Lab Worksheet #1: MATH 495

Materials: This worksheet accompanies the m-files found on the web site for Day 3.

Discussion: In this lab, I will give you some tools for estimating errors in parameter estimates obtained by solving nonlinear least squares problems. In particular, suppose we have a statistical model of the form

$$\mathbf{y} = \mathbf{X}(\boldsymbol{\beta}) + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\mathbf{X}(\boldsymbol{\beta}) = (\mathbf{X}_1(\boldsymbol{\beta}), \ldots, \mathbf{X}_n(\boldsymbol{\beta}))^T$, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ with $\epsilon_i \sim N(0, \sigma^2)$ for all i. Note that (1) is equivalent to writing

$$y_i = X_i(\boldsymbol{\beta}) + \epsilon_i, \quad i = 1, \ldots, n.$$

Our goal in this worksheet will be to estimate, via sampling, the distribution for $\hat{\boldsymbol{\beta}}$, which is defined by

$$\hat{\beta} = \arg\min_{\boldsymbol{\beta}} \left\{ f(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{X}(\boldsymbol{\beta}) - \mathbf{y}\|^2 \right\}. \tag{2}$$

Linear Example: The first example is linear and so can be analyzed using the statistical methods that you have under your belt:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1, \\ y_2 &= \beta_0 + \beta_1 x_2. \end{aligned}$$

This corresponds in (1) to $\mathbf{X}(\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$ with

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \quad \text{and} \quad \mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \end{bmatrix},$$

so that (2) takes the form

$$\hat{\beta} = \arg\min_{\boldsymbol{\beta}} \left\{ f(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{2}(y_i - (\beta_0 - \beta_1 x_1)) \right\}. \tag{3}$$

As you have learned, in this case the $\hat{\boldsymbol{\beta}}$ can be written analytically as follows:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \sim N(\boldsymbol{\beta}_{\text{true}}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}).$$

This analytic form allows for the computation of confidence intervals for the elements of $\hat{\boldsymbol{\beta}}$.

Another approach, which does not require distributional assumptions about $\hat{\boldsymbol{\beta}}$, is to sample from $\hat{\boldsymbol{\beta}}$ and determine confidence intervals using MATLAB's `quantile` function. To see the similarity in the results of the two approaches in the linear case, at the MATLAB prompt type

```
>> LinearDemo
```

Note the output: All of the normality tests pass and the confidence intervals based the assumption of normality are very similar to those obtained using the nonparametric quantile approach.

Nonlinear Example: In the nonlinear example, we have

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix}, \quad \mathbf{X}(\boldsymbol{\beta}) = \begin{bmatrix} \beta_1 - \beta_1 e^{-t_1\beta_2} \\ \beta_1 - \beta_1 e^{-t_2\beta_2} \\ \beta_1 - \beta_1 e^{-t_3\beta_2} \\ \beta_1 - \beta_1 e^{-t_4\beta_2} \\ \beta_1 - \beta_1 e^{-t_5\beta_2} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

so that (2) takes the form

$$\hat{\beta} = \arg\min_{\boldsymbol{\beta}} \left\{ f(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{5}(y_i - (\beta_1 - \beta_1 e^{-\beta_2 t_i})^2 \right\}. \tag{4}$$

But this has no closed form solution as was the case in (3). Still, given a data vector $\mathbf{y}$ we can compute a parameter estimate $\hat{\boldsymbol{\beta}}$ and approximate confidence intervals using the variance approximation

$$Var(\hat{\boldsymbol{\beta}}) \approx \hat{\sigma}^2(\nabla^2 f(\hat{\boldsymbol{\beta}}))^{-1} \approx \hat{\sigma}^2(J(\hat{\boldsymbol{\beta}})^T J(\hat{\boldsymbol{\beta}}))^{-1},$$

where "∇^2" denotes the Hessian (second derivative) operator, and $J(\boldsymbol{\beta})$ is the Jacobian of the residual vector $\mathbf{y} - \mathbf{X}(\hat{\boldsymbol{\beta}})$.

Note: `nlinfit` computes $\hat{\boldsymbol{\beta}}$ and outputs $\mathbf{r}(\hat{\boldsymbol{\beta}}) = \mathbf{y} - \mathbf{X}(\hat{\boldsymbol{\beta}})$, numerical approximations $J(\hat{\boldsymbol{\beta}})$ and $\hat{\sigma}^2\nabla^2 f(\hat{\boldsymbol{\beta}})^{-1}$, and the mean squared error.

However, just as above, we can also sample the probability density of $\hat{\boldsymbol{\beta}}$ by taking realizations of the random variable $\boldsymbol{\epsilon}$ in (1) and then computing the corresponding estimate $\hat{\boldsymbol{\beta}}$ defined in (4). You can do this by typing, at the MATLAB prompt

```
>> NonlinDemo_MonteCarlo
```

Note the output: the normality test no longer passes for both parameters and the confidence intervals based the assumption of normality and the nonparametric quantile approach give noticeably different results.

This Problem Will Appear on Your Next Homework: Modify the code `NonlinDemo_MonteCarlo.m` so that the AIDS data on the web site is used and the nonlinear model is the logistic equation

$$X(P_0, M, k, t) = \frac{M P_0 e^{kt}}{M + P_0(e^{kt} - 1)}.$$

The data and model vectors are then given by

$$\mathbf{y} = \begin{bmatrix} 260 \\ 992 \\ 2717 \\ 5341 \\ 8224 \\ 13195 \\ 21355 \\ 32196 \\ 35230 \\ 43352 \\ 45524 \\ 47572 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} P_0 \\ M \\ k \end{pmatrix}, \quad \mathbf{X}(\boldsymbol{\beta}) = \begin{bmatrix} X(P_0, M, k, 0) \\ X(P_0, M, k, 1) \\ X(P_0, M, k, 2) \\ X(P_0, M, k, 3) \\ X(P_0, M, k, 4) \\ X(P_0, M, k, 5) \\ X(P_0, M, k, 6) \\ X(P_0, M, k, 7) \\ X(P_0, M, k, 8) \\ X(P_0, M, k, 9) \\ X(P_0, M, k, 10) \\ X(P_0, M, k, 11) \end{bmatrix}$$