STAT 8230 — Applied Nonlinear Regression
Homework 4 – Due Tuesday, Oct. 18

Homework Guidelines:

- Homework is due by 4:30 on the due date specified above. You may turn it in at the beginning of class or place it in my mailbox in the Statistics Building. **No late homeworks will be accepted without permission granted prior to the due date.**

- Use only standard (8.5 × 11 inch) paper and use only one side of each sheet.

- Homework should show enough detail so that the reader can clearly understand the procedures of the solutions.

- Problems should appear in the order that they were assigned.

Assignment:

1. Consider again problem 1 from homework 3 and refer to the homework #3 solution set. In this problem we seek to compare the performance of the various different NLS algorithms for fitting the model

$$y_i = \frac{1+x}{\rho_a^{-1} + \beta x/1000} + e_i \qquad (*)$$

where y = apparent density, x = moisture, and we assume independent, homoscedastic errors.

a. Using the model above and the starting values $\hat{\boldsymbol{\theta}}^0 = (1148, 1.02)^T$, fit the model using each of the four NLS algorithms implemented in PROC NLIN (GN, NR, SD, and Marquardt). For each algorithm, record the following information:

– Whether or not the algorithm converged.

– The number of iterations required for convergence.

– whether or not the procedure converged to the NLS estimates (you may assume that the ones obtained in the solution to homework #3 are correct).

b. Repeat part (a) using starting values corresponding to the final parameter estimates, $\hat{\boldsymbol{\theta}} = (1148.5, 1.013)^T$, perturbed by t standard errors, for $t = 1$ and $t = 5$. Also try each method using $\hat{\boldsymbol{\theta}}^0 = (10, 10)^T$ (a really bad set of starting values).

c. Summarize your findings concerning the convergence properties of the 4 algorithms considered here and draw any conclusions that you believe are valid. Compare your results and conclusions with the guidelines/advantages/ disadvantages concerning these algorithms given in the lecture notes.

d. Use the QN method as implemented in PROC NLMIXED to fit model (*) using the correct NLS estimates plus 5 standard errors as starting values. Use both the BFGS and DFP quasi-Newton updates to do this (see the UPDATE option on the PROC NLMIXED statement) and report how many iterations each version of QN took to converge, and whether the algorithm converged to the correct solution. You'll also need a starting value for σ^2; use $\hat{\sigma}_{20} = 3500$.

2. Included in the nlme library for R is a data set containing the circumference (mm) of 5 orange trees measured repeatedly through time (days). The data are contained in the *grouped data set* Orange and are also reproduced in the table below:

	Tree No.				
Time (days)	1	2	3	4	5
118	30	33	30	32	30
484	58	69	51	62	49
664	87	111	75	112	81
1004	115	156	108	167	125
1231	120	172	115	179	142
1372	142	203	139	209	174
1582	145	203	140	214	177

These data are *grouped* in the sense that observations from the same tree "belong together" (are in the same group) and may be more similar to one-another than observations from different trees. The grouping structure of the data can be encoded as part of the data set by making Orange a "groupedData object" rather than simply a data frame. The groupedData() function constructs a groupedData object from a data frame. The groupedData object is similar to a data frame, but it also contains a formula attribute and a group structure. If you simply type Orange in R (you must execute the command *library(nlme)* first), you'll see the Orange groupedData object. Notice it has the formula, circumference $\sim$ age | Tree, which indicates that the response variable in the data set is circumference, the primary covariate is age, and the data are grouped by Tree. Having Orange as a groupedData object rather than simply a data frame, simplifies much of the data analysis. See Pinheiro and Bates (2000, §3.2) for more on creating groupedData objects.

Plot the circumference growth curves separately by tree by using the command, plot(Orange, outer = $\sim$ 1), to get a feel for the data. Notice that the individual-tree growth curves have the familiar sigmoidal form. We will consider logistic models in this problem to capture this sigmoidal shape.

a. Fit a logistic model to data from all 5 trees ignoring the grouping structure of the data. That is, fit a model of the form

$$y_{ij} = \frac{\theta_1}{1 + \exp\{-(\text{age}_{ij} - \theta_2)/\theta_3\}} + e_{ij}$$

where y_{ij} represents the j^{th} response (circumference) on the i^{th} tree. Utilize whatever heteroscedasticity model you feel is appropriate for these data if they exhibit non-constant variance. For now, restrict attention to models with uncorrelated errors.

b. Assess the adequacy of your model from part (a)? Does it fit well? Are model assumptions supported by the data? What flaws does it apear to have? To answer these questions you may want to examine residual plots for all trees combined and also separately by tree (e.g., you can examine standardized residuals separately by tree with the commands, plot(fittedmodel, Tree ∼ resid(., type="p"), abline=0) and plot(fittedmodel, resid(.,type="p") ∼ fitted(.) | Tree,abline=0)) and any other diagnostic measures of your choosing.

c. The growth curves do not appear to be the same for the five trees. Therefore, now consider the model

$$y_{ij} = \frac{\phi_i}{1 + \exp\{-(\text{age}_{ij} - \theta_2)/\theta_3\}} + e_{ij} \quad i = 1, \ldots, 5, j = 1, \ldots, 7.$$

Fit this model using nls(). For now, assume spherical (mean zero, constant variance, uncorrelated) errors. As in part (b), examine residuals separately by tree for this model. Is this model an improvement from part (b)?

d. Refit the model from part (c) using gnls() with the GN algorithm. Is the assumption of spherical errors appropriate here? If not, add heteroscedasticity and/or correlation to your model as appropriate to obtain a suitable model. Describe and summarize your final model and support it with appropriate model diagnostics.

3. The file cleavers.dat on the course website contains data from an experiment designed to measure the effect of two formulations of the herbicide phenmedipham. Herbicide treatment 1 is phenmedipham sprayed alone, and treatment 2 is phenmedipham sprayed with an adjuvant that is expected to enhance its effectiveness. The type of plant used to evaluated these treatments was cleavers (*Galium aparine*). The same ten nonzero doses were used for both formulations, with ten replicates per dose. In addition, 20 replicates of an untreated condition (dose=0, labelled treatment 0 in the dataset) were collected. The response was the dry matter weight of the plant following treatment.

a. Fit the following four-parameter logistic regression model to these data:

$$y_{ijk} = \theta_1 + \frac{\theta_{2i} - \theta_1}{1 + \exp\{(\theta_{3i} - x_{ijk})/\theta_{4i}\}} + e_{ijk},$$

where y_{ijk} is the dry weight for the k^{th} replicate at the jth dose under the ith treatment. Assume the e_{ijk}s are independent normal with mean 0. Use whatever variance model you think is appropriate for these data, and support your choice with empirical results. Here x_{ijk} is the $\log(\text{dose}+1)$ for the i, j, kth measurement. You may re-group the 0 dose measurement into either treatment 1 or 2 (it doesn't matter which) to fit this model. Interpret the parameters of this model and summarize the fitted model with appropriate graphical and numerical summaries.

b. Reduce the model you fit in part (a) by allowing θ_{2i}, θ_{3i}, and/or θ_{4i} to be constant across treatments (across i) as supported by the data. Use appropriate inferential tools (e.g., extra sum of squares tests, likelihood ratio tests, AIC) to determine a simple adequate model. Summarize your findings.