Homework no. 3

This should be turned in by Friday, May 29

Use R to answer these questions. Cut-and-paste material from your R session into a
text file and email it to me (my email address is on my department faculty web page) The
homework should be turned in by email, not on paper. 1. Recall the private-school donation
numbers from the previous homework. Put them into a table with rows being years and
columns being the numbers of parents who donated and who did not donate (I gave you the
numbers who donated and the total number in that grade, so a little subtraction was needed).

Now imagine that we are unsatisfied with the analyses we did last time, and want to try
out a model in which the probability of donation is linear with an intercept (a) and a slope (b).
We want to test whether the slope b is zero. We decide to do this using maximum likelihood
and likelihood ratio tests.

The students are K-12 in reverse order. So if we call the grade levels 0 to 12, iand the
table of numbers is called parents, the R command grade <- 13-row(parents) gives a
table of the grade levels. Once a and b have had values assigned to them the array
expfreq <- cbind(a+b*grade[,1],1-a-b*grade[,2])
makes an array of the expected frequencies for each cell of the table. (What is going on: you
are making two vectors of frequencies, and then binding them side-by-side as columns into a
table).

Each parent is assumed to be independently deciding whether or not to give, so each row
of the data table reports the results of binomial trials. The log likelihood for a row would be

$$\log L = \log \left( \binom{n}{m} p^m (1-p)^{n-m} \right) = K + m \log(p) + (n-m) \log(1-p)$$

The K being some factorials that we can drop as they don’t change when we change a and b,
and they also cancel out of any likelihood ratio. That means the log-likelihood for a row is
just the sum over that row of the product of the observed number and the log of the expected
frequency. Warning: in R, log(x) is minus infinity when x = 0, so you will have to use
max(0,x) to set it to 0 to avoid things blowing us.

• Estimate a and b by maximum likelihood.

• Do a likelihood ratio test of whether b = 0. How many degrees of freedom does this
have?

These may require you to write short functions, even to apply a function to a vector of
values using something like sapply. Don’t be afraid to work together to do this.

2. In lecture I gave an example of Bayesian inference of tossing a coin 11 times and getting
5 heads, with a prior probability on heads which is a truncated exponential distribution,
where the exponential distribution had a mean of 1. Use a fine grid of values of $p$ with values from 0.001 to 0.999 (we’re avoiding 0 and 1 because of overflow/underflow issues). You can approximate integration by just summing up probabilities in intervals in the arrays.

- Compute the densities of the prior distribution, compute the likelihood curve, and the posterior distribution. Please *don't* send me them! (too many numbers!) Instead,
  - ... get the 95% credible interval for the prior and for the posterior.
  - ... and also the mode (highest point) for the posterior
  - ... and the maximum likelihood estimate too.