

# Markdown example: multiple linear regression

*Open Intro (modified by Nick Horton and Caroline Kusiak)*

*June 20, 2016*

## 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.

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

## The 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. (This is a slightly modified version of the original data set that was released as part of the replication data for *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Gelman and Hill 2007).) The result is a data frame where each row contains a different course and columns represent variables about the courses and professors.

```
library(mosaic)
evals <- read.csv("http://www.amherst.edu/~nhorton/workshop/evals.csv",
  stringsAsFactors=FALSE)
```

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.
age	age of professor.
cls_students	total number of students in class.

variable	description
cls_level	class level: lower, upper.
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, can you rephrase the question?

SOLUTION: YOUR SOLUTION GOES HERE

2. 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?

```
favstats(~ score, data=evals)
```

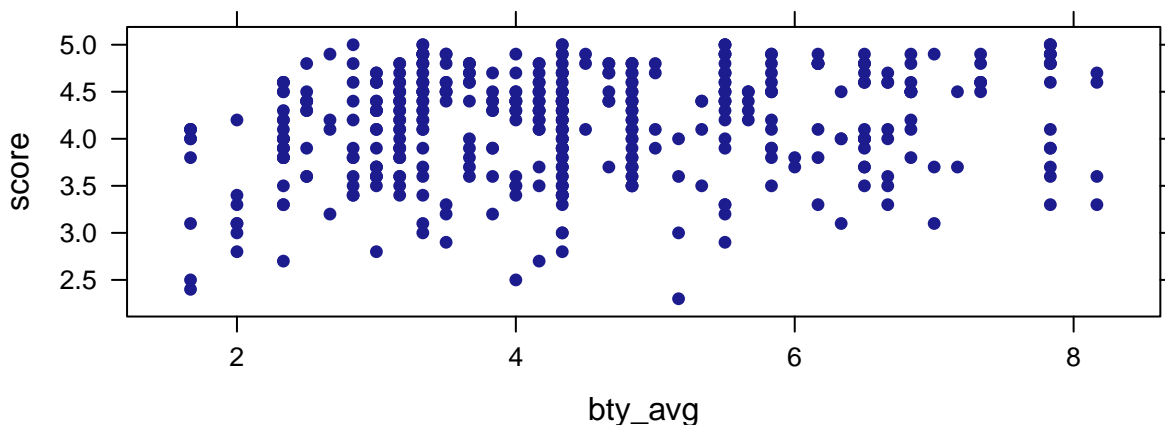
```
## min Q1 median Q3 max mean sd n missing
## 2.3 3.8 4.3 4.6 5 4.17473 0.5438645 463 0
```

SOLUTION: YOUR SOLUTION GOES HERE. Note that there are n=463 observations in the dataset.

## 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:

```
xyplot(score ~ bty_avg, data = evals)
```



Before we draw conclusions about the trend, compare the number of observations in the data frame with the approximate number of points on the scatterplot. Is anything awry?

3. Replot the scatterplot, but this time use the function `jitter()` on the  $y$  or the  $x$ -coordinate. (Use `?jitter` to learn more.) What was misleading about the initial scatterplot? (Note that this code block won't run until you change `eval=FALSE` to `eval=TRUE`.)

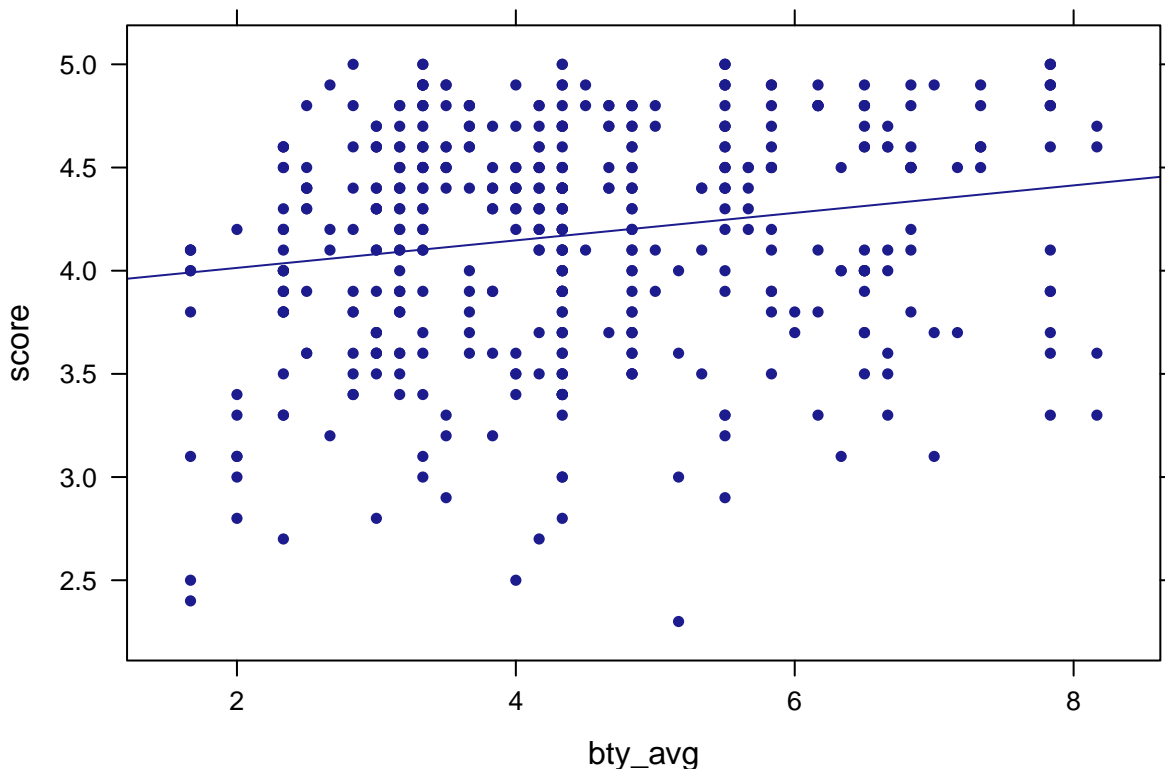
```
xyplot(jitter(score) ~ jitter(bty_avg), data=evals)
```

4. 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 and add the line to your plot using the `type=c("p", "r")` argument or the `plotModel()` function. 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?

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

```
##           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 ***
##
## 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
```

```
plotModel(mod = m_bty)
```



5. Use residual plots to evaluate whether the conditions of least squares regression are reasonable. Provide plots and comments for each one.

```
# mplot(m_bty, which=1) # remove first comment character to run command
```

## Multiple linear regression

The data set contains several variables pertaining to a professor's beauty score: This is based on the average of six students' ratings.

In order to see if beauty is still a significant predictor of professor score after we've accounted for the gender of the professor, we can add the gender term into the model.

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

```
##           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 ***
##
## 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
```

Let's also display this in a better way using the `xtable` package.

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	3.7473	0.0847	44.27	0.0000
bty_avg	0.0742	0.0163	4.56	0.0000
gendermale	0.1724	0.0502	3.43	0.0007

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

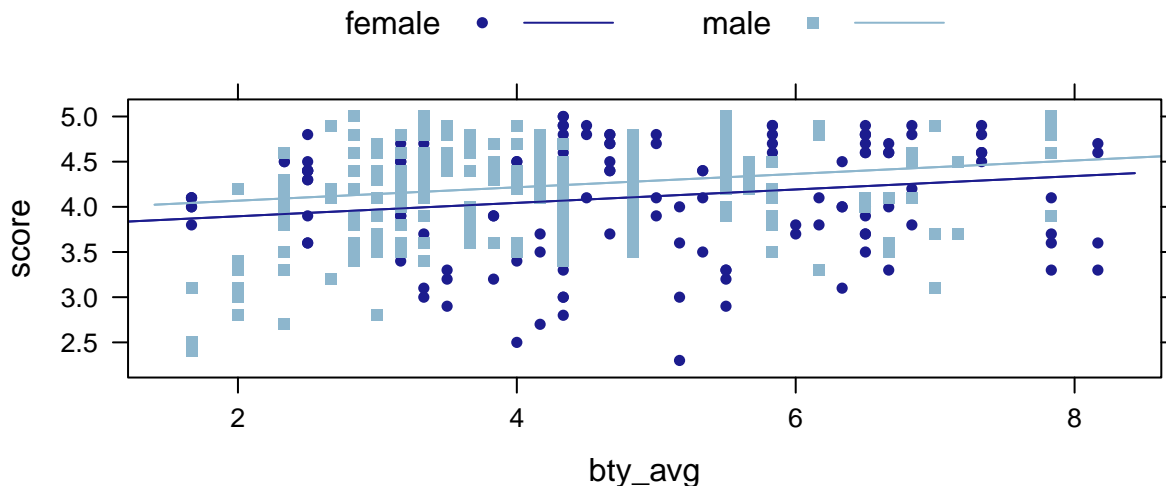
Note that the estimate for `gender` is now called `gendermale`. You'll see this name change whenever you introduce a categorical variable. The reason is that R recodes `gender` from having the values of `female` and `male` to being an indicator variable called `gendermale` that takes a value of 0 for females and a value of 1 for males. (Such variables are often referred to as "dummy" variables.)

As a result, for females, the parameter estimate is multiplied by zero, leaving the intercept and slope form familiar from simple regression.

$$\begin{aligned}\widehat{score} &= \hat{\beta}_0 + \hat{\beta}_1 \times bty\_avg + \hat{\beta}_2 \times (0) \\ &= \hat{\beta}_0 + \hat{\beta}_1 \times bty\_avg\end{aligned}$$

We can plot this data and the line corresponding to males with the following code.

```
plotModel(m_bty_gen)
```



SOLUTION:

7. What is the equation of the line corresponding to males? (*Hint:* For males, the parameter estimate is multiplied by 1.) For two professors who received the same beauty rating, which gender tends to have the higher course evaluation score?

SOLUTION:

The decision to call the indicator variable `gendermale` instead of `genderfemale` has no deeper meaning. R simply codes the category that comes first alphabetically as a 0. (You can change the reference level of a categorical variable, which is the level that is coded as a 0, using `relevel` function. Use `?relevel` to learn more.)

8. 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`.

SOLUTION:

The interpretation of the coefficients in multiple regression is slightly different from that of simple regression. The estimate for `bty_avg` reflects how much higher a group of professors is expected to score if they have a beauty rating that is one point higher *while holding all other variables constant*. In this case, that translates into considering only professors of the same rank with `bty_avg` scores that are one point apart.

## The search for the best model

We will start with a full model that predicts professor score based on rank, ethnicity, gender, language of the university where they got their degree, 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. Let's find the coefficients for the following model...

```
m_full <- lm(score ~ rank + ethnicity + gender + age
             + cls_students + cls_level + bty_avg
             + pic_outfit + pic_color, data = evals)
# msummary(m_full)
```

SOLUTION:

10. Interpret the coefficient associated with the ethnicity variable.

SOLUTION:

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

SOLUTION:

12. 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?

SOLUTION:

13. 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.

SOLUTION:

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

SOLUTION:

This is a product of OpenIntro that is released under a Creative Commons Attribution-ShareAlike 3.0 Unported. This lab was written by Mine Çetinkaya-Rundel and Andrew Bray (and updated by Caroline Kusiak and Nicholas Horton).

## References

- Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. 1st ed. Cambridge University Press.
- Hamermesh, Daniel S., and Amy Parker. 2005. "Beauty in the Classroom - Instructors' Pulchritude and Putative Pedagogical Productivity" 24 (4). *Economics of Education Review*: 369–76. doi:10.1016/j.econedurev.2004.07.013.