Statistics 347, Homework 1, due January 17

Please turn in problems 1,2 in one file and problem 3 in another file, stapled separately.

Discussion of homework problems among students is encouraged. However, all material handed in for credit must be your own work. **Keep the R output brief. Always write a mathematical formula for the model you are estimating. The narrative part of the answers to each item should be no longer than a short paragraph.**

1. Neurons in the rat's hippocampus respond to the rat's location. Neurons tend to have preferred location in a given environment and fire at hire rates when the rat is in that location. The firing rate falls off as the rat moves away from that location. Experimenters recorded voltage from an electrode implanted in a rat's brain over a time period of about 42 seconds and isolated the activity of two neurons. For each millisecond (1/1000 of a second) a post-processing algorithm decides whether each neuron fired or not (1/0) resulting in two binary vectors of length 41328. These can be found in the data set HIPP.Rdata linked in your assignment. They are listed in HIPP$spikes, HIPP$spikes2. For the same time points HIPP$xN and HIPP$yN denote the x,y coordinates of the rat in a small cage.

   (a) Using all the time points fit a glm for the binary data HIPP$spikes with predictors xN and yN. Show a summary of the results and comment on the appropriateness of the model assumptions with respect to this data set.

   (b) Can you improve the fit with interactions or non-linear functions of xN and yN.

   (c) What happens to the fitted parameters if you sample the data every 20 points.

   (d) In the subsampled data, take all time points where the neuron fired (HIPP$spikes=1). There are 100 such points. Take a random sample of 100 points where the neuron did not fire (HIPP$spikes=0). Fit the model in 1b to this data set. Compare the results to those of 1b. Are the coefficients the same? Is anything the same? How does this relate to the different study types discussed in Faraway 4.3.

Faraway Chapter 4.3

   (e) In HIPP$spikes.hist you have the 20 ms history of spike activity of the neuron, in reverse, i.e. starting from the most recent time. Specifically: For each time $k > 20$, HIPP$spikes.hist[k,i]=HIPP$spikes[k-i]. Does this history improve the prediction of the spikes relative to the location covariates? What can you learn from the coefficients of the history?

21: previous 20~1

2. Estimating the LD_{50}. In toxicology, the LD_{50} is the dose that causes a 50% mortality rate (lethal dose 50%). Experiments are often carried out at a sequence of dose levels, $x_0, x_1, \ldots$ each dose being twice the preceding dose. The model most commonly used in toxicology is linear in log dose. Suppose that the following results have been obtained in an experiment at various multiples of the baseline dose.

Faraway Chapter 4.4

Dose $\log_2(x)$	0	1	2	3	4	5
Mortality y/m	0/7	2/9	3/8	5/7	7/9	10/11

Here y/m is the number of deaths occurring in a sample of m individuals.

   (a) Plot the data, i.e. the mortality fraction against log dose.
   Fit the linear logistic model in which the logit of the mortality rate is linear in log dose. Superimpose the fitted probabilities on the plot.

   (b) Obtain the estimate of the $\log_2 LD_{50}$, using the delta method proposed in Faraway 2.10. 4.4

   (c) An alternative confidence interval is obtained as follows: We are interested in a CI for the ratio $\rho = \beta_0/\beta_1$. Assume

$$C = \begin{pmatrix} C_{00} & C_{10} \\ C_{01} & C_{11} \end{pmatrix}$$

   is the covariance matrix of β_0, β_1, and define:

$$T_\rho = \frac{\beta_0 - \rho\beta_1}{\sqrt{C_{00} - 2\rho C_{01} + \rho^2 C_{11}}}.$$

Assuming β_0, β_1 are approximately jointly normal then $T \sim N(0,1)$. Define the $1-\alpha$ confidence interval for ρ as

$$\{\rho : |T_\rho| \leq z_{1-\alpha/2}\}.$$

Solve for ρ and compute the confidence interval for the data above. (This is called Fieller's method)

(d) Use the parametric and non-parametric bootstrap to estimate this confidence interval. $\log_2(LD_{50})$

(e) Consider the null hypothesis that $\log_2 \mathrm{LD}_{50} = 4$ as a sub-model or restriction of the linear logistic model. Fit the sub-model and compute the log likelihood ratio statistic LR(4), i.e. twice the difference between the unrestricted and the restricted log likelihood. If the null hypothesis is correct, this difference should be distributed approximately as χ^2_1. Compute the p-value.

Faraway Chapter 3.1

3. The material in Faraway Chapter 2 on the O-rings analysis from the Challenger crash was obtained from the paper Dalal, Fowlkes, and Hoadley (1989), which you can find on the homework web page.

(a) Load the oring dataset from the Faraway package in R and confirm the output shown on Faraway page 30. The estimates in the paper (equation 3.2) are obtained using the data in table 1 columns Joint-Temperature and Erosion-or-blowby. Why are they different?

(b) Do the binary analysis - event/no-event - performed in the paper using the oring data you have in R. How does it compare to that of the paper.

(c) There is clearly a sparsity problem with the binary analysis. Is there a sparsity problem with the binomial original data? Explain.

(d) Propose how you would remedy the sparsity problem for the binary data.

(e) In the paper they compare the expected number of events from the binomial model and the binary model. How do they derive the expected number from the probability estimated from the binary model.

1. In problem 2d,

Parametric bootstrap refers to sampling from the estimated model probability for each of the dose levels.

Non-parametric bootstrap means sampling from each dose level based on the observed proportions.

In each dose level you sample as many samples as there are original observations.

2. In problem 3,

the data from the paper is the one in column Erosion-or-blowby in the Field part of table 1.