Psychology 202a

Advanced Psychological Statistics

Fifth homework assignment, 10/29/2019 (due 11/5/2020).

In keeping with the spirit of the holiday of Halloween, let's work with some data involving poison this week. Here is a data set adapted from Karagas, M.R., Morris, J.S., Weiss, J.E., Spate, V., Baskett, C., & Greenberg, E.R., "Toenail samples as an indicator of drinking water arsenic exposure," *Cancer Epidemiology, Biomarkers and Prevention, 5*, 849-852. The first column represents amount of arsenic in private wells in New Hampshire, measured in parts per million. The second column is amount of arsenic found in the toenails of persons drinking from those wells, also measured in parts per million. Cases in which the well contained no measurable arsenic have been deleted. Our interest is in whether we can predict the amount of arsenic present in the body from the amount of arsenic in the drinking water. (More precisely, we want to develop a model for the conditional mean of the distribution of arsenic in toenails, given the amount of arsenic in drinking water.)

Part One

Your first task is to investigate the relationship graphically. Read the data into *R* and use everything you have learned about graphics to produce a good scatterplot of arsenic in toenails versus arsenic in water. Be careful of the following points:

- use an appropriate aspect ratio (no square plots)
- plot each variable on the appropriate axis
- add appropriate white space to the plot

Once you are satisfied with your plot, comment on the relationship. Does it appear that linear regression will be appropriate to describe the relationship? Why, or why not? (What must be true for linear regression to be appropriate?)

Regardless of whether you think it is appropriate, use R to estimate the regression of arsenic in toenails on arsenic in drinking water. Answer the following questions based on your output:

- What is the correlation between the two variables?
- Identify and interpret the estimated slope and intercept. ("Interpret" means that you should state what each number says about the conditional distribution of toenail arsenic given drinking water arsenic.)
- Test the null hypothesis that the slope is equal to zero.

Next, use *R* to produce a plot of residuals against fitted values. As before, take care with this plot, so that the result is publication quality, following the guidelines we have developed in class. Interpret the residuals plot. Do you see any problems with assumptions needed for the inference about the slope? Be as specific as you can. (You may find additional graphics useful in your assumption checks.)

Part Two

Sometimes variables that are problematic with respect to assumptions become better behaved when they are subjected to a nonlinear transformation. For some data, assumptions may be better satisfied when we take the natural logarithm. Note that such a transformation will also impact the linearity of the relationship; but often, there will be sufficient noise in the relationship that both the transformed and untransformed plots appear linear. Our focus here will be on other assumptions.

Your task in Part Two is to repeat all of the steps you did in Part One, but working with transformed data: You will use the natural logarithm of each arsenic variable. Produce all the same analyses, plots, and interpretations, but working with the transformed data.

The function to calculate a natural logarithm in *R* is log(), as in

```
lwater <- log(water)
ltoe <- log(toe)
```

(assuming that "water" and "toe" contain the untransformed data).

Part Three

If you had been charged with investigating the relation between arsenic in drinking water and arsenic in the body, which analysis would you prefer? Be as specific as you can about why you have that preference.