Grading the Professor

Group Names:

Due: Monday, April 23 (in class)

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 end to 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).

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 course and its associated professors.

```
evals <- read.csv("evals.csv", header = TRUE)
```

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
cls_did_eval	number of students in class who completed evaluation
cls_students	total number of students in class

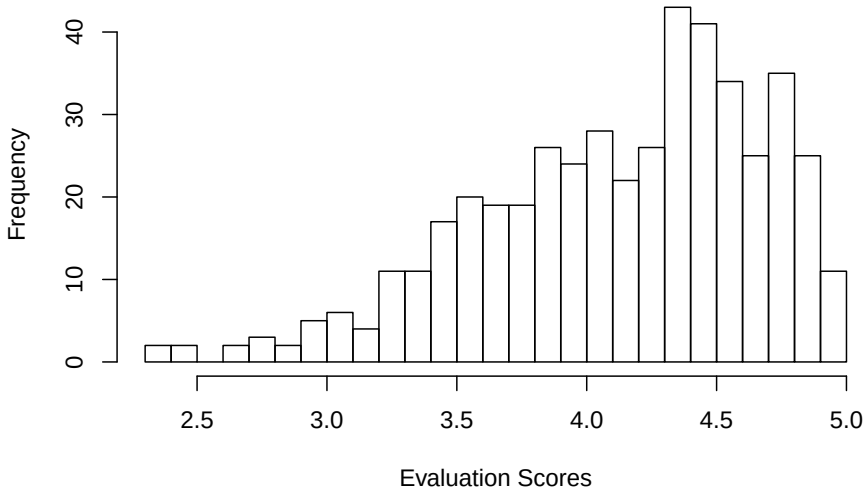
variable	description
cls_level	class level: lower, upper
cls_profs	number of professors teaching sections in course in sample: single, multiple
cls_credits	number of credits of class: one credit (lab, PE, etc.), multi credit
bty_f1lower	beauty rating of professor from lower level female: (1) lowest - (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 differences in course evaluations. Given the study design, is it possible to answer this question as it is phrased? If not, rephrase the question.**

2. **Based on the plot below, describe the distribution of evaluation scores. 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?**

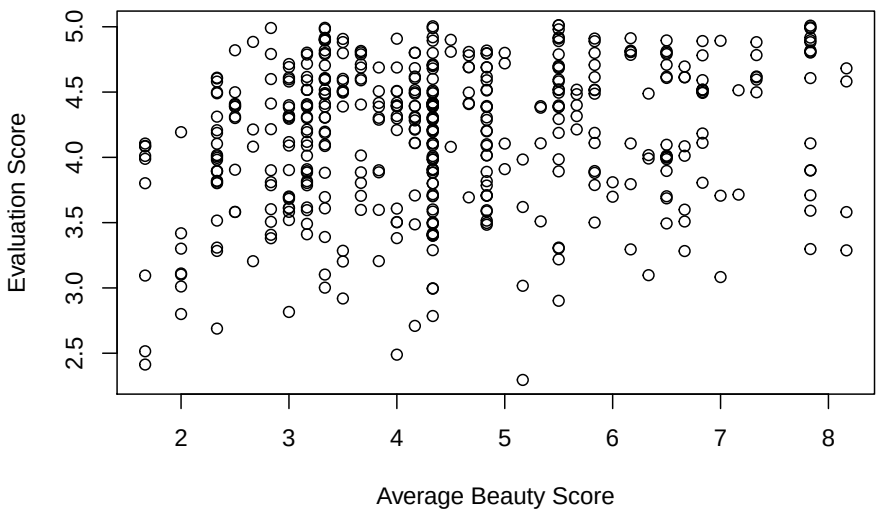
```
hist(evals$score, xlab = "Evaluation Scores", main = "", nclass = 25)
```



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, xlab = "Average Beauty Score", ylab = "Evaluation Score")
```



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

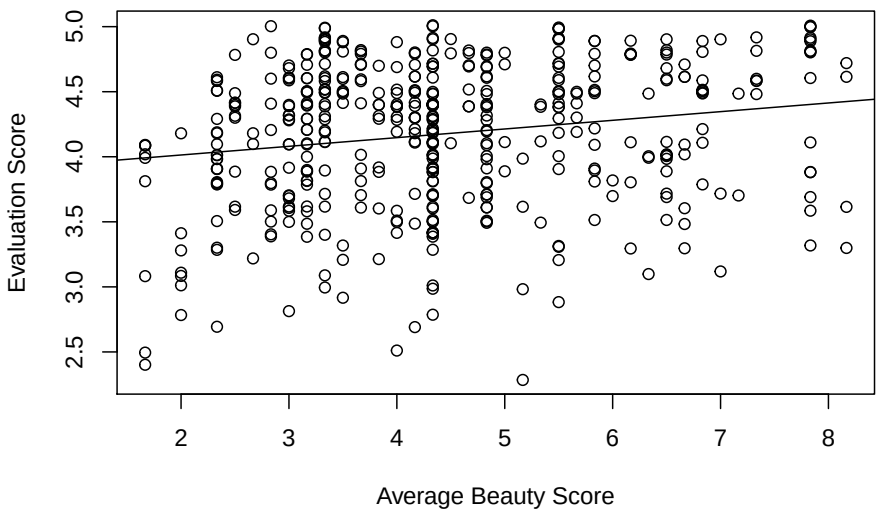
Let's see if the apparent trend in the plot is something more than natural variation. Fit the linear model called `m_bty` to predict average professor score by average beauty rating.

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

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

```
plot(jitter(evals$score) ~ evals$bty_avg, xlab = "Average Beauty Score", ylab = "Evaluation Score")
abline(m_bty)
```

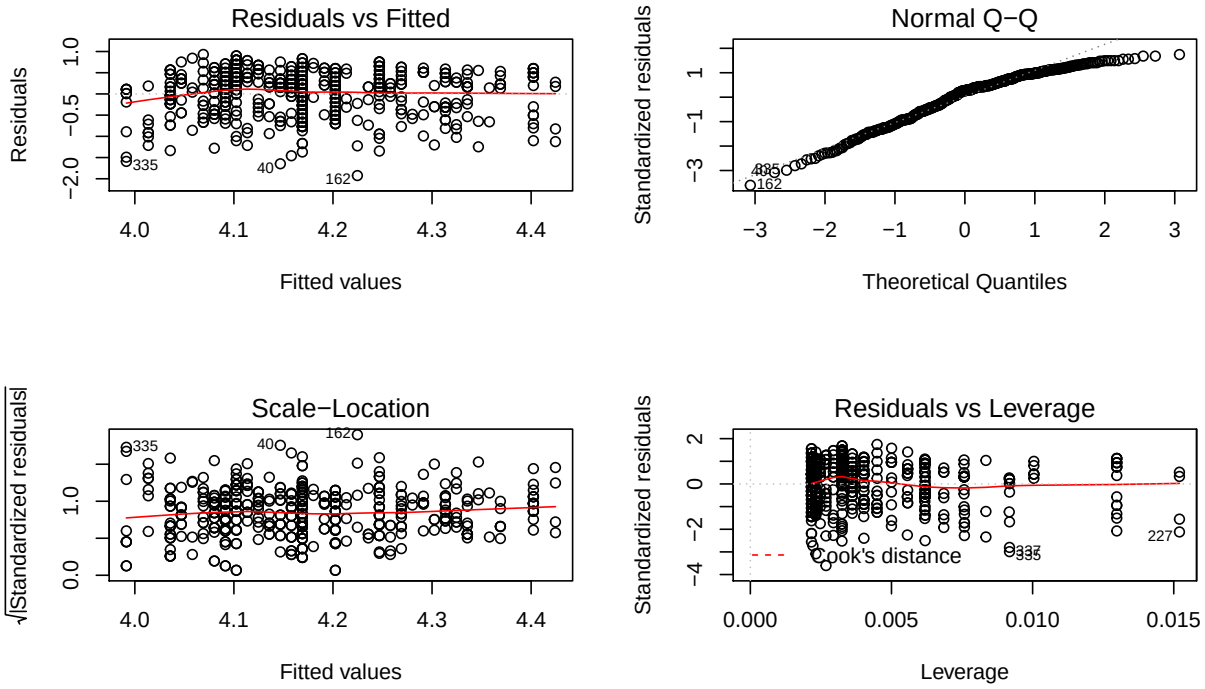


4. **Write out the *estimated* equation for the linear model *and* interpret the slope.**

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)**

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

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

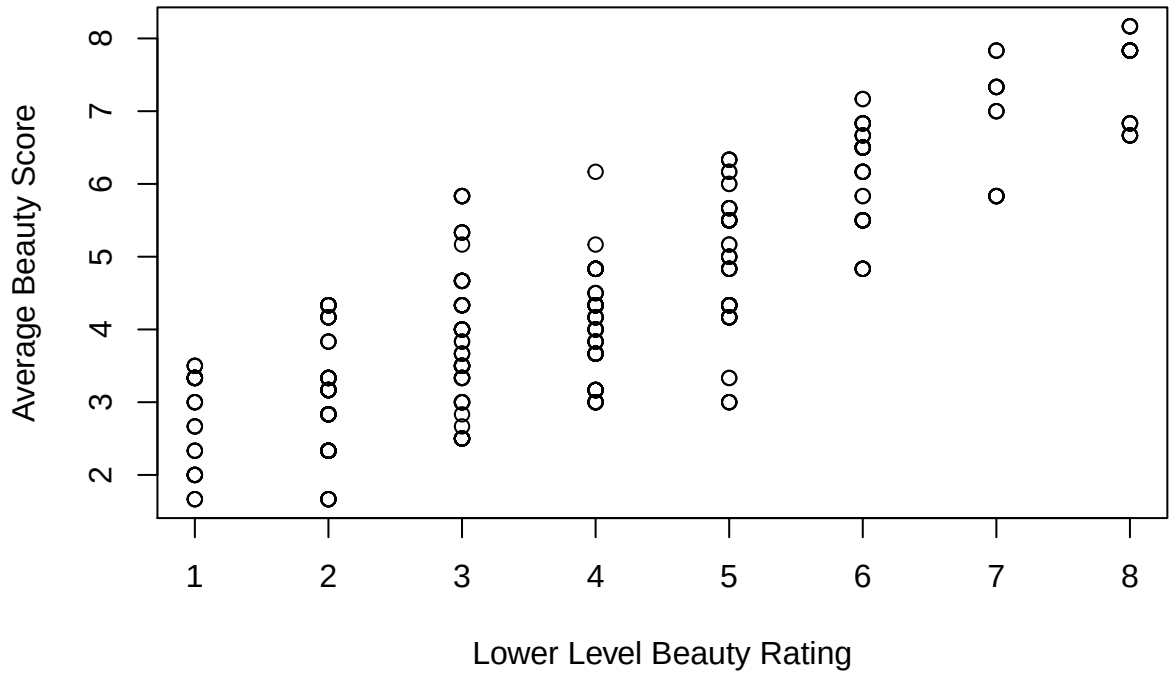


- **Independence:**
- **Normality of Residuals:**
- **Constant Variance:**
- **Linear Relationship:**
- **No Influential Observations:**
- **No Multicollinearity:**

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, xlab = "Lower Level Beauty Rating", ylab = "Average Beauty Score
```

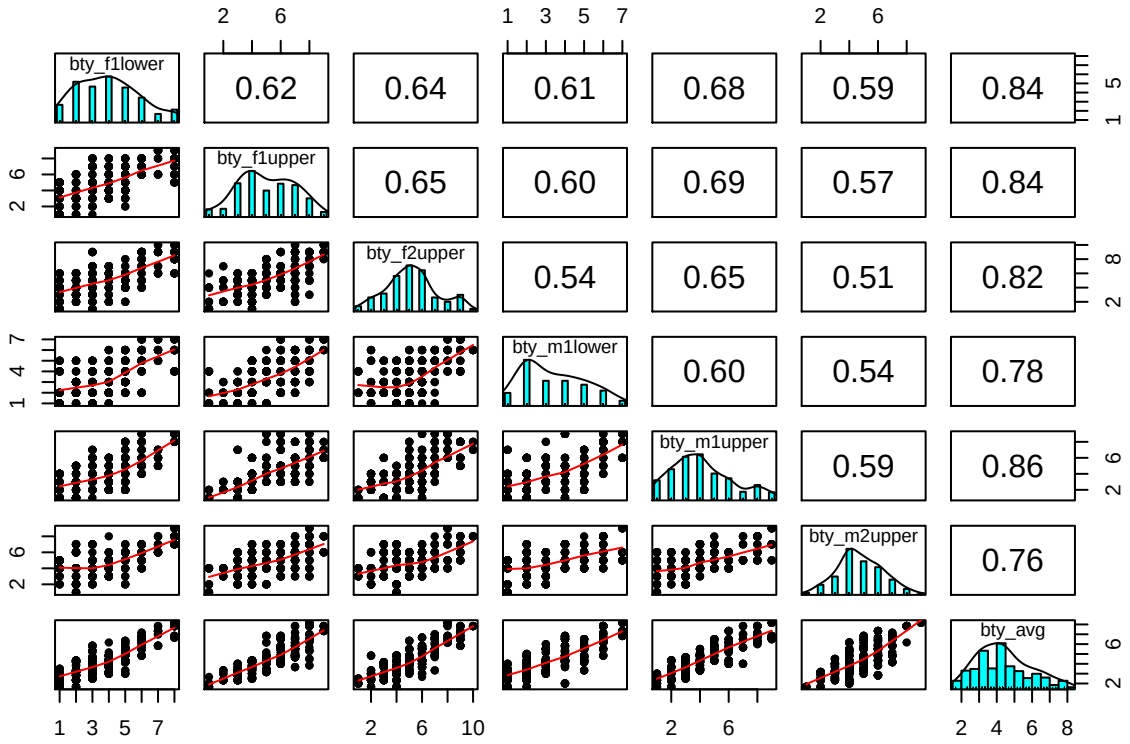



```
cor(evals$bty_avg, evals$bty_f1lower)
```

```
## [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?**

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?**

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.

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?**

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
```

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

12. **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?**

```
# run a model with the variables you left in
```

```
# run a summary of your new 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.**

```
## code for model fitting here!

## include your code for the FINAL model you choose!
```

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.**

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?**