Statistics 520, Fall 2017

Assignment 4

Private and self-supplied drinking water systems are not federally regulated, and owner's are responsible for monitoring their wells for contaminants and pollutants. The US Environmental Protection Agency (EPA) estimates that 18% of the population of Iowa gets drinking water from groundwater sources.

While nitrates occur naturally in groundwater, concentrations greater than 3 mg/L are usually considered to indicate contamination by non-natural sources. The EPA has set a maximum contaminant level for nitrate to protect against blue baby syndrome (which is about a bad as it sounds) of 10 mg/L, and estimates areas of various states with concentrations of greater than 5 mg/L as an indication of risk (https://www.epa.gov/nutrient-policy-data/estimated-nitrate-concentrations-groundwater-used-drinking).

Suppose that a study was conducted in which drinking water was sampled from individual wells in three counties. In each county 25 wells were sampled. The nitrate concentration was measured in each sample, and the objective of an analysis is to determine whether there are differences in the distributions (across wells) of nitrate concentrations in the three counties.

On the course web page in the Data folder is a file called `Nitrates in Wells` that contains hypothetical (i.e., simulated) data from this study. The file contains two columns titled "County" (1, 2 or 3) and "y" (nitrate concentration in mg/L).

We might formulate a model for these data based on random variables that have gamma distributions (partly because of the problem, partly from examination of the data which is not shown). The basic structure of the problem is that of a three group comparison.

1. (20 pt) Consider this problem within the context of generalized linear models.

You know how to formulate a design matrix for the comparison of groups in linear models. You now know how to conduct an analysis with a generalized linear model that has a gamma random component. So, conduct an analysis to answer the question of interest using a basic generalized linear model with a gamma random component. In your answer include:

- Definition (formal definition) of random variables and covariates.
- Present the model used.
- Present parameter estimates and standard errors.
- Present intervals or other quantities used for inference. For the purpose of deciding whether the three counties are different, make use of simultaneous confidence intervals produced using the Bonferroni method.
- Conclusion regarding the question of interest, including an estimate of the probability that a randomly chosen well from each county has nitrate concentration less than 3 mg/L and the probability such a sell has nitrate concentration greater than 10 mg/L. You might want to include a graph that contains the estimated distribution (if common) or distributions (if different) for the three counties.

2. (20 pt) Now consider this problem within the context of a likelihood comparison of three groups of data presumed to be from gamma distributions. Conduct a straightforward likelihood analysis (just extend what you did for Assignment 2). In your answer include:

   - Definition(formal definition) of random variables and covariates.
   - Present the model used.
   - Present parameter estimates and standard errors.

- Present intervals or other quantities used for inference. While your basic method of inference may (should) differ from the previous glm analysis, produce simultaneous intervals analogous to those from the glm analysis for the purpose of comparison.
- Conclusion regarding the question of interest, including an estimate of the probability that a randomly chosen well from each county has nitrate concentration less than 3 mg/L and greater than 10 mg/L. You might want to include a graph that contains the estimated distribution (if common) or distributions (if different) for the three counties.

3. (10 pt) Briefly contrast and compare your two analyses.