Lecture Notes 7: Non-Linear Regression

In the previous lecture note I have discussed about different non-linear regression models, and talked about the identifiability issues of the different parameters. It is very clear if the parameters are not identifiable, then even if some estimators, may be the least squares estimators or maximum likelihood estimator, but they need not be unique. Hence, it is important that when we try to find the least squares estimators or the maximum likelihood estimator, the parameters are identifiable. In this lecture note we will discuss two different numerical methods to find the least squares estimators or the maximum likelihood estimators, when they are identifiable. So it is assumed that the parameters are identifiable, and later we will discuss how to identify the parameters in a non-linear regression model.

We consider the general non-linear regression model, namely

$$y_i = f(\mathbf{x}_i, \boldsymbol{\theta}^*) + \epsilon_i; \qquad i = 1, \ldots, n.$$

Here $f(\cdot)$ is a non-linear function, the parameter vector $\boldsymbol{\theta}$ is identifiable, i.e. if $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$, then there exists at least one $\mathbf{x}_i$ such that

$$f(\mathbf{x}_i, \boldsymbol{\theta}_1) \neq f(\mathbf{x}_i, \boldsymbol{\theta}_2).$$

Further, the error random variables ϵ_i's are assumed to be independent and identically distributed random variables, with mean zero, and variance σ^2. Therefore, the least squares estimators of the unknown parameters $\boldsymbol{\theta}^*$ can be obtained by minimizing the residual sums of squares, i.e.

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(y_i - f(\mathbf{x}_i, \boldsymbol{\theta})\right)^2,$$

with respect to the unknown parameter vector $\boldsymbol{\theta}$. Here it is assumed that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top \in \mathbb{R}^p$. It is assumed that $Q(\boldsymbol{\theta})$ is a nice differentiable function. Hence,

$$\widehat{\boldsymbol{\theta}} = \arg\ \min Q(\boldsymbol{\theta}).$$

can be obtained as a solution of the equations

$$\frac{\partial}{\partial \theta_j} Q(\boldsymbol{\theta}) = 0; \qquad j = 1, \ldots, p. \tag{1}$$

Since (1) is a set of non-linear equations, the solutions cannot be obtained in explicit forms. One needs some iterative technique to solve these equations.

Now I will discuss two different techniques to solve these non-linear equations. It may be mentioned that to solve (1) one needs to use some iterative technique, and for any iterative technique one needs some initial guess to start the process. It is always a challenging issue to choose the initial guesses, and there is no fixed rule/ prescription for that. It depends on the problem, and also on the dimension, i.e. p in this case, of the problem. It is a different issue, and we will discuss about it in some specific cases with examples. At this moment we are assuming that, we have an initial guess of $\widehat{\boldsymbol{\theta}}$, say $\boldsymbol{\theta}^{(0)}$.

Now under this assumption we will describe the first method. It is also known as the Gauss-Newton method. In this case we make first order Taylor series approximation of the non-linear function $f(\boldsymbol{\theta}) = (f(\mathbf{x}_1, \boldsymbol{\theta}), \ldots, f(\mathbf{x}_n, \boldsymbol{\theta}))^\top$ around a point $\boldsymbol{\theta}^{(0)}$ as follows

$$f(\boldsymbol{\theta}) \approx f(\boldsymbol{\theta}^{(0)}) + \mathbf{F}_{\bullet}(\boldsymbol{\theta}^{(0)})(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}).$$

Here $\mathbf{F}_{\bullet}(\boldsymbol{\theta})$ is a $n \times p$ derivative matrix as follows

$$\mathbf{F}_{\bullet}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} f(\mathbf{x}_1, \boldsymbol{\theta}) & \ldots & \frac{\partial}{\partial \theta_1} f(\mathbf{x}_1, \boldsymbol{\theta}) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial \theta_1} f(\mathbf{x}_n, \boldsymbol{\theta}) & \ldots & \frac{\partial}{\partial \theta_1} f(\mathbf{x}_n, \boldsymbol{\theta}) \end{bmatrix},$$

and $\boldsymbol{\theta}$ is a $p \times 1$ vector. Therefore, for $\mathbf{Y} = (y_1, \ldots, y_n)^\top$, and $r(\boldsymbol{\theta}^{(0)}) = \mathbf{Y} - f(\boldsymbol{\theta}^{(0)})$, we have the following approximation of the residual sums of square $Q(\boldsymbol{\theta})$,

$$\begin{aligned} Q(\boldsymbol{\theta}) &= \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i, \boldsymbol{\theta}))^2 \\ &= (\mathbf{Y} - f(\boldsymbol{\theta}))^\top (\mathbf{Y} - f(\boldsymbol{\theta})) \\ &\approx \Big(\mathbf{Y} - f(\boldsymbol{\theta}^{(0)}) - \mathbf{F}_{\bullet}(\boldsymbol{\theta}^{(0)})(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)})\Big)^\top \Big(\mathbf{Y} - f(\boldsymbol{\theta}^{(0)}) - \mathbf{F}_{\bullet}(\boldsymbol{\theta}^{(0)})(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)})\Big) \\ &= \Big(r(\boldsymbol{\theta}^{(0)}) - \mathbf{F}_{\bullet}(\boldsymbol{\theta}^{(0)})(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)})\Big)^\top \Big(r(\boldsymbol{\theta}^{(0)}) - \mathbf{F}_{\bullet}(\boldsymbol{\theta}^{(0)})(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)})\Big) \\ &= r^\top(\boldsymbol{\theta}^{(0)}) r(\boldsymbol{\theta}^{(0)}) - 2(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)})^\top \mathbf{F}_{\bullet}^\top(\boldsymbol{\theta}^{(0)}) r(\boldsymbol{\theta}^{(0)}) \\ &\quad + (\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)})^\top \mathbf{F}_{\bullet}^\top(\boldsymbol{\theta}^{(0)}) \mathbf{F}_{\bullet}(\boldsymbol{\theta}^{(0)})(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}) \end{aligned}$$

The right hand side of $Q(\boldsymbol{\theta})$ can be minimized to obtain the next iterate say $\boldsymbol{\theta}^{(1)}$. If we differentiate the right hand side of $Q(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, and put it equal to zero, we obtain the following equation

$$-2\mathbf{F}_{\bullet}^{\top}(\boldsymbol{\theta}^{(0)})r(\boldsymbol{\theta}^{(0)}) + 2\mathbf{F}_{\bullet}^{\top}(\boldsymbol{\theta}^{(0)})\mathbf{F}_{\bullet}(\boldsymbol{\theta}^{(0)})(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}) = 0. \tag{2}$$

Hence, $\boldsymbol{\theta}^{(1)}$ can be obtained as

$$\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)} + (\mathbf{F}_{\bullet}^{\top}(\boldsymbol{\theta}^{(0)})\mathbf{F}_{\bullet}(\boldsymbol{\theta}^{(0)}))^{-1}\mathbf{F}_{\bullet}^{\top}(\boldsymbol{\theta}^{(0)})r(\boldsymbol{\theta}^{(0)}),$$

assuming that $(\mathbf{F}_{\bullet}^{\top}(\boldsymbol{\theta}^{(0)})\mathbf{F}_{\bullet}(\boldsymbol{\theta}^{(0)}))^{-1}$ exists. Therefore, in general from the k-th step, the $(k+1) - th$ step can be obtained as follows:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + (\mathbf{F}_{\bullet}^{\top}(\boldsymbol{\theta}^{(k)})\mathbf{F}_{\bullet}(\boldsymbol{\theta}^{(k)}))^{-1}\mathbf{F}_{\bullet}^{\top}(\boldsymbol{\theta}^{(k)})r(\boldsymbol{\theta}^{(k)}).$$

Here $r(\boldsymbol{\theta}^{(k)}) = \mathbf{Y} - f(\boldsymbol{\theta}^{(k)})$. Let us denote $\boldsymbol{\delta}^{(k)} = (\mathbf{F}_{\bullet}^{\top}(\boldsymbol{\theta}^{(k)})\mathbf{F}_{\bullet}(\boldsymbol{\theta}^{(k)}))^{-1}\mathbf{F}_{\bullet}^{\top}(\boldsymbol{\theta}^{(k)})r(\boldsymbol{\theta}^{(k)})$. Hence $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \boldsymbol{\delta}^{(k)}$. The iterative process continues, until some convergence criterion has been made. Then it is stopped.

Alternatively, we can expand the function $Q(\boldsymbol{\theta}$ also directly, and proceed. In this case we make 2nd-order Taylor series approximation of $Q(\boldsymbol{\theta})$ around $\boldsymbol{\theta}^{(0)}$ as follows

$$Q(\boldsymbol{\theta}) \approx Q(\boldsymbol{\theta}^{(0)}) + (\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)})^{\top}Q'(\boldsymbol{\theta}^{(0)}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)})^{\top}Q''(\boldsymbol{\theta}^{(0)})(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}). \tag{3}$$

Here $Q'(\boldsymbol{\theta})$ is the $p \times 1$ gradient vector and $Q''(\boldsymbol{\theta})$ is the $p \times p$ Hessian matrix defined as follows:

$$Q'(\boldsymbol{\theta}) = \left(\frac{\partial}{\partial \theta_1}Q(\boldsymbol{\theta}), \ldots, \frac{\partial}{\partial \theta_p}Q(\boldsymbol{\theta})\right)^{\top},$$

and

$$Q''(\boldsymbol{\theta}) = \left[\begin{array}{ccc} \frac{\partial^2}{\partial \theta_1^2}Q(\boldsymbol{\theta}) & \cdots & \frac{\partial^2}{\partial \theta_1 \partial \theta_p}Q(\boldsymbol{\theta}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial \theta_p \partial \theta_1}Q(\boldsymbol{\theta}) & \cdots & \frac{\partial^2}{\partial \theta_p^2}Q(\boldsymbol{\theta}) \end{array}\right].$$

Therefore, following the same manner, if we want to minimize the right hand side of (3), then we obtain the following equation

$$Q'(\boldsymbol{\theta}^{(0)}) + Q''(\boldsymbol{\theta}^{(0)})(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}) = \mathbf{0}.$$

Hence,

$$\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)} - \left(Q''(\boldsymbol{\theta}^{(0)})\right)^{-1} Q'(\boldsymbol{\theta}^{(0)}).$$

Since,

$$Q'(\boldsymbol{\theta}^{(0)}) = -2(\mathbf{F}_{\bullet}^{\top}(\boldsymbol{\theta}^{(0)})r(\boldsymbol{\theta}^{(0)}),$$

hence,

$$\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)} + 2\left(Q''(\boldsymbol{\theta}^{(0)})\right)^{-1} \mathbf{F}_{\bullet}^{\top}(\boldsymbol{\theta}^{(0)})r(\boldsymbol{\theta}^{(0)}).$$

Hence, in general

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + 2\left(Q''(\boldsymbol{\theta}^{(k)})\right)^{-1} \mathbf{F}_{\bullet}^{\top}(\boldsymbol{\theta}^{(k)})r(\boldsymbol{\theta}^{(k)}).$$

In this case also the iterative process continues, until some convergence criterion has been made. Then it is stopped.

Problem

1. If $f(x;\theta) = e^{\theta}$, find the iterative process of both the methods.

2. If $f(x;\theta) = \sin(x\theta)$, find the iterative process of both the methods.