STAT 538 - Generalized Linear Models (Term 1, 2010-11)

ASSIGNMENT

NOTE: More questions will be added as we cover material.

NOTE: Some of the problems are deliberately open-ended, and give you the opportunity to investigate as you best see fit. One tradeoff is as follows. I will not assign a large number of problems, but in return I will be looking for well-documented solutions. At a minimum, you should be reporting your findings in complete sentences/paragraphs (not computer-code fragments!), and using tables and/or figures to summarize empirical work as appropriate. One guideline that cuts across all scientific work is that you should provide enough detail so that an interested reader could replicate what you have done. In terms of using mathematical notation versus words, I'm not fussed. That is, some of you will be more comfortable using less/more mathematical notation than others, and that's fine. Clarity can be achieved either way!

1.[posted Oct. 27] From text, Exercise 3 in Ch. 2.

2.[posted Oct. 27] *Is the deviance-based goodness-of-fit test better than your eyes?*

Simulate some datasets with a single continuous explanatory variable and a binomial response, under conditions where this test can reliably be applied. In some instances, make sure the model you are fitting is indeed incorrect. One recipe for this would be fitting the logistic regression model having only a main effect of the predictor, when in fact the true relationship involves both linear and quadratic terms.

One 'by eye' way to screen for a problem with the fitted model would be to look at a scatterplot of the data (e.g. Y_i/n_i versus X_i), with the fitted relationship superimposed. Another way would be to rely on the deviance-based goodness-of-fit test.

Presumably when the model is only wrong 'by a little bit,' neither method will ring the model-wrong alarm bell. What happens though, as the model becomes increasingly wrong? Does one method ring the alarm before the other? Or are the test and your eyes about equally powerful?

3.[posted Nov. 5] From text, Exercise 3 in Ch. 3. To make the data available in R and look at a description, use:

```
> library(MASS)
> help(ships)
```

4.[posted Nov. 5] In lecture 5 we will do some simulation work to assess when the chi-squared approximation to the sampling distribution of the deviance is reasonable, in the case of logistic regression. Try investigating this issue in the case of Poisson regression. Does the approximation tend to improve by adding more datapoints? What about a fixed number of datapoints but increasing means of the responses?

5.[posted Nov. 9] Carry out some simulation work to assess whether standard errors obtained by quasi-likelihood seem to work well, i.e., if we repeatedly simulate data and use the SEs obtained to form confidence intervals, what proportion of the intervals actually contain the true

parameter values. Try both using quasi-Poisson when the data are generated from a negative binomial distribution, and using quasi-Binomial when the data are generated from a beta-binomial distribution (the distribution arising by first generating a success probability ffrom a Beta distribution and then generating a Binomial observation given this success probability). In the latter case, how well does the quasi idea of introducing a multiplicative constant to the variance function match with the actual mean-variance relationship?

6.[posted Nov. 12] Think about the material we have covered in the course so far. Pick one thing (idea/fact/technique) that made an impression on you (i.e., maybe you found it to be particularly interesting or innovative or helpful or surprising). *In half a page or less*, describe what this thing is and why it particularly caught your attention. This is not meant to be a time-consuming question, hence the modest space limit. I am curious, though, to know what caught your interest.

This is the cutoff for Assignment 1. Problems 1 through 6 are due Monday November 22. Subsequent problems are part of Assignment 2.

7.[posted Nov. 25] From text, Exercise 2 in Ch. 5.

8.[posted Nov. 25] Carry out an AIC versus BIC stepwise regression competition. That is, repeatedly simulate datasets (it doesn't have to be a huge number of times) from some true model, and apply stepAIC and stepBIC each time. Which method leads you closer to the true model, on average? Try to make the simulation scenario challenging for any method (i.e., perhaps a reasonably large number of predictors, perhaps some high correlatons between predictors, perhaps some small effects in the true model, etc.). Try several different true models - does the relative performance of the two methods seem to depend on whether the true model is simple or complex?

9.[posted Nov. 30] From text, Exercise 3a in Ch. 4. You can just focus on the relevant concepts that we discussed in class, i.e., independence, symmetry, quasi-symmetry. But do be clear about how the models you are fitting are interpreted in the subject-area problem.

10.[posted Nov. 30] From text, Exercise 4 in Ch. 4.

11.[posted Nov. 30] Repeat Question 6, but this time with reference to material covered in the latter part of the course. The half-page limit applies again.

THE END. Questions 7 through 11 are due Wednesday December 15.