# Multiple Linear Regression Lab - KEY

*Allison Theobold*

## My answers are in blue!

## Grading the Professor

Many college courses conclude by giving students the opportunity to evaluate the course and the instructor anonymously. However, the use of these student evaluations as an indicator of course quality and teaching effectiveness is often criticized because these measures may reflect the influence of non-teaching related characteristics, such as the physical appearance of the instructor. The article titled, "Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity" (Hamermesh and Parker, 2005) found that instructors who are viewed to be better looking receive higher instructional ratings. (Daniel S. Hamermesh, Amy Parker, Beauty in the classroom: instructors pulchritude and putative pedagogical productivity, Economics of Education Review, Volume 24, Issue 4, August 2005, Pages 369-376, ISSN 0272-7757, 10.1016/j.econedurev.2004.07.013. (http://www.sciencedirect.com/science/article/pii/S0272775704001165.)

In this lab we will analyze the data from this study in order to learn what goes into a positive professor evaluation.

### Data

The data were gathered from end of semester student evaluations for a large sample of professors from the University of Texas at Austin. In addition, six students rated the professors' physical appearance. The result is a data frame where each row contains a different course and columns represent variables about the courses and professors.

```
download.file("http://www.openintro.org/stat/data/evals.RData", destfile = "evals.RData")
load("evals.RData")
```

| variable | description |
|---|---|
| score | average professor evaluation score: (1) very unsatisfactory - (5) excellent |
| rank | rank of professor: teaching, tenure track, tenured |
| ethnicity | ethnicity of professor: not minority, minority |
| gender | gender of professor: female, male |
| language | language of school where professor received education: English or non-English |
| age | age of professor |
| cls_perc_eval | percent of students in class who completed evaluation |

| variable | description |
|---|---|
| cls_did_eval | number of students in class who completed evaluation |
| cls_students | total number of students in class |
| cls_level | class level: lower, upper |
| cls_profs | number of professors teaching sections in course in sample: single, multiple |
| cls_credits | number of credits of class: one credit (lab, PE, etc.), multi credit |
| bty_f1lower | beauty rating of professor from lower level female: (1) lowest - (10) highest |
| bty_f1upper | beauty rating of professor from upper level female: (1) lowest - (10) highest |
| bty_f2upper | beauty rating of professor from second upper level female: (1) lowest - (10) highest |
| bty_m1lower | beauty rating of professor from lower level male: (1) lowest - (10) highest |
| bty_m1upper | beauty rating of professor from upper level male: (1) lowest - (10) highest |
| bty_m2upper | beauty rating of professor from second upper level male: (1) lowest - (10) highest |
| bty_avg | average beauty rating of professor |
| pic_outfit | outfit of professor in picture: not formal, formal |
| pic_color | color of professor's picture: color, black & white |

## Exploring the Data

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.**

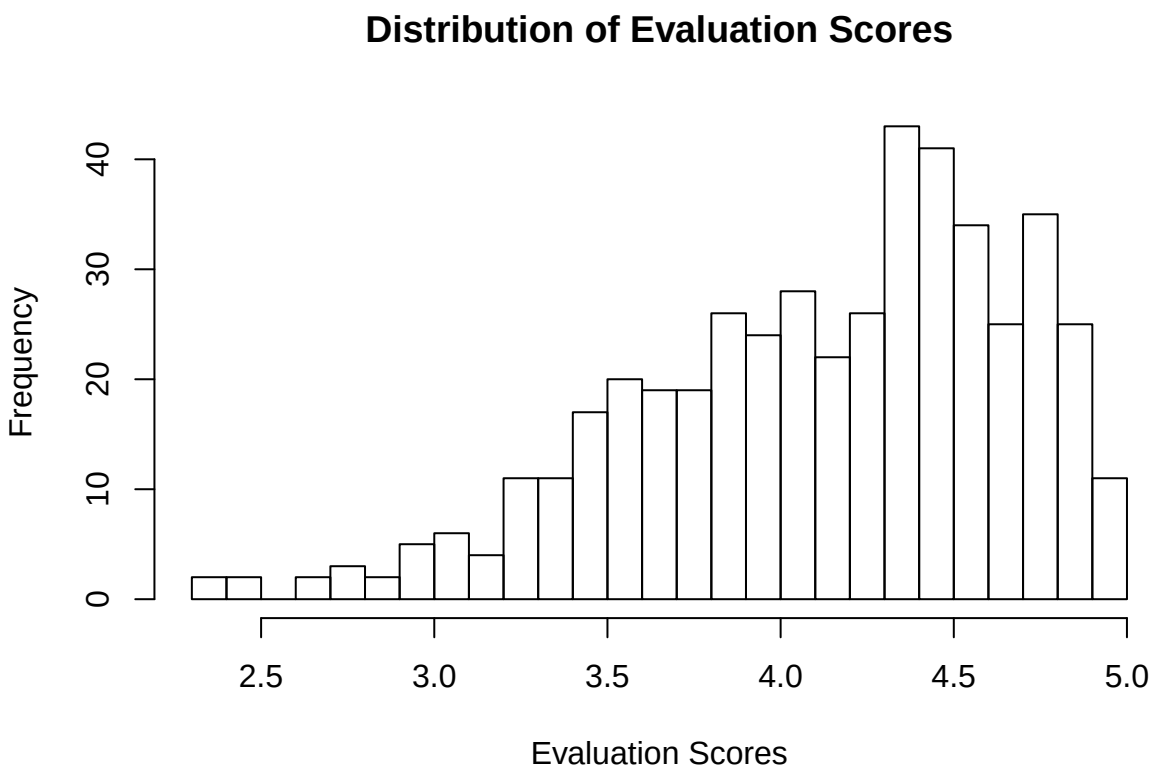
This is an observational study, as the professors at the University of Texas as Austin were observed. The teacher's age, gender, rank, ethnicity, language, and evaluation scores were not randomly assigned, so this is an observational study.

The phrase "whether beauty leads directly to the difference in course evaluations" sounds to me like the authors are seeking to make a *causal* statement. This is not possible, as none of the professor's characteristics (explanatory variables) were randomly assigned. I propose the authors rephrase their research question to:

What is the *relationship* between beauty and course evaluations, for professors at the University of Texas at Austin?

2. **Based on the plots below, 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?**

```
hist(evals$score, xlab = "Evaluation Scores", main = "Distribution of Evaluation Scores", nclass = 25)
```

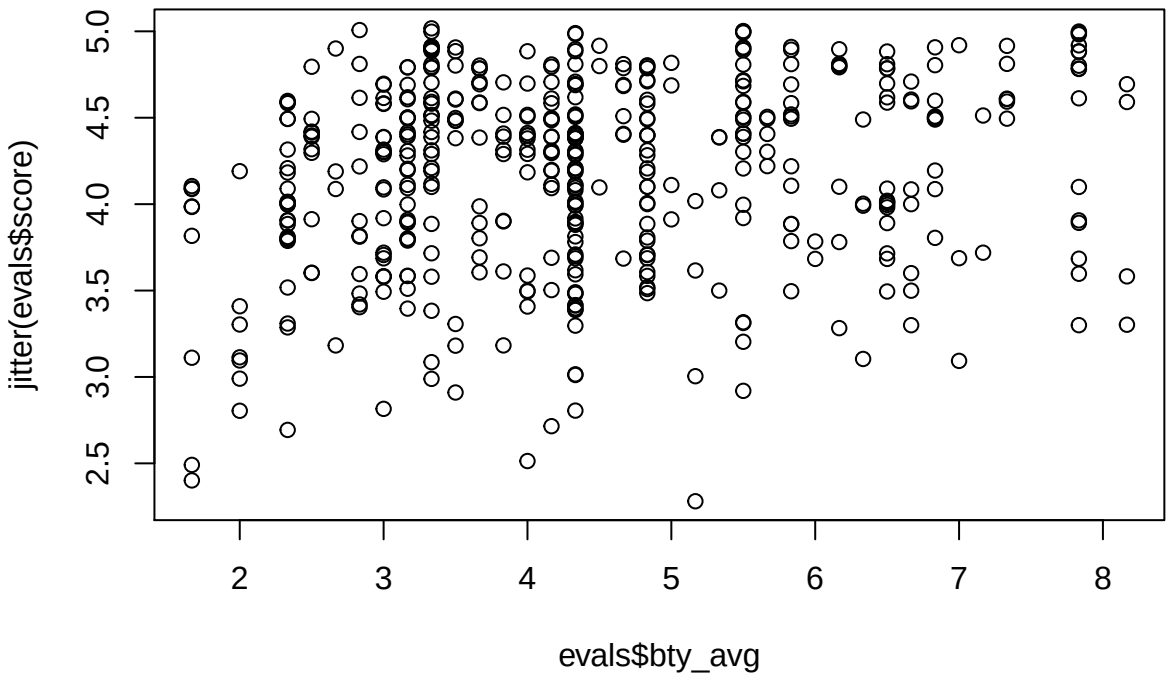



The distribution of evaluation scores is left-skewed (long left tail), so the lower evaluation scores (2-3) occur less frequently and higher evaluation scores (4-5) occur more frequently. This is not what I expected to see, as I thought there would be a more bell shaped distribution of scores.

## 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:

```
plot(jitter(evals$score) ~ evals$bty_avg)
```




3. **What relationship do you see in the scatterplot above?**

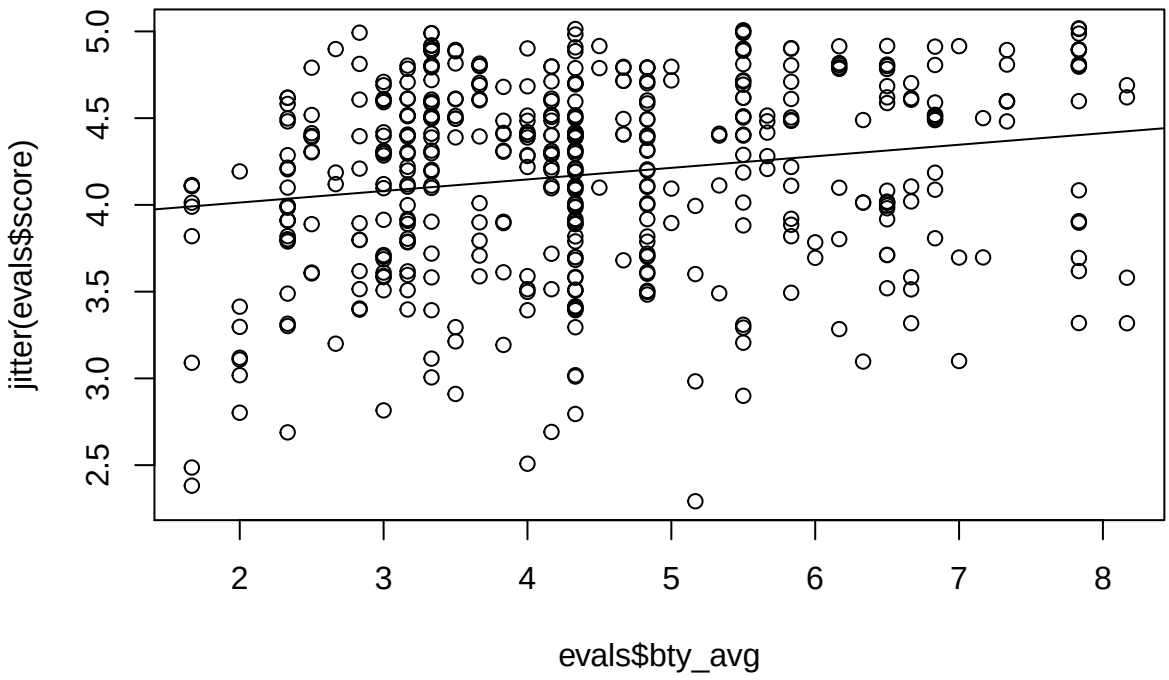
There appears to be a moderate, positive, semi-linear relationship between the professor's evaluation score and their average beauty score. I suggest a semi-linear relationship, as if you construct a line of best fit there appears to be a spike in the evaluation scores when going from an average beauty score of 2 to 3.

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

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

Now, we can add this regression line the scatterplot using abline(m_bty).

```
plot(jitter(evals$score) ~ evals$bty_avg)
abline(m_bty)
```




4. **Write out the equation for the linear model *and* interpret the slope.**

```
summary(m_bty)$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 3.88033795 0.07614297 50.961212 1.561043e-191
## bty_avg     0.06663704 0.01629115  4.090382  5.082731e-05
```

$$\widehat{\text{eval score}} = 3.88 + 0.067 * (\text{avg beauty})$$

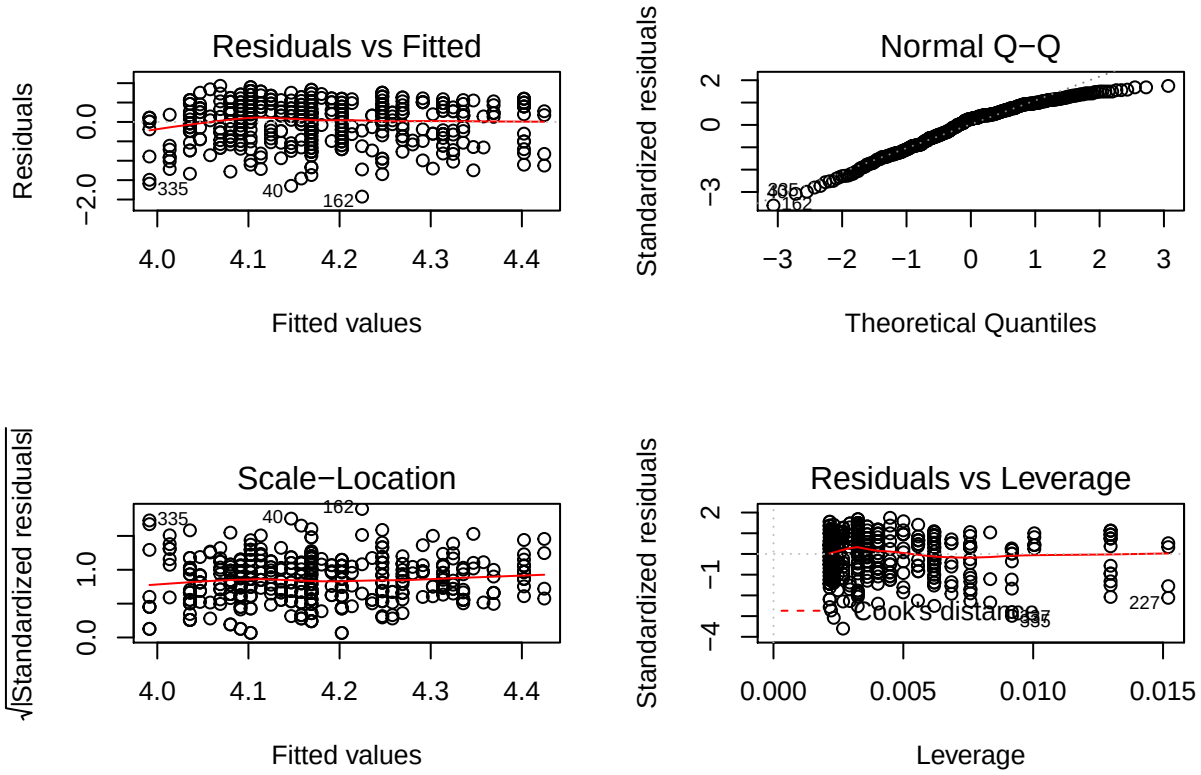
As we increase average beauty rating by 1, on average, we expect the evaluation score to increase by 0.067.

5. **Is average beauty score a statistically significant predictor? Does it appear to be a practically significant predictor? (hint: we describe predictors as "practically significant" if they have a "large" estimated effect)**

Average beauty is a "statistically significant" predictor, as the p-value is quite small (0.0000508). However, the coefficient estimate of average beauty is quite small (0.067), so changes in average beauty rating have a very small impact on the evaluation score. For this reason, I would say that average beauty is not a "practically significant" predictor.

6. **Use diagnostic plots and critical thinking to evaluate whether the conditions of simple linear regression are reasonable. Provide plots and comments for each one.**

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

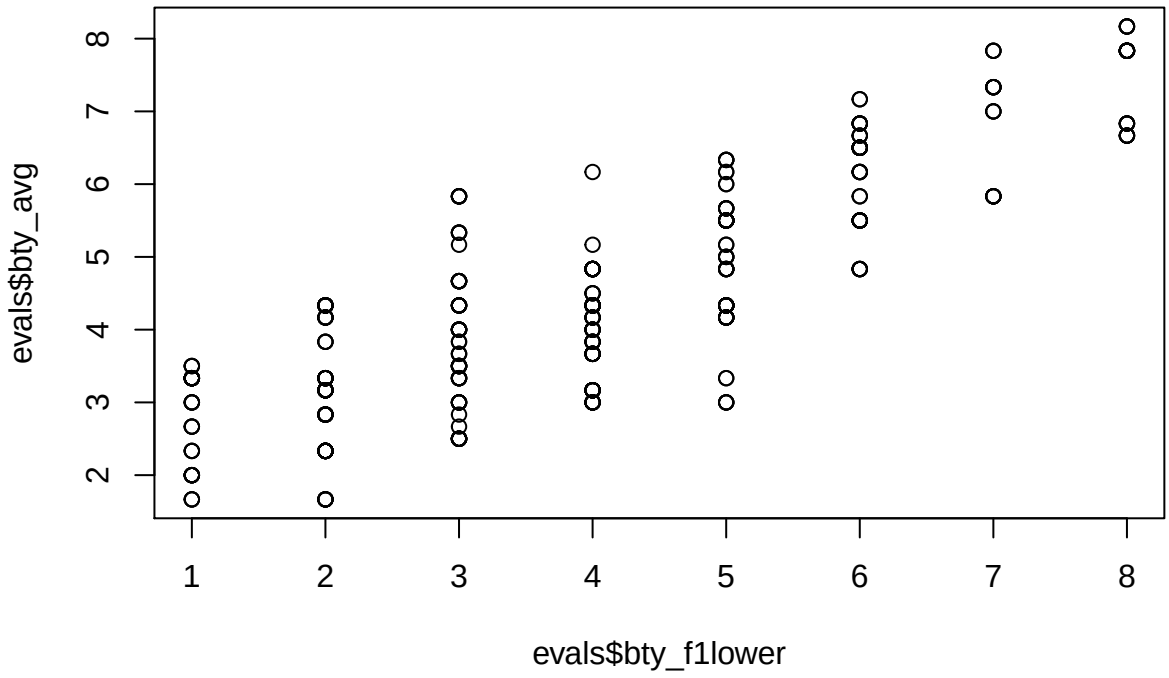



- Independence: As stated in the introduction, in the dataset "each row contains a different course and columns represent variables about the courses and professors." This means that our observations (rows in the dataset) are the courses, which we want to be independent of each other. **If** a professor teaches more than one course, they would be included in multiple rows of our dataset, violating our assumption of independence.
- Linear Relationship: As seen in the scatterplot, it appears that the relationship between evaluation score and average beauty is linear, with the exception of the jump in evaluation scores when going from beauty scores of 2 to 3.
- Constant Variance: In the Residuals vs Fitted plot, we see a fairly uniform spread of residuals across the range of fitted values.
- Normality of Residuals: In the Normal Q-Q plot, we see strong evidence of a short, right tail (dots are below the line). As a short tail is a "more conservative" distribution (Greenwood, 2017, pg. 97) which is "relatively harmless" so we can "proceed with methods that assume normality."
- No Influential Points: In the Residuals vs Leverage plot, we see no points with large Cook's Distance (greater than 5 or 10).

## Multiple Linear Regression

The data set contains several variables on the beauty score of the professor: individual ratings from each of the six students who were asked to score the physical appearance of the professors and the average of these six scores. Let's take a look at the relationship between one of these scores and the average beauty score.

```
plot(evals$bty_avg ~ evals$bty_f1lower)
```

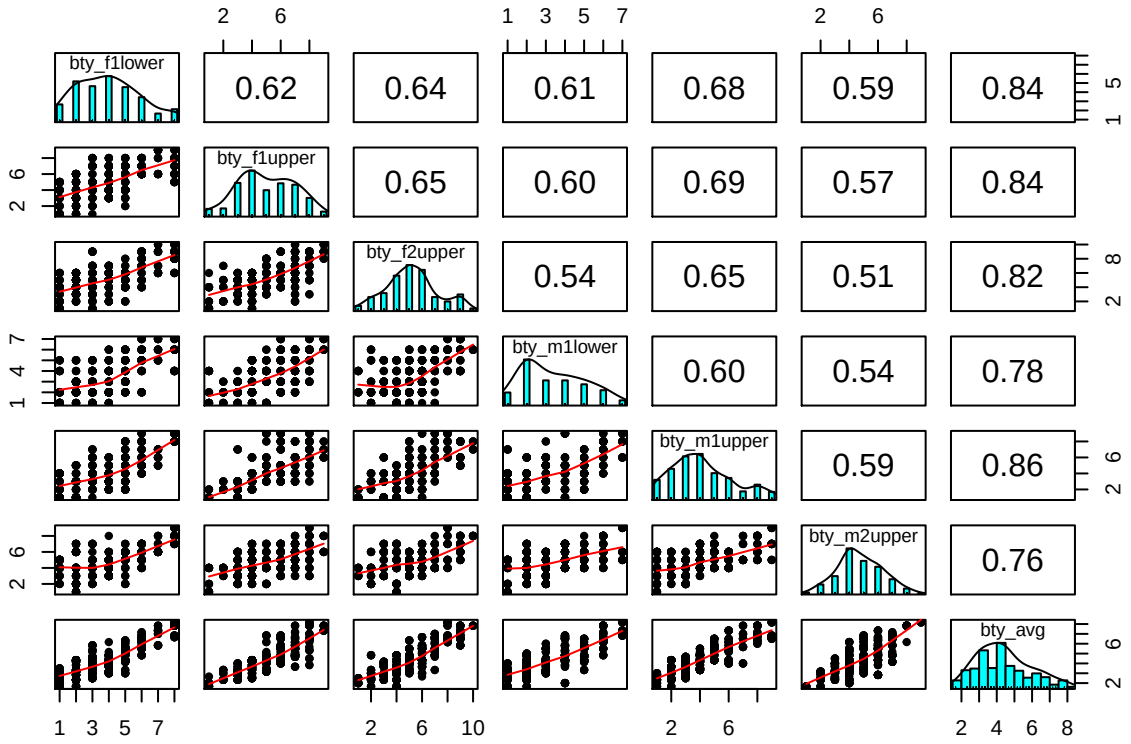



```
cor(evals$bty_avg, evals$bty_f1lower)
```

```
## [1] 0.8439112
```

As expected the relationship is quite strong - after all, the average score is calculated using the individual scores. We can actually take a look at the relationships between all beauty variables (columns 13 through 19) by making a scatterplot matrix.

```
library(psych)
pairs.panels(evals[,13:19], ellipses = FALSE)
```




7. **What statistical term do we use to describe when there are "large" correlations between explanatory variables?**

**Multicollinearity!**

## The Search for the Best Model

We will start with a full model that predicts professor score based on all of the available quantitative predictors: age, proportion of students that filled out evaluations, the number of students that did the evaluation, class size, and all 7 of the available beauty ratings.

Let's run the model.

```
m_full <- lm(score ~ age + cls_perc_eval + cls_did_eval + cls_students + bty_f1lower +
    bty_f1upper + bty_f2upper + bty_m1lower + bty_m1upper + bty_m2upper + bty_avg,
    data = evals)

summary(m_full)
```

```
##
## Call:
## lm(formula = score ~ age + cls_perc_eval + cls_did_eval + cls_students +
##     bty_f1lower + bty_f1upper + bty_f2upper + bty_m1lower + bty_m1upper +
##     bty_m2upper + bty_avg, data = evals)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9130 -0.3466  0.1252  0.3938  1.1157
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.668826   0.240403  15.261   <2e-16 ***
## age            -0.002010   0.002971  -0.677   0.4990
## cls_perc_eval   0.004479   0.002224   2.014   0.0446 *
## cls_did_eval    0.002321   0.003235   0.717   0.4735
## cls_students   -0.000801   0.002022  -0.396   0.6922
## bty_f1lower     6.247973  15.764776   0.396   0.6921
## bty_f1upper     6.271251  15.760392   0.398   0.6909
## bty_f2upper     6.239297  15.761817   0.396   0.6924
## bty_m1lower     6.208149  15.761460   0.394   0.6939
## bty_m1upper     6.250523  15.762688   0.397   0.6919
## bty_m2upper     6.244747  15.758114   0.396   0.6921
## bty_avg       -37.421076  94.572327  -0.396   0.6925
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5291 on 451 degrees of freedom
## Multiple R-squared:  0.07622,    Adjusted R-squared:  0.05369
## F-statistic: 3.383 on 11 and 451 DF,  p-value: 0.0001624
```

8. **What do you notice about the standard errors for the beauty variables?**

The standard errors for the beauty variables are all incredibly large (over 15), with average beauty having a standard error off 95!

When we believe that the explanatory variables are highly correlated with each other, the standard errors are inflated. We can actually measure how much each of the standard errors are inflated because of multicollinearity with other variables in the model, using what are called *variance inflation factors* (or VIFs). VIFs provide a way to assess the multicollinearity in the MLR model that is caused by including specific explanatory variables. The amount of information that is shared between a single explanatory variable and the others can be found by regressing that variable on the others and calculating the $R^2$ for that model. More information about VIFs can be found in the textbook on pages 302 to 304.

The easy way to obtain VIFs is using the vif function from the car package (Fox, 2003). Run the following code!

```
library(car)
vif(m_full) ## VIFs
```

```
##           age cls_perc_eval  cls_did_eval  cls_students   bty_f1lower
##  1.400105e+00  2.291530e+00  3.501024e+01  3.803838e+01  1.440490e+06
##   bty_f1upper   bty_f2upper   bty_m1lower   bty_m1upper   bty_m2upper
##  1.534151e+06  1.670230e+06  1.098928e+06  1.826803e+06  1.017048e+06
##       bty_avg
##  3.443876e+07
```

Basically, large VIFs are bad, with the rule of thumb that values over 5 or 10 are considered "large" values indicating high multicollinearity in the model for **that particular variable**. We use this scale to determine if multicollinearity is a problem for a variable of interest.

```
sqrt(vif(m_full)) ## square root of VIFs
```

```
##           age cls_perc_eval  cls_did_eval  cls_students   bty_f1lower
##      1.183260      1.513780      5.916945      6.167527   1200.203991
##   bty_f1upper   bty_f2upper   bty_m1lower   bty_m1upper   bty_m2upper
##   1238.608524   1292.373932   1048.297773   1351.592807   1008.488006
##       bty_avg
##   5868.454387
```

If we take the square root of the VIF numbers next each variable, we can interpret these numbers as "the number of times larger the standard error for the slope for **that variable** is, due to collinearity with other variables in the model."

9. **Based on the above VIF output, which variables have "large" multicollinearity problems? (hint: remember that 1.82e+06 in R means 1820000)**

The beauty variables all have very high VIFs, all greater than 5. In the square-root of the VIFs, we see that the standard errors for **all** of the beauty variables are over 1200 times larger, due to multicollinearity problems.

Having more than one of the beauty variables in the model seems like a bad choice, since they all are highly correlated with each other. In this application and with these highly-correlated predictors, I would choose the average beauty score as the single representative of these variables. Since the correlations between the beauty variables are the highest when comparing with average beauty score (all correlations $> 0.75$), it seems to be a reasonable choice.

10. **Drop all of the beauty variables except the average beauty score from the model.**

```
## new MLR model with ONLY bty_avg, cls_perc_eval, cls_did_eval,
## cls_students, age AS EXPLANATORY VARIABLES

model2 <- lm(score ~ age + cls_perc_eval + cls_did_eval + cls_students + bty_avg,
    data = evals)

summary(model2)$coefficients
```

```
##                    Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)    3.6484628036 0.230526187 15.8266740 2.667851e-45
## age           -0.0024555046 0.002636873 -0.9312182 3.522324e-01
## cls_perc_eval  0.0053003790 0.002135951  2.4815078 1.344129e-02
## cls_did_eval   0.0017099142 0.003135162  0.5453989 5.857453e-01
## cls_students  -0.0004919387 0.001970245 -0.2496840 8.029439e-01
## bty_avg        0.0486779854 0.017242151  2.8231967 4.962363e-03
```

11. **Did the standard errors of the explanatory variables change? How did they change?**

The sign of beauty average changed (negative to positive) and the standard error went from 94.5 to 0.02! The signs of the other coefficients and the size of the coefficients did not change.

12. **Add `Rank`, `Ethnicity`, `Gender`, and `Language` into your model (additively). Then, with this new model, drop the variable with the highest p-value (as long as it is above 0.05) and re-fit 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?**

```
model3 <- lm(score ~ rank + ethnicity + gender + language + age + cls_perc_eval + cls_students +
               cls_did_eval + bty_avg, data = evals)

summary(model3)$coefficients
```

```
##                             Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)            3.7769046642 0.260689549 14.4881323 2.414360e-39
## ranktenure track      -0.1589578489 0.082610170 -1.9241923 5.495800e-02
## ranktenured           -0.1412748175 0.062299872 -2.2676582 2.382132e-02
## ethnicitynot minority  0.0937442681 0.074413554  1.2597741 2.084009e-01
## gendermale             0.2116848809 0.052025897  4.0688368 5.575838e-05
## languagenon-english   -0.1714174620 0.108663645 -1.5775052 1.153789e-01
## age                   -0.0067240067 0.003134440 -2.1452022 3.246891e-02
## cls_perc_eval          0.0052177034 0.002109074  2.4739313 1.372981e-02
## cls_students          -0.0009161855 0.001945081 -0.4710269 6.378488e-01
## cls_did_eval           0.0021438583 0.003091974  0.6933623 4.884384e-01
## bty_avg                0.0516205928 0.016857931  3.0620954 2.329051e-03
```

```
# delete cls_students from the model (p-value = 0.637)

model4 <- lm(score ~ rank + ethnicity + gender + language + age + cls_perc_eval + cls_did_eval + bty_avg
             data = evals)

summary(model4)$coefficients
```

```
##                             Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)            3.7268689067 0.2378602953 15.668310 1.592534e-44
## ranktenure track      -0.1617767621 0.0823223147 -1.965163 5.000599e-02
## ranktenured           -0.1418762040 0.0622332688 -2.279749 2.308669e-02
## ethnicitynot minority  0.0974285029 0.0739377530  1.317710 1.882668e-01
## gendermale             0.2089733786 0.0516620100  4.045011 6.150085e-05
## languagenon-english   -0.1671727540 0.1081962831 -1.545088 1.230233e-01
## age                   -0.0067398098 0.0031315670 -2.152216 3.190798e-02
## cls_perc_eval          0.0059177497 0.0014951641  3.957927 8.775255e-05
## cls_did_eval           0.0007118999 0.0005636362  1.263049 2.072214e-01
## bty_avg                0.0515556351 0.0168428823  3.060975 2.337313e-03
```

With the inclusion of rank, ethnicity, gender, and language the coefficients for age, percent eval, and average beauty differ from `model2`. The slope on age nearly tripled (before: -0.0024555046, after: -0.0067240067), the slope on percent eval stayed about the same, and the estimated slopes on total students (in the course) and students doing the eval nearly doubled.

When deleting `cls_students` from the model, the coefficients did not change substantially. This indicates to us that age, percent evals, number of evaluations completed, rank, ethnicity, gender, language, and average beauty are not correlated with the number of student enrolled that was removed from the model.

A model selection method used in statistics is called "backward-selection". The process is as follows:

- Fit the full model (all possible quantitative variables)
- Find the p-values of each variable in the model
- Delete the **one** variable with the **largest** p-value, as long as the p-value is larger than your specified significance level (say, $\alpha = 0.05$)
- Re-fit the model without that deleted variable
- Find the p-values of each variable in the **new** model
- Delete the **one** variable with the **largest** p-value, as long as the p-value is larger than your specified significance level (say, $\alpha = 0.05$)

This process continues until **all** of the variables included in the model have p-values less than your specified significance level (say, $\alpha = 0.05$).

13. **Using this procedure as the selection method, determine the best model. You do not need to show all steps in your answer, just the output for the final model.**

```
library(car)
Anova(model4)
```

```
## Anova Table (Type II tests)
##
## Response: score
##                Sum Sq  Df F value    Pr(>F)
## rank            1.617   2  3.0688  0.047444 *
## ethnicity       0.458   1  1.7364  0.188267
## gender          4.311   1 16.3621 6.150e-05 ***
## language        0.629   1  2.3873  0.123023
## age             1.221   1  4.6320  0.031908 *
## cls_perc_eval   4.128   1 15.6652 8.775e-05 ***
## cls_did_eval    0.420   1  1.5953  0.207221
## bty_avg         2.469   1  9.3696  0.002337 **
## Residuals     119.362 453
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Delete cls_did_eval deleted (p-value = 0.207)

model5 <- lm(score ~ rank + ethnicity + gender + language + age + cls_perc_eval + bty_avg,
             data = evals)

Anova(model5)
```

```
## Anova Table (Type II tests)
##
## Response: score
##                Sum Sq  Df F value    Pr(>F)
## rank            1.446   2  2.7413 0.0655559 .
## ethnicity       0.492   1  1.8629 0.1729660
## gender          4.564   1 17.2983 3.821e-05 ***
## language        0.693   1  2.6250 0.1058881
## age             1.334   1  5.0566 0.0250116 *
## cls_perc_eval   3.766   1 14.2743 0.0001791 ***
## bty_avg         2.929   1 11.1008 0.0009334 ***
## Residuals     119.782 454
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Delete ethnicity (p-value = 0.173)

model6 <- lm(score ~ rank + gender + language + age + cls_perc_eval + bty_avg, data = evals)

Anova(model6)
```

```
## Anova Table (Type II tests)
##
## Response: score
##                Sum Sq  Df F value    Pr(>F)
## rank            1.533   2  2.9004 0.0560187 .
## gender          4.826   1 18.2574 2.352e-05 ***
```

```
## language        1.110   1  4.1993 0.0410132 *
## age             1.315   1  4.9731 0.0262319 *
## cls_perc_eval   3.480   1 13.1648 0.0003176 ***
## bty_avg         2.964   1 11.2142 0.0008792 ***
## Residuals     120.274 455
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## All p-values are below 0.1, so this is my final model!

summary(model6)$coefficients
```

```
##                          Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept)          3.869781442 0.223576813 17.308510 6.178844e-52
## ranktenure track    -0.168650493 0.082171491 -2.052421 4.070029e-02
## ranktenured         -0.129888496 0.061498183 -2.112070 3.522480e-02
## gendermale           0.219708514 0.051419471  4.272866 2.351568e-05
## languagenon-english -0.213899702 0.104380852 -2.049224 4.101319e-02
## age                 -0.006975629 0.003128030 -2.230039 2.623192e-02
## cls_perc_eval        0.005276418 0.001454226  3.628335 3.175828e-04
## bty_avg              0.055612913 0.016606982  3.348767 8.791897e-04
```

14. **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 professor with a high evaluation score would be a young, male, teaching (non-tenure) faculty, who went to an English speaking university. They also would have a large percentage of their students complete the course evaluation and would have a high average beauty score.

15. **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 is bad!** Our original assumption of independence was not violated if each professor taught only one course and was only included in one row of our dataset. Since a professor can teach more than one course, they would be included in multiple rows of our dataset, violating our assumption of independence.

This violation is because knowing what a professor teaches (e.g. statistics 217 and psych 101) would make it so knowing one course's evaluation scores **would** give you information about what you would expect another course's evaluation scores to be.