Bio 754: Homework 1

March 31, 2011

See the class site for due date, guidance on how to answer, and notice of any corrections or clarifications. For problems which require computation, please append neat and commented R code as an appendix to your homework.

1. [**Sandwich-based inference**] The table below gives a small dataset, from a lab-based experiment. Under controlled conditions, 15 rats were each given a contaminant, at different concentrations. The outcome of interest Y_i, is time until death, in days. x_i denotes the concentration level, in grams.

Rat ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x	6.1	4.2	0.5	8.8	1.5	9.2	8.5	8.7	6.7	6.5	6.3	6.7	0.2	8.7	7.5
y	0.8	3.5	12.4	1.1	8.9	2.4	0.1	0.4	3.5	8.3	2.6	1.5	16.6	0.1	1.3

   (a) Using the estimating equations $n^{-1}\sum_i\{1, x_i\}^T\left(1 - Y_i e^{\beta_0+\beta_1 x_i}\right) = 0$, calculate estimates of β_0, β_1, using your own Newton-Raphson code. Report the line you fit, to an appropriate number of decimal places.

   (b) Interpreting the solution to these estimating equations as a weighted fit to a line of the form $y = h(\mathbf{x}^T\beta)$, give the form of $h(\cdot)$ and the corresponding weights. Writing for a non-statistician, i.e. in simple but precise language, describe the parameter β_1 that you estimated. You may instead describe some 1-1 transformation of β_1, if you find this easier.

   (c) Calculate the form of the sandwich estimate of covariance for $\hat{\beta}$. By writing your own 'by hand' code, compute this estimated covariance matrix, and give standard sandwich-based 95% intervals for β_0 and β_1. Give brief descriptions of these intervals (or equivalent transformations) in language appropriate for a non-statistical co-author.

   (d) We will use simulation to assess the accuracy of both types of interval, in finite samples. Using `R`, write code to generate samples of independent data, where the density of Y_i is

$$\lambda_i e^{-\lambda_i y_i}, \;\; y_i > 0,$$

where $\lambda_i = e^{\beta_0+\beta_1 x_i}$, using the same values of $\mathbf{x}$ as above. Use $\beta_0 = -2.82, \beta_1 = 0.3$ in your simulations. Under these assumptions, what is the actual coverage of the approximate 95% intervals you used in c)?

(e) Under the same assumptions, investigate the coverage of β_1 for larger sample sizes. In all your simulations, use repeated copies of the original contaminant concentrations (i.e. make a large vector of covariates by writing out the short one several times, $(x, x, ..., x)$). Your report should describe the behavior of both $\hat{\beta}_1$ and its estimated standard error. Conclude your report with a brief explanation of your findings (and why they are relevant) for a non-statistician.

(f) What happens if you consider use of the same sandwich intervals on data where the x_i are in fact random draws from a distribution? Illustrate what happens using the 'empirical distribution', of the data, i.e. where the values $\{6.1, 4.2, 0.5, ...7.5\}$ all occur with probability 1/15. (For the few datasets where all X_i are identical, report the whole real line as your interval)

2. [**Speeding Up Computation**] In Lecture 2 it is noted that `crossprod()` can save computational time, compared to constructing very large diagonal matrices. Based on your work in Q1e, illustrate this for analysis of one very large dataset, and for repeated use on smaller datasets; you may find the `system.time()` command helpful.

3. [**Comparing super-populations**] Consider the setting of a simple case-control study, investigating a binary risk factor X (having completed a postgraduate degree) and binary disease outcome Y (being myopic, i.e. shortsighted). Without using mathematical formulae, describe the super-population in which you would:

   (a) Most plausibly *like* to assess association between Y and X.

   (b) *Actually* assess association, when using a standard case-control sample in which equal numbers of subjects with $Y = 0$ and $Y = 1$ are recruited into the study and recruitment is independent of X.

   (c) *Actually* assess association, if your recruitment for cases and controls in b) unwisely used an advert that was printed in a very small font in an alumni magazine, but a regular font elsewhere.

   In (c), what would be the consequences of naïve analysis, i.e. ignoring the imperfect recruitment method? (Hint: use a bit of simulation to help, if you find it difficult to intuit what will happen)

4. [**Quasi-likelihood**] A common set of parametric assumptions that justify linear regression is that

$$\mathbf{Y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_p)$$

   where $\mathbf{X}$ is an $n \times p$ matrix of covariates, $\boldsymbol{\beta}$ is a $p \times 1$ vector of coefficients, σ^2 is a positive scalar, and I_p is the $p \times p$ identity matrix.

(a) For $\boldsymbol{\beta}$, state the score equations, and describe how they give the $\boldsymbol{\beta}$ component of the MLE for this model.

(b) Describe the steps you would go through to implement a quasi-likelihood version of this model. Give mathematical formulae for any calculations this would involve, and state explicitly what assumptions this approach requires.

(c) Both `family=quasipoisson` and `family=quasibinomial` are available. Why is there no `family=quasinormal`?