Exercise #1 Due: January 13, 2010	Biostat/Stat 571 Winter 2010 P. Heagerty

Reading:
- Lindsey and Jones (1998) *Statistics in Medicine*
- McCullagh & Nelder, Chapter 5, "Polytomous Data"
- ○ Akaike (1973) *Second Int. Symp. on Inference Theory*

Generalized linear models

1. The use of generalized linear models in specific applications requires identification of an appropriate statistical model that can be used to answer focused scientific questions. The process of choosing a regression model, and/or assessing the adequacy of the regression model is often not discussed in the presentation of results. For example, Bagley, White & Golomb (2001) *J. Clin. Epi.* "Logistic regression in the medical literature: standards for use and reporting" discuss the selection of predictor variables: "Does the article explain how variables were selected for inclusion into the model? Usually the variables are chosen based on earlier research; sometimes they are selected by virtue of significant association in bivariate analysis with the outcome variable [4]. (Problems associated with use of bivariate analysis for variable selection are discussed in [15].)" These authors also discuss methods for validation of the model including cross-validation techniques. The paper by Lindsey and Jones (1998) discusses the use of one criterion that can be used to compare different models. In addition, these authors recommend that for a "phase III" study "... an appropriate, completely specified, model (is) available, both distribution and covariates; only the final, crucial, confirmatory test of a treatment difference should be left." (page 67).

The goal of this exercise is to survey and then synthesize recommendations regarding regression model specification (selection). On the course web page are a number of recently published papers in applied statistics and substantive fields that discuss model choice. Please choose (1) article and then write a brief summary of the main points contained in the paper (1 page). If you wish to search `PubMed` and/or `Current Index to Statistics` and select a different article then please let me know what you have chosen. Be sure to identify what the author(s) state as their goal for use of regression models. Summarize the main conclusions and critique the paper by describing its strengths and weaknesses. Also list any questions that you have after reading the paper.

Please send me your comments electronically as a text file and I will post them on the course web page after some editing (to combine similar comments and to make anonymous).

2. The data `KingCounty2001.data` on the class web page contain a sample of data taken from birth certificates for children born in King County WA in 2001. The data are restricted to singleton births (ie. not twins or triplets). A number of additional covariate measurements are also available.

(a) The public program *First Steps* was implemented in the early 1990's to try and reduce the number of low birth weight infants. Babies that are born small are known to have a number of additional medical and developmental complications. Specifically, the goal of the *First Steps* program, authorized by the Maternity Care Access Act of 1989, was to provide "maternity care necessary to ensure healthy birth outcomes for low-income families." The legislation called for removal of unnecessary barriers to receiving prenatal care. Additional information about the program can be found in the report from DSHS that is linked on the course web page.

Use the King County data to summarize the evidence for the effectiveness of the program. Justify the methods that you use, summarize your conclusions, and state any limitations of your analysis.

(b) The covariates are all easily obtained, and therefore might be used for a clinical prediction rule that identifies women who are at high risk of delivering a low birth weight child. The standard criterion that is used to indicate a poor birth weight outcome is based on a birth weight of less than 2500 grams, `lbw`=**1**(`bwt`<2500). Use these data to develop a clinical prediction criterion and then summarize the accuracy of the predictive model.

(c) How would your analysis differ for (a) and (b) if the data were collected by oversampling women who are enrolled in the First Steps program? For example, what if women with `firsteps`=0 were sampled with probability 1/10 while women with `firsteps`=1 were sampled with probability 1/3? Discuss how you may modify your analysis in (a) and (b), but you do not need to perform a reanalysis.