Math 58 / 58B - Introduction to (Bio)Statistics

your name here

due Friday, May 1, 2020

Homework 11

Assignment Summary (Goals)

- inference on linear model
- residual plots
- models with multiple variables

1. **Breaking Ice**[1]

R: `summary(lm(response ~ explanatory, data=dataset))`

Nenana is a small, interior Alaskan town that holds a famous competition to predict the exact moment that "spring arrives" every year. The arrival of spring is defined to be the moment when the ice on the Tanana River breaks, which is measured by a tripod erected on the ice with a trigger to an official clock. The minute at which the ice breaks has been recorded in every year since 1917. For example, the dates and times for the years 2000-2004 were:

2000	2001	2002	2003	2004
May 1, 10:47am	May 8, 1:00pm	May 7, 9:27pm	April 29, 6:22pm	April 24, 2:16pm

The data file NenanaIceBreak.txt contains all of the data since 1917. Scientists have examined these data for evidence of global warming, which would suggest that the ice break day should be tending to occur earlier as time goes on.

(a) Examine a scatterplot of the day in which the ice broke (coded in column 7 with April 1 = 1) vs. year. Does it reveal any association between the two variables? In other words, is there any indication that the day on which spring begins is changing over time? Explain.

```
ice <- read_delim("http://www.rossmanchance.com/iscam2/data/NenanaIceBreak.txt", "\t")
```

(b) Determine and report the regression line for predicting ice break day from year. Also calculate the correlation coefficient and the value of R^2. Comment on what these reveal, including an interpretation of the slope coefficient.

(c) Conduct a test for whether there is a linear association between ice break day and year. State the hypotheses, and report the test statistic and p-value. Check the technical conditions, and summarize your conclusions.

(d) Would you say that the p-value reveals evidence of a **strong association** or **strong evidence** of an association? Explain.

(e) Do the data suggest that one can make better predictions by taking year into account, rather than simply using the average of the ice break days? Explain.

(f) What date would the regression model predict for the ice break-up in the year 2005? What about 2020? Explain why you should regard these predictions cautiously.

2. **Grading the professor**[2]

[1] From ISCAM, HW 5.39

[2] From OpenIntro Labs, MLR; see lab 12 from class, solutions on Sakai

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.)

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 aslightly 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.

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.
`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.

```
evals <- read_csv("https://www.openintro.org/data/csv/evals.csv")
```

(a) Excluding `score`, select two other variables and describe their relationship using an appropriate visualization (scatterplot, side-by-side boxplots, barplot, or histogram). Try using the function `ggpairs` in the `GGally` R package.

(b) Searching for a good model...

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.

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.

```
m_full <- lm(score ~ rank + ethnicity + gender + language + age + cls_perc_eval
             + cls_students + cls_level + cls_profs + cls_credits + bty_avg
             + pic_outfit + pic_color, data = evals)
```

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

(d) 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?

(e) Drop a few more variables, see if you can find a model with only significant coefficients.

(f) Choose a set of variables and create a confidence interval at those values. Interpret the CI (what is the value inside the CI???).

The code looks like this except you have to use your own model from (e) and then set **all** the variables that are in (e). Make different choices than what I did!

```
predict(your.model.from.e.,
        newdata=data.frame(ethnicity="minority", gender="female", age=47,
        cls_perc_eval = 85, bty_avg = 4),
        interval="confidence", level=.95)
```

(g) Using the same set of variables in (f), create a prediction interval. Interpret the PI (what is the value inside the PI???). [Also, is a prediction interval reasonable to create for these data? Why or why not?]

(h) Based on your final model (the coefficients), describe the characteristics of a professor and course at University of Texas at Austin that would be associated with a high evaluation score.