STA 2101/442f12 Assignment Eight[1]

Please bring your R printout for Question 4 to the quiz, and also your SAS log file and list file from the Question 5. I really mean the log file, not just a listing of the SAS program. The log file and the list file *must be from the same run of SAS.* It is very common to print the good list file, but bring an older log file that still has errors. This could lose you substantial marks. The other questions are just practice for the quiz. Any necessary formulas will be provided.

1. Show that for Poisson regression, the natural link is the log link. What is the dispersion parameter? What is the variance function?

2. Consider a generalized linear model in which the response variable has a gamma distribution with unknown parameter α, and known $\beta = 1$. What is the natural link function? What is the dispersion parameter? What is the variance function?

3. In the Logistic regression with R slide show, I reproduced the standard error for an estimated probability of passing the course. How did I do it? Show your work.

4. People who raise large numbers of birds inhale potentially dangerous material, especially tiny fragments of feathers. Can this be a risk factor for lung cancer, controlling for other possible risk factors? From the Data Sets link on the course home page, you can find the Bird Lung Cancer data. For a sample of birdkeepers and non-birdkeepers, it has Whether they got lung cancer (1=Yes, 0=No), Gender (0=M, 1=F), Socioeconomic Status (0=Low, 1=High), Whether they are birdkeepers (1=Yes, 0=No) Age, How many years they have been smoking (including zero), and Cigarettes per day. `help(colnames)` may be useful.

   First, make tables of the binary variables using `table`, Use `prop.table` to find out the percentages. What proportion of the sample had cancer. Any comments?

   There is one primary issue in this study: Controlling for all other variables, is birdkeeping significantly related to the chance of getting lung cancer? Perform a likelihood ratio test to answer the question.

   (a) In symbols, what is the null hypothesis?

   (b) What is the value of the likelihood ratio test statistic G^2? The answer is a number.

   (c) What are the degrees of freedom for the test? The answer is a number.

   (d) What is the p-value? The answer is a number.

   (e) What do you conclude? Presence of a relationship is not enough. Say what happened.

[1] Copyright information is at the end of the last page.

(f) For a non-smoking, bird-keeping woman of average age and low socioeconomic status, what is the estimated probability of lung cancer? The answer (a single number) should be based on the full model.

(g) For a non-smoking, non-bird-keeping woman of average age and low socioeconomic status, what is the estimated probability of lung cancer? The answer (a single number) should be based on the full model.

(h) Obtain a 95% confidence interval for that last probability. Please don't use the delta method, or anyway don't calculate g-dot yourself; it's too much work! Your answer is a pair of numbers.

(i) Naturally, you should be able to interpret all the Z-tests too. Which one is comparable to the main likelihood ratio test you have just done?

(j) Also, are *any* of the explanatory variables related to getting lung cancer? Carry out a single likelihood ratio test. You could do it from the default outut with a calculator, but use R. Get the p-value, too.

(k) Now please do the same as the last item, but with a Wald test.

5. This question uses the `sat.data` file you first saw in Assignment 5. There is a link on the course web page in case the one in this document does not work. This is just to get your feet wet with SAS and unix, in case you have not used these tools before. Using SAS, get ns, means and standard deviations for all three variables. That's it. Bring your log file and your list file to the quiz.