Data_606_Lab_9_Multiple_Linear_Regression

Enid Roman

2022-11-27

```
#install.packages('GGally')
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.2.2

## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.2 ──
## ✓ ggplot2 3.3.6     ✓ purrr   0.3.4
## ✓ tibble  3.1.8     ✓ dplyr   1.0.9
## ✓ tidyr   1.2.0     ✓ stringr 1.4.1
## ✓ readr   2.1.2     ✓ forcats 0.5.2
## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## ✗ dplyr::filter() masks stats::filter()
## ✗ dplyr::lag()    masks stats::lag()

library(openintro)

## Loading required package: airports
## Loading required package: cherryblossom
## Loading required package: usdata

library(GGally)

## Warning: package 'GGally' was built under R version 4.2.2

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

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, …
```

```
?evals
```

```
## starting httpd help server ... done
```

Exploring the data

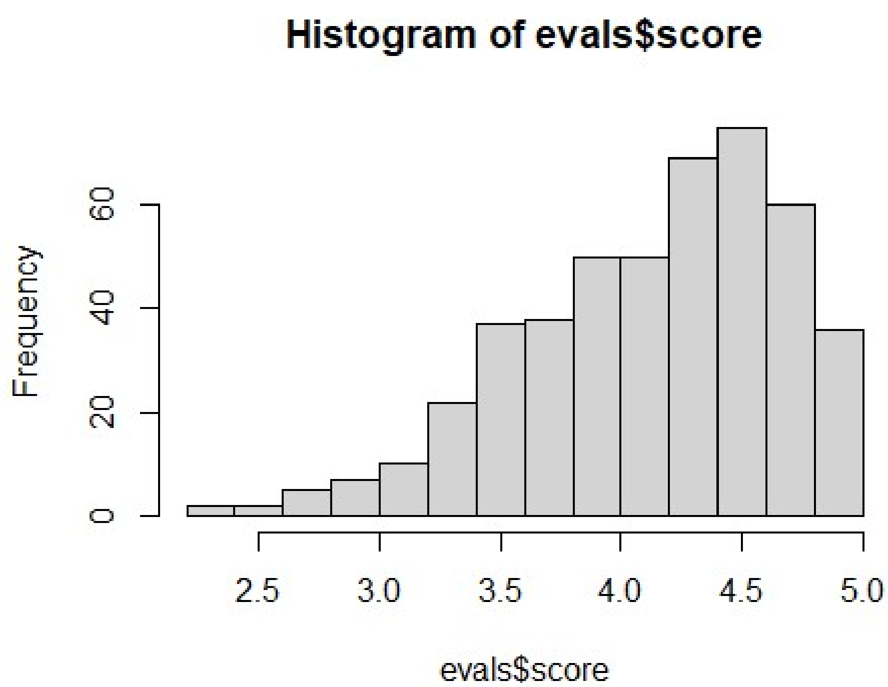
Exercise 1

This is an observational study. The original research question posed in the paper is whether beauty leads directly to the differences in course evaluations. Given the study design, we cannot easily answer the question as phrased. To isolate and test whether beauty (farily subjective) causes changes in course evalutaion would require randomized trials. In this case, we can more appropriately ask and answer is there a correlation between beauty and evaluations and/or how much of the variablility in course evaluations might be explained by beauty.

Exercise 2

Scores are left skewed where the most scores are between 4.0 and 5.0 and a long tail going down to 2.0. In general students give courses above average scores >= 3.5. An ideal scoring prototcal would have resulted in a more normalized curve with the bulk in the middle and symmetic tails in both directions. I would have expected more courses to receive good score than bad score.

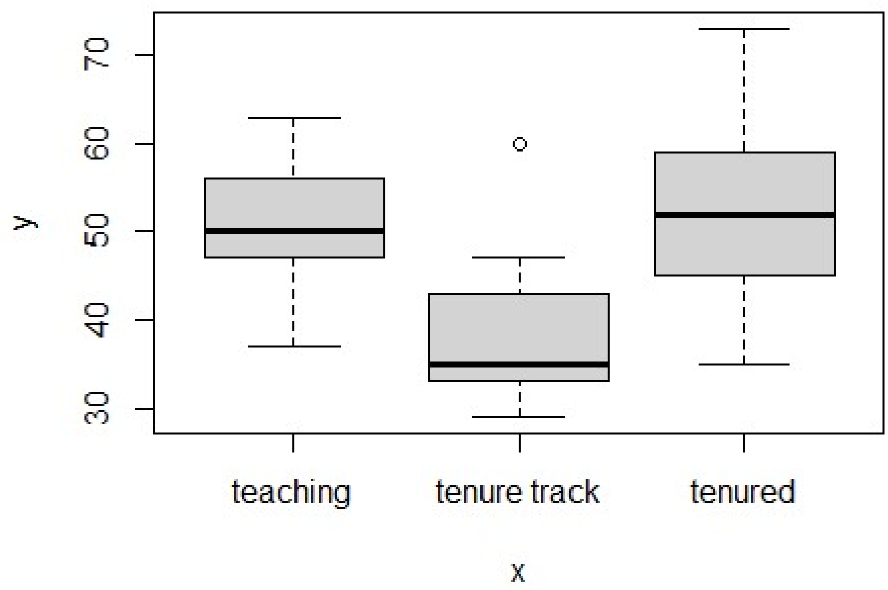
```
hist(evals$score)
```



Exercise 3

The two other variables I selected was rank and age. Here you can see the rank of teaching is expected to be close to age 50 to about age 55. Tenured Rank is more younger group from approximately age 35 to 45. Which is expected because that this age the tenure track is a professor's pathway to promotion and academic job security. It's the process by which an assistant professor becomes and an associate professor and then a professor. Tenured group is from approximately 45 to almoast 60. Which is expected. Tenured is having or denoting a permanent post.

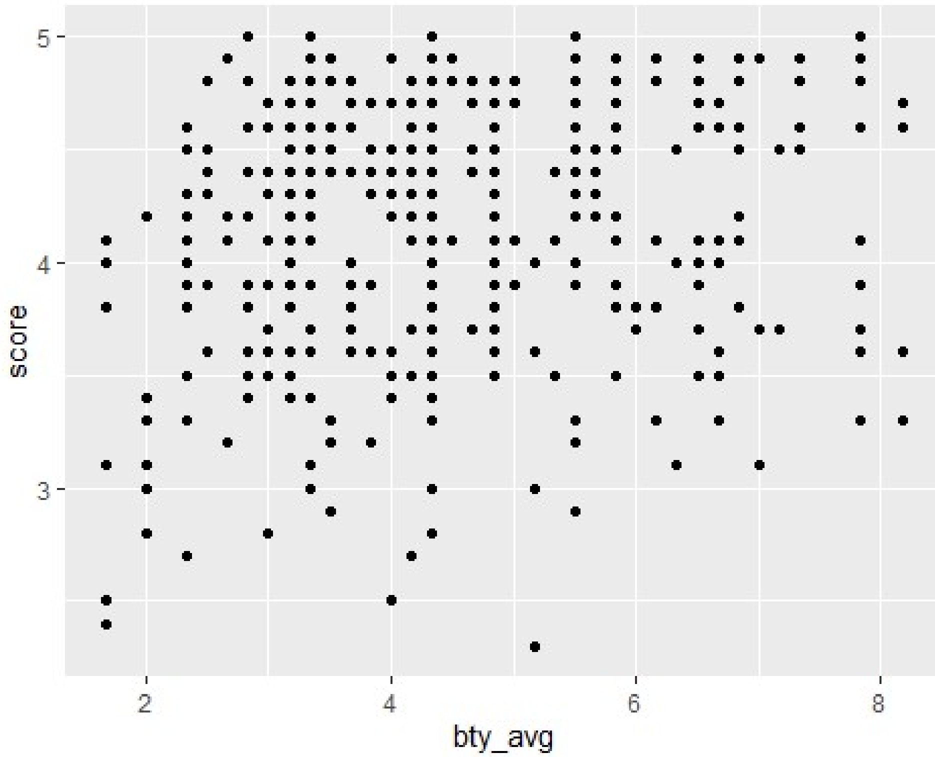
```
plot(evals$rank, evals$age)
```



Simple linear

regression

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



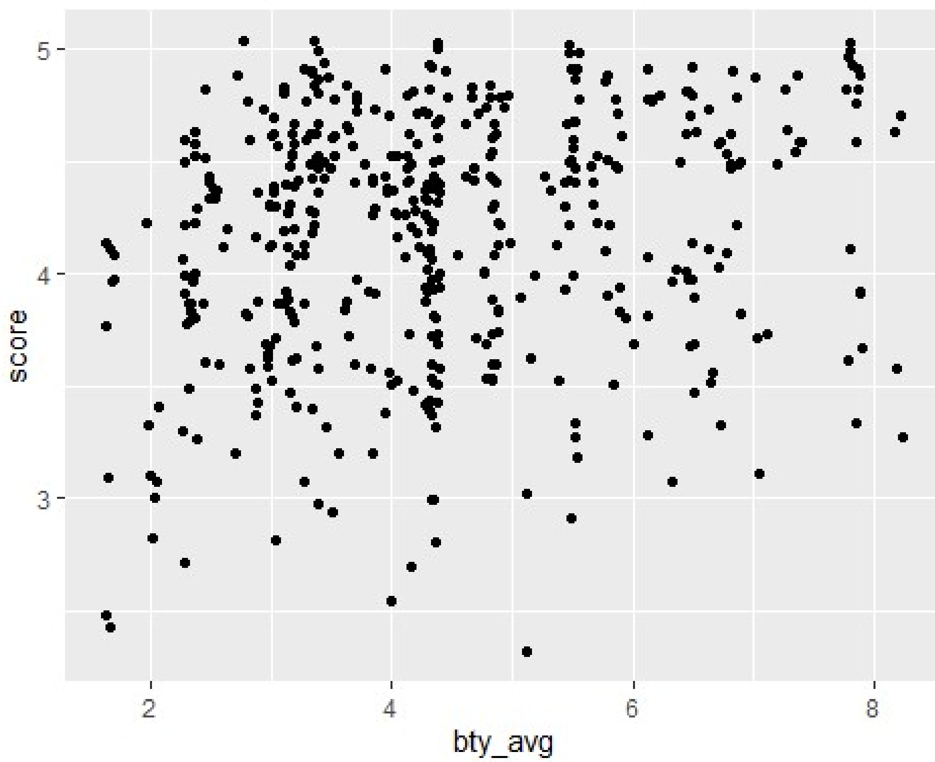
There are 463 rows in the data set, but fewer points on the scatter plot.

```
nrow(evals)

## [1] 463
```

Exercise 4

The initial scatter plot had overlapping points. With jitter we can more clearly see where the bulk of points are landing.

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



Exercise 5

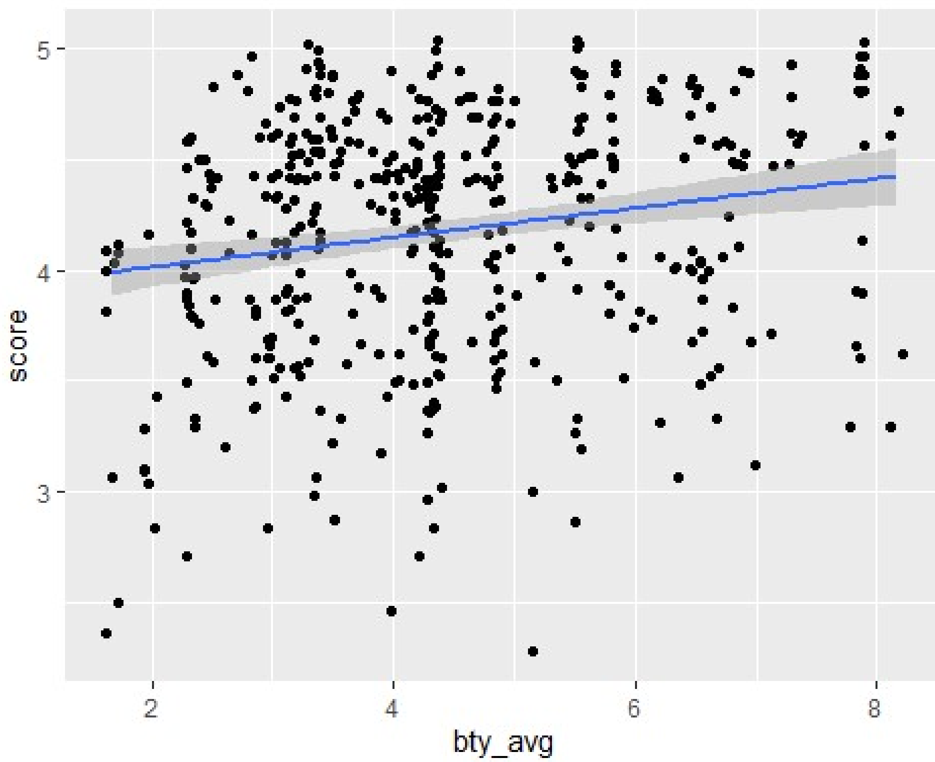
*The equation for the linear model is score = 3.88034+0.06664 *btyavg*

The slope of the line is 0.0664. The interpretation of the slope is, as avg_bty increases, the scores also increaeses; while the slope and intercept are "significant" (ie there is a positive correlation), the R2 is ~0.033 meaning only about 3.5% of the variation in score can be explained by beauty. ~96.5% of the variation is due to other factors and/or randomness.

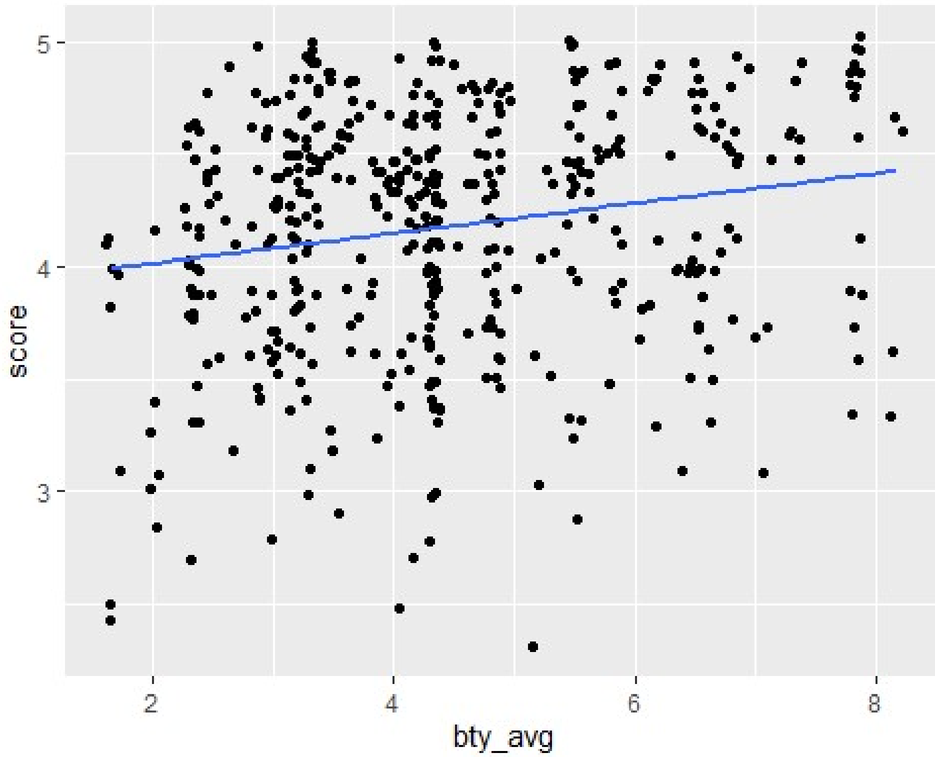
```
m_bty <- lm(evals$score ~ evals$bty_avg)
summary(m_bty)

## 
## 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
```

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

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



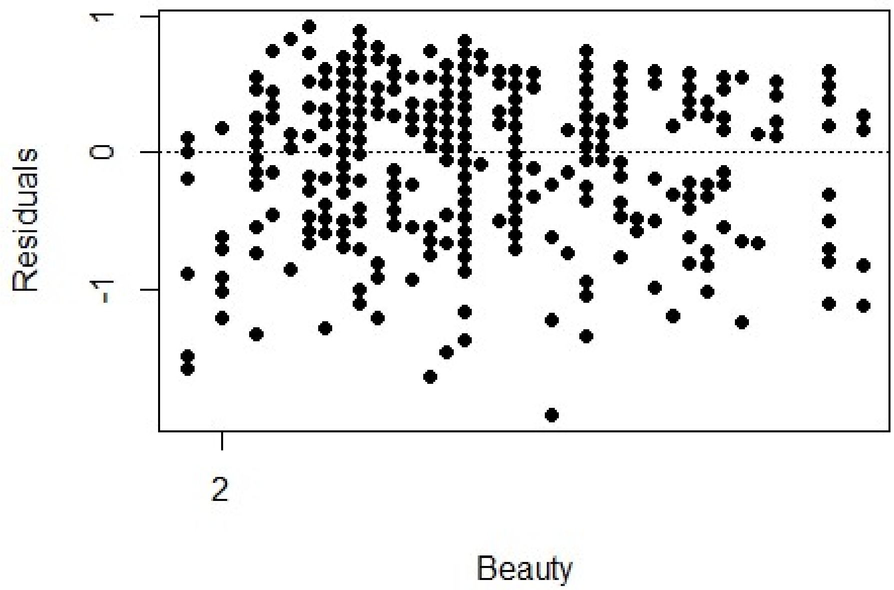

```
ggplot(data = evals, aes(x = bty_avg, y = score)) +
  geom_jitter() +
  geom_smooth(method = "lm", se = FALSE)

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



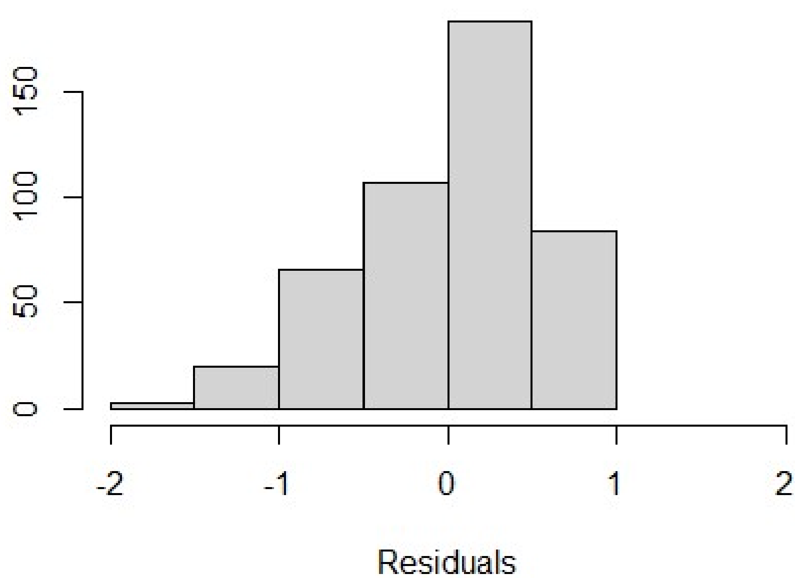
Exercise 6

Residuals while not perfectly normally distributed (left skewed a little), do appear to be overall mostly normal. There so not appear to be any trends. The assumptions are broadly met.

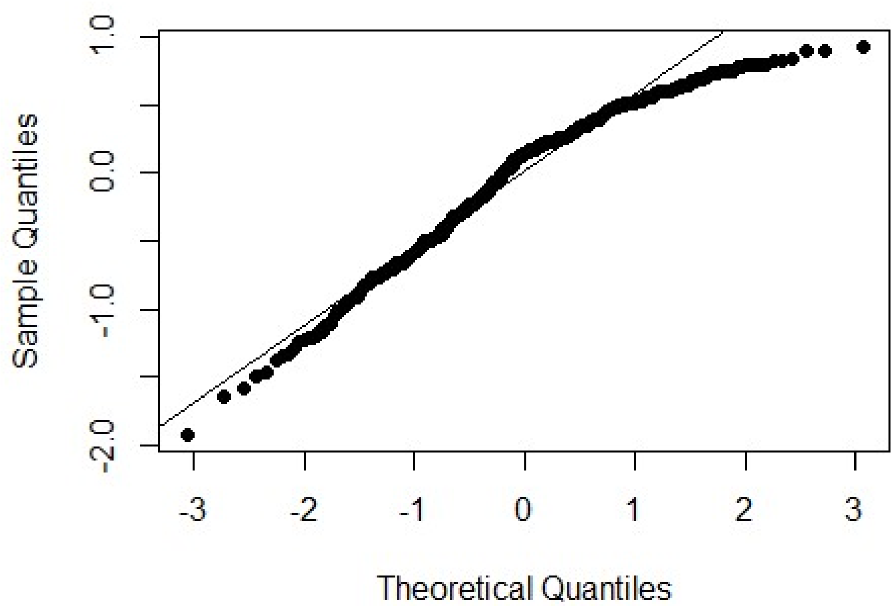
```
plot(m_bty$residuals ~ evals$bty_avg,
     xlab = "Beauty", ylab = "Residuals",
     pch = 19,
     axes = FALSE)
axis(1, at = seq(-1, 2, 1))
axis(2, at = seq(-1, 1, 1))
box()
abline(h = 0, lty = 3)
```



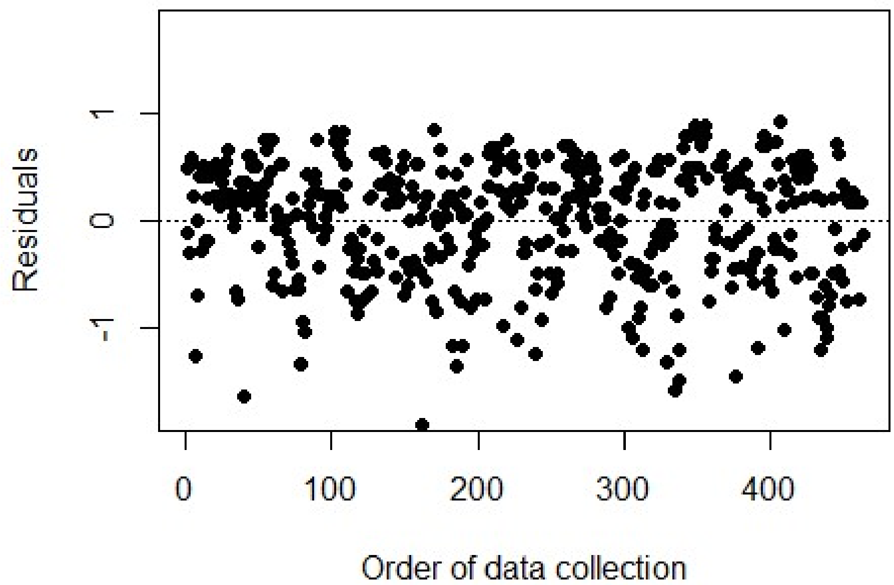

```
hist(m_bty$residuals,
     xlab = "Residuals", ylab = "", main = "",
     xlim = c(-2,2))
```



```r
qqnorm(m_bty$residuals,
       pch = 19,
       main = "", las = 0)
qqline(m_bty$residuals)
```

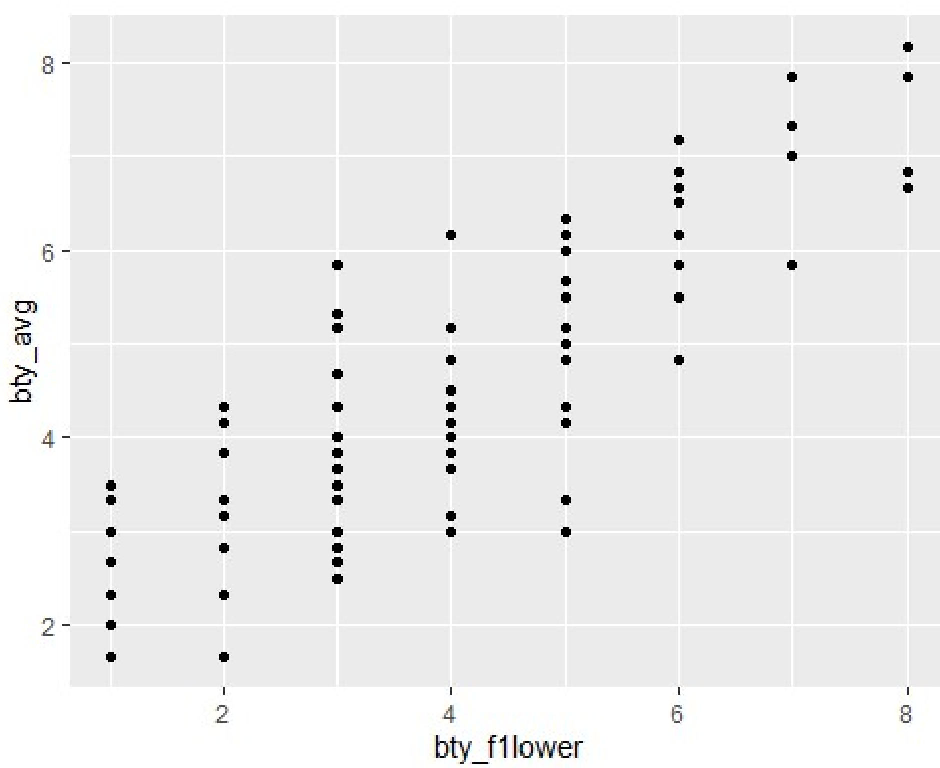



```r
plot(m_bty$residuals,
     xlab = "Order of data collection", ylab = "Residuals", main = "",
     pch = 19,
     ylim = c(-1.82, 1.82), axes = FALSE)
axis(1)
axis(2, at = seq(-1, 1, 1))
box()
abline(h = 0, lty = 3)
```



Multiple linar regresssion

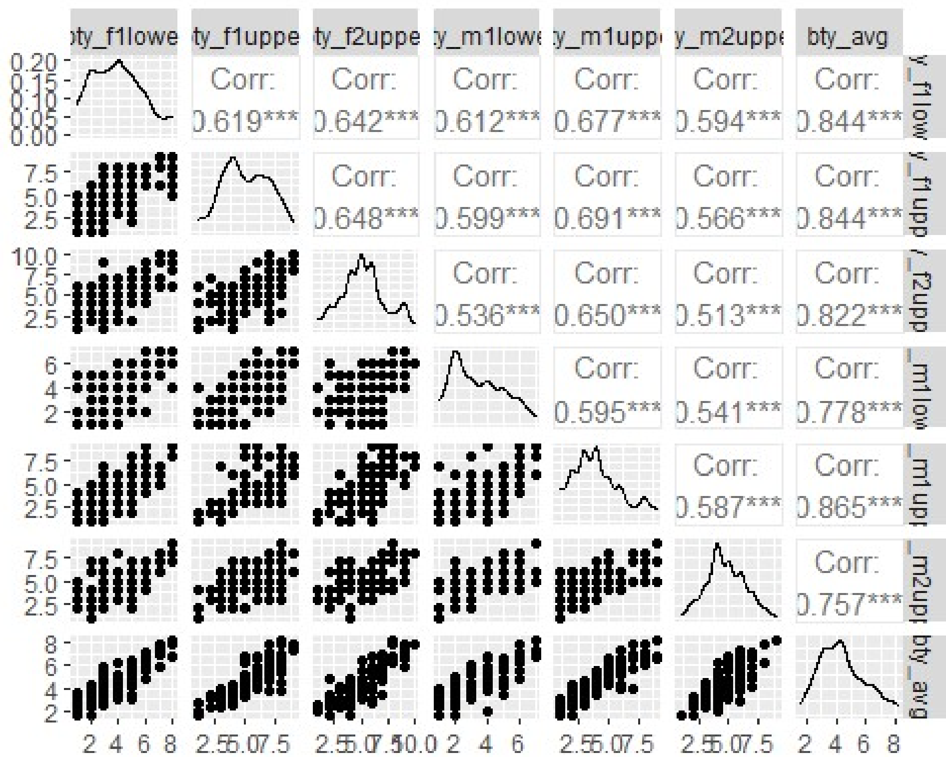
```
ggplot(data = evals, aes(x = bty_f1lower, y = bty_avg)) +
  geom_point()
```



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

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

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




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

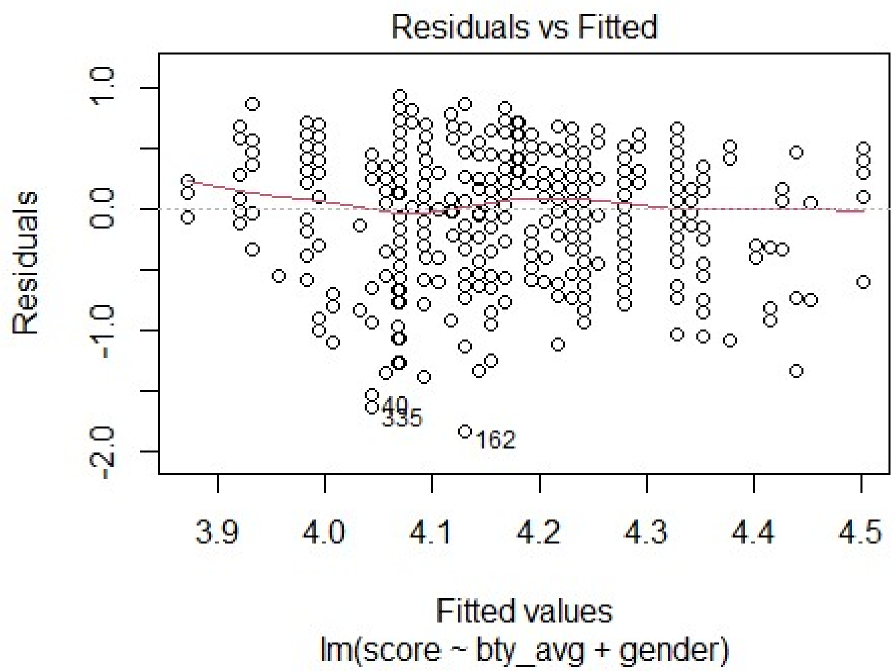
```
##
## Call:
## lm(formula = score ~ bty_avg + gender, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8305 -0.3625  0.1055  0.4213  0.9314
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.74734    0.08466  44.266  < 2e-16 ***
## bty_avg      0.07416    0.01625   4.563 6.48e-06 ***
```

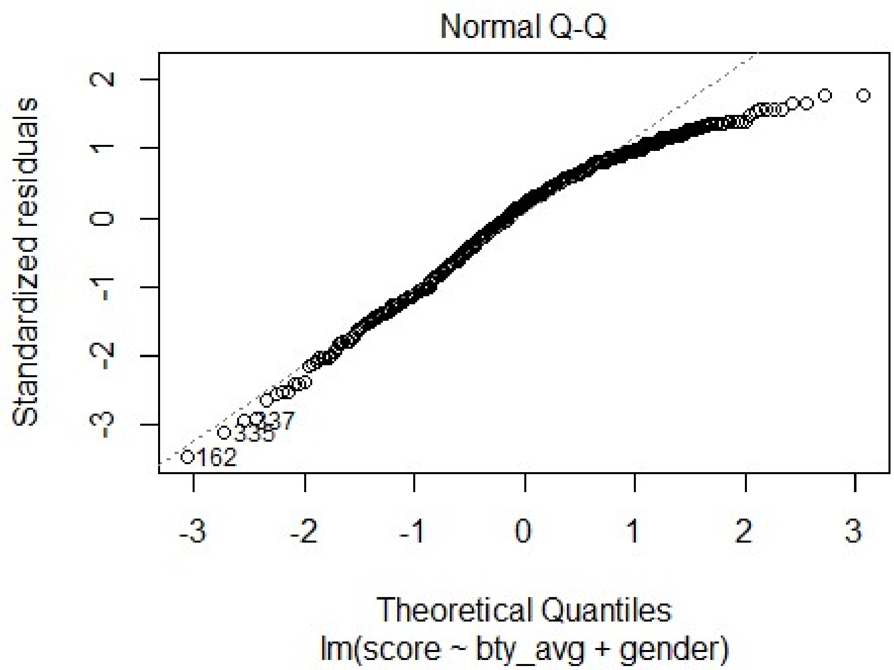
```
## gendermale   0.17239    0.05022   3.433 0.000652 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5287 on 460 degrees of freedom
## Multiple R-squared:  0.05912,    Adjusted R-squared:  0.05503
## F-statistic: 14.45 on 2 and 460 DF,  p-value: 8.177e-07
```

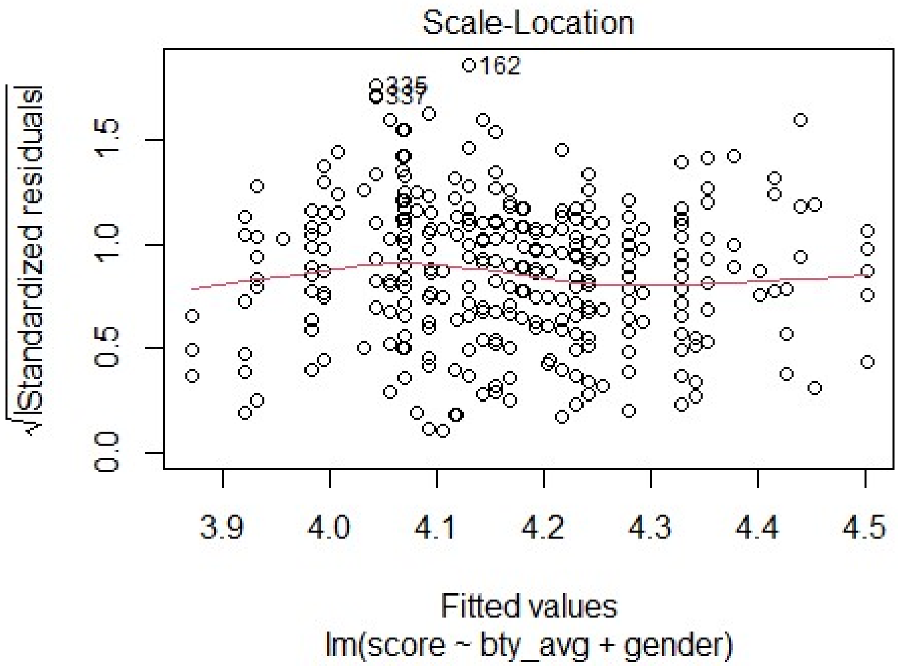
Exercise 7

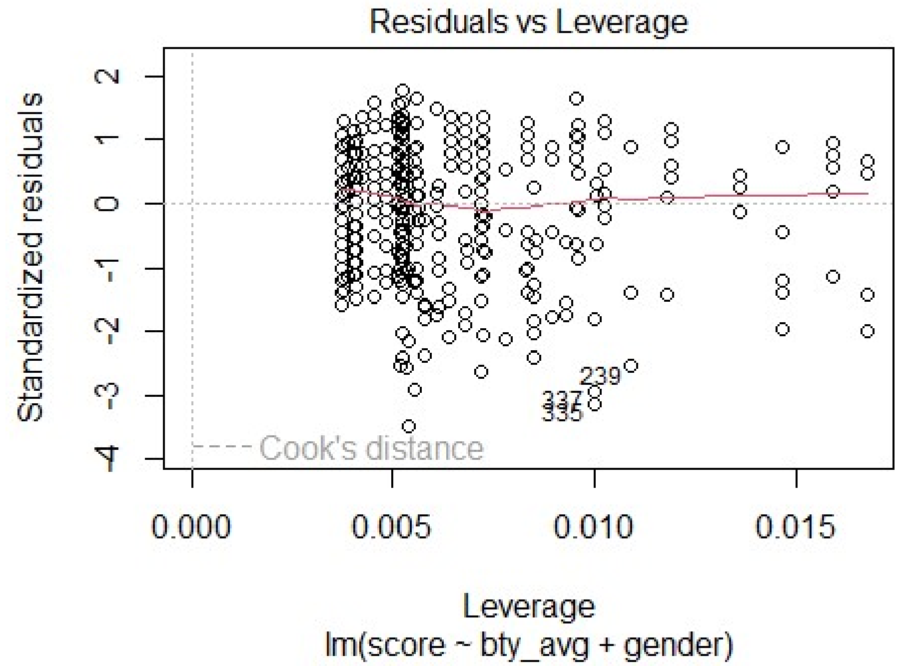
The conditions of the regression are reasonable therefore, P-values and parameter estimates could be trusted.: linearity of data, residuals are normal, no patterns in residuls, no strong leverage points.

```
plot(m_bty_gen)
```


Residuals vs Fitted
Residuals
1.0
0.0
-1.0
-2.0
40
335
162
3.9
4.0
4.1
4.2
4.3
4.4
4.5
Fitted values
lm(score ~ bty_avg + gender)

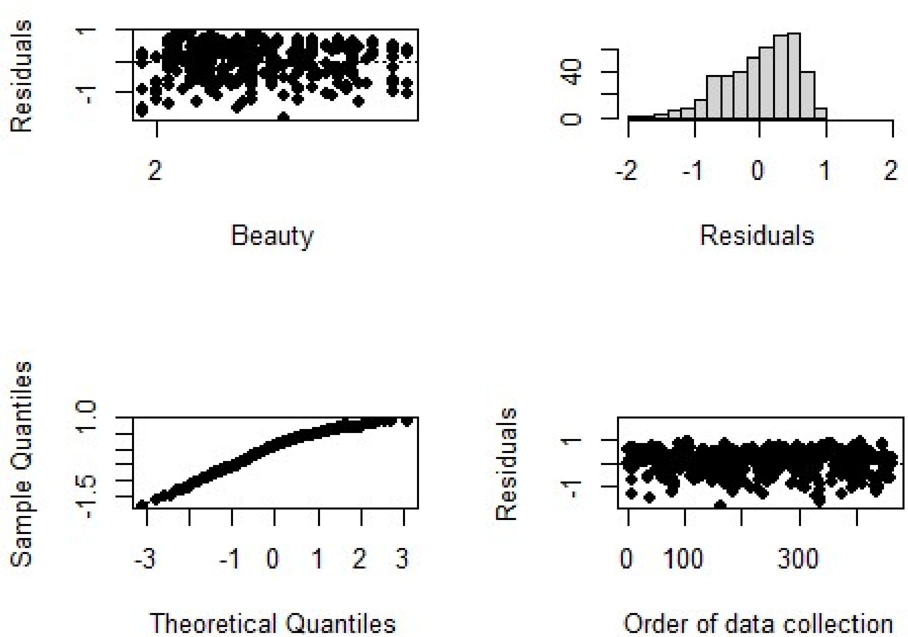

Normal Q-Q
Standardized residuals
2
1
0
-1
-2
-3
337
335
162
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(score ~ bty_avg + gender)






```
par(mfrow = c(2, 2))

plot(m_bty_gen$residuals ~ evals$bty_avg,
     xlab = "Beauty", ylab = "Residuals",
```

```
     pch = 19,
     axes = FALSE)
axis(1, at = seq(-1, 2, 1))
axis(2, at = seq(-1, 1, 1))
box()
abline(h = 0, lty = 3)

hist(m_bty_gen$residuals,
     xlab = "Residuals", ylab = "", main = "",
     xlim = c(-2,2))

qqnorm(m_bty_gen$residuals,
       pch = 19,
       main = "", las = 0)
qqline(m_bty_gen$residuals)

plot(m_bty_gen$residuals,
     xlab = "Order of data collection", ylab = "Residuals", main = "",
     pch = 19,
     ylim = c(-1.82, 1.82), axes = FALSE)
axis(1)
axis(2, at = seq(-1, 1, 1))
box()
abline(h = 0, lty = 3)
```



Exercise 8

Bty_avg is still significant predictor of score. Together with gender we now explain 5.5% of the variation in scores. The presence of gender has improved our model slightly, but while these are significant features, they offer low explanatory value.

Exercise 9

score^=β0+β1×bty_avg+β^2×(1)

score=(3.74734+0.17239)+0.07416∗btyavg

Between two professors who received the same beauty rating, the gender that tends to have the higher course evaluation score is male.

Exercise 10

R appear to handle categorical variables that have more than two levels by it creates a separate value for each rank, again leaving off the first alphabetic category which is treated as 0. Depending on which rank we are interested in, we use that value and the other is multiplied by zero so it drops out.

```
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
```

The search for the best model

Exercise 11

Either cls_level or cls_profs likely do not have much association with professor score and thus have a high p-value.

```
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
```

Exercise 12

My suspicions were correct. cls_level(upper in this case) has a p-value of .29369 and cls_profs has a p-value of 0.77806. These are indeed the highest p-values based on the model output.

Exercise 13

The coefficient associated with the ethnicity varuable, the score is increased by 0.12 points if the professor is ethnicity notminority.

Exercise 14

Yes, the coefficients and significance of the other explanatory variables changed meaning that the drop of the variable is dependent on the other variables.

```
drop_cls_profs <- 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(drop_cls_profs)

##
## 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

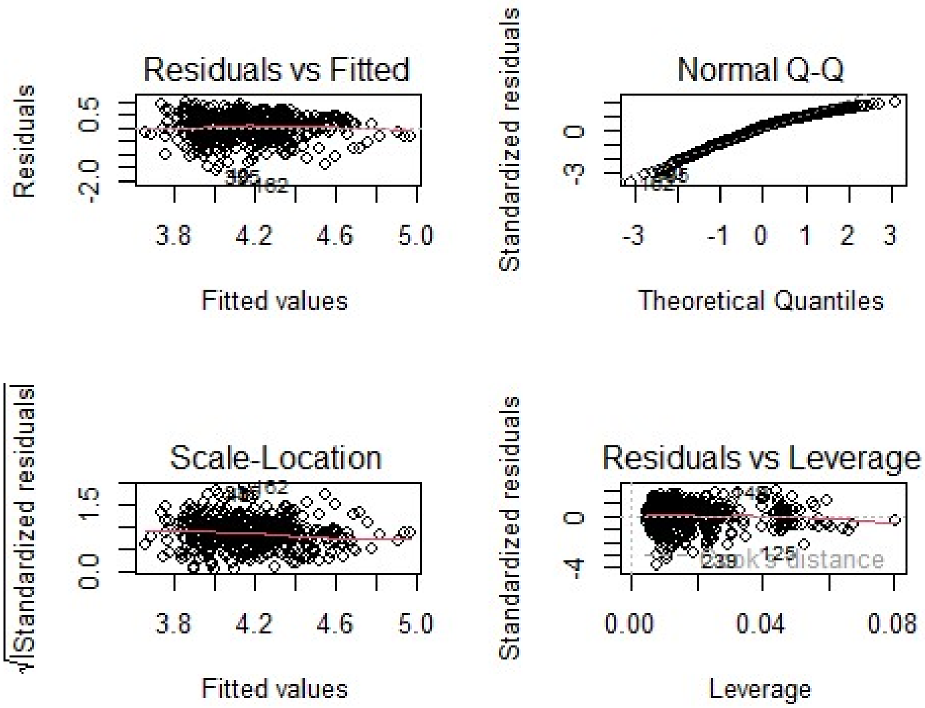
$\hat{score}=3.771922+(ethnicity\times0.167872)+(gender\times0.207112)+(language\times-0.206178)+(age\times-0.006046)+(clsperceval\times0.004656)+(clscreditsone\times0.505306)+(btyavg\times0.051069)+(piccolor\times-0.190579)=3.91973+0.07416\times bty\_avg$

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

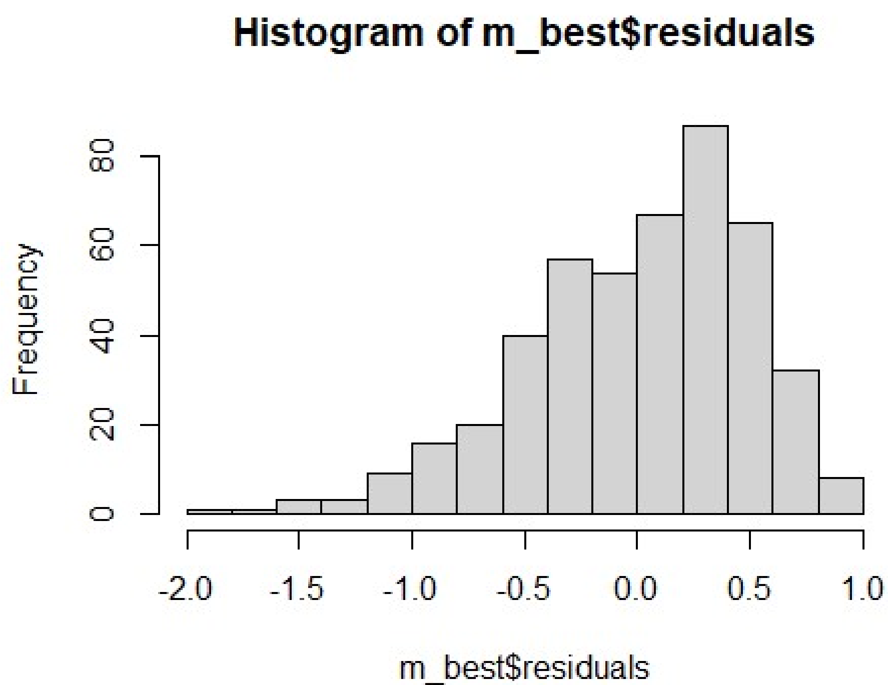
Exercise 16

The conditions for this model are reasonable. The residuals look good, the linear model fits well and there's no problem with the leverage points.

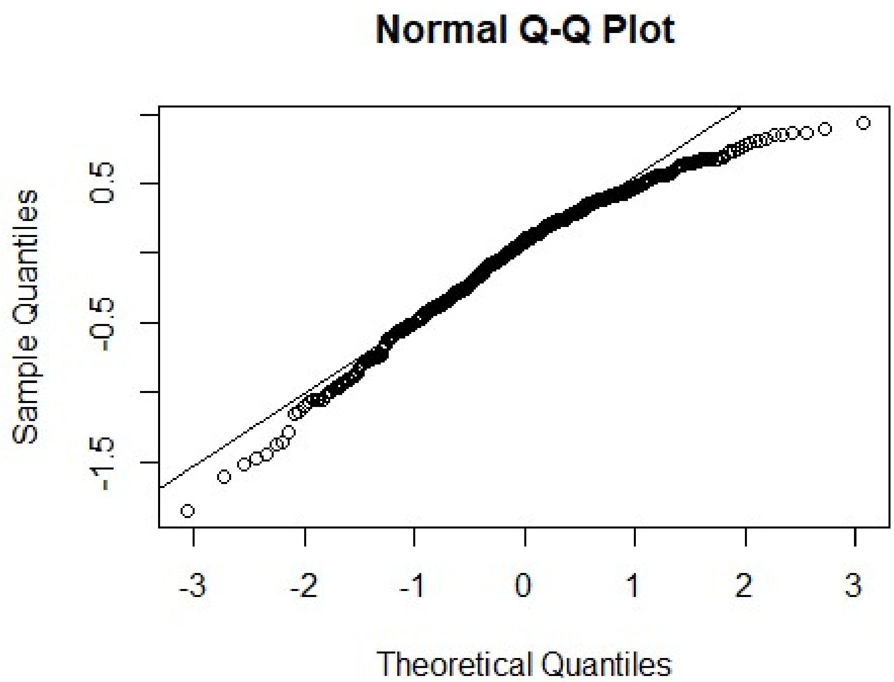
```
par(mfrow = c(2, 2))
plot(m_best)
```




```
hist(m_best$residuals)
```



```
# Normal Probability Plot
qqnorm(m_best$residuals)
qqline(m_best$residuals)
```



Exercise 17

No, considering that each row represents a course, this new information could not have an impact on any of the conditions of linear regression. The class courses are independent from each other therefore, the scores would also be independent.

Exercise 18

The classifications for highest ranked professors based on my final model would be: non-minority, male, young, speaks English, high number of evaluations, higher amount of credits being taught, percieved as beautiful, and picture is colored.

Exercise 19

I would not feel comfortable generalizing these conclusions because because other universities have different cultures. Other universities would have different results depending on their culture.