STA 410/2102, Fall 2014 — Assignment #1

Due at the start of class on October 16. Please hand it in on 8 1/2 by 11 inch paper, stapled in the upper left, with no other packaging.

This assignment is to be done by each student individually. You may discuss it in general terms with other students, but the work you hand in should be your own. In particular, you should not leave any discussion with someone else with any written notes (either paper or electronic).

In this assignment, you will use R's `nlm` and `optim` functions to find maximum likelihood estimates for parameters of a model involving counts of births, along with standard errors for these estimates, and investigate by simulation whether these standard errors are a good indication of the accuracy of the estimates. Graduate students in STA 2102 will also investigate how robust results from the model are when its assumption of a linear trend in birth rate is not true. (Undergrads in STA 410 can do this as a bonus question.)

The problem. Let's suppose that we have data on the number of children born each day in some region for a fairly large number of days (hundreds or thousands). Some of these births are of single children, others are of twins (we assume there are no triplets or higher multiple births). All we know, however, is the total number of children born each day, not how many of these were twins. Our primary interest is in estimating what proportion of pregnancies that produce at least one live birth produce two live births (ie, twins).

The model. The model we will use assumes that each day the number of women giving birth to at least one live child has a Poisson distribution, with the number for one day being independent of the number for other days. We will allow for the mean of this Poisson distribution to change over time according to a linear trend. Furthermore, we assume that each woman who gives birth has some fixed probability of giving birth to twins, and that whether twins are born is independent for each woman.

We will number the days from 1 to n, and denote the day number by t. We denote the number of births on day t by y_t. Our model has parameters a and b defining a linear trend and a parameter w giving the logit of the probability of a twin birth. To avoid having to constrain the parameters, we will take the mean for the Poisson distribution for the number of women giving birth to be $|a+bt|$ (the sign of the a and b parameters is therefore indeterminant). The w parameter is also unconstrained, the probability of a twin birth being $p = 1 / (1+\exp(-w))$, which is always between 0 and 1. Note that the number of women giving birth to a single child each day will have a Poisson distribuiton with mean $(1-p)\,|a+bt|$ and the number of women giving birth to twins will have a Poisson distribution with mean $p\,|a+bt|$.

With this model, the likelihood function will be the product of probabilities of the counts of children born for each day, with

$$P(y_t = y) \;=\; \sum_{i=0}^{\lfloor y/2 \rfloor} \text{Pois}(i,\, p\,|a+bt|)\ \text{Pois}(y-2i,\, (1-p)\,|a+bt|)$$

where $\text{Pois}(i, \lambda)$ is the probability mass function for the Poisson distribution with mean λ. The sum accounts for all possibilities for how many of the children born on day t are twins.

Finding the maximum likelihood estimates. You should write an R function that tries to find the maximum likelihood estimates for the parameters a, b, and w given a vector of counts, y. This function should be of the following form:

```
mle <- function (y, initial, method) { ... }
```

with `y` being the vector of counts (its length is n), `initial` being a vector of intial values for the parameters (in the order a, b, and w), and `method` being either `"nlm"` or `"optim"`, specifying which of R's built-in optimization functions should be used.

The value returned by this function should be a list with elements `estimate`, a vector of maximum likelihood estimates, `logl`, the log likelihood for this estimate, and `se`, a vector of standard errors for the estimates. The standard errors should be found from the square roots of the diagonal elements of the inverse of the observed information matrix at the MLE. (The observed information matrix is minus the Hessian matrix for the log likelihood.)

For details on the `nlm` and `optim` functions, see `help(nlm)` and `help(optim)`. Note that the first two arguments of these functions are the same, but in opposite order. Both functions will return the Hessian matrix if given an argument `hessian=TRUE`. You should use the default method for both of these functions, when no derivatives are supplied. For `nlm`, this is a Newton-like method using numerical derivatives; for `optim`, it is the Nelder-Mead method.

You should maximize the log likelihood, not the likelihood itself, and compute the log likelihood by summing log probabilities for each count. However, you needn't try to avoid underflow in computation of the probability for a single count (although this can lead to some warning messages).

Trying out the model and MLE function. You should first try out your function for finding the MLE on simulated data sets with $n = 200$ and $n = 600$, produced as follows (with `n` set to either 200 or 600):

```
set.seed(1)
a <- 2.3; b <- 0.02; p <- 0.05
y <- rpois(n,(1-p)*(a+b*(1:n))) + 2*rpois(n,p*(a+b*(1:n)))
```

You should try both `nlm` and `optim` for both sizes of data set. Be sure to set the seed as above (each time) so that your data sets are the same as everyone else's.

You should try to think of a reasonable scheme for setting the initial parameter estimates for the methods, and also try other initial values. You should comment on which method seems to produce more reliable results, and on how much computation time they take (you can measure computation time with the `system.time` function).

Checking whether the standard errors are reliable. You should next look at whether the standard errors found from the observed information are reliable when n is 600, by finding the MLE for many simulated data sets, with the same parameters as above, but setting the seed to values from 1 to K, where K is the number of data sets you simulate. (The bigger K is the better, but you'll be limited by the amount of computation time you can tolerate.) You can try this for both `nlm` and `optim` methods, to see if there is any difference.

For each optimization method, you should produce (and hand in) three scatterplots, one for each parameter, with a dot for each simulated dataset giving the estimated standard error (horizontal axis) and the actual error (vertical axis). Comment on whether the standard errors are generally a good guide to the accuracy of the estimates, and whether the variation in the standard error from one dataset to another reflects actual differences in accuracy.

Checking whether the model is robust to non-linear trends. Graduate students in STA 2102 should also check whether this model is robust to the trend not being linear. Undergrad students in STA 410 can do this for bonus marks (a maximum of 10 marks in addition to the 100 for the main part of the assignment).

To check this, you should simulate data sets in which the Poisson mean for the number of women giving birth does not vary linearly, as $a+bt$, but by some non-linear function. Focus on fairly minor deviations from linearity, since the model would likely not be used if the trend is drastically non-linear. The concern is instead that a misleading estimate for the probabilty of a twin birth might be obtained as a result of deviations from linearity that might not be obvious.

Explain the details of how you went about testing robustness, and give your conclusions.

What to hand in. You should hand in a paper printout of all your functions, the R scripts used to apply them to the datasets, and the output or plots from these scripts, along with your discussion of what it means. Your discussion should aim to provide insight into the results, not just summarize the numbers you obtained.