

Stat 215B (Spring 2005): Lab 4

GSI: Victor Panaretos
victor@stat.berkeley.edu

Due March 29, at Lab Section

Part I - Non-linear regression

In this part of the lab we consider a non-linear regression problem. In particular, a sample of blood is taken, labelled radioactively and re-injected. Then, further samples, all of the same size, are taken at specific times after the re-injection, and a radioactive count is taken of each. Such procedures are used to assess the relative lifetimes of blood cells treated differently.

Thus there are two variables:

- **time**: The time (in days) at which a sample was retaken - time 0 is the time of re-injection of the initial sample.
- **count**: The radioactive count for a sample.

We want to consider the nonlinear model

$$\frac{1}{1000} \mathbb{E}[\text{count}] = \alpha + \beta \exp(-\gamma t)$$

where $t \geq 0$ refers to **time**, and α, β, γ are unknown parameters. The parameter α is the background radiation level.

1. Plot the data and comment on their form.
2. Estimate the parameters using nonlinear regression. Try several sets of initial values. Does the program converge to the same place each time? Can you make it diverge?
3. Assume that the model errors are IID Normal. Use the estimated information matrix to estimate the variance/covariance matrix of the estimator. Why is this reasonable under the distributional assumption for the errors?
4. Is there sufficient evidence to suggest that background radiation is present (i.e. that $\alpha > 0$)? Give a 95% confidence interval for α .
5. Assess the fit of the model using the usual plots.

Part 2: Maximum Likelihood Mixture Model Estimation

This section of the lab considers the problem of maximum likelihood estimation in a case when the likelihood equations are non-linear, and no closed-form solution may be found.

In particular, we consider the “classic” data set of the eruption durations of the Old Faithful geyser¹ of Yellowstone National Park, Wyoming. The data consist of the the recordings of the duration (in minutes) of 299 eruptions of the geyser. Although the eruption durations occur serially in time, we shall assume them to be independent realisations from a common underlying distribution.

You are asked to perform the following analyses:

1. Construct a histogram for the data set, and also an estimate of the underlying density function. What do you observe?
2. Assume that the data originate from a mixture of two normal distributions with density:

$$f(x; p, \mu_1, \mu_1, \sigma) = \frac{p}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x - \mu_1)^2}{2\sigma^2}\right\} + \frac{1-p}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x - \mu_2)^2}{2\sigma^2}\right\}$$

Provide the maximum likelihood estimate for the parameter $\tilde{\theta} = (p, \mu_1, \mu_1, \sigma)^T$.

3. Provide an estimate for the variance-covariance matrix of your estimator.
4. Test the hypothesis $\mu_1 = \mu_2$ using a likelihood ratio test.

Notice that it has been assumed that the two components of the mixture distribution have the same variance.

Note: The data are contained in the files `lab4a.txt` and `lab4b.txt` on the section webpage.

Useful Commands: `nls()`, `deriv()`, `hist()`, `density()`, `nlm()`.

Some fun: you can watch the Old Faithful geyser at:

<http://www.nps.gov/yell/oldfaithfulcam.htm>

¹*type of hot spring that periodically erupts hot water and steam*