Biostatistics 515/518, Winter 2020
HW4 (9 problems)

Researchers at General Motors collected data on 60 U.S. Standard Metropolitan Statistical Areas (SMSA's) in a study of whether air pollution contributes to mortality. This is the 'SMSA' dataset on the course website. The response variable for analysis is age-adjusted-mortality (called "Mortality"). The data include variables measuring demographic characteristics of the cities, variables measuring climate characteristics, and variables recording the pollution potential of three different air pollutants. (There is not a codebook for this dataset, but fortunately most of the variable names are self-explanatory. Assume 'Rain' is inches per year.)

The pollution variables are highly skewed. A logarithm transformation makes them much more nearly symmetric. Use the natural log transform ('Log base e') to work with these variables.

First, consider log(NOx) as a predictor of mortality.

1. Show a scatterplot of mortality against log(NOx) and comment on the trends that you see.

2. Fit a simple linear regression model to use log(NOx) to predict mortality and interpret the results, including your inference on the regression parameter(s) of interest.

3. For your regression in (2), state your reasoning for your choice of model-fitting method (e.g., robust or non-robust). State the assumptions that are necessary for the inferences you made in (2).

Next, consider adjusting the above model for rainfall using multiple linear regression. Your scientific collaborator informs you whether or not a city is "low rainfall" may be an important variable to consider. "Low rainfall" is defined to be under 20 inches of rain per year. You need to create a variable that indicates, for each city, whether it gets under 20 inches of rain per year. The variable should equal 1 if 'low rainfall,' 0 otherwise.

4. Produce the scatterplot in (1) again, but use different symbols or some other way to distinguish cities with low rainfall from those that do not have low rainfall using the definitions of 'low' and 'high' above. Do you think rainfall should be a precision variable, confounder, or effect modifier? (Do not worry about significance here, just describe what you see.)

5. Fit a multiple regression model with log(NOx) as your predictor of interest and include rainfall (as a <u>continuous</u> variable) as a covariate. If you find statistical evidence for an interaction between Rainfall and log(NOx), include this in the model as well. Write out the fitted model (as always, do NOT give raw software output).

6. Interpret the model in (5) in terms of the scientific question about the relationship between log(NOx) and mortality. (Caution: with interactions in a model, this is trickier than with simple linear regression.)

7. Suppose you are a consultant for two communities who are considering their Mortality risks as related to NOx. One community is Aridia, which gets about 15 inches of rainfall per year. The other community is Seattle, which gets about 36 inches of rainfall per year. You don't know the NOx levels of these communities – each community plans to measure NOx in the near future and wants to be prepared for the implications of the results.

For each community, use your analysis in (5) to advise them on the potential for improved mortality if the community introduces regulation to lower the NOx in their environment. Write a short paragraph for each community, explaining the implications of your statistical analysis in (5).

8. For this problem, use the 'MRI2' dataset on the class website. This dataset is slightly different from the 'MRI' dataset we used before.

a) Make a scatterplot for predicting weight from height for this population. Add three lines to your plot: the simple regression line only for females, the simple regression line only for males, and the simple regression line for males and females combined.

b) Scientifically, do you think that 'sex' is an important variable to consider for predicting weight from height? Why or why not?
c) Are the slopes from the first two simple regressions in part (a) very similar or very different? Do you think there is an interaction between 'height' and 'sex'?
d) Altogether, what role does 'sex' play: precision variable, confounding variable, or effect modifier?
e) How does the slope from the 3rd regression compare to the first two? Why does this make sense?
f) Compare the regression coefficient for 'height' from the simple regressions above to the regression coefficient for 'height' adjusted for 'sex' from the multiple regression model $E[\text{weight} \mid \text{height}] = \beta_0 + \beta_1 \text{ height} + \beta_2 \text{ male}$, where male=1 for males and male=0 for females. How does the estimate of β_1 from the multiple regression compare to the estimated slopes from the simple regressions?

9. Problem 4.8 (page 137). The problem says "using (4.4)", which refers to a mathematical expression in the book. You can alternatively refer to the expression at the bottom of Lecture 8, slide 34.