

## ST 762, HOMEWORK 2, FALL 2009

These two problems are to be turned in on the due date.

1. Using standard nonlinear regression software to implement GLS. The data in the file `trees.dat`, available on the class web page, were collected by forest science researchers interested in the effects of two different ways of preparing sites for planting of slash pine trees (*Pinus elliotii* Englem) on the growth of the trees. One way of preparing the site is to do nothing (the control); the other way is to carry out what is referred to as “chop-burn-bed,” where the plot is subjected to a pass with a rolling drum chopper that chops any materials remaining on the plot from a previous planting, followed by a broadcast burn of the plot, followed by bedding of the plot for planting.

The researchers have data from a study in which plots were randomly allocated to receive either the control treatment or the chop-burn-bed treatment. Subsequently, at each of several time points (starting at 4 years post-planting and then at 2-year intervals until age 18), 2 plots were selected at random from those for each treatment, and for each, a measure of the size of trees on the plot, dominant height (meters), was recorded. In the data set, the plots selected at each observation time were different for each treatment, so that each measurement of dominant height in the data set corresponds to a different plot.

In the data set, the columns are (1) treatment indicator (0 = control, 1 = chop-burn-bed); (2) age of the trees in years,  $t$ , (time since planting), and (3) dominant height (m). A popular model for the relationship between dominant height (the response) and age of the trees  $t$  is the so-called modified Chapman-Richards model given by

$$f(t, \boldsymbol{\beta}) = \beta_1 \left\{ \frac{1 - \exp(-\beta_2 t)}{1 - \exp(-\beta_2 t_0)} \right\}^{\beta_3},$$

where  $\beta_1$  is the expected dominant height at a chosen reference age  $t_0$ ,  $\beta_2$  is the growth rate parameter, and  $\beta_3$  is the curve shape parameter allowing flexibility in the relationship.  $\beta_1$  is often referred to as the “site index” corresponding to the chosen reference age. The researchers are interested in a site index corresponding to  $t_0 = 20$  years.

The researchers believe that the shape parameter is similar for the two treatments; however, they believe that the site indices and growth rate parameters may differ between the two treatments.

(a) Using your favorite software, plot the data on the same graph, using a different symbol for each site preparation treatment. Does the plot suggest that the data exhibit nonconstant variance? From your visual inspection, do you think it is likely that site index and/or growth rate might be different for the two treatments?

(b) With the examples in Section 3.7 as a guide, write a program to implement the 3-step GLS algorithm using either SAS `proc nlin` or the R function `nls()` to fit the modified Chapman-Richards model *simultaneously* to the data from both treatments, allowing the possibility that both site index and growth rate parameter differ across treatments. Thus, you will need to assume that  $E(Y_j|x_j) = f(x_j, \boldsymbol{\beta})$ , where  $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T$  is such that  $\beta_1$  and  $\beta_2$  are different across treatments but  $\beta_3$  is the same (You’ll need to figure out how to represent and fit this model to all the data simultaneously.) Assume also that  $\text{var}(Y_j|x_j) = \sigma^2 f^{2\theta}(x_j, \boldsymbol{\beta})$ ,  $j = 1, \dots, 32$  (so the pattern of variance is the same – same  $\sigma^2$  and  $\theta$  – for both treatments).

Under these conditions, obtain estimates of  $\sigma$  and  $\boldsymbol{\beta}$  when  $\theta$  is known and such that the data are assumed to have constant coefficient of variation. Be sure you do all of the following:

- Iterate the 3-step GLS algorithm at least  $C = 10$  times
- Determine suitable starting values for  $\beta$
- Estimate  $\sigma$  using the final estimate of  $\beta$  from the 3-step procedure

*Note:* Both SAS `proc nlin` and R `nls()` have options that control the “*tol*” (see Homework 1) for the convergence criterion; moreover, they use different convergence criteria. For both `nls()` and `proc nlin`, the convergence criterion used by default is different from the relative criterion given in (3.11) on page 59 of the notes.

For `nls()`, something called the *relative offset criterion* is used, which is discussed in the book by Bates and Watts listed in the syllabus (the `help` for `nls()` in R is pretty vague on this). The *tol* for this criterion is by default set to  $10^{-5}$ . To change it, add the following to the call to `nls`:

```
nls.control(tol=1e-8)
```

(this changes the tolerance to  $10^{-8}$ , for example).

For `proc nlin`, the various convergence criteria, including the default, are described in the SAS documentation for `proc nlin` under the syntax for the `proc nlin` statement (there is link to the on-line documentation in the gray box on the class web page). The criterion (3.11) in the notes can be invoked by adding the option `convergeparm=c` to the `proc nlin` statement, where `c` equals what we called *tol* above and in the notes.

You may want to try fooling around with changing the criterion and *tol* to get a sense of the effect on the reported estimates.

(c) Make another plot of the data, again using different symbols for each treatment, and superimpose the OLS and GLS fits for each treatment on the plot, using a different line type for each fit. Does the fit change appreciably depending on whether or not nonconstant variance is taken into account?

2. *Poisson regression.* Recall that in Homework 1, you wrote a program to implement iteratively reweighted least squares. In the file `skincancer.dat`, available on the class web page, you will find data from an observational study that collected information on  $n = 954$  skin cancer patients. Recorded on each subject are characteristics thought to increase the risk of developing larger numbers of skin cancers along with the response, the number of skin cancers experienced by the subject in the past five years. In particular, each line of the data set corresponds to a different subject, and the columns are as follows:

In the data set, each line corresponds to a different subject, and the columns are as follows:

column	variable
1	subject ID number
2	sunburn (1=no history of sunburn, 2=three or more major sunburn episodes, 3=one or two major sunburn episodes)
3	age (years)
4	skin type (1=fair skin, 0 = other)
5	gender (0=female, 1=male)
6	number of skin cancers

The investigators wanted to know which of the variables in columns 2–5 are associated with having greater numbers of skin cancers in subjects with skin cancer.

Let  $Y_j$  denote the number of skin cancers observed on subject  $j$ , and consider the following model for mean number of skin cancers given the variables in columns 2–5 of the form

$$\log\{E(Y_j|\mathbf{x}_j)\} = \beta_1 + \beta_2x_{j1} + \beta_3x_{j2} + \beta_4x_{j3} + \beta_5x_{j4} + \beta_6x_{j5},$$

where, for subject  $j$ ,

$x_{j1}$	sunburn indicator (= 0 if no history, = 1 if three or more episodes)
$x_{j2}$	sunburn indicator (= 0 if no history, = 1 if one or two episodes)
$x_{j3}$	age
$x_{j4}$	skin type (= 0 if other, = 1 if fair)
$x_{j5}$	gender (= 0 if female, = 1 if male)

(a) Let  $\mathbf{x}_j = (1, x_{j1}, x_{j2}, x_{j3}, x_{j4}, x_{j5})^T$ . We include the “1” for the intercept as a covariate here for reasons that will be clear below. Write down the assumed model  $f(\mathbf{x}_j, \boldsymbol{\beta})$  for  $E(Y_j|\mathbf{x}_j)$  implied by the above. Assume that, conditional on  $\mathbf{x}_j$ ,  $Y_j$  has a Poisson distribution with this mean. Write down the maximum likelihood estimating equation for  $\boldsymbol{\beta}$  under these conditions and show that it is of the form of equation (4.8) on page 89 of the notes.

(b) *Starting values.* Note that the model in (a) is of the generalized linear model form  $f(\mathbf{x}_j, \boldsymbol{\beta}) = f(\mathbf{x}_j^T \boldsymbol{\beta})$ ; thus, page 92 of the notes is applicable. A standard technique for finding starting values for generalized linear models may be based on these developments as follows. Let  $\mathbf{X}_*$  be defined on page 92. Show that  $\mathbf{W}_*$  and  $\mathbf{Z}_*$  on page 92 may be expressed as functions of  $f$ . Then, taking as an initial guess for the value of  $f(\mathbf{x}_j^T \boldsymbol{\beta})$  for each  $j$ ,

$$\frac{Y_j + \bar{Y}}{2}, \quad \bar{Y} = n^{-1} \sum_{j=1}^n Y_j, \tag{1}$$

use the update equation at the bottom of page 92 evaluated at  $f(\mathbf{x}_j^T \boldsymbol{\beta}) = (1)$  to derive a starting value for  $\boldsymbol{\beta}$ .

(c) Fit this model to the skin cancer data using the IRWLS program you developed in Homework 1. Use the starting values you found in (b), and take  $tol = 10^{-8}$  as in Homework 1.

(d) Fit this model to the skin cancer data using the 3-step GLS algorithm implemented using either SAS proc `nlin` or the R function `nls()`, as in the examples in Section 3.7.

*Note:* See the remarks on convergence criteria in Problem 1(b). Note that, as both this implementation and that in (c) are solving the same set of estimating equations, we would expect them to yield identical results, as you saw in Homework 1. However, it is entirely possible that, due to differences in convergence criteria and choice of  $tol$  used that your results in (c) and (d) may not agree precisely. You may wish to fool around with these to see what you need to do so that the numerical results from (c) and (d) agree precisely (to some number of decimal places).

(e) From either fit in (c) or (d), provide an estimate of the probability that a 55 year old male subject in this population with fair skin and no history of sunburn would have two or fewer skin cancers.