STA 410S/2102S: Homework #3
Due April 7, 2005

1. The data in Table 1 show the calcium uptake of cells y (`cal`) as a function of x (`time`), after being suspended in a solution of radioactive calcium. The model suggested for this data is a nonlinear regression model

$$y_i = \beta_0\{1 - \exp(-\beta_1 x_i)\} + \epsilon_i$$

where $i = 1, \ldots, 27$, and we assume $\epsilon_i \sim N(0, \sigma^2)$.

   (a) Calcium is absorbed into cells through the cell walls. The calcium molecules in this experiment were labelled with a radioactive dye, so that at each time point the amount of calcium absorbed could be measured. The measurement of calcium absorbed is moles/microgram. What interpretation can you give to the parameters β_0 and β_1?

   (b) Fit this model using either `nls` or `optim`, and report the estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and their estimated standard errors. (Provide your code as an appendix.)

   (c) Plot the data and the fitted model.

   (d) Define the parameter $\pi(x_0) = 1 - \exp(-\beta_1 x_0)$; it is the proportion of the maximum value reached by time x_0. Estimate $\pi(15)$ and obtain an estimate of its standard error. Use these to construct an approximate 95% confidence interval for $\pi(15)$.

2. Nonlinear mixed effects models: In §10.3 the book describes extending mixed effects models to nonlinear least squares. This is implemented in R in `nlme`, in the library `nlme`. This question concerns the book's analysis of the data set `Sitka` in library `MASS`, described on pp. 286 – 288.

   (a) Each tree is measured at 5 different time points. Adapt the code on p.272 to plot the curves of size against time for several trees; some from treatment (`treat = ozone`) and some from control (`treat=control`).

   (b) Fit the nonlinear model adopted in the book, using the code near the top of p.287.

   (c) Give the mathematical formulation of this model, in the form of equation (10.4), but giving the specific form of the nonlinear and random effects parts of the model.

   (d) Choose 5 trees and plot the observed and fitted equations for each of these trees.

   (e) What does the statement (top of p.288) "The t value for a difference in slope by treatment is convincing" mean, and what part of the `summary` on the bottom of p.287 does it refer to?

3. **2102 only; optional for 410** the EM algorithm for a mixture of Gaussians: Suppose we have data that we expect might be modelled as a mixture of two Gaussian distributions:

$$f(y; \mu_1, \sigma_1, \mu_2, \sigma_2, \pi) = \pi \frac{1}{\sqrt{2\pi}\sigma_1} \exp\{-\frac{1}{2\sigma_1^2}(y_i - \mu_1)^2\} + (1-\pi)\frac{1}{\sqrt{2\pi}\sigma_2} \exp\{-\frac{1}{2\sigma_2^2}(y_i - \mu_2^2)\}$$

with all five parameters $\theta = (\pi, \mu_1, \mu_2, \sigma_1, \sigma_2)$ unknown. We considered fitting this model using maximum likelihood in class on March 10. Another way of fitting it is described in the book by Hastie, Tibshirani and Freedman used as the 450 text this year (§8.5.1). We imagine y has having been generated in a two-stage process: there is a latent variable Δ that follows a Bernoulli distribution:

$$\begin{aligned} \Delta &= 1 \quad \text{with probability } \pi \\ &= 0 \quad \text{with probability } 1-\pi. \end{aligned}$$

If $\Delta = 1$ then we generate an observation from a $N(\mu_2, \sigma_2^2)$, and if $\Delta = 0$ then we generate an observation from a $N(\mu_1, \sigma_1^2)$.

(a) Suppose we actually observed these latent variables: then our data set would be $(y, \delta) = (y_1, \delta_1, \ldots y_n, \delta_n)$. Show that the likelihood based on this data is

$$\ell(\theta; y, \delta) = \sum_{i=1}^{n} \{(1-\delta_i) \log \phi_{\theta_1}(y_i) + \delta_i \log \phi_{\theta_2}(y_i)\} + \sum (1-\delta_i) \log(1-\pi) + \sum \delta_i \log(\pi)$$

where ϕ_{θ_1} is shorthand for the density of a $N(\mu_1, \sigma_1^2)$ and similarly ϕ_{θ_2}. Show also that the maximum likelihood estimates of μ_1 and σ_1^2 are the sample mean and variance of the y_i's for which $\delta_i = 0$, and that the maximum likelihood estimates of μ_1 and σ_1^2 are the sample mean and variance of the y_i's for which $\delta_i = 1$.

(b) Since the δ_i are not in fact known, these estimates are not available. The EM algorithm proceeds in steps to first estimate the δ_i based on a current guess for θ, and then to update the estimates of the mean and variance parameters using the easier likelihood based on y and the estimates of δ. Argue that a reasonable estimate of δ_i is

$$\hat{\delta}_i = \frac{\hat{\pi}\phi_{\hat{\theta}_2}(y_i)}{(1-\hat{\pi})\phi_{\hat{\theta}_1}(y_i) + \hat{\pi}\phi_{\hat{\theta}_2}(y_i)};$$

and that given these estimates of $\hat{\delta}_i$ estimates of θ should be updated as follows:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^n (1-\hat{\delta}_i) y_i}{\sum_{i=1}^n (1-\hat{\delta}_i)}, \qquad \hat{\sigma}_1^2 = \frac{\sum_{i=1}^n (1-\hat{\delta}_i)(y_i - \hat{\mu}_1)^2}{\sum_{i=1}^n (1-\hat{\delta}_1)}$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^n \hat{\delta}_i y_i}{\sum_{i=1}^n \hat{\delta}_i} \qquad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^n \hat{\delta}_i (y_i - \hat{\mu}_2)^2}{\sum_{i=1}^n \hat{\delta}_i}$$

$$\hat{\pi} = \sum_{i=1}^{n} \hat{\delta}_i / n.$$

These new estimates are now used to update $\hat{\delta}_i$ and we continue until convergence. The estimation of δ_i is called the *Expectation* step as $E(\Delta_i) = \text{pr}(\Delta_i = 1)$. The update of the parameters is called the *Maximization* step.

(c) Apply the EM algorithm to the `geyser` data in the `MASS` library, and compare the results with those obtained by directly maximizing the log-likelihood.

(d) This algorithm can be applied in many similar contexts: the basic idea is that there is a likelihood that is easy to maximize, and we can define some latent variables whose values, if they were known, would enable us to calculate the easy likelihood. For more discussion see HTF Ch. 8 and Thisted "Elements of Statistical Computing" Ch. 4.7.

Table 1: Calcium uptake data for exercise 1

	time (minutes)	cal (moles/mg)
1	0.45	0.34170
2	0.45	-0.00438
3	0.45	0.82531
4	1.30	1.77967
5	1.30	0.95384
6	1.30	0.64080
7	2.40	1.75136
8	2.40	1.27497
9	2.40	1.17332
10	4.00	3.12273
11	4.00	2.60958
12	4.00	2.57429
13	6.10	3.17881
14	6.10	3.00782
15	6.10	2.67061
16	8.05	3.05959
17	8.05	3.94321
18	8.05	3.43726
19	11.15	4.80735
20	11.15	3.35583
21	11.15	2.78309
22	13.15	5.13825
23	13.15	4.70274
24	13.15	4.25702
25	15.00	3.60407
26	15.00	4.15029
27	15.00	3.42484