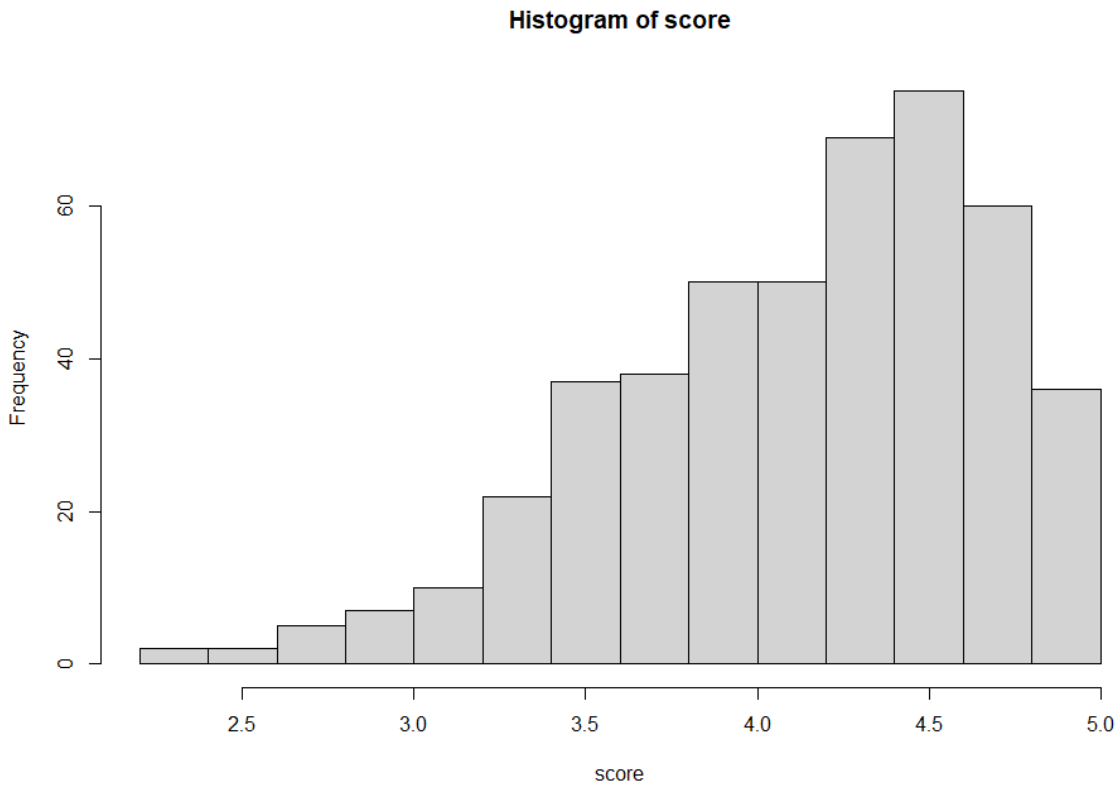
R Assignment 14

Code segments provided for each question, full R script for the assignment at the bottom

> 1: Describe the distribution of score. Is the distribution skewed? What does that tell you about how the students rate the courses? Is this what you expected to see? Why or why not?

```
hist(score)
```



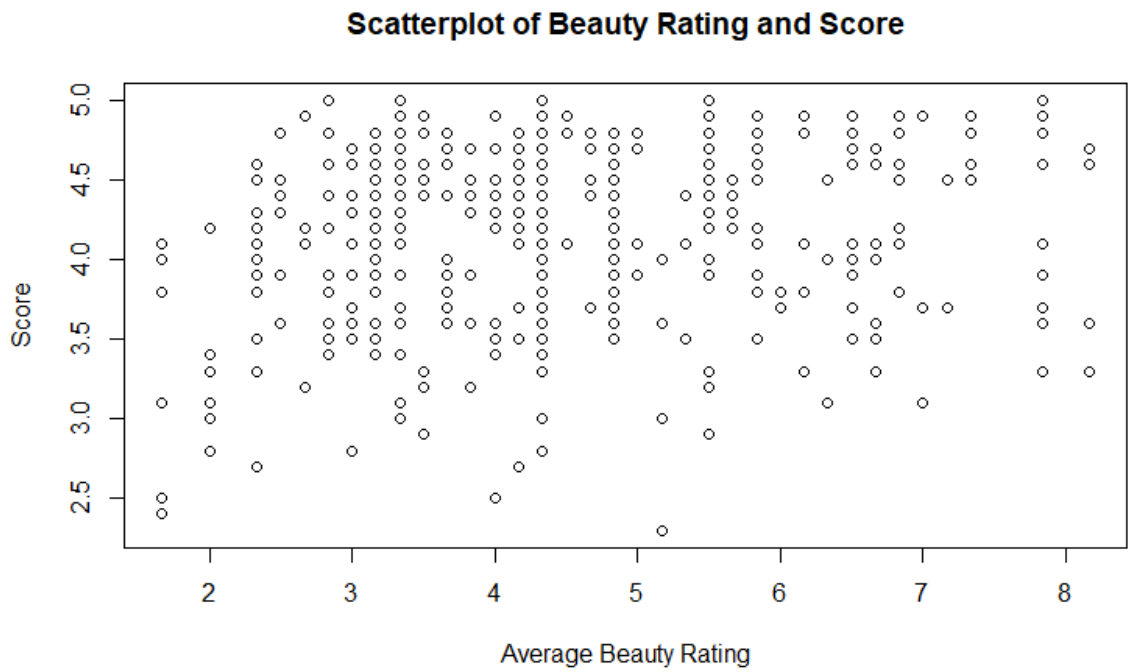
The distribution of the score is strong left-skewed. This means, that students tend to rate their professors higher than the average possible score. This is expected, because people seem to have a tendency to not give bad ratings, unless something is really bad (with exceptions to everything).

Simple Linear Regression

The fundamental phenomenon suggested by the study is that better looking teachers are evaluated more favorably. Let's create a scatterplot to see if this appears to be the case:

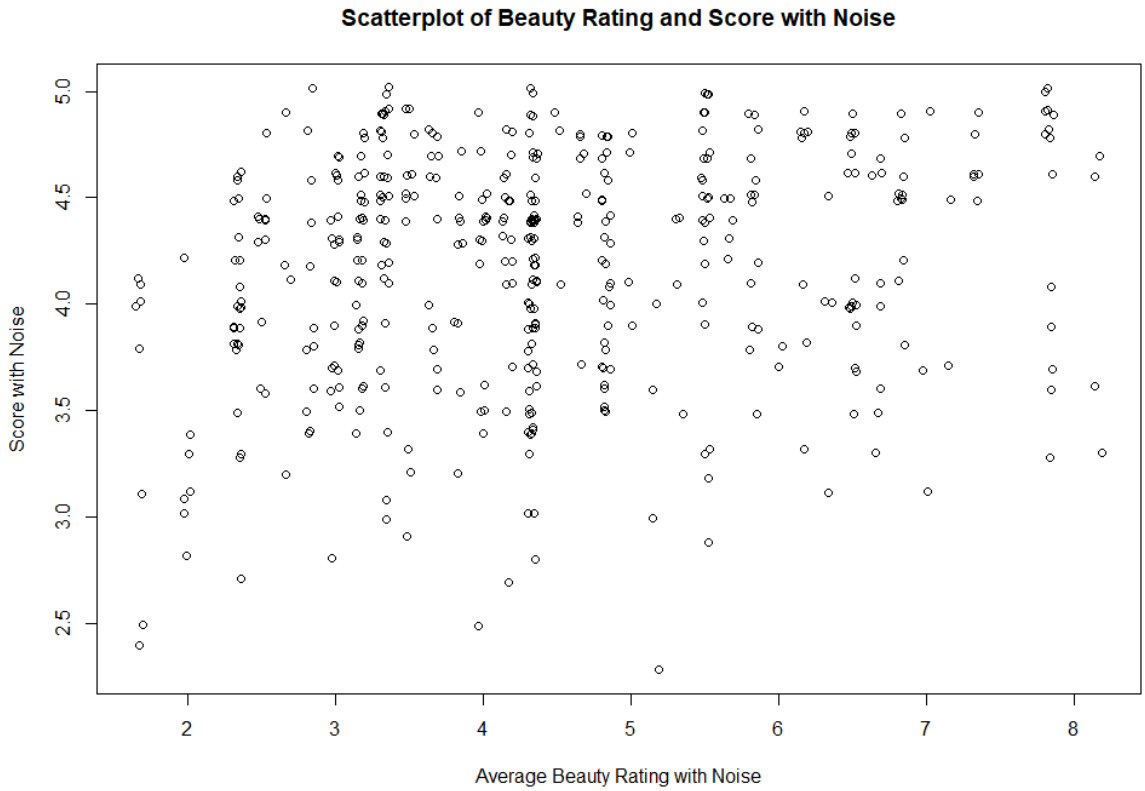
> 2: Create a scatterplot of bty_avg on the x-axis and score on the y-axis. Do you notice anything misleading? Now, remake the plot using the jitter() function (see ?jitter for help). What was misleading about the initial plot?

```
plot(bty_avg, score)
```



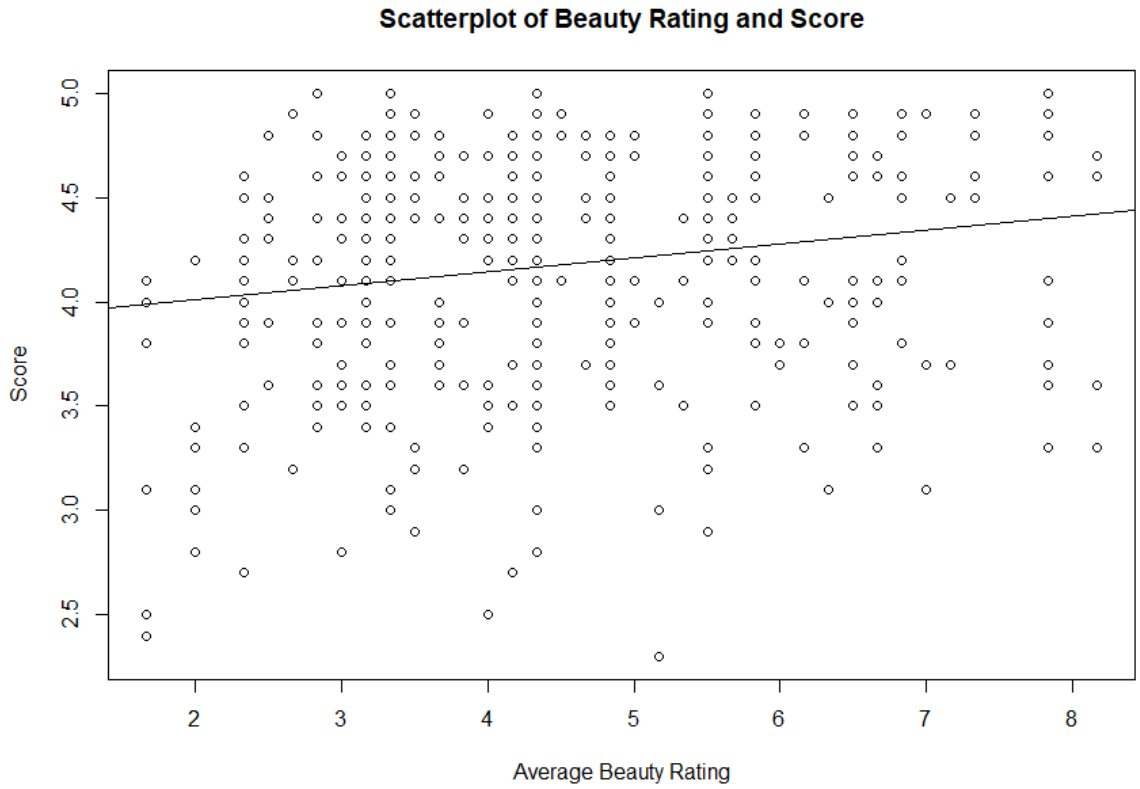
There are clear "lines" detectable for the average beauty rating, indicating that there might be some influence from the evaluation on how the answers for that questions where collected.

```
plot(jitter(bty_avg), jitter(score), main = "Scatterplot of Beauty Rating and
    Score with Noise", xlab = "Average Beauty Rating with Noise",
    ylab = "Score with Noise")
```



> 3: Fit a simple linear regression model with bty_avg as the predictor and score as the response. Call this model m_bty and overlay a line on the scatterplot you made in Question 2 by running abline(m_bty). Share your plot here.

```
m_bty = lm(score ~ bty_avg, data = evals)
abline(m_bty)
```



> 4: Interpret the intercept and slope in the model you fit in Question 3.

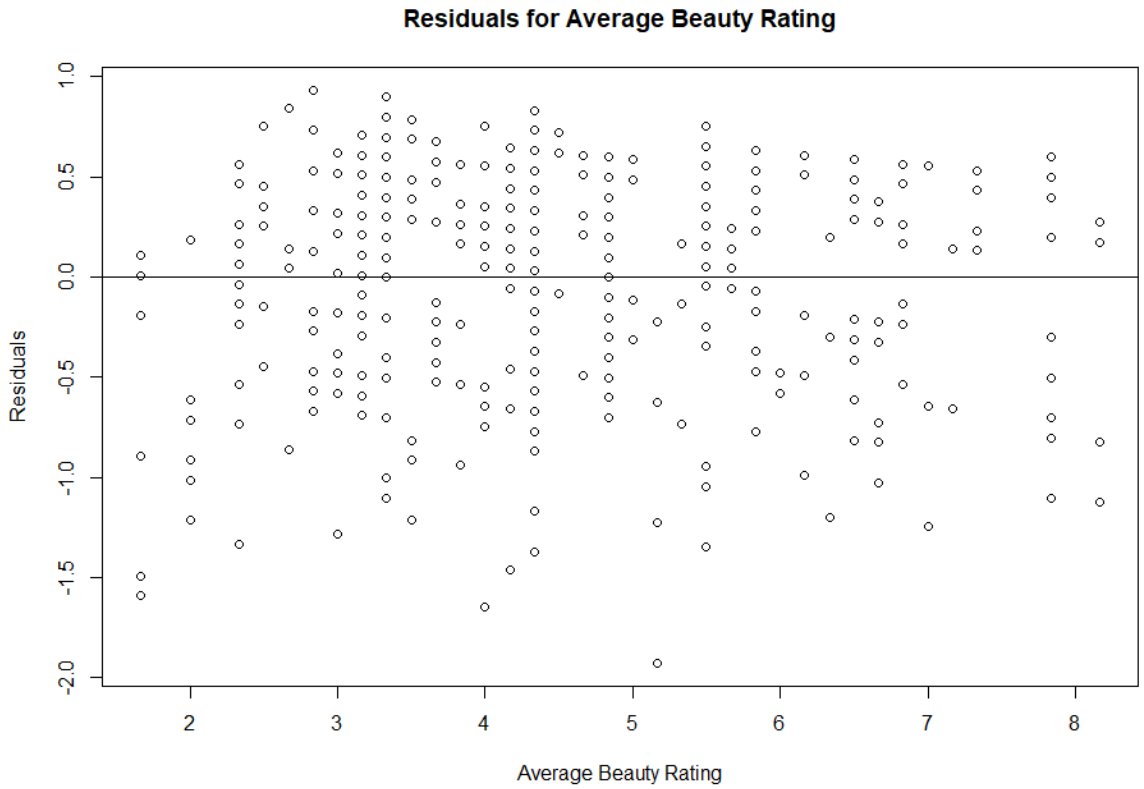
Intercept: 3.88034
Slope: 0.06664

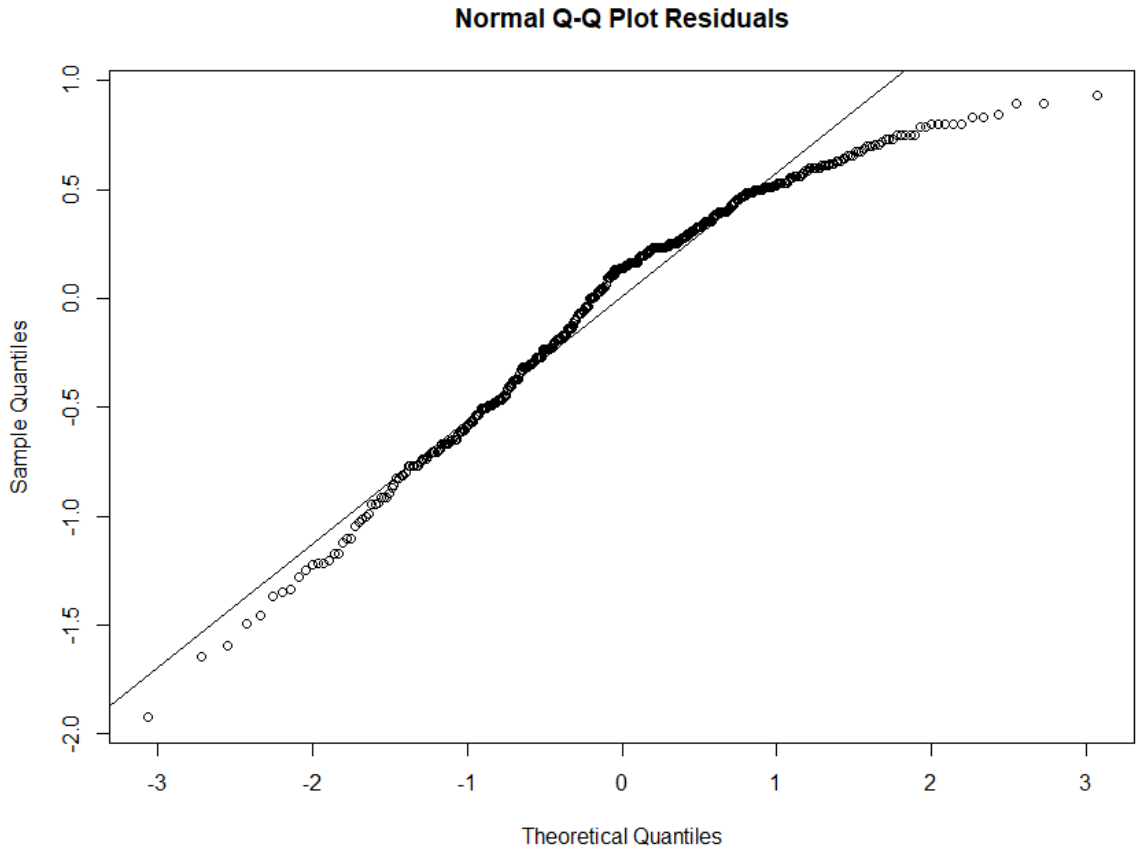
The intercept means, that a professor with an average beauty rating of 0 would have a score of 3.88 and the slope means that for the increase of one unit of the average beauty rating, the score would increase by 0.067.

> 5: Check the residual plot from m_bty. Does it appear that the constant variance and linearity assumption are satisfied? In addition, create a normal probability plot to assess the normality of the residuals and assess the condition here.

```r
# plotting residuals
residualsModel = resid(m_bty)
plot(bty_avg, residualsModel, ylab = "Residuals",
     xlab = "Average Beauty Rating",
     main = "Residuals for Average Beauty Rating")
abline(0, 0)

# normality plot
qqnorm(residualsModel, main = "Normal Q-Q Plot Residuals")
qqline(residualsModel)
```





The variance of the residuals seems to be wider on the left side of the plot and the mean of the residuals seems to be centered around a negative value, so one might argue, that both criteria are not met. The normality plot shows, that the residuals only partly follow the normal distribution but not overall, meaning that this criteria is not met.

Multiple Linear Regression

We will start with a full model that predicts professor score based on rank, gender, age, proportion of students that filled out evaluations, class size, course level, number of professors, number of credits, average beauty rating, outfit, and picture color.

> 6: Interpret the coefficient associated with the gender variable.

	Estimate	Std. Error	t-value	Pr(>|t|)
gendermale	0.2171212	0.0515111	4.215	3.02e-05

The gender variable has female as base. The estimate of 0.217 means, that professors have a 0.217 higher score, if they are male.

> 7: Which predictor had the highest p-value? Drop this variable from the model and refit the model. Did the coefficients and significance of the other explanatory variables change? If not, what does this say about whether or not the dropped variable was collinear with the other explanatory variables?

The highest p-value of 0.553845 has the predictor *number of professors*.

After fitting a new model without this predictor, the coefficients and significance level of the other explanatory variables changed minimally. The adjusted R^2 increases from 0.1471 to 0.1483, meaning that there is a slight influence of that variable on the score.

Now, we are going to implement the backward selection algorithm using the olsrr package in R. To start, install this package and load it in using the library() function.

> 8: Using backward selection and the ols_step_backward_p() function (this function drops variables iteratively based on their p-value), obtain the best model starting with the model in Question 6. Report your final model here and state which variables remain in the final model. Were you surprised by any variables chosen from the backward selection? **Think about goal of the analysis (do you want a causal effect? Are you primarily interested in prediction?) as well to make a final decision on which variables to include in your model. Notice how including different variables changes the coefficients!**

```
dropped_model = ols_step_backward_p(m_multi)
summary(dropped_model$model)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.71937 -0.34425  0.08138  0.38321  0.98082

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.1229960  0.2463891  16.734  < 2e-16 ***
ranktenure track     -0.1995388  0.0786523  -2.537 0.011517 *
ranktenured          -0.1016310  0.0649933  -1.564 0.118584
gendermale            0.2171547  0.0514420   4.221 2.94e-05 ***
age                  -0.0090783  0.0030803  -2.947 0.003372 **
cls_perc_eval         0.0050096  0.0015363   3.261 0.001195 **
cls_students          0.0007086  0.0003593   1.972 0.049235 *
cls_levelupper        0.0695964  0.0559979   1.243 0.214571
cls_creditsone credit 0.4647415  0.1129776   4.114 4.63e-05 ***
bty_avg               0.0408853  0.0174269   2.346 0.019402 *
pic_colorcolor       -0.2386909  0.0689001  -3.464 0.000582 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5019 on 452 degrees of freedom
Multiple R-squared:  0.1668,    Adjusted R-squared:  0.1484
F-statistic: 9.051 on 10 and 452 DF,  p-value: 1.171e-13
```

The variables remaining in the model are rank, gender, age, proportion of students that filled out evaluations, class size, course level, number of credits, average beauty rating, and picture color. I was surprised by the predictor picture color, but overall, these ratings are always influenced by factors not related to the teaching itself, like age, gender or the average beauty rating.

> 9: Verify that the conditions for the final model are reasonable using diagnostic plots.

> 10: Would you be comfortable generalizing your conclusions to apply to professors generally (at any university)? Why or why not?

I would not feel comfortable generalizing my conclusions to apply to professors generally. I think these results very much depend on the demographics of the university the data was taken from, both for students taking part in the evaluation and professors being rated. It might be possible to generalize the results to professors from universities with very similar demographics.

Complete Script for R Assignment 14

```
install.packages("olsrr")
library(olsrr)

load("C:/Users/Caro/Documents/R_631/evals.rda")
attach(evals)

hist(score)

# visualize avg beauty rating and score
plot(bty_avg, score, main = "Scatterplot of Beauty Rating and Score", xlab =
"Average Beauty Rating", ylab = "Score")
plot(jitter(bty_avg), jitter(score), main = "Scatterplot of Beauty Rating and
Score with Noise", xlab = "Average Beauty Rating with Noise", ylab = "Score with
Noise")

# LR
m_bty = lm(score ~ bty_avg, data = evals)
plot(bty_avg, score, main = "Scatterplot of Beauty Rating and Score", xlab =
"Average Beauty Rating", ylab = "Score")    # create plot again to draw line in
plot without noise
abline(m_bty)

# plotting residuals
residualsModel = resid(m_bty)
plot(bty_avg, residualsModel, ylab = "Residuals", xlab = "Average Beauty
Rating", main = "Residuals for Average Beauty Rating")
abline(0, 0)

# normality plot
qqnorm(residualsModel, main = "Normal Q-Q Plot Residuals")
qqline(residualsModel)


# multiple linear regression
# predicts professor score based on rank, gender, age, proportion of students
# that filled out evaluations, class size, course level, number of professors,
# number of credits, average beauty rating, outfit, and picture color

m_multi = lm(score ~ rank + gender + age + cls_perc_eval + cls_students +
               cls_level + cls_profs + cls_credits + bty_avg +
               pic_outfit + pic_color, data = evals)
summary(m_multi)

m_multi_noClsProfs = lm(score ~ rank + gender + age + cls_perc_eval +
               cls_students + cls_level + cls_credits + bty_avg +
               pic_outfit + pic_color, data = evals)
```

```r
summary(m_multi_noClsProfs)

# OLS
dropped_model = ols_step_backward_p(m_multi)
summary(dropped_model$model)

# diagnostic plot for dropped model
```