STAT 8230 — Applied Nonlinear Regression
Homework 3 – Due Thursday, Oct. 6

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.

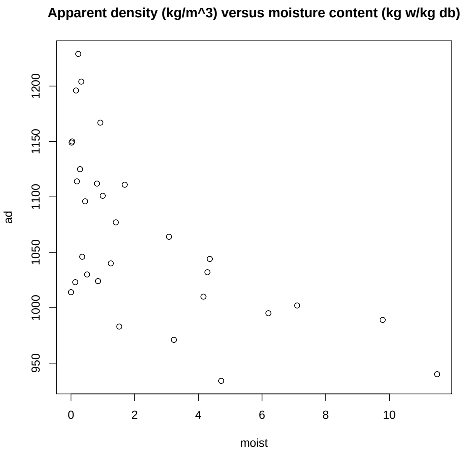
- Problems should appear in the order that they were assigned.

Assignment:

From the book by Bates & Watts, do the following problems: 3.4, 3.5. These problems are appended to this assignment sheet. When they ask that you plot the data on semi-log paper, instead plot the log response variable versus x.

In addition, do the following problems:

1. The plot below shows the apparent density (also known as bulk density) versus moisture content for a sample of 30 onions. These data can be found on the course website in the file onions2.dat. For now, ignore the variables "source" and "variety" in this dataset.



These data were analyzed by researchers who used the following function as the deterministic component of a nonlinear regression model relating apparent density

(y) to moisture content (x):

$$\frac{1+x}{\rho_a^{-1}+\beta x/\rho_w},$$

where ρ_a and β are unknown parameters, identified by the researchers as the *dry solid apparent density* and the *volume shrinkage coefficient*, respectively. In addition, ρ_w is the *enclosed water density*, which is a constant known to be equal to 1000.

a. Add an appropriate random component to this model and fit this model with ordinary least squares. Summarize the resulting fitted model and plot the fitted curve on a scatterplot of the data. Compute 95% confidence intervals for each regression parameter. Does the model seem to fit these data adequately? Why or why not?

b. To fit your model you will need starting values. Interpret the regression parameters in your model and describe a method for obtaining starting values for each one. Create a self starting function in R to implement your method automatically for models like the one you chose to use for these data. Make sure that your self starting function returns analytic derivatives as well as starting values.

c. Suppose that the investigators were concerned that the enclosed water density may not have been exactly 1000 and they wanted to estimate this parameter instead of treating it as a known constant. However, when they tried to do this they ran into difficulty fitting the model. They suspected that the problem was poor starting values or not having enough data. What do you think the trouble is that they encountered, and what should they do about it?

2. The data below come from an experiment in which turf grass was grown in acidic soil with differing amounts of calcium added to the soil. The calcium was added for the purpose of reducing soil acidity and improving plant growth. The response variable measured was the root mass of grass grown over a fixed period of time. These data are also contained in the file grass.dat on the course web site.

Calcium (x)	Root Mass (y)
0	5.5
0	2.1
0	2.0
1250	7.7
2500	30.1
2500	33.9
2500	23.8
5000	57.2
5000	70.4
5000	62.9
10000	85.5
10000	92.5
10000	86.2

a. Fit the Gompertz model

$$y_i = \theta_1 \exp\{-\exp[-(x-\theta_3)/\theta_2]\} + e_i \qquad (*)$$

to these data, where $e_1, \ldots, e_n$ are mean zero, constant variance errors. Summarize the fitted model and comment on the adequacy of the assumption of additive, homoscedastic errors.

b. With reference both to the context of this particular data set and the general mathematical properties of this model, provide interpretations for the parameters $\theta_1, \theta_2, \theta_3$ in model (*).

c. The scientist who collected these data was interested in determining how much calcium was necessary to fully neutralize soil acidity and achieve "optimal" growth. In an asymptotic model for these data, mean growth would increase asymptotically with calcium and maximal growth would never be achieved with a finite vlaue of calcium. So, trying to identify the calcium level that achieves maximal growth according to model (*) is clearly not an appropriate way to try to answer the question of scientific interest. Instead, one way to answer the scientific question is to identify the calcium level that a given high percentage of maximal growth (80% of maximal growth say). Model (*) can be reparameterized to facilitate identification of this "effective dose" value of calcium. In particular, model (*) can be reparameterized so that the ED80 (effective dose at which 80% of the maximal response is achieved) can be directly estimated. Find such a reparameterization and fit the resulting model to the grass data set. Report an estimate and 95% confidence interval for the ED80 for this data set. *Hint:* consider a transformation of the form $(\theta_1, \theta_2, \theta_3)^T \rightarrow (\phi_1, c(\phi_2 - \phi_3), \phi_3)^T$ for some appropriately chosen constant c.

3. For 31 black cherry trees the following measurements were obtained:

$$\begin{aligned} V &= \text{Volume of usable wood (cubic feet)} \\ H &= \text{Height of tree (feet)} \\ D &= \text{Diameter at breast height (inches)} \end{aligned}$$

These data are contained in the file cherrytree.dat. Before analyzing these data, transform D from inches to feet by the replacement $D \leftarrow D/12$. We consider a model for volume based on the formula for the volume of a cone: $V = \pi H D^2/12$. Of course, cherry trees aren't exactly shaped like cones, so we consider a model of the form

$$V_i = \theta_1 H_i^{\theta_2} D_i^{\theta_3} + e_i, \qquad i = 1, \ldots, 31.$$

a. Fit this model using both homoscedastic errors and heteroscedastic errors. Which assumption appears to be more appropriate here? (You may want to make use of the AIC and BIC information criteria here).

b. Describe both a theoretical approach and an empirical approach to obtaining starting values in this problem.

c. Summarize and interpret the fitted model that you feel is most appropriate for these data.

d. Test the null hypothesis $H_0 : \theta_1 = \frac{\pi}{12}$ versus $H_1 : \theta_1 \neq \frac{\pi}{12}$ at significance level $\alpha = .05$. State and interpret your conclusions from this test.

(c) Use the starting values in a nonlinear least squares routine to find the least squares estimates for the parameters for each data set.

(d) Use incremental parameters and indicator variables to fit all of the data sets together.

(e) Simplify the model by letting some of the parameters be common to all of the data sets. Use extra sum of squares analyses to determine a simple adequate model.

(f) Write a short report about this analysis and your findings.

3.2 Use the data from Appendix 1, Section A1.14 to determine an appropriate sum of exponentials model.

(a) Plot the data on semilog paper and use the plot to determine the number of exponential terms to fit to the data.

(b) Use curve peeling to determine starting estimates for the parameters.

(c) Use the starting estimates from part (b) to fit the postulated model from part (a).

3.3 (a) Use the plot from Problem 2.6 and sketch in the curve of steepest descent from the point $\boldsymbol{\theta}^0$. Hint: The direction of steepest descent is perpendicular to the contours.

(b) Is the direction of the Gauss–Newton increment close to the initial direction of steepest descent?

(c) Calculate and plot the Levenberg increment using a conditioning factor of $k=4$.

(d) Calculate and plot the Marquardt increment using a conditioning factor of $k=4$.

(e) Comment on the relative directions of the Gauss–Newton, Levenberg and Marquardt increment vectors.

3.4 Use the data from Appendix 4, Section A4.3 to determine an appropriate model and to estimate the parameters.

(a) Plot the concentration versus time on semilog paper, and use the plot to determine the number of exponential terms necessary to fit the data.

(b) Use the plot and the method of curve peeling to determine starting values for the parameters.

(c) Use a nonlinear estimation routine to estimate the parameters.

3.5 Use a nonlinear estimation routine and the data and model from Appendix 4, Section A4.4 to estimate the parameters. Take note of the number of iterations required and any difficulties you encounter in each attempt.

(a) Use any approach you think is appropriate to obtain starting values for the parameters in the model.

(b) Use your starting values in a nonlinear estimation routine to estimate the parameters. If you achieve convergence, examine the parameter approximate correlation matrix, and comment on the conditioning of the model.

(c) Reparametrize the model by centering the factor $1/x_3$, and use the equivalent starting values from part (a) to estimate the parameters. If you achieve convergence, examine the parameter approximate correlation matrix, and comment on the conditioning of the model. What effect does this reparametrization have on the number of iterations to convergence?

(d) Reparametrize the model in part (a) using $\theta_1 = e^{\phi_1}$ and $\theta_2 = e^{\phi_2}$ and the equivalent starting values from part (a) to estimate the parameters. If you achieve convergence, examine the parameter approximate correlation matrix, and comment on the conditioning of the model. What effect does this reparametrization have on the number of iterations to convergence?

(e) Reparametrize the model in part (b) using the same parametrization as in part (c) and the equivalent starting values from part (a) to estimate the parameters. If you achieve convergence, examine the parameter approximate correlation matrix, and comment on the conditioning of the model. What effect does this reparametrization have on the number of iterations to convergence?

3.6 Use a nonlinear estimation routine and the data and model from Appendix 4, Section A4.5 to estimate the parameters. Take note of the number of iterations required and any difficulties you encounter in each attempt.

(a) Use any approach you think is appropriate to obtain starting values for the parameters in the model.

(b) Use your starting values in a nonlinear estimation routine to estimate the parameters. If you achieve convergence, examine the parameter approximate correlation matrix, and comment on the conditioning of the model.

(c) Reparametrize the model in part (a) using $\theta_2 e^{-\theta_3 x} = e^{-\phi_3(x-\phi_2)}$. If you achieve convergence, examine the parameter approximate correlation matrix, and comment on the conditioning of the model. What effect does this reparametrization have on the number of iterations to convergence?

3.7 (a) Show that the theoretical D-optimal starting design for the logistic model of Problem 3.1 consists of $\mathbf{x} = (-\infty, \theta_3 - 1.044/\theta_4, \theta_3 + 1.044/\theta_4, +\infty)^{\mathrm{T}}$.

(b) Interpret the choice of the design points graphically by plotting the logistic function versus x and plotting the location of the design points on the x-axis.

(c) Plot the derivatives with respect to the parameters versus x and use these plots to help interpret the choice of the design points.

The model is an empirical generalization of two models based on theory. It is written

$$f(x,\boldsymbol{\theta}) = \theta_2 + \frac{\theta_1 - \theta_2}{\left[1 + \left(i2\pi x\, e^{-\theta_3}\right)^{\theta_4}\right]^{\theta_5}}$$

where f is predicted relative complex impedance and x is frequency.

A1.14 Tetracycline

Data on the metabolism of tetracycline were presented in Wagner (1967). In this experiment, a tetracycline compound was administered orally to a subject and the concentration of tetracycline hydrochloride in the serum in micrograms per milliliter (μg/ml) was measured over a period of 16 hours. (See Table A1.14.)

A 2-compartment model was proposed, and dead time was incorporated as

$$f(x,\boldsymbol{\theta}) = \theta_3[e^{-\theta_1(x-\theta_4)} - e^{-\theta_2(x-\theta_4)}]$$

where f is predicted tetracycline hydrochloride concentration and x is time.

Table A1.14 Tetracycline concentration versus time.

Time (hr)	Tetracycline Conc. (μg/ml)	Time (hr)	Tetracycline Conc. (μg/ml)
1	0.7	8	0.8
2	1.2	10	0.6
3	1.4	12	0.5
4	1.4	16	0.3
6	1.1		

From "Use of Computers in Pharmacokinetics," by J.G. Wagner, in *Journal of Clinical Pharmacology and Therapeutics*, 1967, 8, 201. Reprinted with permission of the publisher.

A4.2 Nitrendipene

Data on binding of [^{3}H] nitrendipine to sites in rat heart homogenate were obtained by Abdollah (1986). In this study, experiments were performed to investigate the competition for binding to the sites between nitrendipene (NTD), a calcium channel antagonist, and nifedipine (NIF), another calcium channel antagonist. Heart tissue was homogenated and incubated with radioactively tagged NTD at molar concentration $\approx 5\times10^{-10}$ in the presence of different concentrations of NIF, which are given in Table A4.2 as $x = \log_{10}$(NIF concentration), except for the rows with (0), for which the actual concentration was 0. The NIF has greater binding ability and so displaces the NTD. Counts on radioactive material were obtained to determine how much material was bound under different conditions. When the NIF concentration is 0, all of the radioactive NTD is bound to the sites, and so a large count is recorded: as the NIF concentration increases, it displaces NTD and so lower counts are recorded. Although the nominal NTD concentration was 5×10^{-10}, the actual concentrations were 4.76, 5.11, 4.78, and 5.02 $\times10^{-10}$ respectively, for the four tissue samples.

The proposed model is

$$f(x, \boldsymbol{\theta}) = \theta_1 + \frac{\theta_2}{1+\exp[-\theta_4(x-\theta_3)]}$$

where f is the predicted total count and x is $\log_{10}$(NIF concentration).

A4.3 Saccharin Data Set 2

Data on the concentration of saccharin in plasma were reported in Renwick (1982) and are reproduced in Table A4.3.

A4.4 Steady State Adsorption

Data on the disappearance of o-xylene as a function of oxygen concentration, inlet o-xylene concentration, and temperature, were obtained by Juusola (1971) and were further analyzed by Pritchard (1972). The data are reproduced in Table A4.4.

The postulated model is a steady state adsorption model written

$$f(\mathbf{x}, \boldsymbol{\theta}) = \frac{f_1 f_2}{f_1 + 2.2788 f_2}$$

$$f_1 = \theta_1 x_1 e^{-\theta_3/x_3}$$

$$f_2 = \theta_2 x_2 e^{-\theta_4/x_3}$$

Table A4.4 Rate of oxidation of o-xylene versus oxygen concentration (gm-mole/l), inlet o-xylene concentration (g-mole/l), and temperature (K). The reaction rate is recorded as (g-mole/g-mole catalyst second) at standard catalyst age.

Oxygen	o-Xylene	Temp.	Rate
0.00502	0.000200	543	116
0.00499	0.000190	543	120
0.00504	0.000200	543	114
0.00505	0.000200	543	117
0.01000	0.000351	543	245
0.01010	0.000351	543	230
0.01030	0.000050	543	106
0.01040	0.000361	543	230
0.01010	0.000049	543	121
0.01010	0.000050	543	115
0.01010	0.000050	543	127
0.00570	0.000201	563	408
0.00552	0.000201	563	380
0.00551	0.000202	563	320
0.00551	0.000186	563	399
0.00554	0.000202	563	371
0.00553	0.000199	563	368
0.00108	0.000051	563	63
0.00707	0.000099	563	333
0.00554	0.000197	563	322
0.00605	0.000351	563	413
0.00552	0.000202	563	344
0.01016	0.000189	563	543
0.00552	0.000200	563	372
0.00603	0.000049	563	229
0.01000	0.000201	563	563
0.01010	0.000151	563	490
0.00805	0.000354	563	595
0.00552	0.000199	563	352
0.00249	0.000198	563	224
0.00571	0.000049	563	198
0.00555	0.000347	563	463
0.00549	0.000274	563	370
0.00554	0.000095	563	258
0.00507	0.000191	573	543
0.00502	0.000187	573	561
0.00505	0.000192	573	560
0.00506	0.000188	573	578
0.00500	0.000201	573	542
0.00100	0.000350	573	197
0.00505	0.000202	573	559
0.00306	0.000349	573	414
0.00502	0.000198	573	467
0.00504	0.000201	573	468
0.01017	0.000245	573	933
0.00499	0.000187	573	509
0.01000	0.000253	573	955
0.00496	0.000346	573	650
0.01000	0.000253	573	902
0.00502	0.000199	573	532
0.00399	0.000357	573	552
0.00107	0.000196	573	184
0.00499	0.000353	573	663
0.00503	0.000100	573	409
0.00251	0.000199	573	326
0.00499	0.000277	573	580
0.00906	0.000205	573	831

Copyright 1971 by J. A. Juusola. Reproduced from "A Kinetic Mechanism for the Vapor-Phase Oxidation of o-xylene," Ph.D. Thesis, Queen's University. Reprinted with permission of the author.

where f is predicted reaction rate, x_1 is oxygen concentration, x_2 is o-xylene inlet concentration, and x_3 is temperature.