STA 2201S Homework 1. *due Wednesday February 4 2015 at 11.59pm on Blackboard*

When answering questions requiring numerical work, the results are to be reported in a narrative summary, in your own words. Tables and Figures may be included, but must be formatted along with the text. DO NOT include in this summary printouts of computer code with the relevant selections highlighted. All computer code used to obtain the results summarized in the response should be provided as an appendix. In this appendix you may highlight the relevant results. The LaTeX code used to produce this HW is available on the course web site. It is not necessary to submit your homework in LaTeX, but it makes the TA much happier.

1. ELM, Problem 2.2 The dataset `wbca` in the library `faraway` comes from a study of breast cancer in Wisconsin (see `?wbca`). There are 681 cases of potentially cancerous tumours, of which 238 are actually malignant. Malignancy of a tumor is traditionally determined by an invasive surgical procedure. The purpose of this study was to determine whether a new procedure called fine needle aspiration, which draws only a small sample of tissue, could be effective in determining tumor status.

   (a) Fit a binomial regression with `Class` as the response and the other nine variables as predictors. Report the residual deviance and associated degrees of freedom. Can this information be used to determine if this model fits the data? Explain.

   (b) Use the `step` function to choose final model.

   (c) For a patient with the following measured values of the predictors: `Adhes` = 1, `BNucl` = 1, `Chrom` = 3, `Epith` = 2, `Mitos` = 1, `NNucl = 1`, `Thick = 4`, `UShap`= 1, `USize` = 1, predict the `Class` of the tumor. Give a confidence interval for your prediction, and explain how you obtained it.

   (d) Suppose that a cancer is classified as benign (`Class` = 1) if $p(\hat{\beta}) > 0.5$, and otherwise is classified as malignant. Give a table showing the misclassification errors if this method is applied to the current data, using the final model from (b). Give a similar table but using the cutoff 0.9 instead.

   (e) The error estimates in (d) are overly optimistic, because they are computed on the same data that was used to fit the model. A better strategy is to fit the model on training data and assess the errors on test data. Choose a random 1/3 of the data to hold out as a test set, and find the best fitting model on the remainder. Construct the tables of misclassification errors for the cutoffs 0.5 and 0.9, and compare to the results in (d).
   Note: If you have issues with convergence on the randomly selected training set, you could try following Faraway's suggestion of choosing each 3rd case to construct the test set.

   (f) How is this data set relevant to determining the effectiveness of fine need aspiration, relative to surgery? (I think you will need to consult the source article Bennett, K.,P., and Mangasarian, O.L., Neural network training via linear programming. In P. M. Pardalos, editor, Advances in Optimization and Parallel Computing, pages 56-57. Elsevier Science, 1992. This is posted on the course web page. You may need to chase the references in that paper as well.)

2. SM, Exercise 10.4.1: Suppose $y_1, \ldots, y_n$ are independent Binomial(n_i, p_i) random variables, with $p_i = \exp(x_i^{\mathrm{T}}\beta)/\{1 + \exp(x_i^{\mathrm{T}}\beta)\}, i = 1, \ldots, n$.

   (a) Derive expressions for the log-likelihood function, the equations defining the maximum likelihood estimator of β, and the residual deviance.

   (b) Use the approximation $\log(1 + x) \simeq x - x^2/2$ to show that the residual deviance is approximately
   $$X^2 = \sum_{i=1}^{n} \frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1 - \hat{p}_i)}.$$

   (c) Show that if $n_i \equiv 1$, that the deviance is a function only of $\hat{\beta}$, and not otherwise a function of y.

   (d) Show that if $n_i \equiv 1$ and $p_i \equiv p$, that $X^2 = n$. If the p_i's are not all equal, X^2 is a function only of $\hat{\beta}$ and not otherwise a function of y.

   (e) Explain why these two last results mean that neither the residual deviance nor Pearson's X^2 can be used to assess goodness of fit with binary data.

3. Find an article about the results of a study, in a scientific journal on a topic of interest to you. The article should discuss a single study, and should provide information enough information on the study methods to answer the questions below.

   (a) Give the complete bibliographic reference, as well as a web link, to the published paper.

   (b) Was the study observational or a designed experiment?

   (c) What was the study population? What is the population of interest for the research?

   (d) If the study was observational, was it a prospective, or a retrospective study? If it was an experiment, was it randomized?

   (e) What were the units of analysis?

   (f) What was the primary endpoint and the main analysis of this endpoint?

   (g) What were the main conclusions of the study, in your own words?

4. *Data Set for Final Project*: You will be required to submit this in HW 2, but are welcome to submit it with HW 1. The data set should have a single response variable of interest, and a number of potential covariates related to the response variable. It should have more than thirty observations, and less than about 1000, although exceptions can be made. It will be helpful if there is not a lot of missing data, as we will not cover techniques for missing data in detail. We will also not cover specialized models for time series data nor for data with strong spatial dependence, but data sets with these features can sometimes be analyzed, at least in part, by other means. Ideally the dataset will not have been analyzed before, but this is not a requirement. Both Statistics Canada, and the City of Toronto, and many other government organizations, make data available online. The UC Irvine Machine Learning Repository collects data sets, but these are a bit tired. The homework questions for Shalizi's Data Analysis course at CMU also includes many data sources.