.404 1.33e-05 ***
## bty_avg                0.0398629  0.0174780   2.281 0.023032 *
## pic_outfitnot formal  -0.1083227  0.0721711  -1.501 0.134080
## pic_colorcolor        -0.2190527  0.0711469  -3.079 0.002205 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4974 on 449 degrees of freedom
## Multiple R-squared:  0.187,  Adjusted R-squared:  0.1634
## F-statistic: 7.943 on 13 and 449 DF,  p-value: 2.336e-14
```

The coefficients barely change which does indeed suggest that cls_profs is highly correlated with another variable in the model.

Exercise 15

Using backward-selection and p-value as the selection criterion, determine the best model. You do not need to show all steps in your answer, just the output for the final model. Also, write out the linear model for predicting score based on the final model you settle on.

```
score_lm5 <- lm(score ~ gender + ethnicity + language + age + cls_perc_eval
             + cls_credits + bty_avg + pic_color, data = evals)
summary(score_lm5)
```

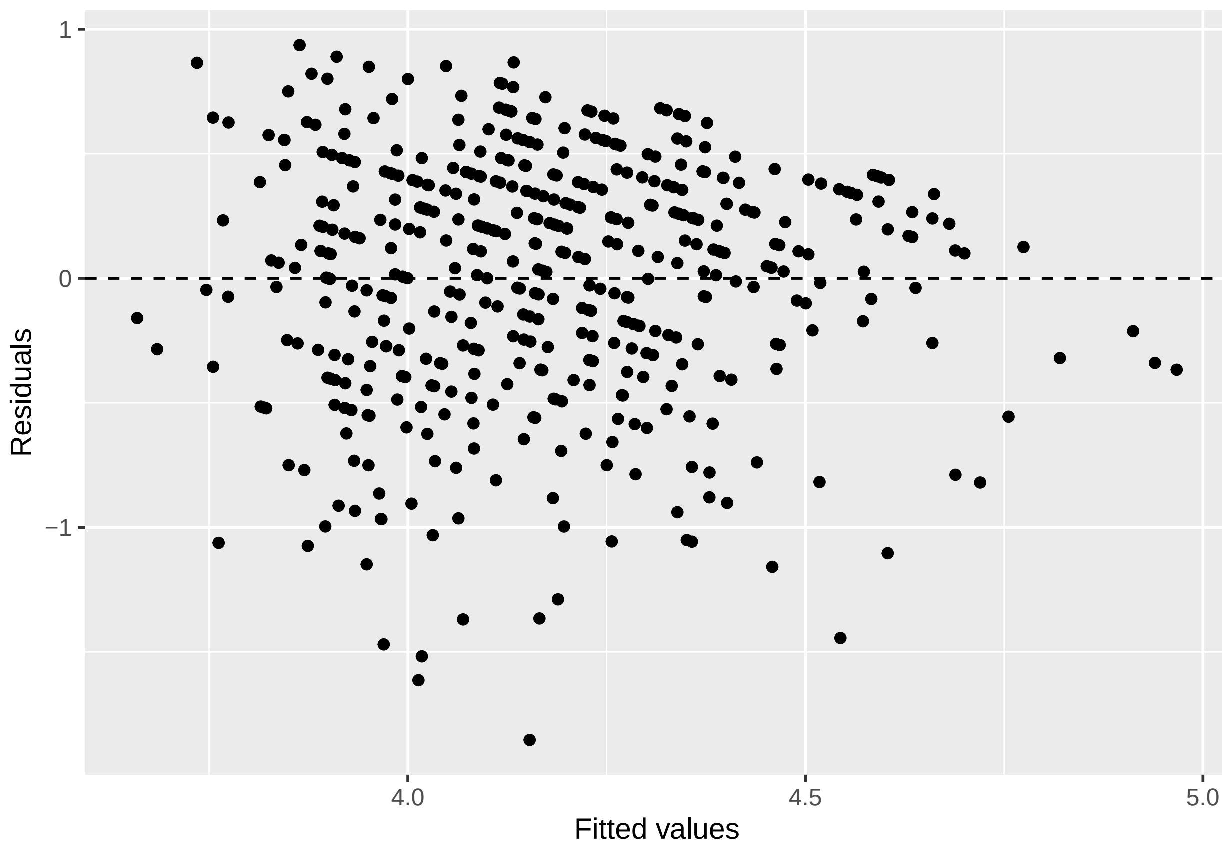
```
##
## Call:
## lm(formula = score ~ gender + ethnicity + language + age + cls_perc_eval +
##     cls_credits + bty_avg + pic_color, data = evals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.85320 -0.32394  0.09984  0.37930  0.93610
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.771922   0.232053  16.255  < 2e-16 ***
## gendermale              0.207112   0.050135   4.131 4.30e-05 ***
## ethnicitynot minority   0.167872   0.075275   2.230  0.02623 *
## languagenon-english    -0.206178   0.103639  -1.989  0.04726 *
## age                    -0.006046   0.002612  -2.315  0.02108 *
## cls_perc_eval           0.004656   0.001435   3.244  0.00127 **
## cls_creditsone credit   0.505306   0.104119   4.853 1.67e-06 ***
## bty_avg                 0.051069   0.016934   3.016  0.00271 **
## pic_colorcolor         -0.190579   0.067351  -2.830  0.00487 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4992 on 454 degrees of freedom
## Multiple R-squared:  0.1722, Adjusted R-squared:  0.1576
## F-statistic:  11.8 on 8 and 454 DF,  p-value: 2.58e-15
```

Stopping backwards elimination once all p values are below an alpha of .05 yields an equation of: Score = 3.771922 + .207112 * male + .167872 * not minority - .206178 * non_english - .006046 * age + 0.004656 * cls_perc_eval + .505306 * one credit + .051069 * bty_avg - 0.190579 * color

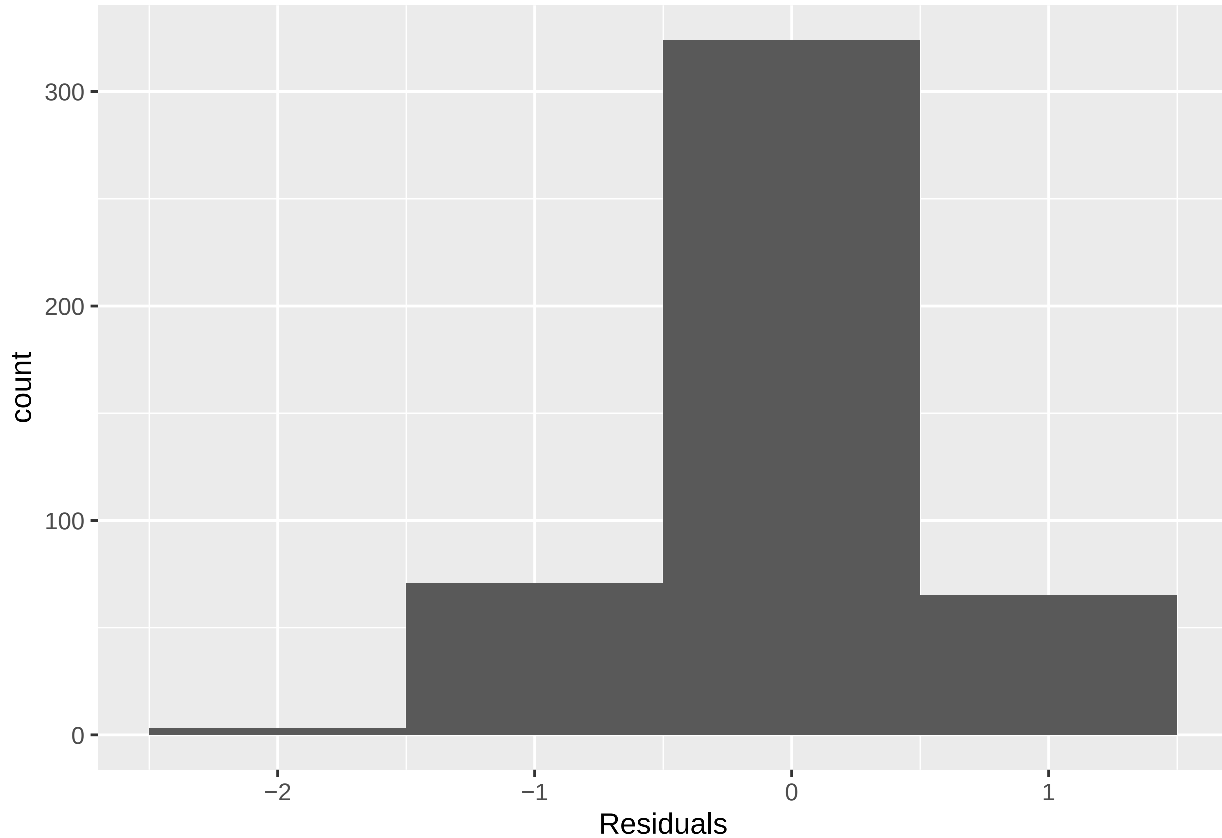
Exercise 16

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

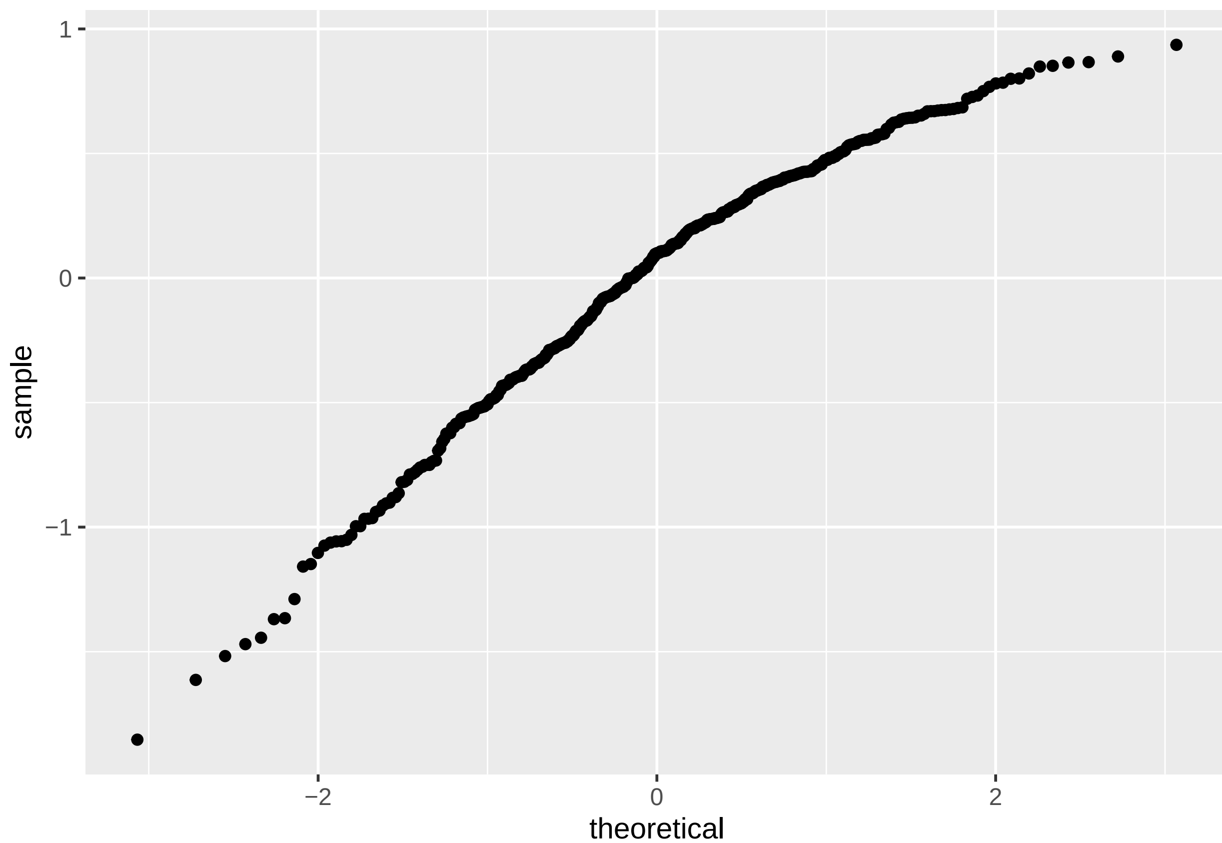
```
ggplot(data = score_lm5, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```




```
ggplot(data = score_lm5, aes(x = .resid)) +
  geom_histogram(binwidth = 1) +
  xlab("Residuals")
```




```
ggplot(data = score_lm5, aes(sample = .resid)) +
  stat_qq()
```



It appears that the variance may be problematic for higher fitted values.

Exercise 17

The original paper describes how these data were gathered by taking a sample of professors from the University of Texas at Austin and including all courses that they have taught. Considering that each row represents a course, could this new information have an impact on any of the conditions of linear regression?

This may present issues with violating the independence requirements for linear regression. A given student may have taken professors for multiple courses.

Exercise 18

Based on your final model, describe the characteristics of a professor and course at University of Texas at Austin that would be associated with a high evaluation score.

A higher score would be associate with a male, non-minority, English speaking, younger, more attractive, having a black and white picture professor and a course that had good voter turnout and was 1 credit.

Exercise 19

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

I would be comfortable generalizing this model to in person courses at US universities. Most of the model seems to be based on common US demographic biases. Even accounting for reformulating language for the native language I would think different cultures around the world have a lot of variance in implicit biases.