Lab 9 - Multiple Linear Regressions

Keith Colella

2023-05-08

Setup

Config

```
library(tidyverse)
library(openintro)
library(StepReg)
library(GGally)
library(cowplot)
library(patchwork)
```

Data

```
data('evals', package='openintro')
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

Question

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.

Response

The description provided indicates an observation study. There does not appear to be any separation between control and experimental groups, as required by experimental design. So, we likely cannot draw conclusions about any causal relationships between variables. Instead, we can only examine correlation, i.e. whether or not a significant relationship exists. We can rephrase our research question accordingly: is there a statistically significant relationship between the physical appearance of an instructor and course evaluations?

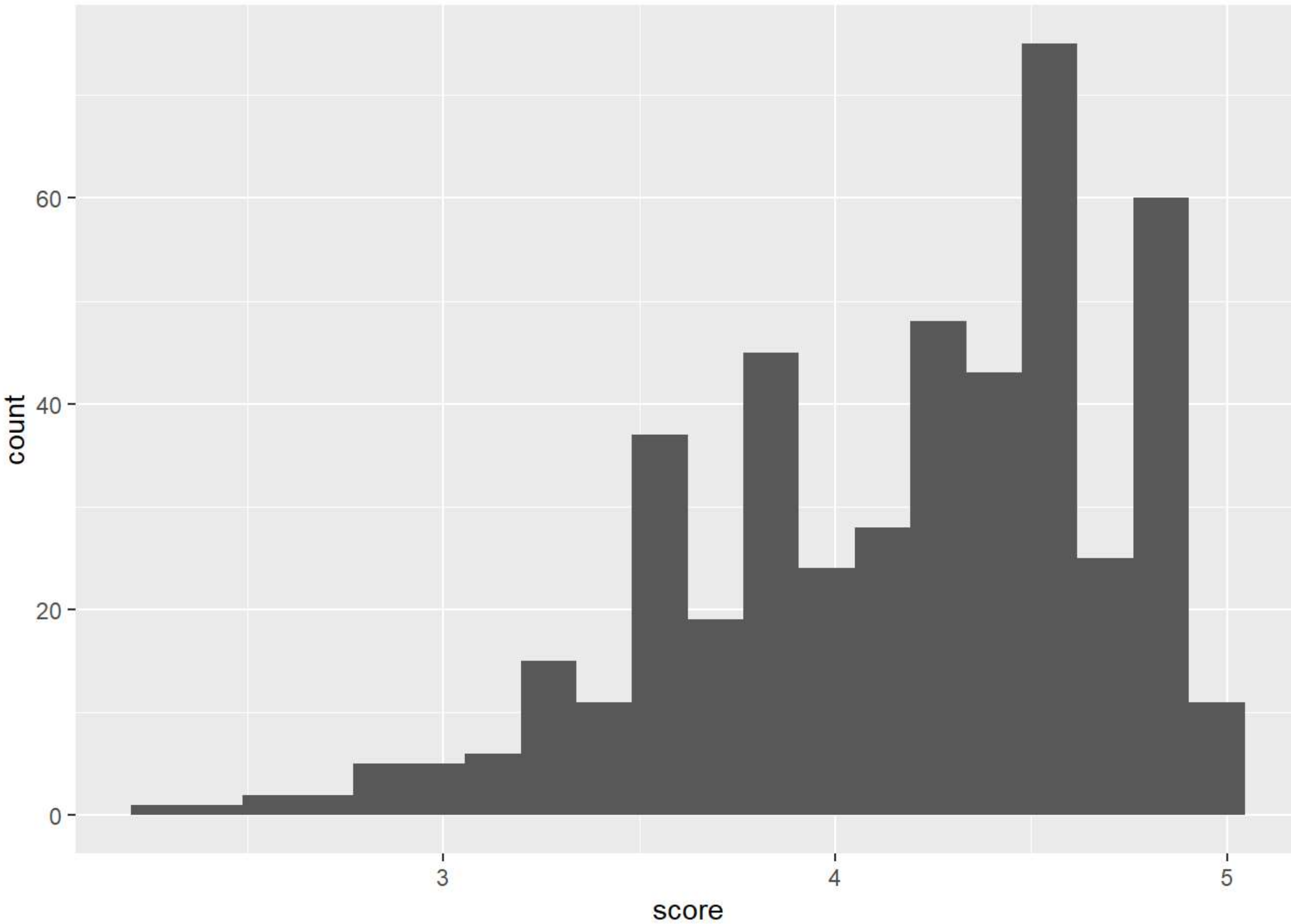
Exercise 2

Question

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?

Response

```
evals %>%
  ggplot(aes(score)) +
  geom_histogram(bins = 20)
```



Yes, the distribution appears negatively skewed. This indicates that students are wary of providing very low ratings. This may be driven by a high level of instruction at UT, or by a general unwillingness to provide instructors very low (1 or 2) ratings. Perhaps students are hesistant to provide low ratings to instructors after having spent months developing a relationship with them.

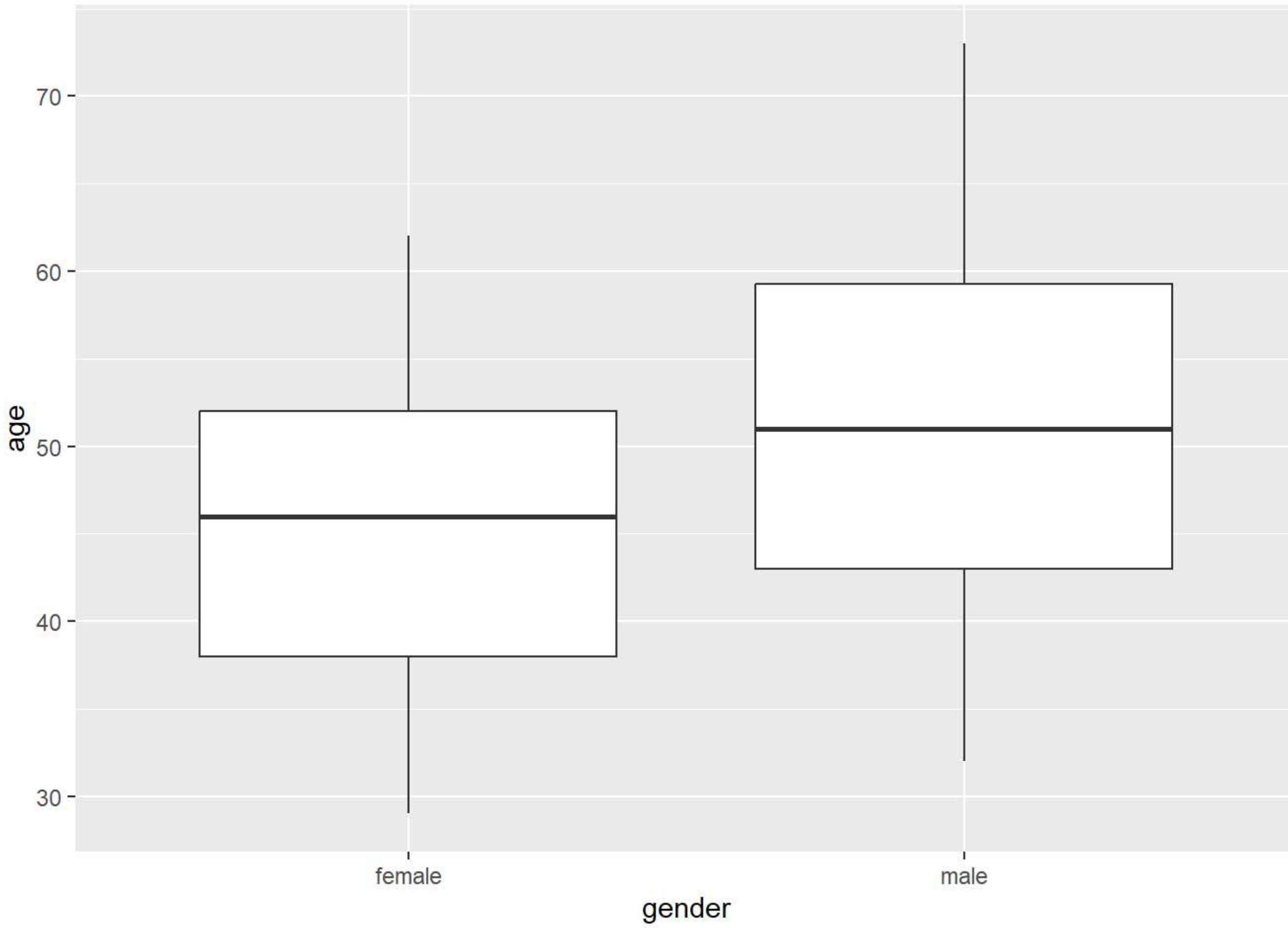
Exercise 3

Question

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

Response

```
evals %>%
  ggplot(aes(gender, age)) +
  geom_boxplot()
```



In examining the relationship between gender and age, it appears male instructors in the survey are generally older than female students. The median age for women appears roughly 46 years old, whereas the median age for men is roughly 51.

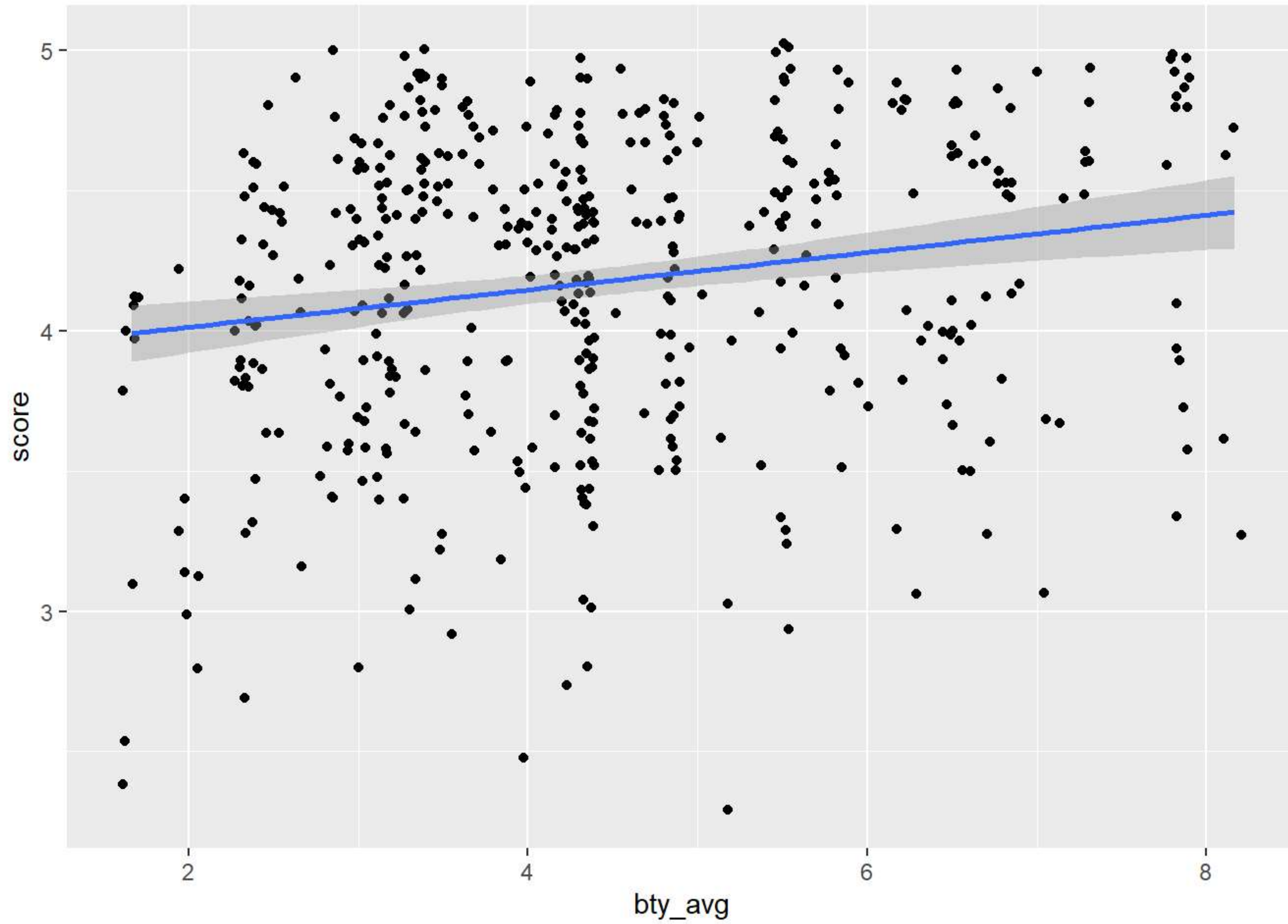
Exercise 4

Question

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

Response

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



The discretized nature of `bty_avg` perhaps gave the indication of a stronger relationship that there really is. When using `geom_jitter` , the relationship appears even more noisy.

Exercise 5

Question

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?

Response

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

```
##
## 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
```

The average beauty rating does appear to be a significant predictor of evaluation score. The positive coefficient indicates that, for each 1-point in crease in beauty score, the model would predict a ~0.07 higher score. The fact that the coefficient is so small shows that, while beauty is statistically significant, it is not practically significant. Moreover, the R^2 is very small, highlight the fact that beauty alone explains almost non of the variability in scores: less than 4%.

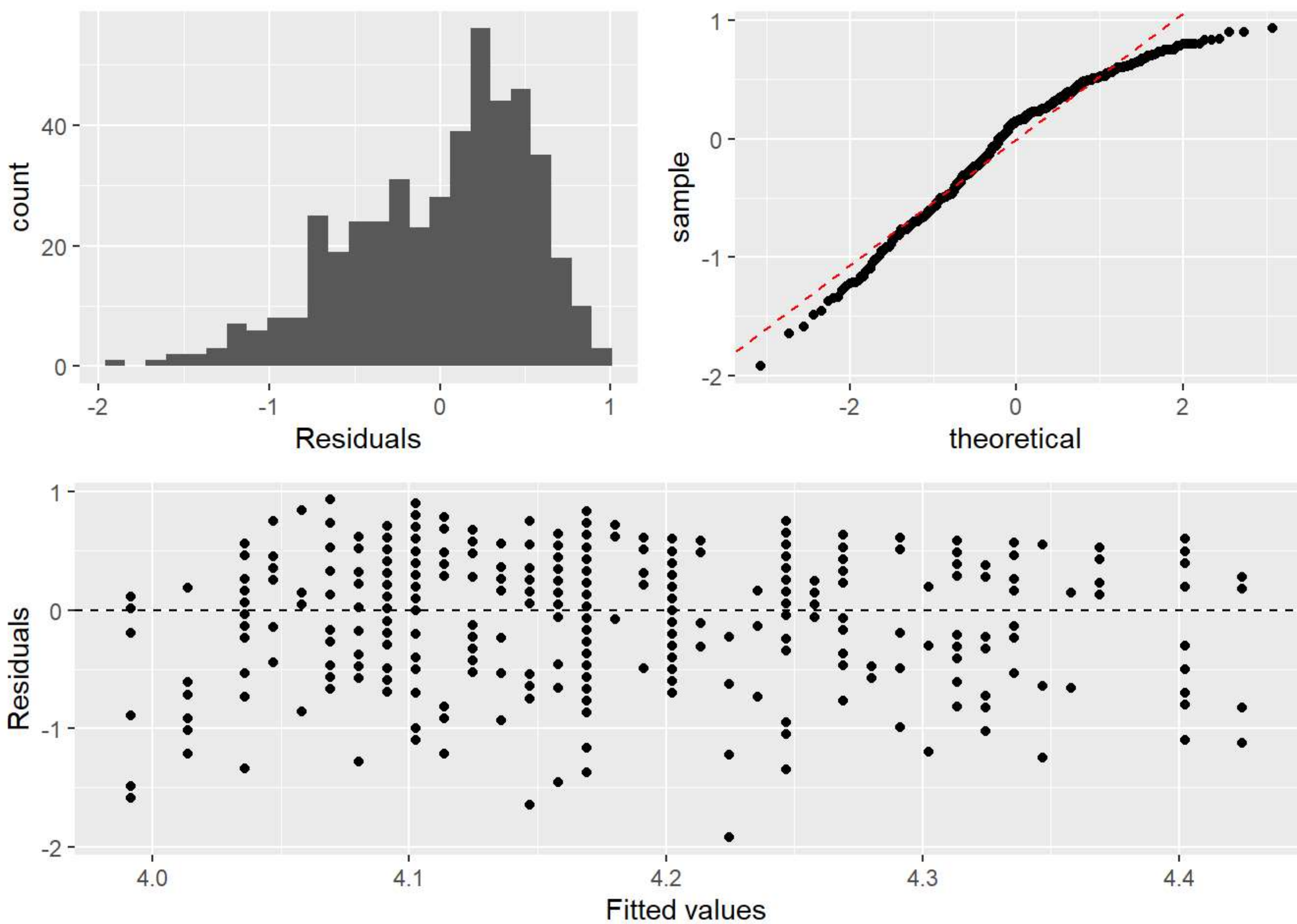
Exercise 6

Question

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).

Response

```
p1 <- ggplot(data = m_bty, aes(x = .resid)) +
  geom_histogram(bins = 25) +
  xlab("Residuals")

p2 <- ggplot(data = m_bty, aes(sample = .resid)) +
  stat_qq() +
  geom_abline(intercept = mean(m_bty$residuals),
              slope = sd(m_bty$residuals),
              color = 'red', linetype = 'dashed')


p3 <- ggplot(data = m_bty, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")

plot_grid(plot_grid(p1, p2), p3, nrow = 2, ncol = 1)
```



The residual plots indicate some non-normality in the residuals and some clear skewness, potentially violating of assumptions. The residual versus fitted plot, however, adheres to regression assumptions, as they show no clear pattern, indicating constant variance (i.e. homoskedasiticity).

Exercise 7

Question

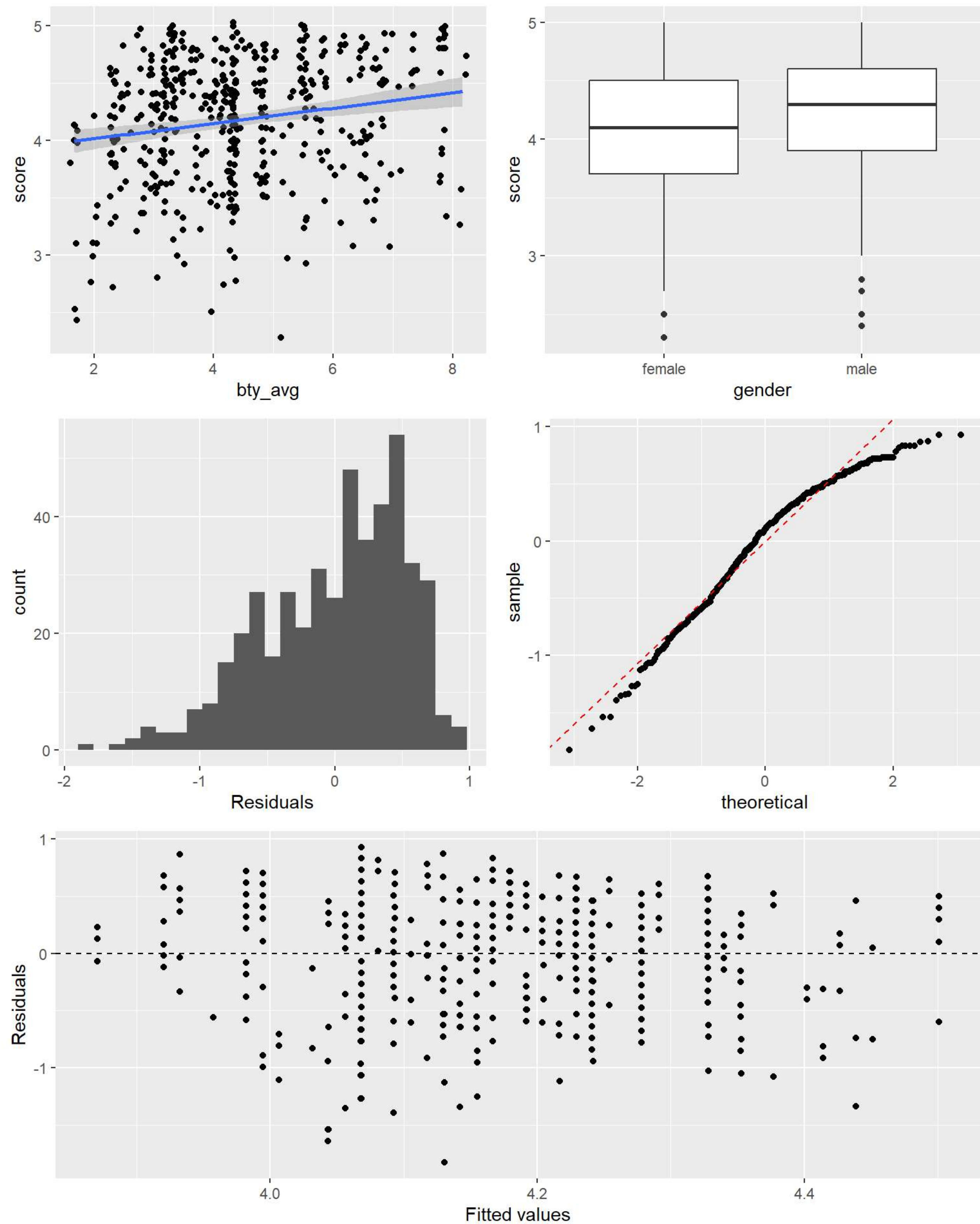
P-values and parameter estimates should only be trusted if the conditions for the regression are reasonable. Verify that the conditions for this model are reasonable using diagnostic plots.

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

Response

```r
p1 <- ggplot(data = evals, aes(x = bty_avg, y = score)) +
  geom_jitter() +
  geom_smooth(method = 'lm', formula = 'y~x')

p2 <- ggplot(data = evals, aes(x = gender, y = score)) +
  geom_boxplot()

p3 <- ggplot(data = m_bty_gen, aes(x = .resid)) +
  geom_histogram(bins = 25) +
  xlab("Residuals")

p4 <- ggplot(data = m_bty_gen, aes(sample = .resid)) +
  stat_qq() +
  geom_abline(intercept = mean(m_bty$residuals),
              slope = sd(m_bty$residuals),
              color = 'red', linetype = 'dashed')

p5 <- ggplot(data = m_bty_gen, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")

plot_grid(plot_grid(p1, p2), plot_grid(p3, p4), p5, nrow = 3, ncol = 1)
```



First, there does appear to be a linear relationship between our independent and target variables. Regarding normal residuals, as above, we do see some skewness in the model residuals, likely as a result of skewness in our target variable (scores). The level of skewness does not appear to be an extreme violation of the normality assumption. Finally, regarding homoskedasticity, residuals again appear random distributed around zero, indicating constant variance.

Exercise 8

Question

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

Response

The average beauty rating remains a significant, though minor, predictor of evaluation scores. The addition of gender did not make beauty any less significant, though the model does have a slightly increased coefficient for this variable. Overall, however, the impact of beauty appears minimal, as does the overall ability of the model to explain the variance in scores (as shown by the low R^2).

Exercise 9

Question

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?

Response

```
m_color <- lm(score ~ pic_color, data = evals)
summary(m_color)
```

```
##
## Call:
## lm(formula = score ~ pic_color, data = evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8369 -0.3615  0.1385  0.4385  0.8631
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.36154    0.06090  71.613  < 2e-16 ***
## pic_colorcolor -0.22466    0.06679  -3.364 0.000833 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5379 on 461 degrees of freedom
## Multiple R-squared:  0.02395,    Adjusted R-squared:  0.02184
## F-statistic: 11.31 on 1 and 461 DF,  p-value: 0.0008334
```

The formula of the line is given as follows:

$$\hat{score} = \hat{\beta}_0 + \hat{\beta}_1 * \text{pic\_color}$$

Exercise 10

Question

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` .

Response

```
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
```

For categorical variables with more than two levels, R automatically one-hot encodes each level beyond the reference level as a separate dummy variable. So, a value of 0 for `ranktenure track` and 0 for `ranktenured` would indicate that rank = teaching. A value of 1 for `ranktenure track` and 0 for `ranktenured` would indicate that rank = tenure track. And finally, a value of 0 for `ranktenure track` and 1 for `ranktenured` would indicate that rank = tenured.

Exercise 11

Question

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.

Response

I would expect that some of the more superficial variables have the weakest relationship with scores. For example, pic_outfit and pic_color should have the weakest relationships.

Exercise 12

Question

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

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
```

As I expected, `pic_outfit` does indeed have an insigificant relationship in the current model. However, the highest p-value is with the `cls_profs` variable.

Exercise 13

Question

Interpret the coefficient associated with the ethnicity variable.

Response

```
m_full$coefficients
```

```
##          (Intercept)      ranktenure track           ranktenured 
##         4.0952140795         -0.1475932457         -0.0973377624 
##           gendermale ethnicitynot minority   languagenon-english 
##         0.2109481296          0.1234929213         -0.2298111901 
##                  age          cls_perc_eval          cls_students 
##        -0.0090071896          0.0053272412          0.0004546339 
##       cls_levelupper       cls_profssingle cls_creditsone credit 
##         0.0605139602         -0.0146619208          0.5020431770 
##              bty_avg  pic_outfitnot formal       pic_colorcolor 
##         0.0400333017         -0.1126816871         -0.2172629964
```

The coefficient for `ethnicitynot minority` is ~0.1234, indicating that when an instructor is not an ethnic minority, their scores are 0.1234 points higher, holding all else equal.

Exercise 14

Question

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?

Response

```
sort(summary(m_full)$coefficients[,4], decreasing = TRUE)
```

```
##       cls_profssingle        cls_levelupper          cls_students 
##          7.780566e-01          2.936925e-01          2.289607e-01 
##           ranktenured  pic_outfitnot formal ethnicitynot minority 
##          1.429455e-01          1.279153e-01          1.169791e-01 
##      ranktenure track   languagenon-english               bty_avg 
##          7.277935e-02          3.965088e-02          2.267440e-02 
##                   age        pic_colorcolor         cls_perc_eval 
##          4.268765e-03          2.516206e-03          5.902546e-04 
##            gendermale cls_creditsone credit           (Intercept) 
##          5.544372e-05          1.839347e-05          1.319326e-37
```

We see that `cls_profs` has the highest p-value.

```
m_fullish <- 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(m_fullish)
```

```
##
## 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
```

After removing `cls_profs` , not much changes. The signs of coefficients are all the same, and significance remains largely unchanged. The `ethnicity` variable has actually increased in significance.

Exercise 15

Question

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.

Response

```r
evals_oh <- mltools::one_hot(data.table::as.data.table(evals)) %>%
  select(-contains('bty_f'),-contains('bty_m'))
colnames(evals_oh) <- colnames(evals_oh) %>%
  str_replace_all(' ', '_') %>%
  str_replace_all('-', '_') %>%
  str_replace_all('&', '_')
exclusions <- c()
model_list <- list()
results <- data.frame(step = as.numeric(),
                      removed_var = as.character(),
                      r_squared = as.numeric(),
                      variables = as.character())

for (i in 1:(length(evals_oh) - 1)) {
  model <- lm(score ~ ., data = select(evals_oh, -all_of(exclusions)))
  worst_pval <- names(sort(summary(model)$coefficients[,4],
                           decreasing = TRUE))[1]
  exclusions <- c(exclusions, worst_pval)
  model_list[[i]] <- model
  result <- data.frame(step = i,
                       removed_var = worst_pval,
                       r_squared = summary(model)$adj.r.squared,
                       variables = paste(colnames(model$model), collapse = ', '))
  results <- rbind(results, result)
}

results %>%
  arrange(desc(r_squared)) %>%
  select(-variables) %>%
  head()
```

```
##   step        removed_var r_squared
## 1    6 cls_profs_multiple 0.1824940
## 2    7   cls_profs_single 0.1824940
## 3    8      cls_perc_eval 0.1824111
## 4    5    cls_level_upper 0.1815861
## 5    4    cls_level_lower 0.1815861
## 6    9 ethnicity_minority 0.1813966
```

This process also requires a ranking metric, for which I've chosen R^2. To facilitate the manual implementation of backward selection, I also one-hot encoded the dataframe before feeding it to the selection loop. I started with all variables (except for redundant beauty variables) and removed the one with the highest p_value at each step until only one variable was left. Then, I ranked all those models by R^2 to choose the "best", given as follows.

$$\hat{score} = \hat{\beta}_0 + \hat{\beta}_1 * \text{course\_id} + \hat{\beta}_2 * \text{rank\_teaching} + \hat{\beta}_3 * \text{ethnicity\_minority} +$$

$$\hat{\beta}_4 * \text{ethnicity\_not\_minority} + \hat{\beta}_5 * \text{gender\_female} + \hat{\beta}_6 * \text{gender\_male} +$$

$$\hat{\beta}_7 * \text{language\_english} + \hat{\beta}_8 * \text{language\_non\_english} + \hat{\beta}_9 * \text{age} +$$

$$\hat{\beta}_{10} * \text{cls\_perc\_eval} + \hat{\beta}_{11} * \text{cls\_did\_eval} + \hat{\beta}_{12} * \text{cls\_students} +$$

$$\hat{\beta}_{13} * \text{cls\_profs\_multiple} + \hat{\beta}_{14} * \text{cls\_profs\_single} + \hat{\beta}_{15} * \text{cls\_credits\_multi\_credit} +$$

$$\hat{\beta}_{16} * \text{cls\_credits\_one\_credit} + \hat{\beta}_{17} * \text{bty\_avg} + \hat{\beta}_{18} * \text{pic\_outfit\_formal} +$$

$$\hat{\beta}_{19} * \text{pic\_outfit\_not\_formal} + \hat{\beta}_{20} * \text{pic\_color\_black\_white} + \hat{\beta}_{21} * \text{pic\_color\_color}$$

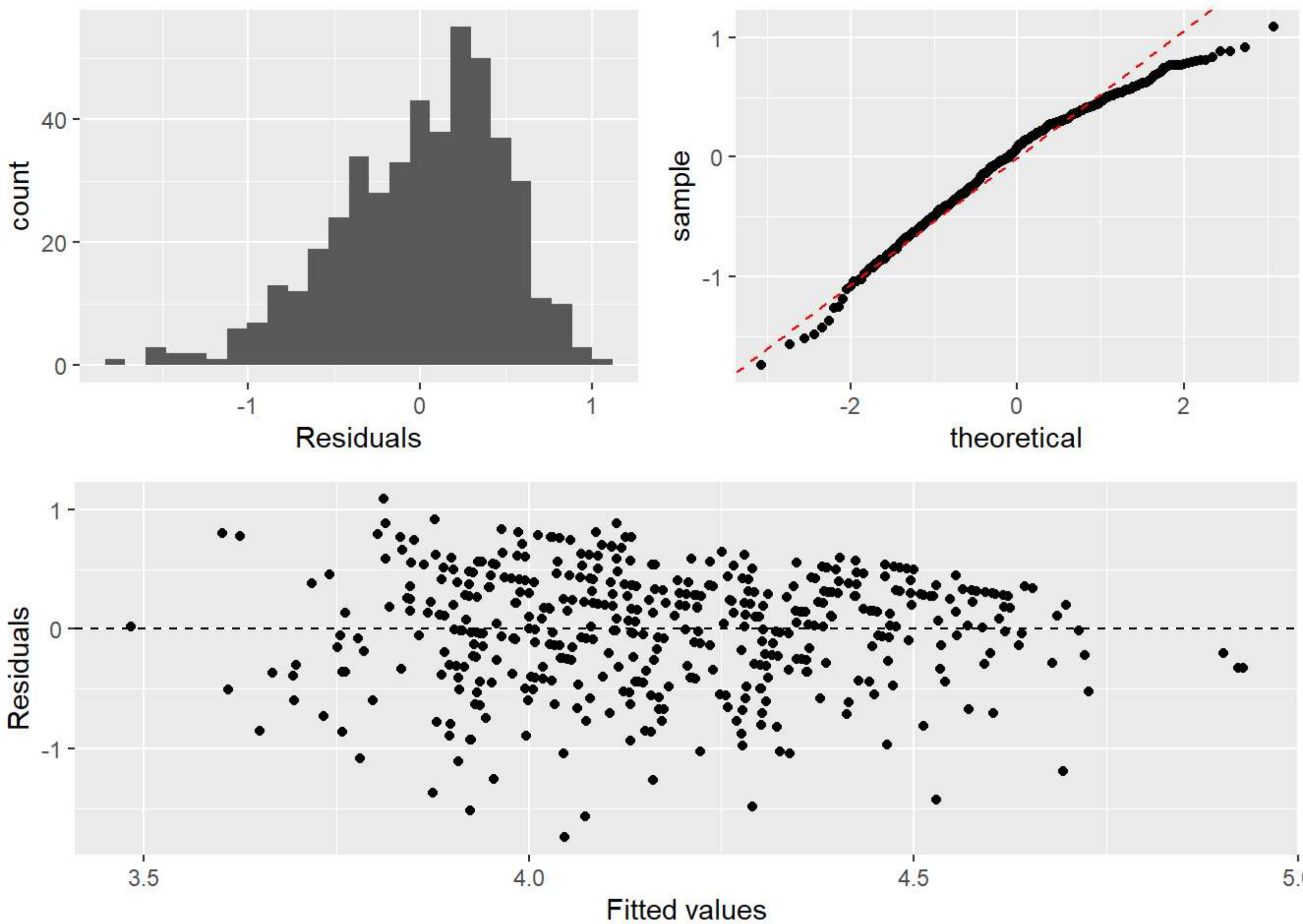
Exercise 16

Question

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

Response

```
p1 <- ggplot(data = model_list[[6]], aes(x = .resid)) +
  geom_histogram(bins = 25) +
  xlab("Residuals")

p2 <- ggplot(data = model_list[[6]], aes(sample = .resid)) +
  stat_qq() +
  geom_abline(intercept = mean(m_bty$residuals),
              slope = sd(m_bty$residuals),
              color = 'red', linetype = 'dashed')


p3 <- ggplot(data = model_list[[6]], aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")

plot_grid(plot_grid(p1, p2), p3, nrow = 2, ncol = 1)
```



Using the top ranked model (i.e. the 6th in our list), we generate the residual plots above. As observed previously, the residuals exhibit some negative skew, but it does not appear sufficient to drive meaningful unreliability in terms of p-values. The residuals also do not appear to exhibit heteroskedasticity.

Exercise 17

Question

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?

Response

The sampling process may introduce some problems with inference. If the unit of observation is a course, rather than an instructor, then the courses were not randomly sampled. Randomly sampling courses would involve starting with the population of all courses scored at UT and then randomly choosing a sample from the course list. Instead, a sample of teachers was chosen. As a result, there may be some violation of the independence condition, which could impact our ability to draw inference from these results.

Exercise 18

Question

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.

Response

```
data.frame(
  variables = names(summary(model_list[[6]])$coefficients[,1]),
  coefficients = summary(model_list[[6]])$coefficients[,1],
  pvals = summary(model_list[[6]])$coefficients[,4]
) %>%
  filter(pvals < 0.1) %>%
  arrange(desc(coefficients))
```

```
##                                          variables  coefficients        pvals
## (Intercept)                            (Intercept)  4.5293682976 1.356415e-46
## pic_color_black_white        pic_color_black_white  0.2947670974 6.011625e-05
## language_english                  language_english  0.2426924284 2.300491e-02
## pic_outfit_formal                pic_outfit_formal  0.1357062427 6.206656e-02
## bty_avg                                    bty_avg  0.0397578528 2.176232e-02
## course_id                                course_id -0.0006587161 9.253912e-04
## age                                            age -0.0072897563 7.274394e-03
## gender_female                        gender_female -0.2351186389 4.971330e-06
## cls_credits_multi_credit cls_credits_multi_credit -0.5383202827 1.988595e-06
```

According to our top-ranked model, the variables that are significant and have the greatest coefficients are listed above. So, we we expect the top-scored instructors to have black-and-white profile pictures of themselves in formal outfits, to speak English, to be physically attractive, to be young, and to be male.

Exercise 19

Question

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

Response

I would not. First, there are potential issues with the sampling approach described above (i.e. the courses were not truly randomly sampled), so inference even within UT may be problematic. More importantly, however, the sample includes no instructors from other universities. Only with conclusions draw from a random sample across many universities would I feel comfortable drawing conslusions about instructors at colleges across the country.
