BST 760: Advanced Regression
Breheny

Assignment 8
Due: Tuesday, April 12

1. Write an `R` function (or a SAS macro, if you know how to do so) called `glm.poisson` that can fit GLMs based on the Poisson distribution with canonical link function. To clarify and provide some hints:

   - The function you're writing should have accept arguments (inputs) `X` and `y`, and return the standard estimate/SE/test statistic/p-value output. Specifically, it should look like:

     ```
     glm.poisson <- function(X,y)
       {
         ...
         return(data.frame(Estimate=b,SE=SE,z=z,p=p))
       }
     ```

     where of course you calculate `b`, `SE`, `z`, and `p` in place of the dots.
   - Please turn in the function electronically as well as print out the code. As part of the grading of the problem, I will check to see that the function actually works correctly when applied to a real data set.
   - It would be a good idea to check that your function works by generating an `X` and `y` and running your function. You can then check your answers against those given by `glm(...,family=poisson)`.

2. The course website contains a data set (`challenger.txt`) which contains information on the first 24 space shuttle launches of the National Aeronautics and Space Administration's Space Shuttle program. Information is recorded on two variables: `Temp`, the outside temperature (in degrees Fahrenheit) at the time of the launch, and `BadRings`, the number of O-rings that showed signs of thermal distress following the launch. An O-ring is a seal that separates the fuel supply from the combustible gases in the rocket's exhaust; if it fails to do so perfectly, it will show signs of thermal distress after the launch. In cold weather, O-rings are less resilient and may be more likely to fail.

   (a) Using logistic regression, model the way in which the probability of an O-ring failure depends on temperature. What is the coefficient for `Temp`?
   (b) Test the null hypothesis that O-ring failures are independent of temperature. Does this seem plausible given the data?
   (c) Estimate the odds ratio for O-ring failure with a 10 degree decrease in the launch temperature.
   (d) The 25th space shuttle launch, involving the space shuttle *Challenger*, took place on January 27, 1986. Seventy-three seconds into the flight, the fuel mixed with the rocket exhaust, resulting in an explosion which destroyed the shuttle and killed all seven astronauts on board. The launch temperature that day was 31 degrees. Based on data from the first 24 launches, estimate the probability of a O-ring failure on the *Challenger* flight.

(e) Provide a confidence interval for the probability in part (d).

3. Duchenne Muscular Dystrophy (DMD) is a sex-linked genetic disease. Boys with the disease usually die at a young age, while affected girls usually do not suffer symptoms and may unknowingly carry the disease and pass it to their offspring. It is desirable to have some kind of test to detect whether or not a woman is a carrier of the disease. The dataset `dystrophy.txt` contains information from a 1981 study attempting to develop such a test based on two serum enzymes, creatine kinase (`CK`) and hemopexin (`H`) for 38 known DMD carriers (`Case`) and 82 women who are not carriers (`Control`). (Note: In the last 30 years, it has become possible to perform genetic testing to provide a definitive answer; however, tests based on the above proteins are still used as rapid and inexpensive alternatives).

   (a) Use logistic regression to model the way in which case/control status depends on creatine kinase and hemopexin. Construct (using the Wald approach) a table containing the estimated odds ratios and p-values for the two enzymes. Provide confidence intervals for the odds ratios, and give some thought as to what would constitute a meaningful difference (δ_j) for the two enzymes when calculating the odds ratios.

   (b) Can you calculate confidence intervals for the odds ratios in part (a) using the likelihood ratio approach? If so, calculate them. If not, explain why you can't do so.

   (c) Can you carry out the hypothesis testing in part (a) using the likelihood ratio approach? If so, perform the tests. If not, explain why you can't do so.

   (d) Describe (quantitatively) the relationship between creatine kinase levels and the likelihood that a woman is a carrier without using the phrase "odds ratio" (you can use "odds", just not "odds ratio").

   (e) Suppose a woman has a hemopexin level of 100 and a creatine kinase level of 150. Can you estimate the probability that she is a carrier? If so, estimate it. If not, explain why you can't do so.

   (f) It is estimated that 1 in 3,300 women are carriers. Treating this as a known constant, calculate the sampling ratio τ_1/τ_0.

   (g) Based on your answer to (f), can you calculate the probability from part (e)? If so, estimate it. If not, explain why you can't do so.