Code ▾

# Lab 9 - Multiple Linear Regression

Julian Adames-Ng

2022-10-24

Hide

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

## 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. No treatment was applied that influenced the outcome of the study. The question on whether beauty leads directly to differences in course evaluations is not easy to answer. Although there may well be a relationship between beauty and course evaluations, there may be other variables at play that influence the outcome of the evaluations. Simply changing the question to, “Is there a correlation or relationship between beauty and course evaluation outcomes?”

Hide

```
# Insert code for Exercise 1 here

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, …
```

Hide

```
?evals
```

# 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?

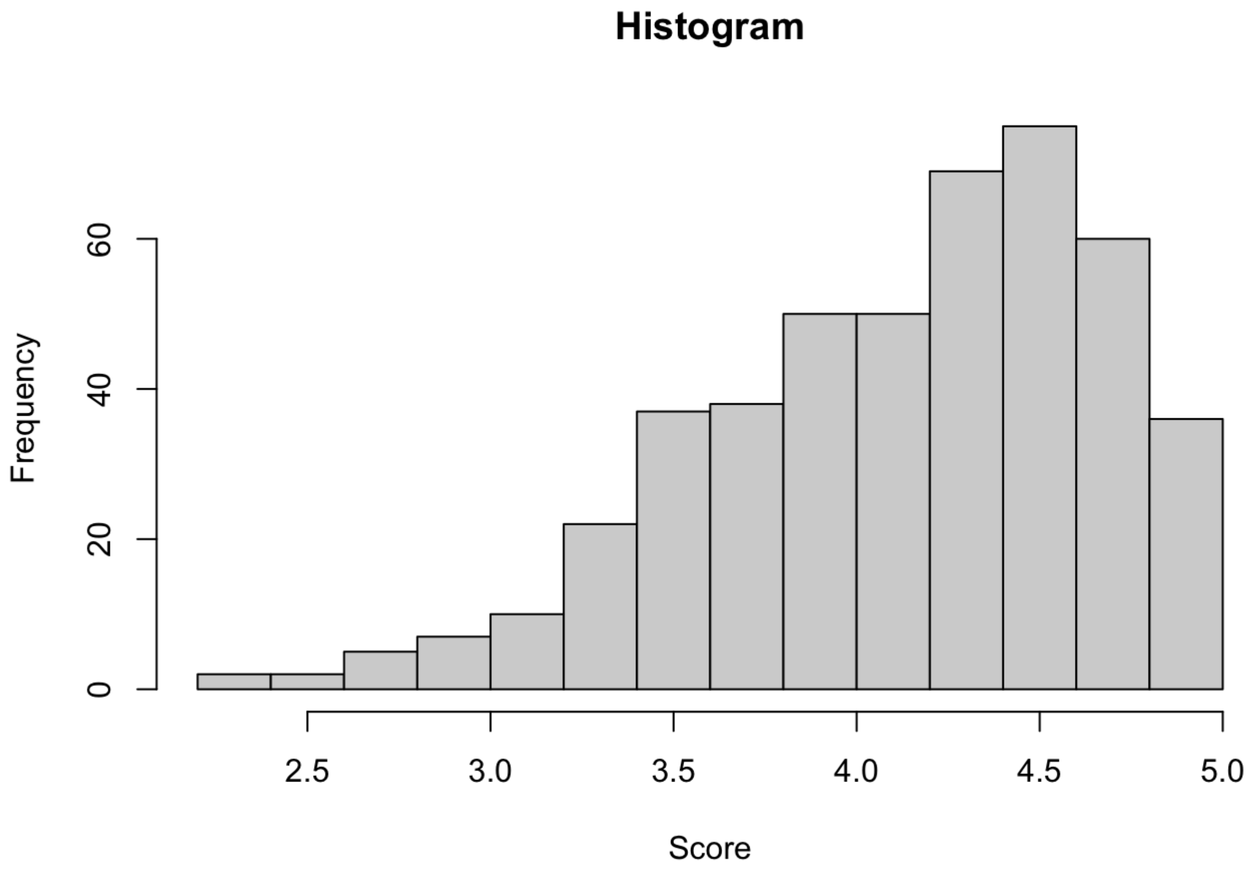
-The distribution appears to be skewed left. More students tended to leave positive evaluation marks. My expectation was for a normal distribution.

…

Hide

```
# Insert code for Exercise 2 here

hist(evals$score, main = "Histogram", xlab = "Score")
```



## Exercise 3

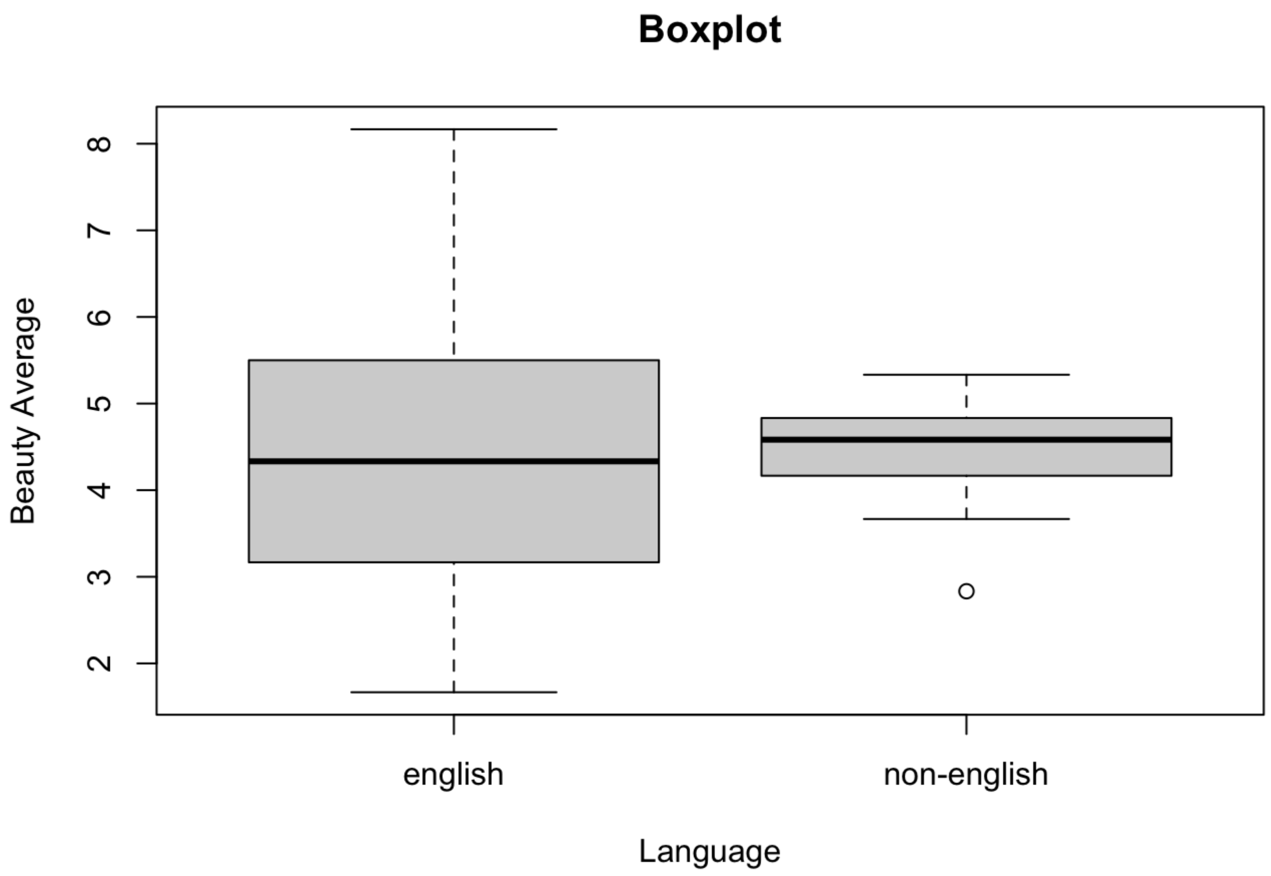
Excluding score, select two other variables and describe their relationship with each other using an appropriate visualization.

…

Hide

```
# Insert code for the exercise here

boxplot(evals$bty_avg ~ evals$language, main = "Boxplot", ylab = "Beauty Average",
        xlab = "Language")
```



## Exercise 4

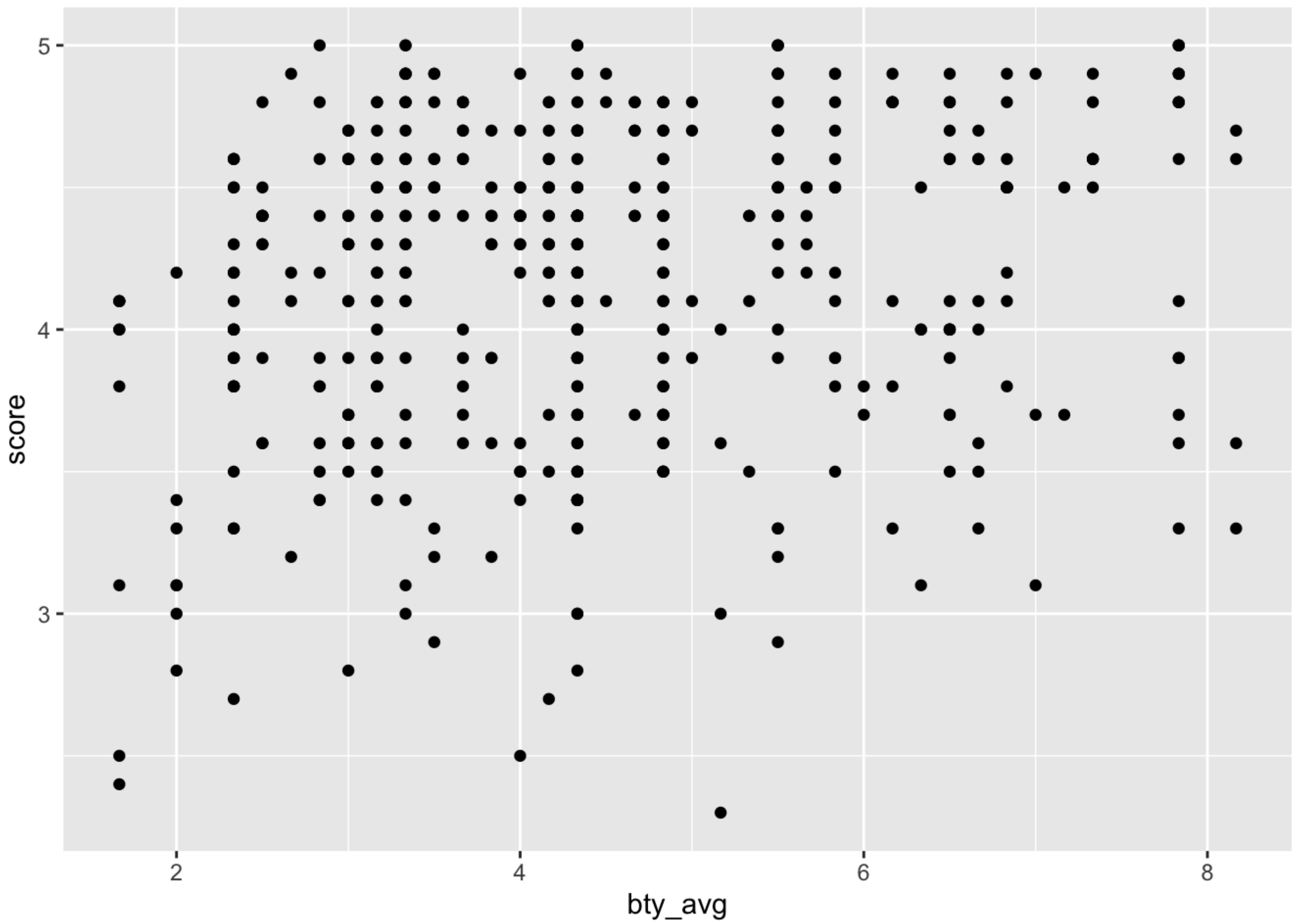
Replot the scatterplot, but this time use geom_jitter as your layer. What was misleading about the initial scatterplot?

-The evaluation marks in the first plot seemed to overlap. This makes it more difficult to show a relationship between the variables. …

Hide

```
# Insert code for the exercise here

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



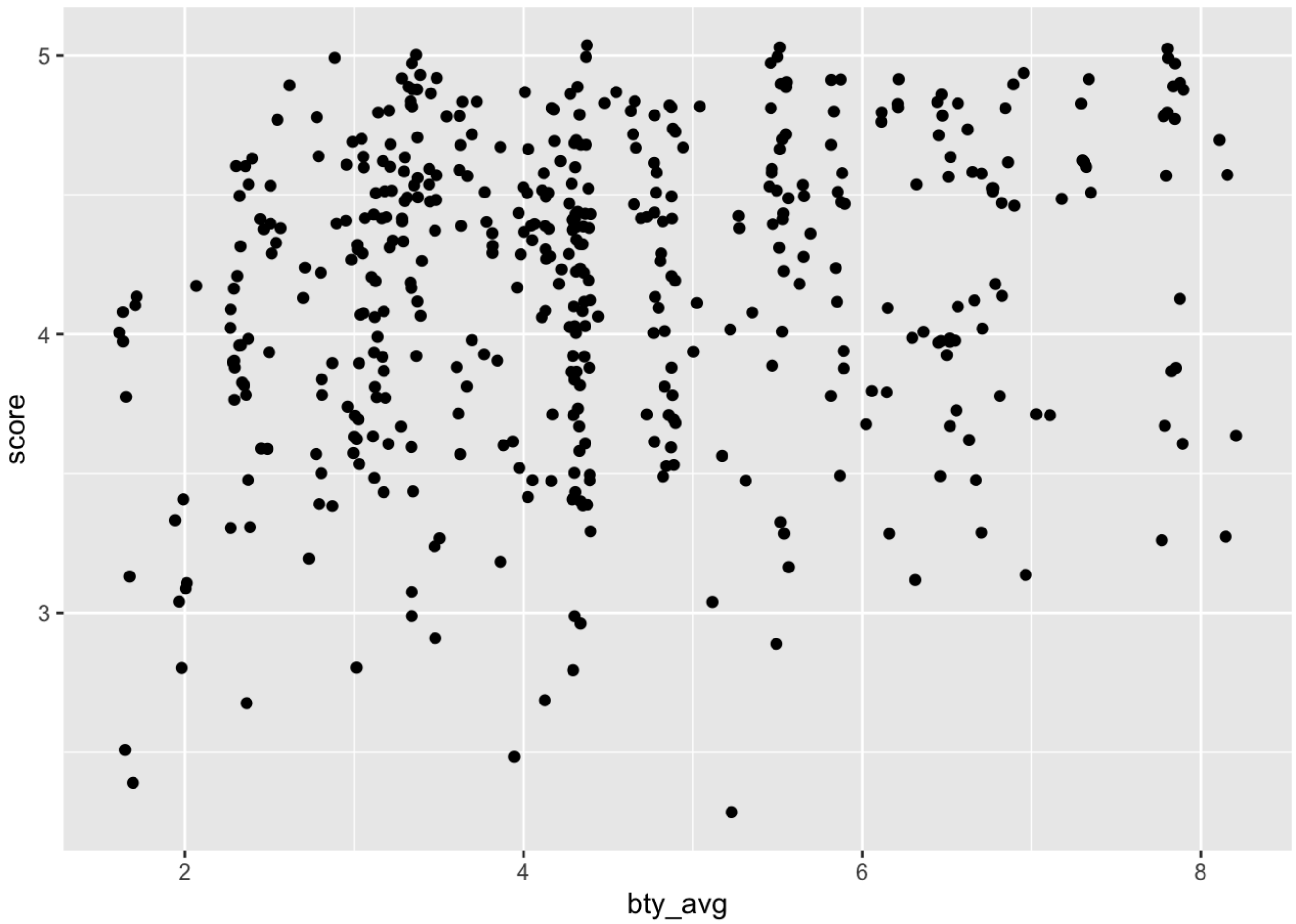
Hide

```
nrow(evals)
```

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

Hide

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



## 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?

-The equation for the linear model is score = 3.88034 + 0.06664*(bty_avg). From the slope, we can say that as the beauty average score increases by one increment, the evaluation rating increases by 0.06664. The average beauty score is a statistically significant predictor because the p_value is an extremely small number, however it fails to show practical significance because a one increment increase amounts to a considerably lesser increase in evaluation score.

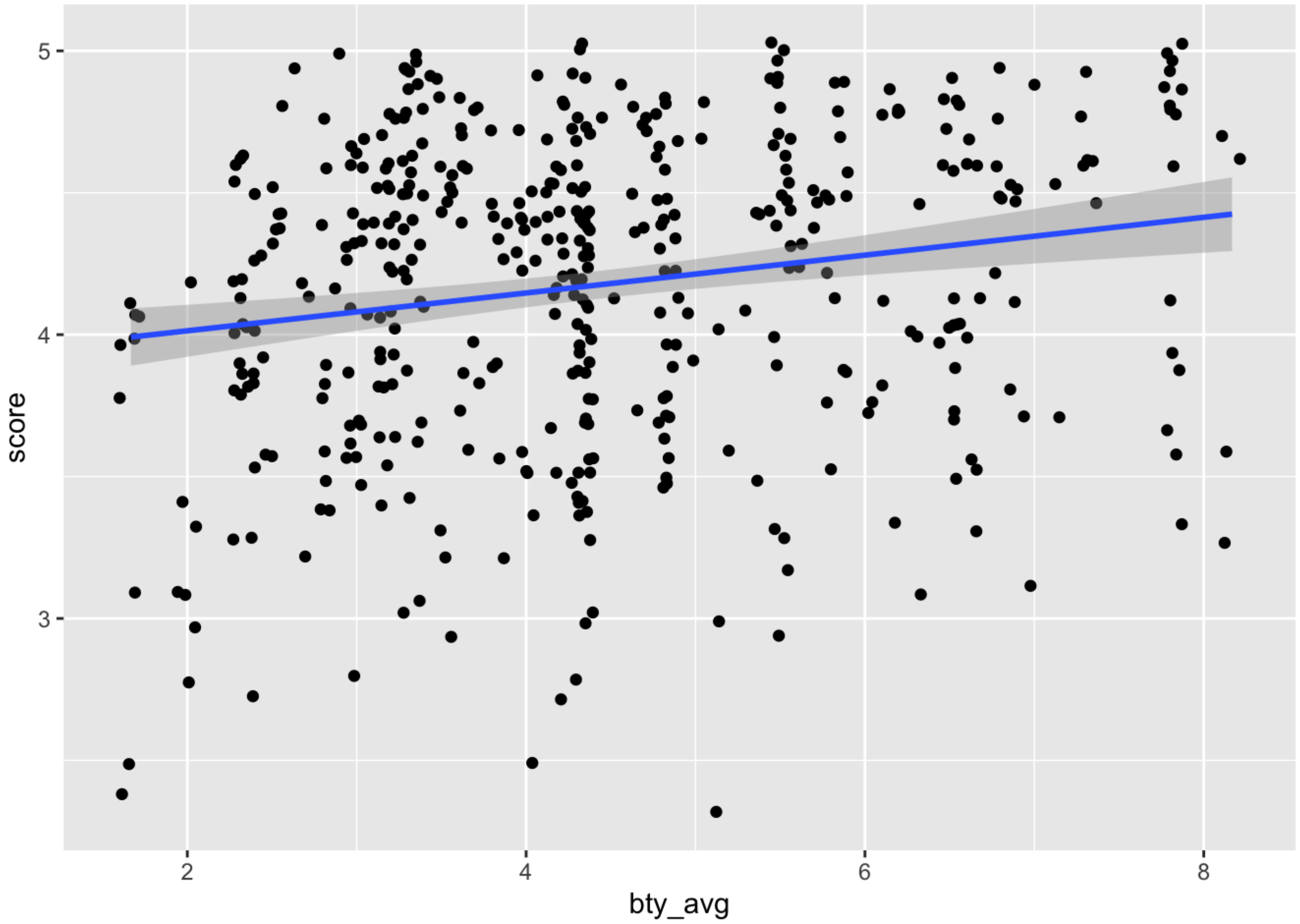
…

Hide

```
# Insert code for the exercise here

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

```
## `geom_smooth()` using formula 'y ~ x'
```



Hide

```
m_bty <- lm(evals$score ~ evals$bty_avg)
summary(m_bty)
```

```
##
## Call:
## lm(formula = evals$score ~ evals$bty_avg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9246 -0.3690  0.1420  0.3977  0.9309
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.88034    0.07614   50.96  < 2e-16 ***
## evals$bty_avg  0.06664    0.01629    4.09 5.08e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5348 on 461 degrees of freedom
## Multiple R-squared:  0.03502,    Adjusted R-squared:  0.03293
## F-statistic: 16.73 on 1 and 461 DF,  p-value: 5.083e-05
```

# 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).

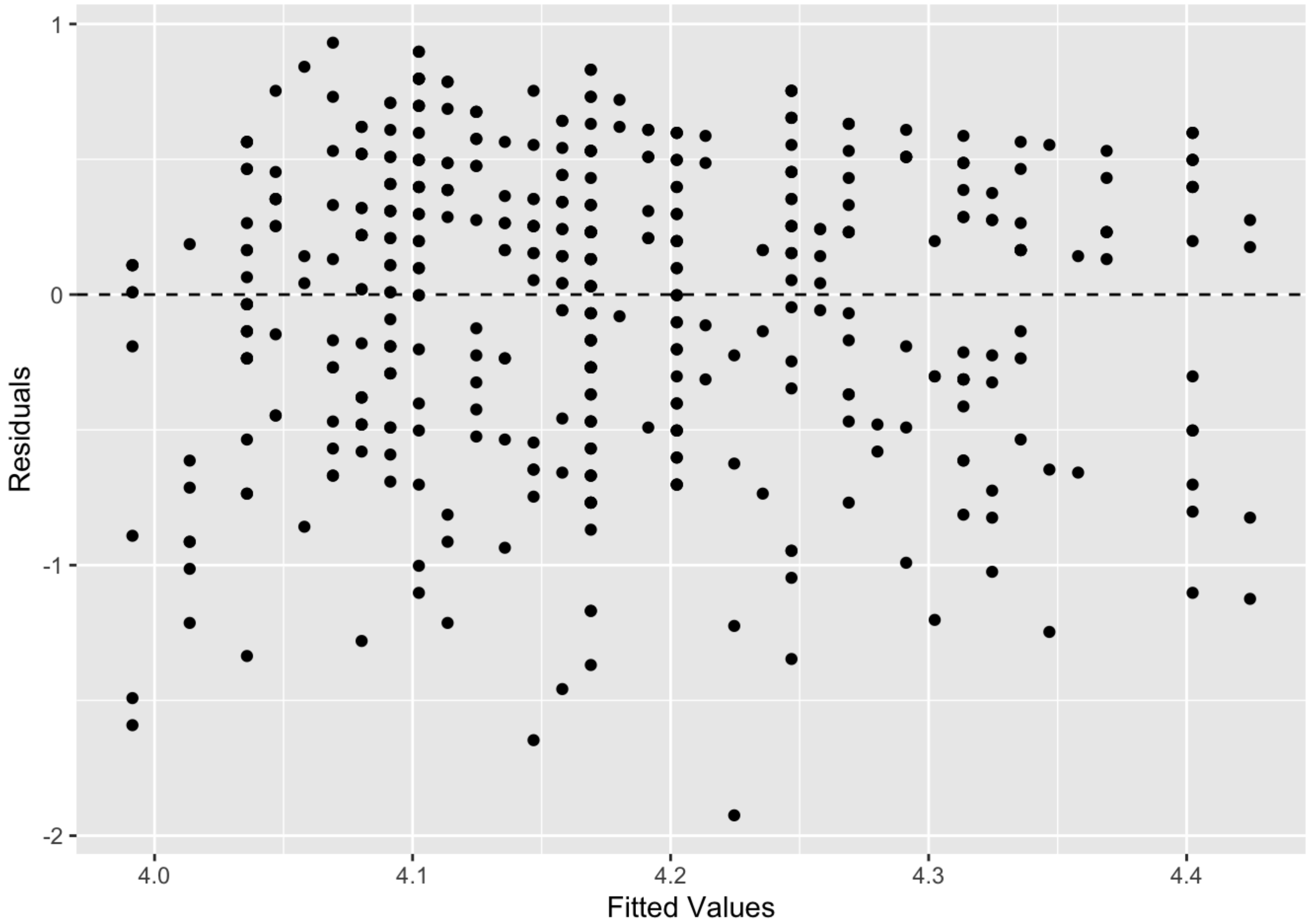
-

…

Hide

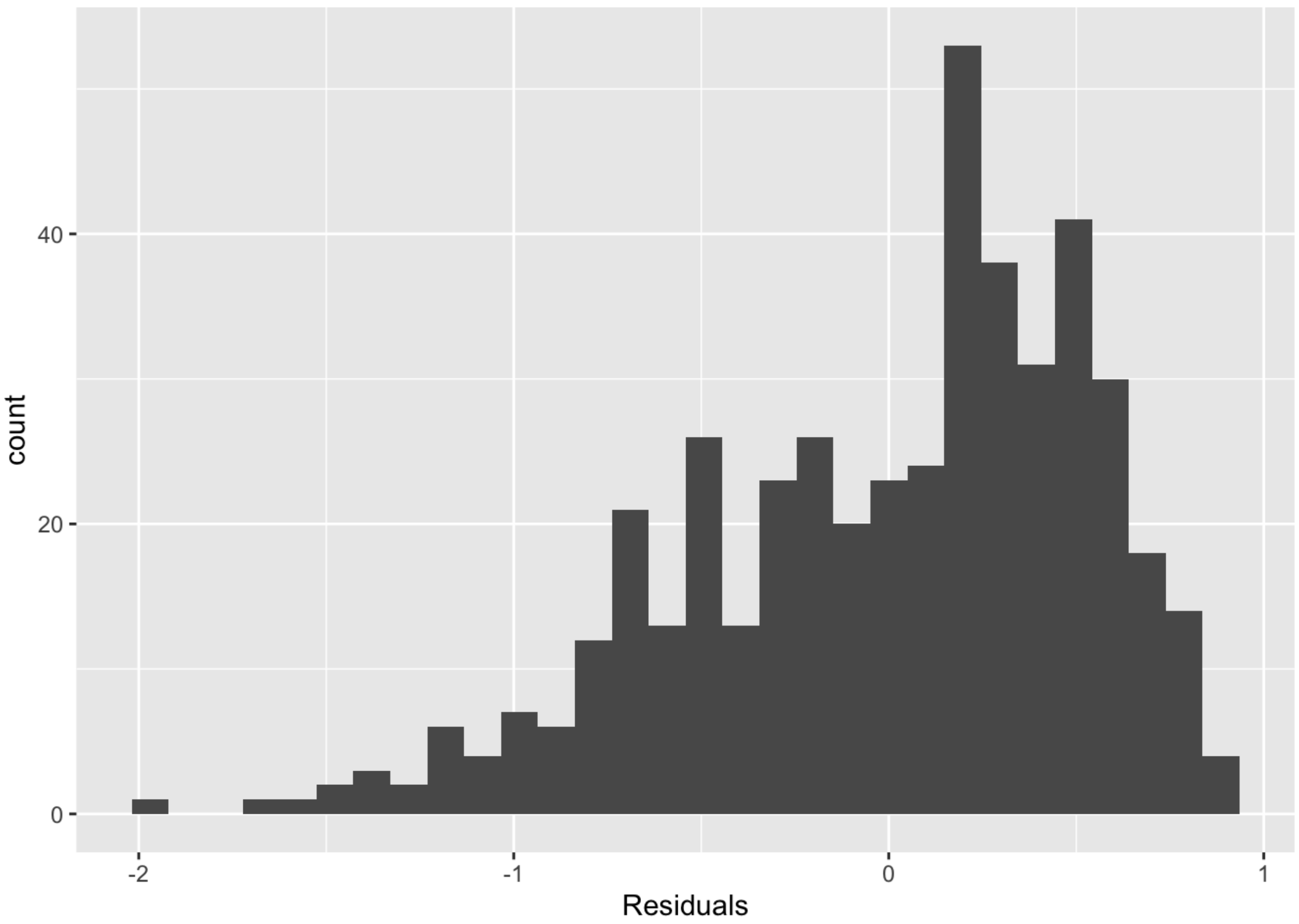
```
# Insert code for the exercise here

ggplot(m_bty, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = 'dashed') +
  xlab('Fitted Values') +
  ylab('Residuals')
```



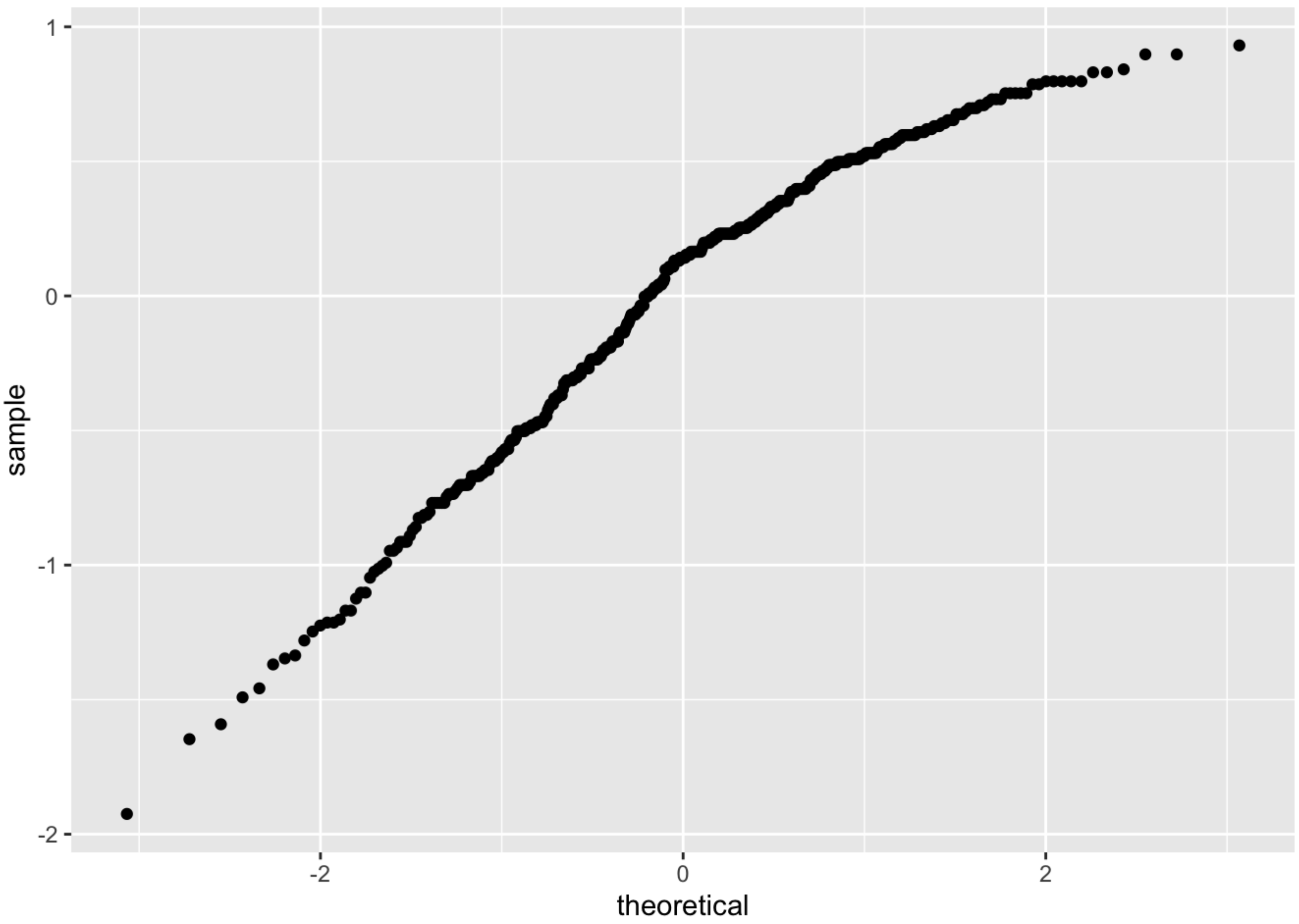
Hide

```
ggplot(m_bty, aes(x = .resid)) +
  geom_histogram(bins = 30) +
  xlab("Residuals")
```



Hide

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



## 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.

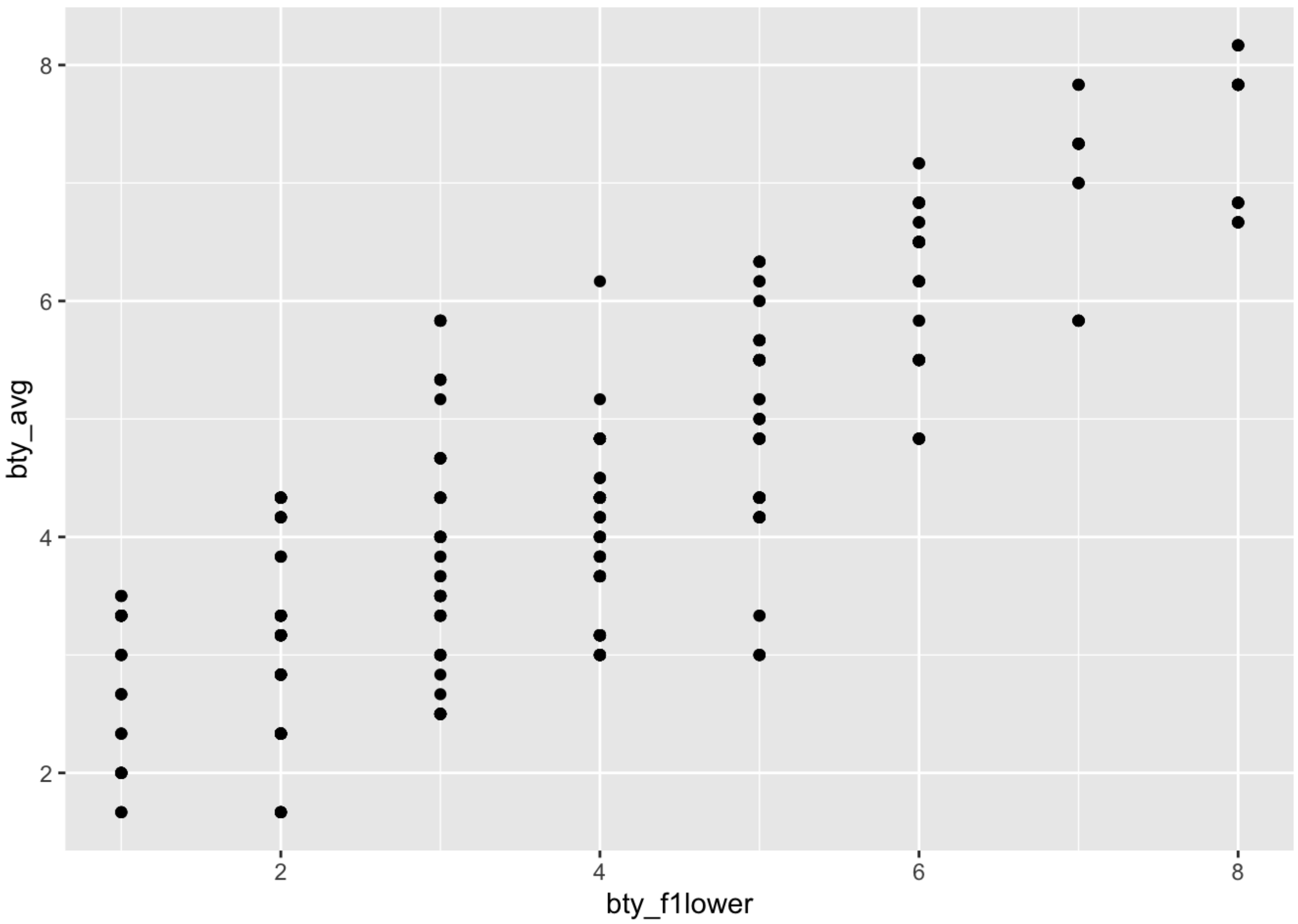
-The conditions for the regression are indeed reasonable based on the following diagnostic plots.

…

Hide

```
# Insert code for the exercise here

ggplot(data = evals, aes(x = bty_f1lower, y = bty_avg)) +
  geom_point()
```



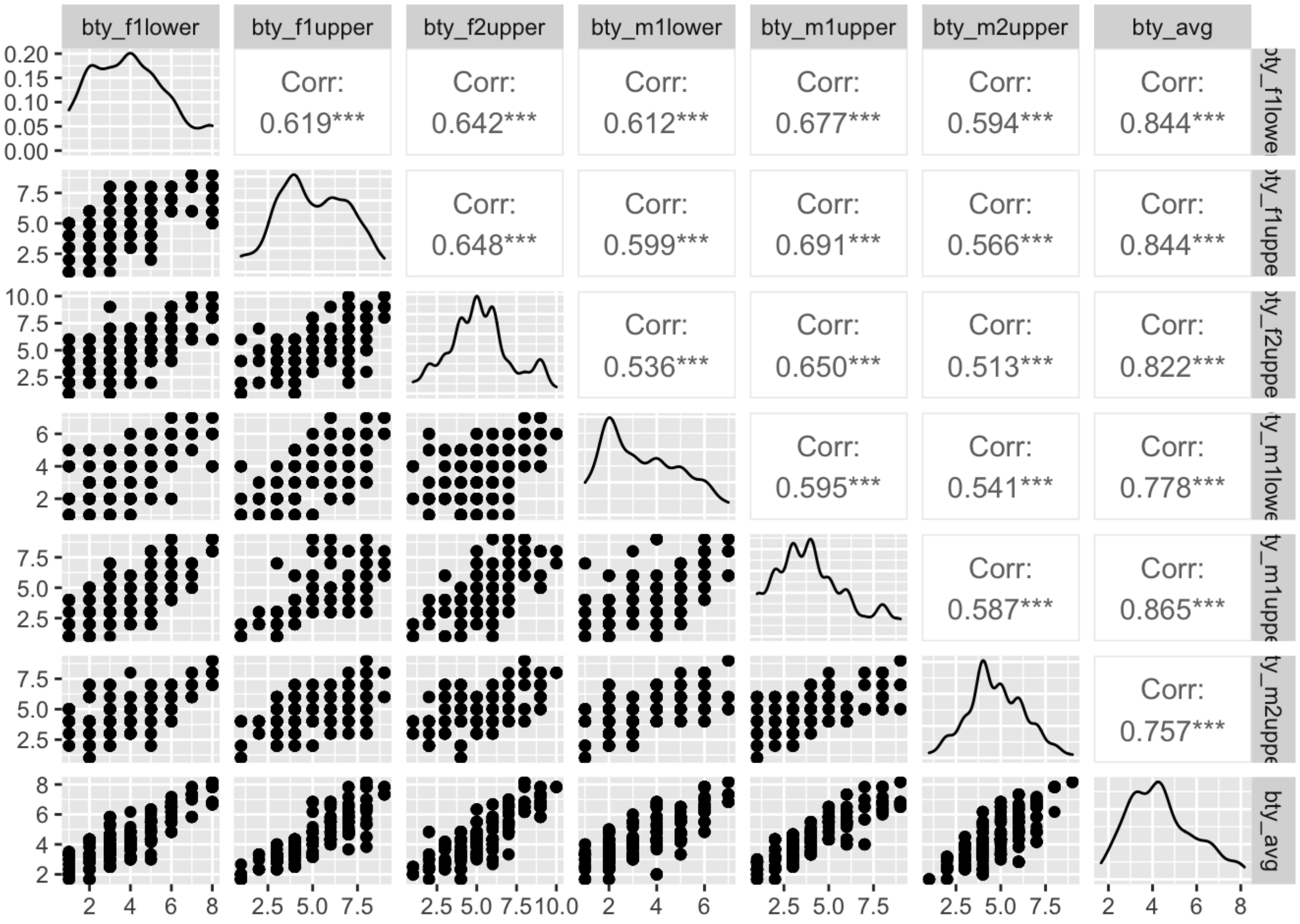
Hide

```
evals %>%
  summarise(cor(bty_avg, bty_f1lower))
```

```
## # A tibble: 1 × 1
##   `cor(bty_avg, bty_f1lower)`
##                         <dbl>
## 1                       0.844
```

Hide

```
evals %>%
  select(contains("bty")) %>%
  ggpairs()
```



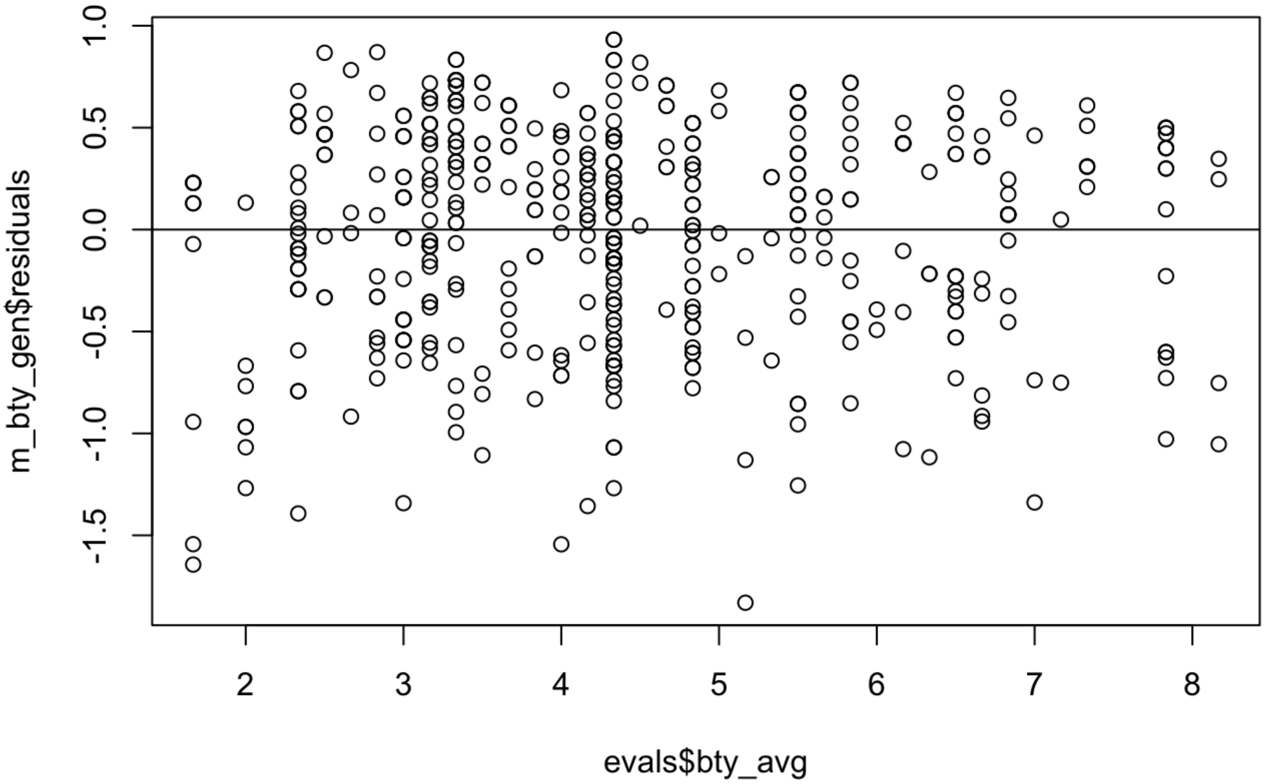
Hide

```
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
```

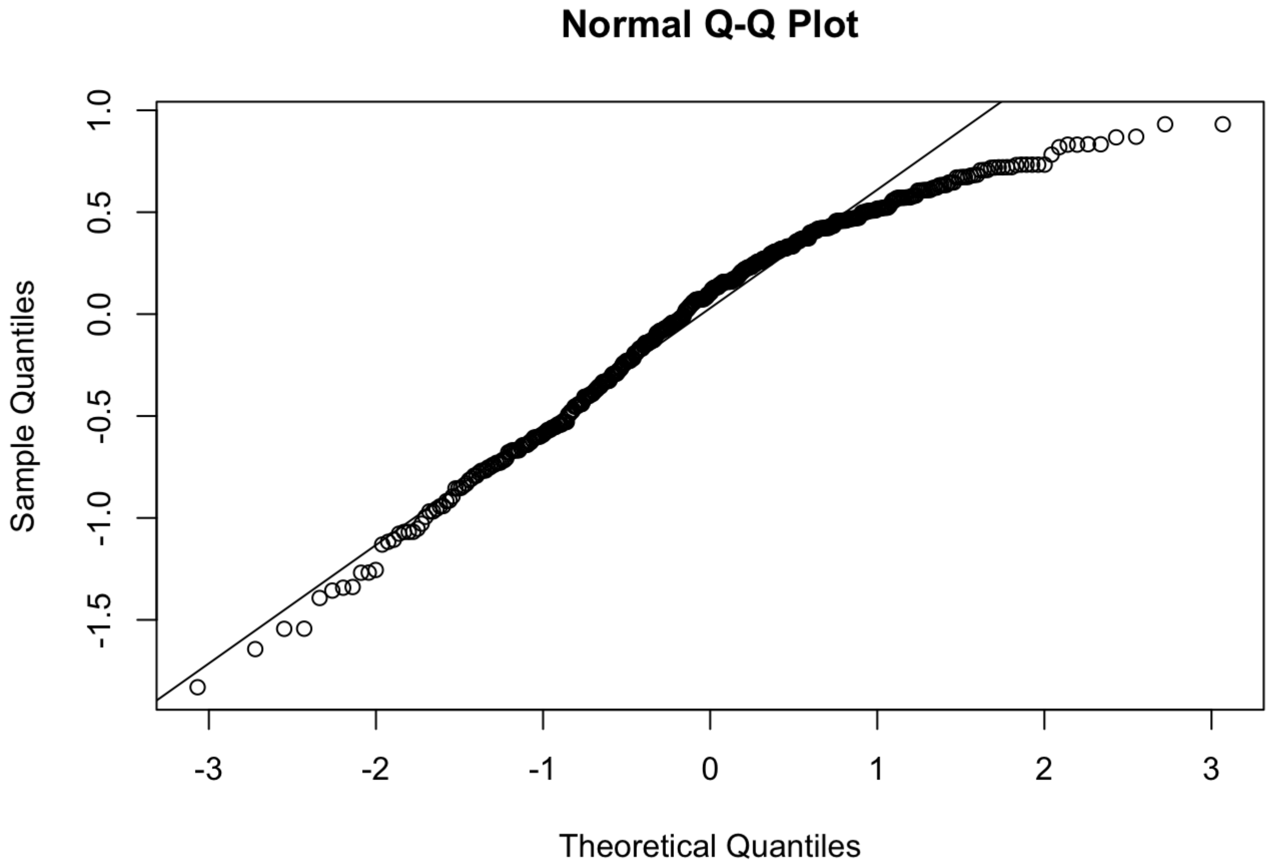
Hide

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



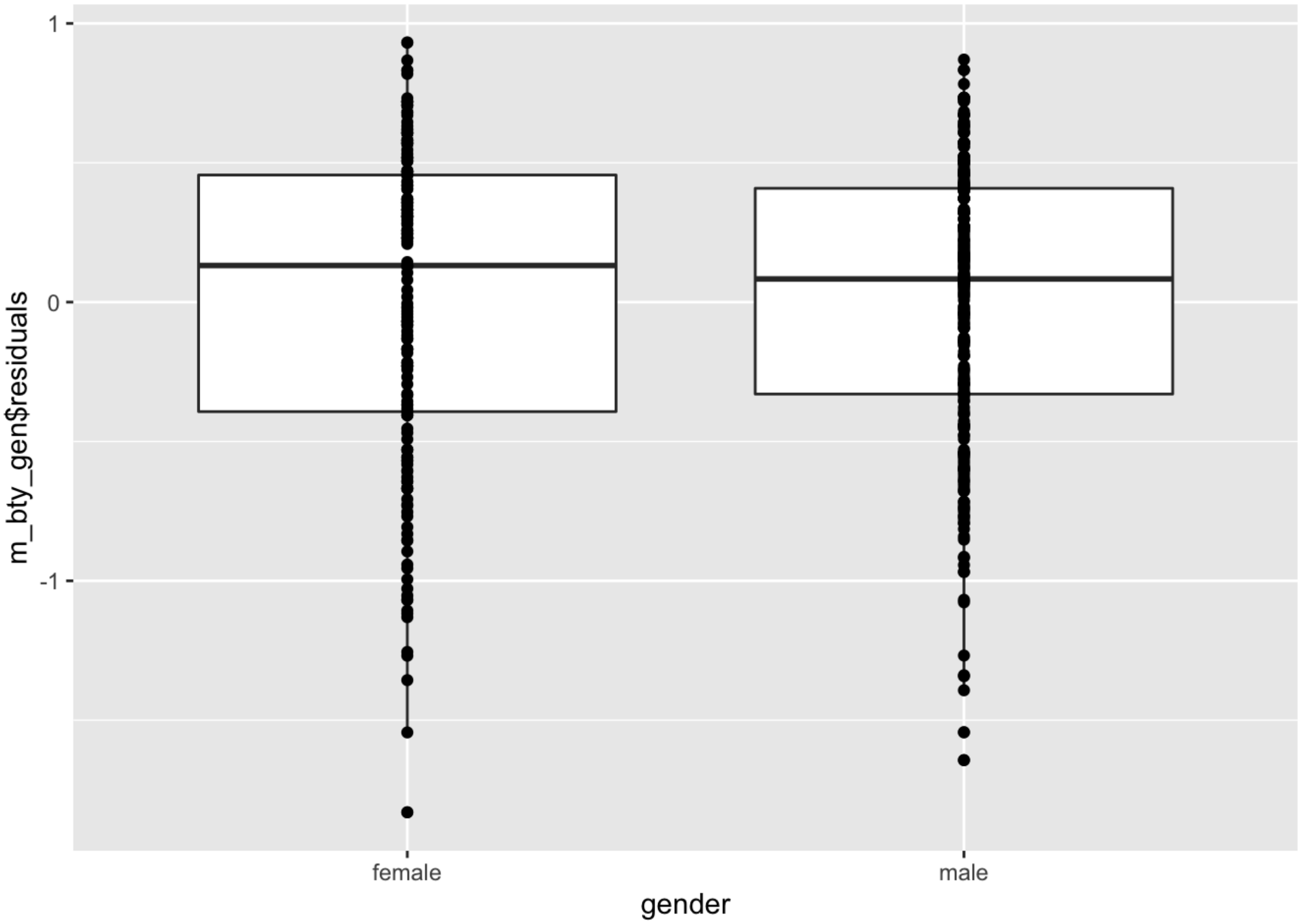
Hide

```
qqnorm(m_bty_gen$residuals)
qqline(m_bty_gen$residuals)
```



Hide

```
ggplot(evals, aes(gender, m_bty_gen$residuals)) +
  geom_boxplot() +
  geom_point()
```



## 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?

-Beauty average is still a significant predictor of score. Adding gender altered the parameter estimate for beauty average.

…

Hide

```
# Insert code for the exercise here

summary(m_bty)
```

```
##
## Call:
## lm(formula = evals$score ~ evals$bty_avg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9246 -0.3690  0.1420  0.3977  0.9309
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.88034    0.07614   50.96  < 2e-16 ***
## evals$bty_avg  0.06664    0.01629    4.09 5.08e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5348 on 461 degrees of freedom
## Multiple R-squared:  0.03502,    Adjusted R-squared:  0.03293
## F-statistic: 16.73 on 1 and 461 DF,  p-value: 5.083e-05
```

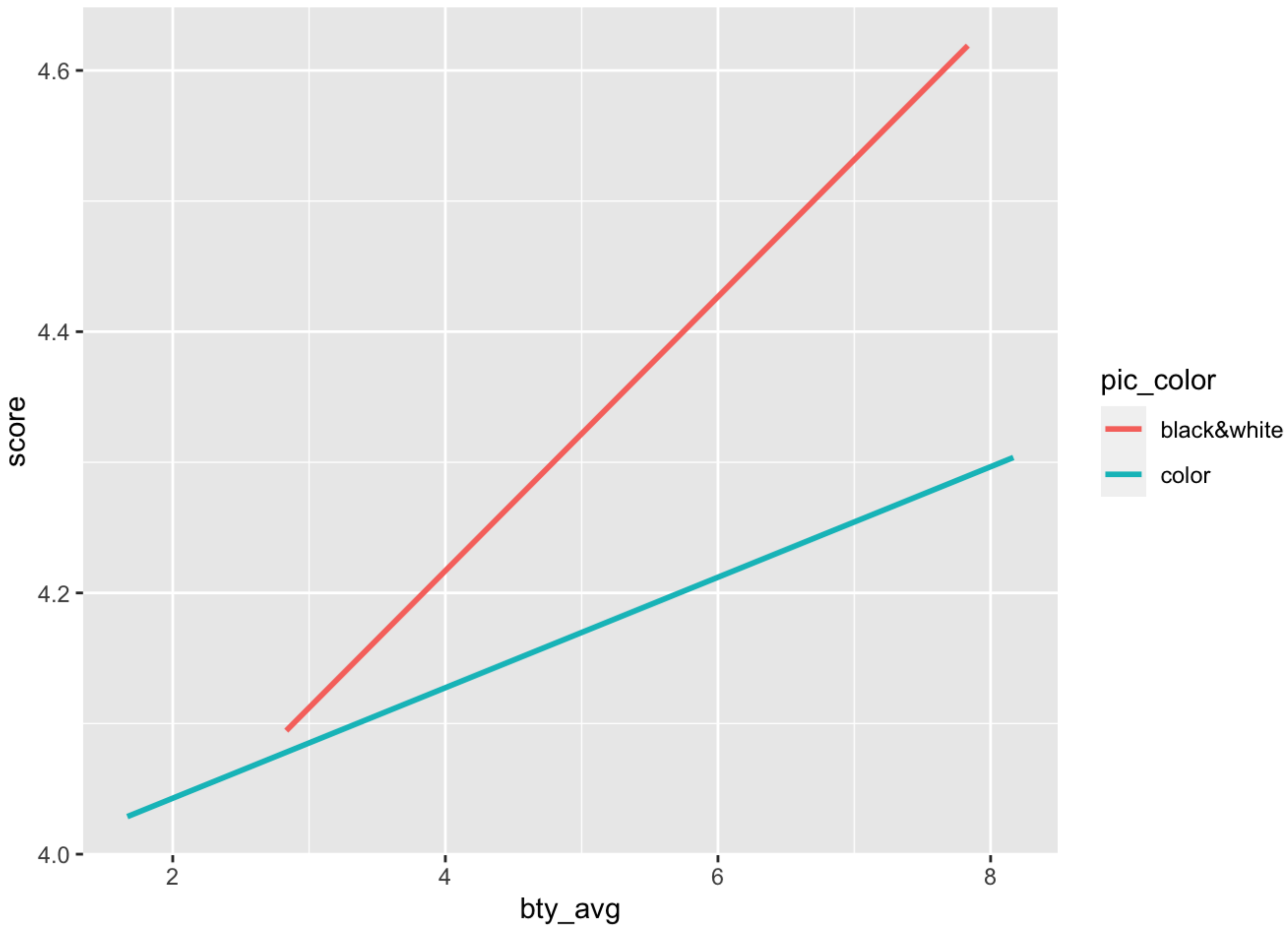
Hide

```
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
```

Hide

```
ggplot(data = evals, aes(x = bty_avg, y = score, color = pic_color)) +
geom_smooth(method = "lm", formula = y ~ x, se = FALSE)
```



## 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?

-The regression line is evaluation score = 4.06318 + 0.05548*(bty_avg) - 0.16059* (pic_color)

black&white color picture tends to have the higher course evaluation score when selecting two professors with the same beauty rating.

…

Hide

```
# Insert code for the exercise here

summary (lm(score ~ bty_avg + pic_color, data = evals   ))
```

```
##
## 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
```

# 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.

-Rank has three levels which are teaching, tenure track, tenured. So R creates two indicator variables:

one variable for tenure track and one variable for tenured.

Teaching is used as a reference level so it does not show up in the output.

…

Hide

```
# Insert code for the exercise here

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
```

# 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.

-The variable with the highest p-value would probably be cls_profs. The number of professors teaching these sections should not be associated with the professor score.

…

Hide

```
# Insert code for the exercise here

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
```

Hide

```
?m_full
```

```
## No documentation for 'm_full' in specified packages and libraries:
## you could try '??m_full'
```

# Exercise 12

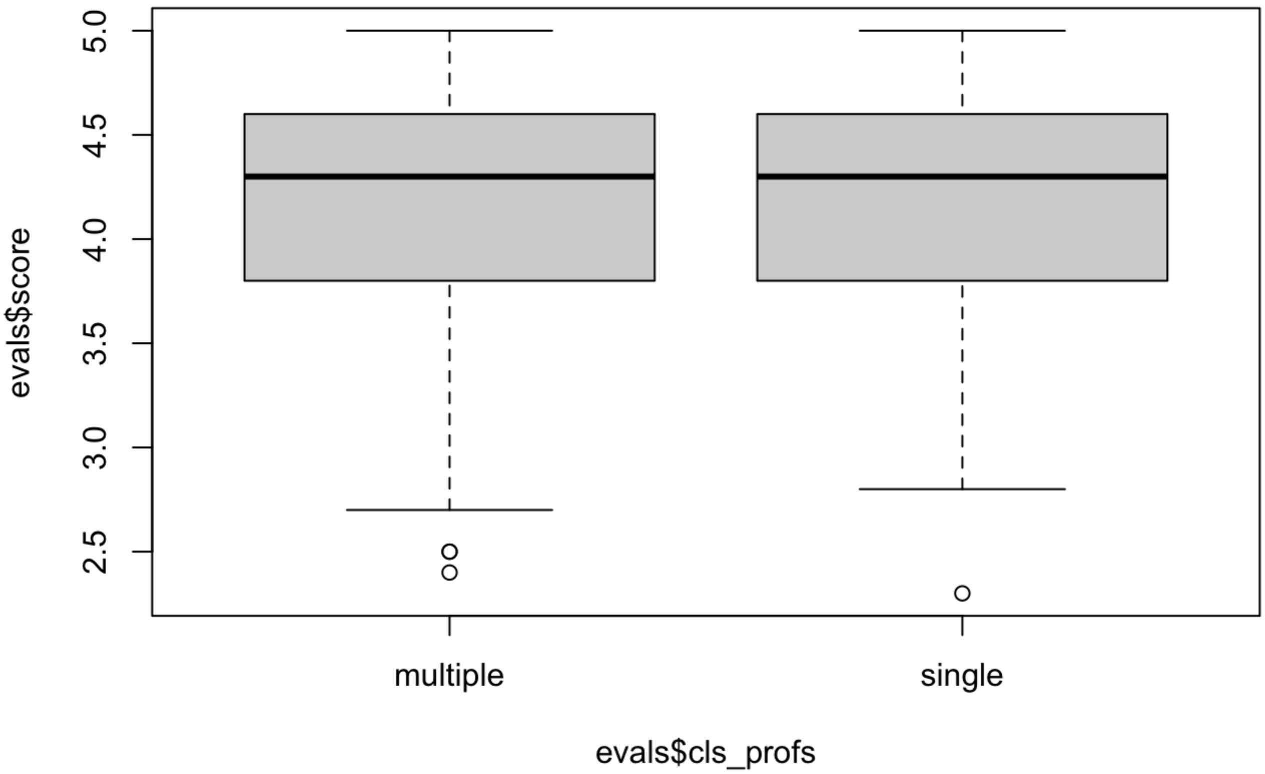
Check your suspicions from the previous exercise. Include the model output in your response.

-As expected, there is no change in the side by side plots. …

Hide

```
# Insert code for the exercise here

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



## Exercise 13

Interpret the coefficient associated with the ethnicity variable.

-With every increment increase for a non-minority professor, the evaluation score tends to increase by 0.1234929.

…

## 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?

-After dropping the variable with the highest p-value, the coefficients and significance changed a bit.

…

Hide

```
# Insert code for the exercise here

m_full_1 <- lm(score ~ rank + ethnicity + gender + language + age + cls_perc_eval
             + cls_students + cls_level + cls_credits + bty_avg
             + pic_outfit + pic_color, data = evals)
summary(m_full_1)
```

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

## 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.

-The linear model is as follows:

eval_score = 3.772 + 0.207(gender) + 0.168(ethnicity) - 0.206(language) - 00.6(age) + 0.005(cls_perc_eval) + 0.505(cls_credits) + 0.051(bty_avg) - 0.191(pic_color)

…

Hide

```
# Insert code for the exercise here

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

```
##
## Call:
## lm(formula = score ~ ethnicity + gender + 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 ***
## ethnicitynot minority  0.167872   0.075275   2.230  0.02623 *
## gendermale             0.207112   0.050135   4.131 4.30e-05 ***
## 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
```

# Exercise 16

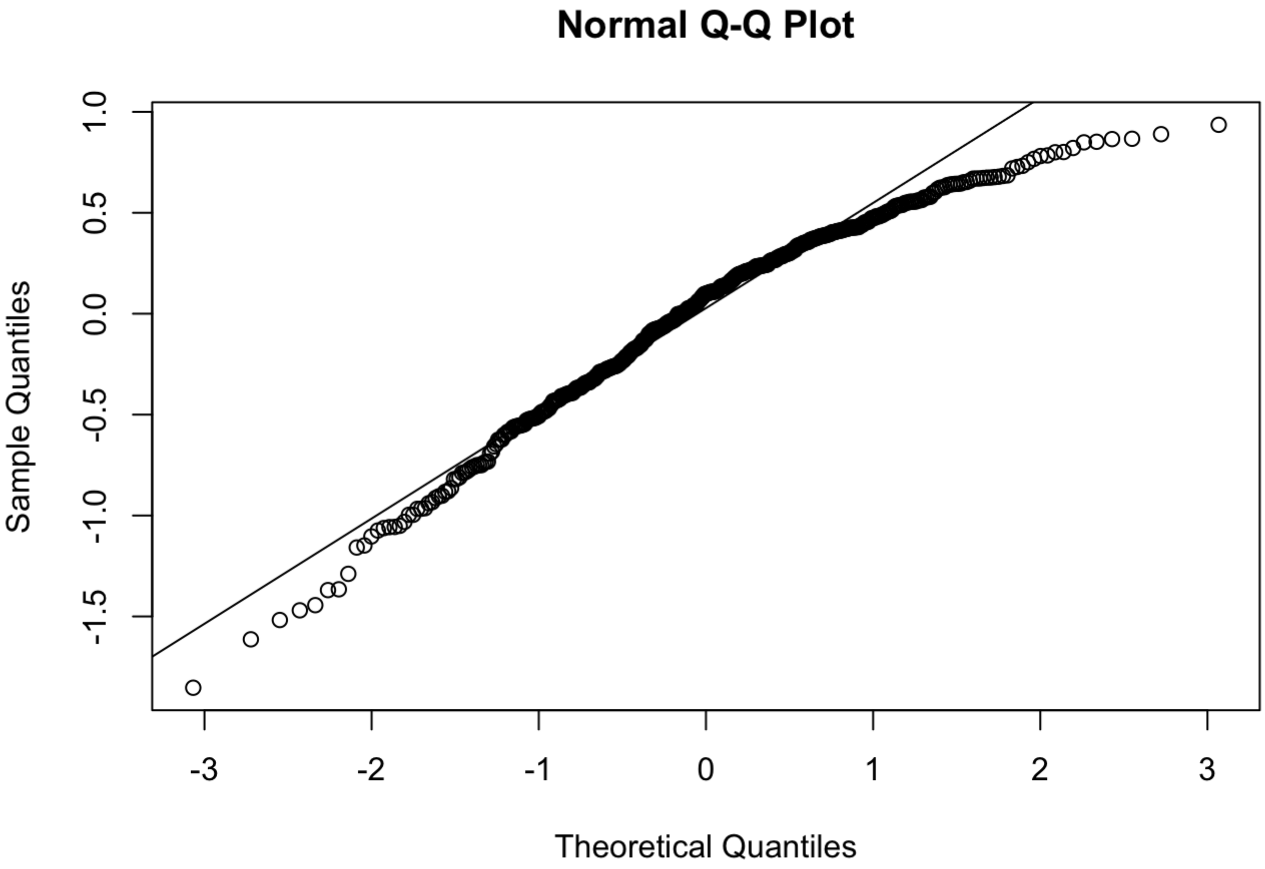
Verify that the conditions for this model are reasonable using diagnostic plots.

…

Hide

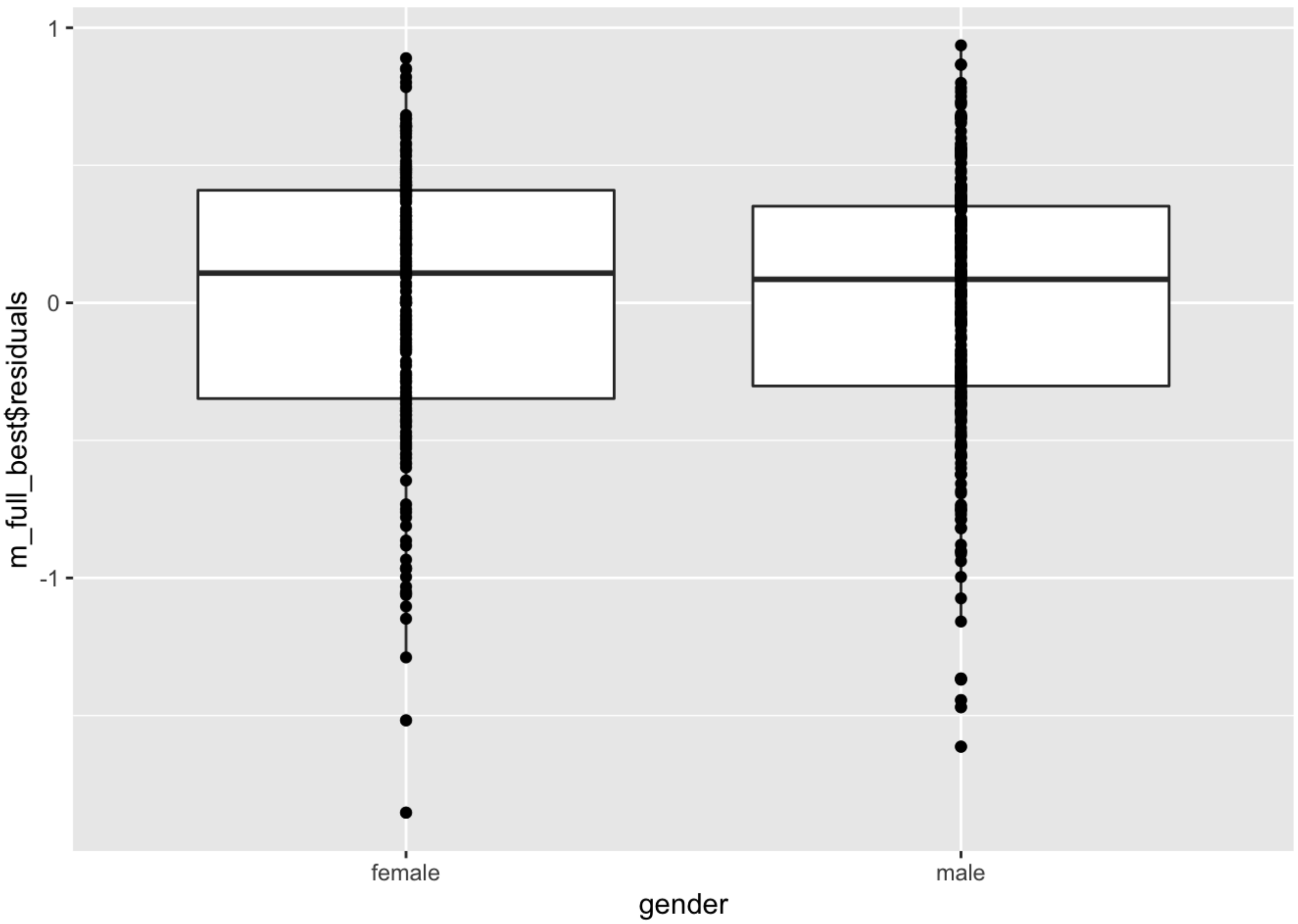
```
# Insert code for the exercise here

qqnorm(m_full_best$residuals)
qqline(m_full_best$residuals)
```



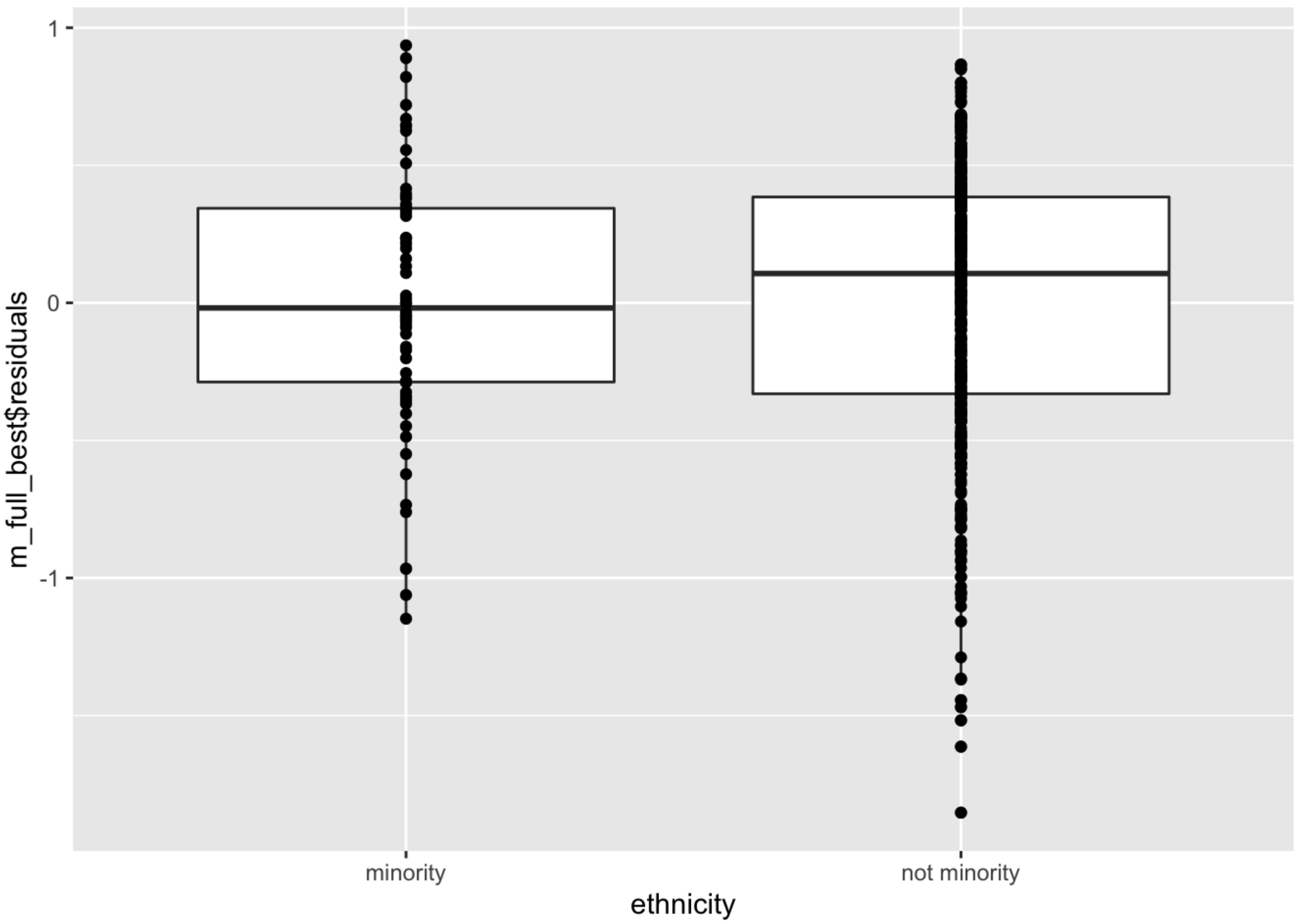
Hide

```
ggplot(evals, aes(gender, m_full_best$residuals)) +
  geom_boxplot() +
  geom_point()
```



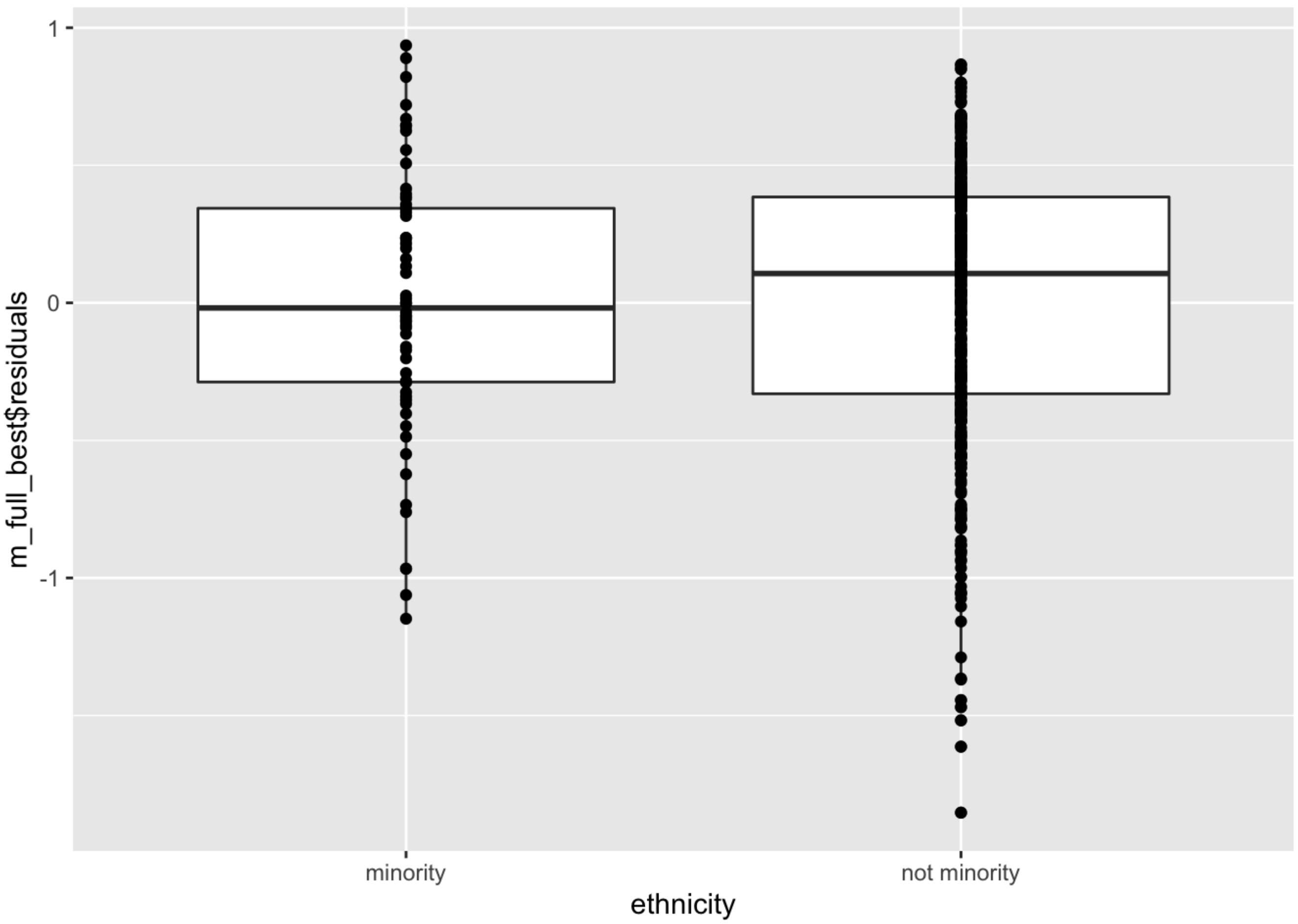
Hide

```
ggplot(evals, aes(ethnicity, m_full_best$residuals)) +
  geom_boxplot() +
  geom_point()
```



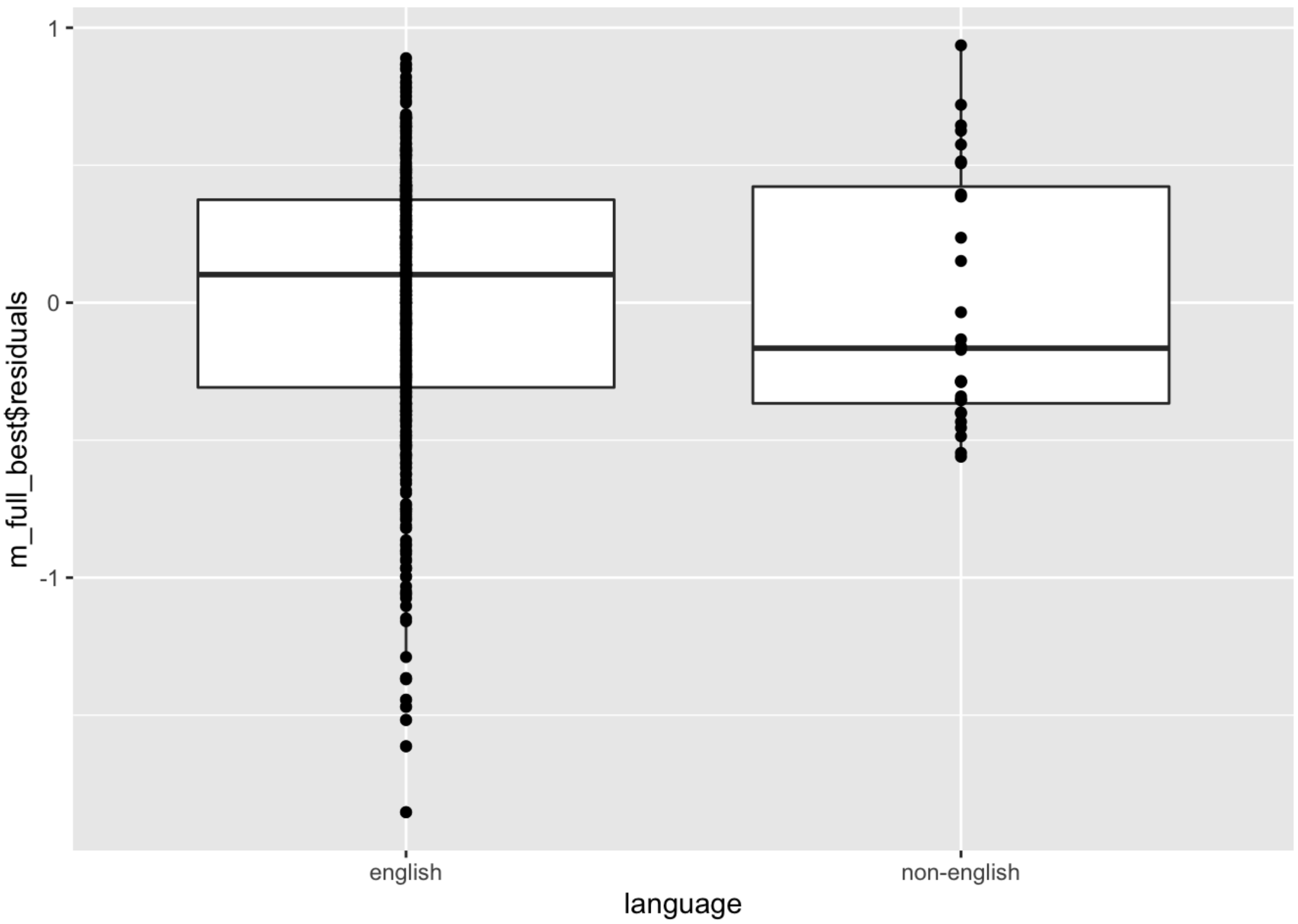
Hide

```
ggplot(evals, aes(ethnicity, m_full_best$residuals)) +
  geom_boxplot() +
  geom_point()
```



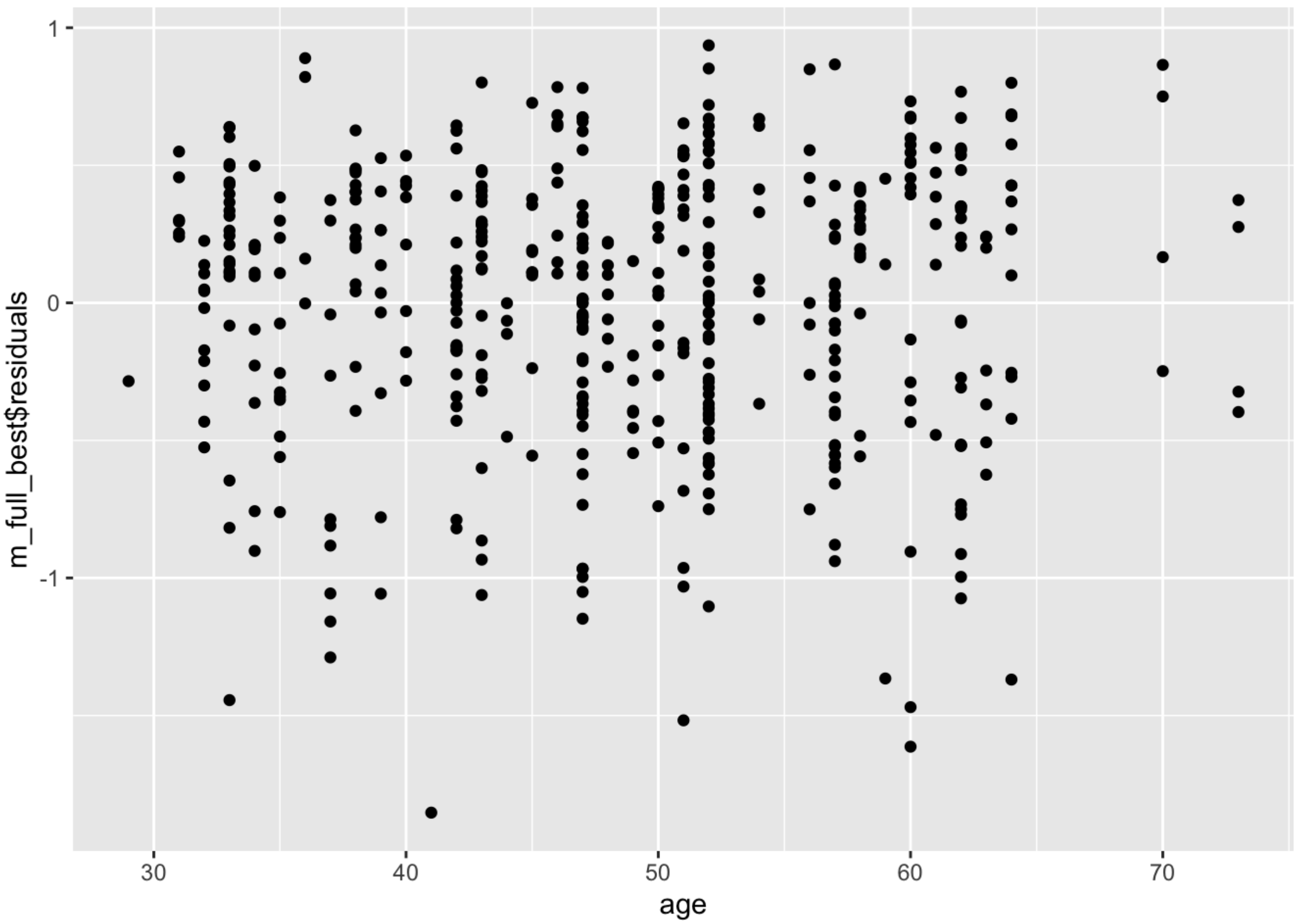
Hide

```
ggplot(evals, aes(language, m_full_best$residuals)) +
  geom_boxplot() +
  geom_point()
```



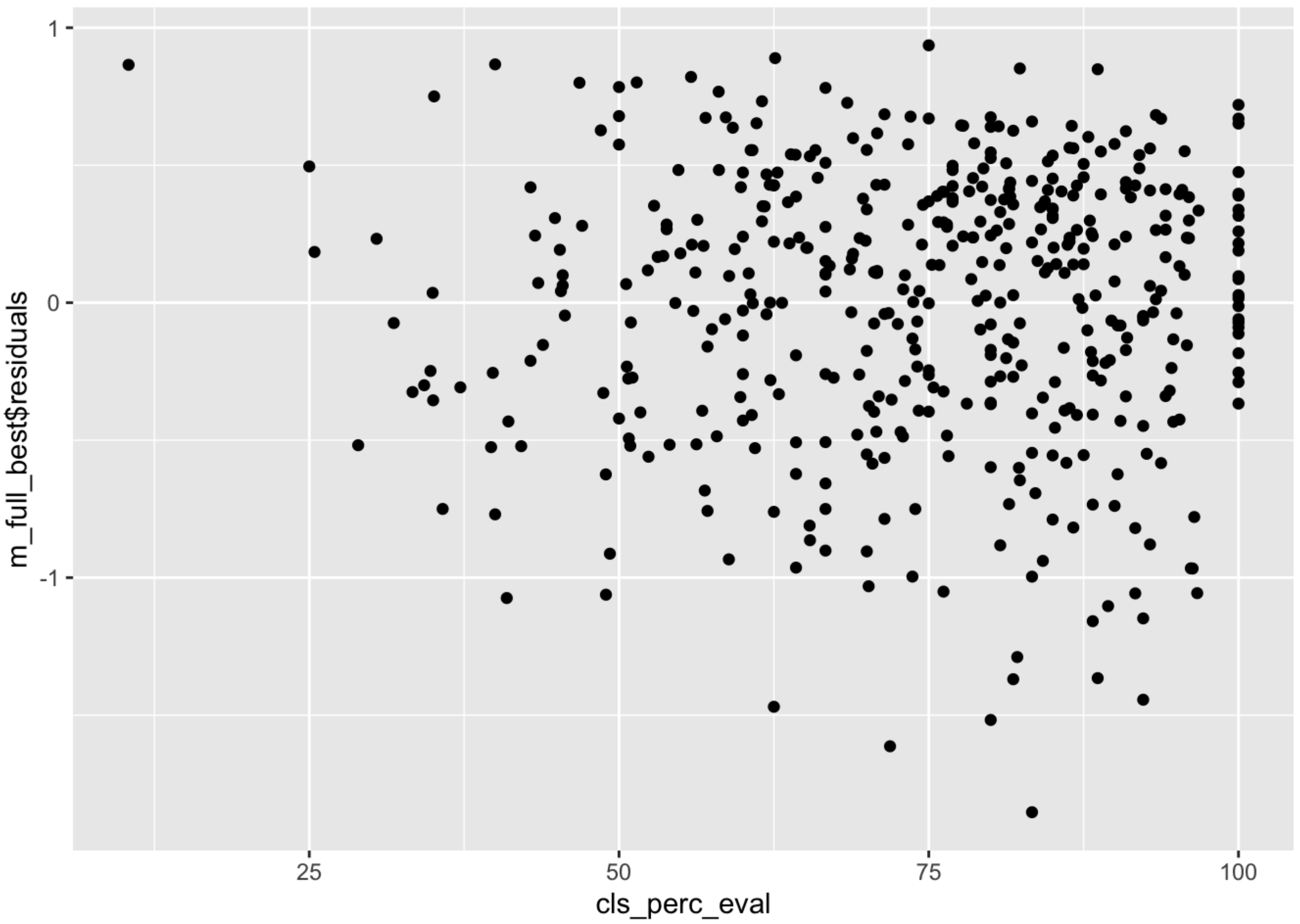
Hide

```
ggplot(evals, aes(age, m_full_best$residuals)) +
  geom_point()
```



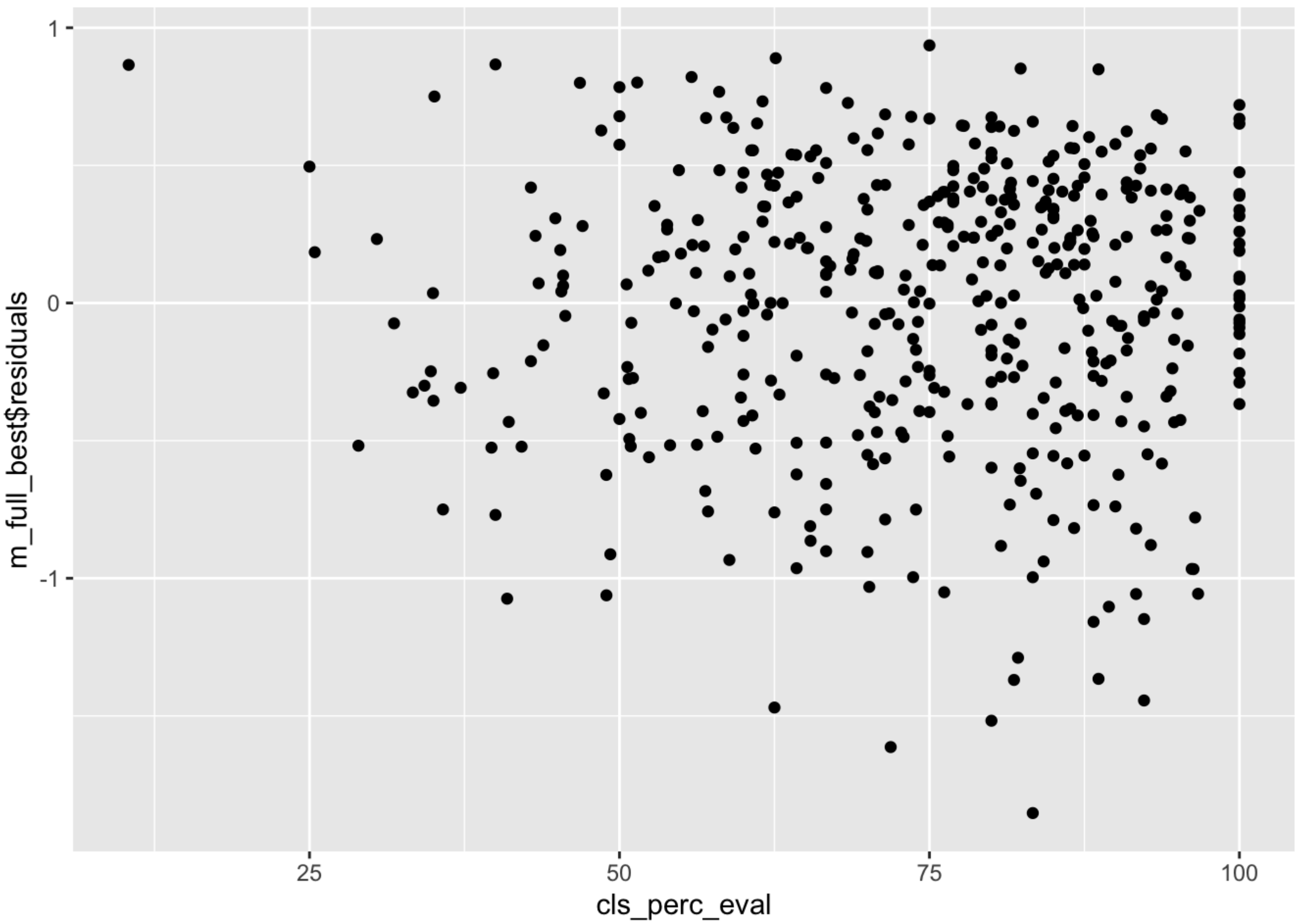
Hide

```
ggplot(evals, aes(cls_perc_eval, m_full_best$residuals)) +
  geom_point()
```



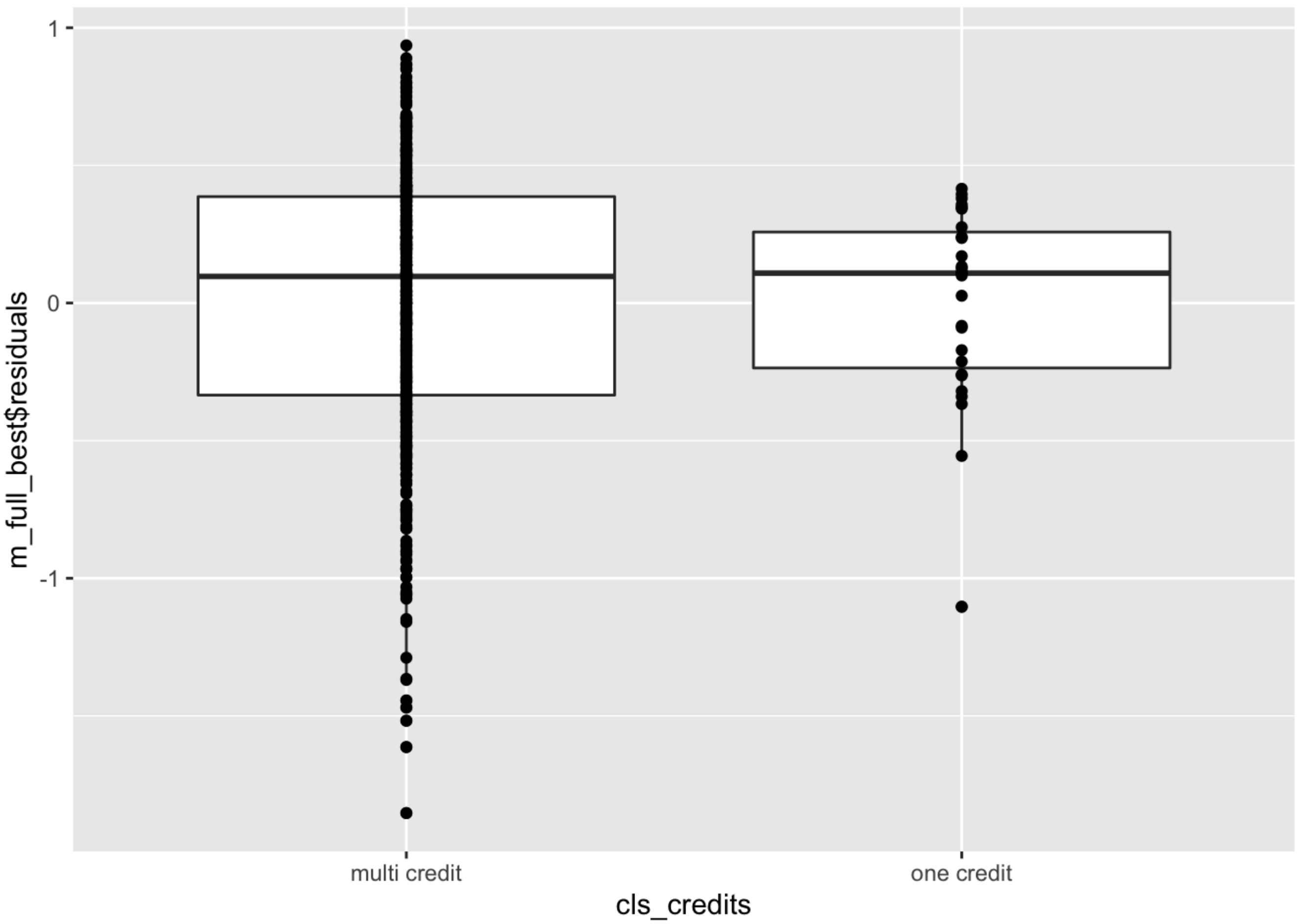
Hide

```
ggplot(evals, aes(cls_perc_eval, m_full_best$residuals)) +
  geom_point()
```



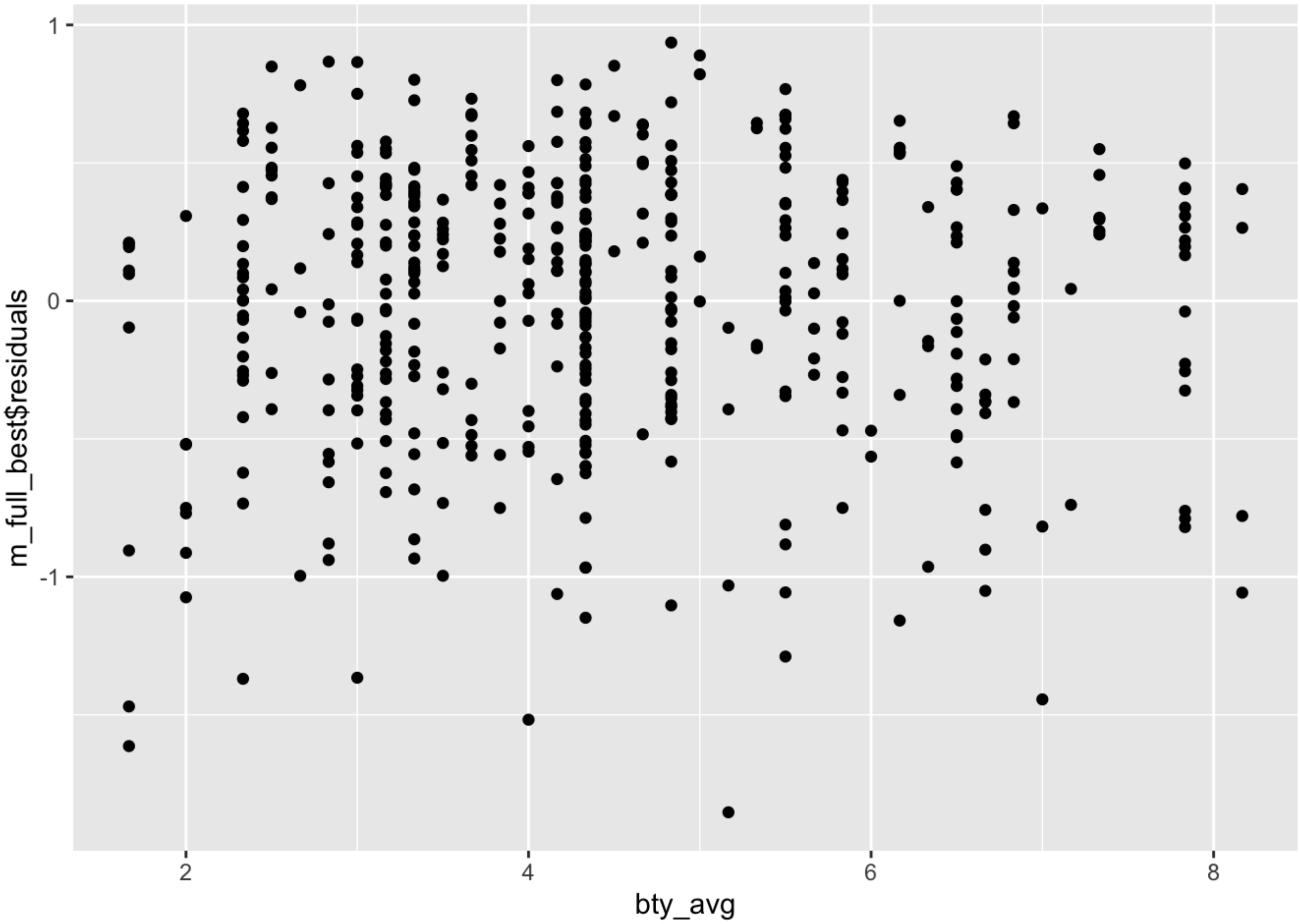
Hide

```
ggplot(evals, aes(cls_credits, m_full_best$residuals)) +
  geom_boxplot() +
  geom_point()
```



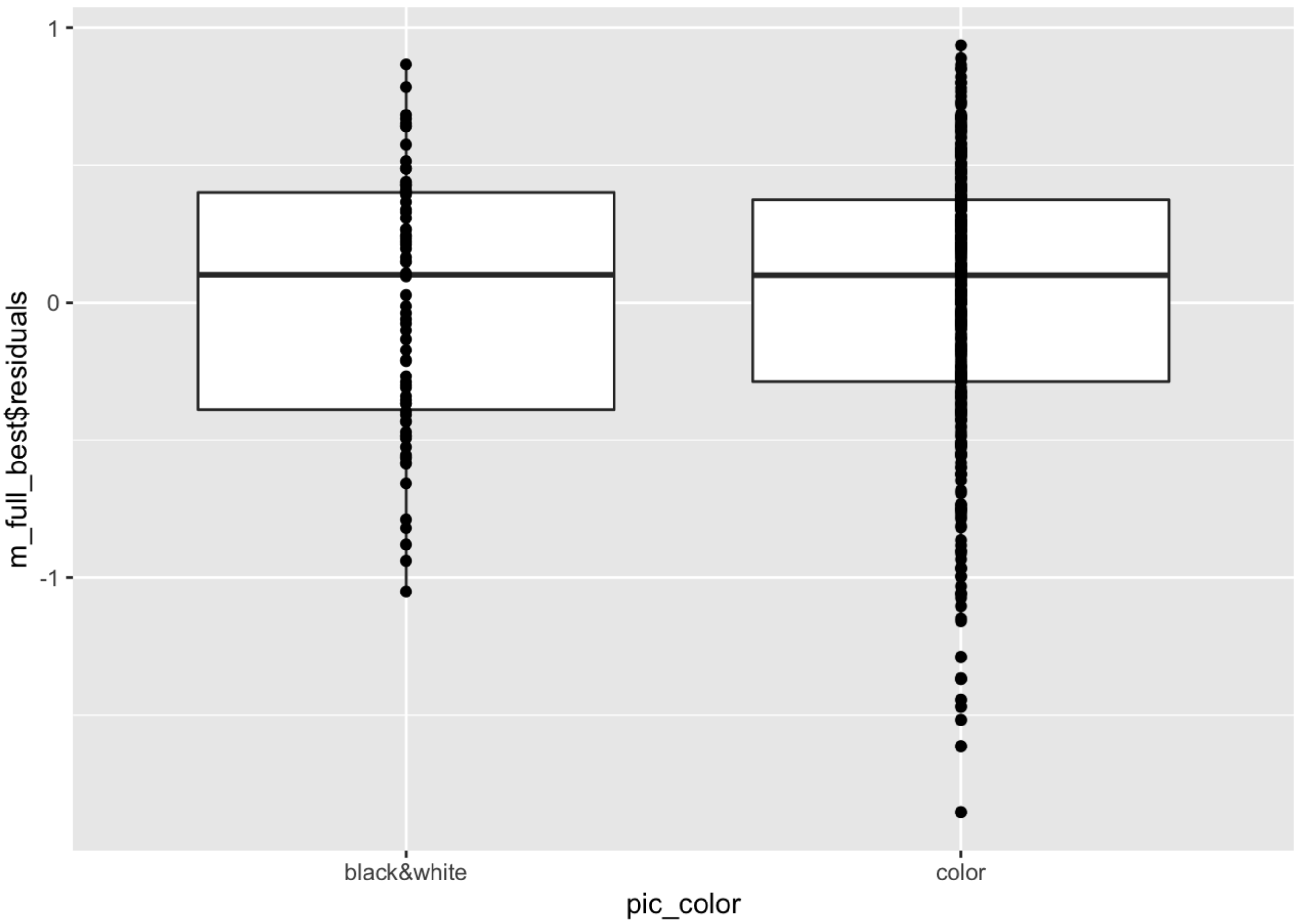
Hide

```
ggplot(evals, aes(bty_avg, m_full_best$residuals)) +
  geom_point()
```



Hide

```
ggplot(evals, aes(pic_color, m_full_best$residuals)) +
  geom_boxplot() +
  geom_point()
```



## 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?

-The new information probably does not have any impact on the conditions of linear regression because the courses are independent from one another.

…

## 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.

-High evaluation scores are associated with graduates from english speaking universities and non-minority males of young age with a higher beauty average score with a black and white class photo. In addition, the fewer the credits, the better and the higher response rate also lends itself to this higher score. …

## Exercise 19

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

-Since there are only 463 evaluations, I’d say that is still too small of a sample size to conclude a generalization to professors teaching at other Universities.

…