Overview and Motivation: Introduction to Nonlinear Regression Models

Evgenia Ageeva

Seminar in Statistics :

Mixed-effects Models

26.04.2010, ETH Zurich

Linear regression is a powerful method for analyzing data described by models which are linear in parameters. Often, however, a researcher has a mathematical expression which relates the response to the predictor variables, and these models are usually nonlinear in the parameters. In such cases, linear regression techniques must be extended, which introduces considerable complexity.

1 Nonlinear Model: Setting

The general normal **nonlinear regression model** can be written

$$\boxed{Y_i = f(\boldsymbol{x_i}, \boldsymbol{\theta}) + \boldsymbol{\epsilon_i}}$$

where f is the **expectation function**, $\boldsymbol{x_i}$ is a vector of associated regressor variables for the ith case, and $\boldsymbol{\epsilon_i}$ the error terms, $\overset{iid}{\sim} N(0, \sigma^2)$.

That is, for nonlinear models, at least one of the derivatives of the expectation function with respect to the parameters depends on at least one of the parameters.

Consider the vectors $\boldsymbol{x}_i$, $i = 1, 2, \ldots, N$ as fixed (observations) and concentrate on the dependance of the expected responses on $\boldsymbol{\theta}$.

We create the N-vector $\boldsymbol{\eta}(\boldsymbol{\theta})$ with ith element

$$\eta_i(\boldsymbol{\theta}) = f(\boldsymbol{x}_i, \boldsymbol{\theta}) \quad i = 1, 2, \ldots, N$$

and write the nonlinear regression model as

$$\boxed{\boldsymbol{Y} = \boldsymbol{\eta}(\boldsymbol{\theta}) + \boldsymbol{\epsilon}}$$

1.1 Transformably Linear Models

Many models can be transformed into a linear model. For example,

$$f(x, \boldsymbol{\theta}) = \theta_1 \ x^{\theta_2}$$

can be tranformed into a linear model:

$$\ln\left(f(x, \boldsymbol{\theta})\right) = \ln \theta_1 + \theta_2 \ln x = \beta_0 + \beta_1 \tilde{x},$$

where $\beta_0 = \ln \theta_1$ and $\tilde{x} = \ln x$.

We call such regression models **transformably linear**. Transformation of the model implies transformation of the disturbance terms, which undesirable.

Examle: If we start with

$$\ln Y_i = \beta_0 + \beta_1 \tilde{x} + \epsilon_i,$$

where ϵ_i additiv, normally distributed. Transform it back, obtain:

$$Y_i = \theta_1 \cdot x^{\theta_2} \cdot \tilde{\epsilon}_i,$$

where $\tilde{\epsilon}_i$ perform as multiplicative and lognormally distributed! Therefore: trasformation of the data is to enjoy with caution.

2 Determining the Least Squares Estimates: The Gauss-Newton Method

2.1 The Geometry of the Expectation Surface

The N-vectors $\boldsymbol{\eta}(\boldsymbol{\theta})$ define a P-dimensional surface called the **expectation surface** in the response space.

The least squares estimates correspond to the point on the expectation surface,

$$\hat{\boldsymbol{\eta}} = \boldsymbol{\eta}(\hat{\boldsymbol{\theta}})$$

which is closest to $\mathbf{y}$. That is, $\hat{\boldsymbol{\theta}}$ minimizes the residual sum of squares

$$S(\boldsymbol{\theta}) = \|\mathbf{y} - \boldsymbol{\eta}(\boldsymbol{\theta})\|^2$$

where $\boldsymbol{\eta}(\boldsymbol{\theta}) = f(\mathbf{x}, \boldsymbol{\beta})$.

- find the point $\hat{\boldsymbol{\eta}}$ on the expectation surface which is closest to $\boldsymbol{y}$, and then

- determine the parameter vector $\hat{\boldsymbol{\theta}}$ which corresponds to the point $\hat{\boldsymbol{\eta}}$.

In the nonlinear case, however, the two steps are very difficult, so we use **iterative methods** to determine the least squares estimates $\hat{\boldsymbol{\theta}}$.

Goal: to iteratively improve an initial guess $\boldsymbol{\theta}^{(0)}$ for $\boldsymbol{\theta}$ and keep improving the estimates until there is no change.

(1) Expand the expectation function $f(\boldsymbol{x_i}, \boldsymbol{\theta})$ in a first order Taylor series about $\boldsymbol{\theta}^{(0)}$ as

$$f(\boldsymbol{x_i}, \boldsymbol{\theta}) \approx f(\boldsymbol{x_i}, \boldsymbol{\theta^{(0)}}) + A_{i1}\langle\theta^{(0)}\rangle(\theta_1 - \theta_1^{(0)}) + \ldots$$
$$\cdots + A_{iP}\langle\boldsymbol{\theta}^{(0)}\rangle(\theta_P - \theta_P^{(0)})$$

where

$$A_{ip}\langle\boldsymbol{\theta}\rangle = \frac{\partial \eta_i\langle\boldsymbol{\theta}\rangle}{\partial \theta_p} \mid_{\boldsymbol{\theta}^{(0)}}, \quad p = 1, 2, \ldots, P$$

Incorporating all N cases, write

$$\boldsymbol{\eta}(\boldsymbol{\theta}) \approx \boldsymbol{\eta}(\boldsymbol{\theta^{(0)}}) + \boldsymbol{A}\langle\theta^{(0)}\rangle(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)})$$

where $\boldsymbol{A}\langle\boldsymbol{\theta}^*\rangle$ is the $N \times P$ derivative matrix with elements A_{ip}.

(2) Adding the error terms:

$$\boldsymbol{Y} - \boldsymbol{\eta}(\boldsymbol{\theta^{(0)}}) = \boldsymbol{A}\langle\theta^{(0)}\rangle \cdot \underbrace{(\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)})}_{:=\boldsymbol{\beta}} + \boldsymbol{E}$$

This is equivalent to the linear regression model.

(3) On this stage do linear regression and obtain $\boldsymbol{\beta}$.

(4) Put $\boldsymbol{\theta}^{(1)} := \boldsymbol{\theta}^{(0)} + \boldsymbol{\beta}$;

(5) The point $\boldsymbol{\eta}(\boldsymbol{\theta}^{(1)})$ should now be closer to $\boldsymbol{y}$ and so we move to this better parameter value $\boldsymbol{\theta}^{(1)}$ and perform another iteration.

(6) Iterate this process until there is no useful change in the elements of the parameter vector.

2.2 Initial Values

Iterative methods need **initial values** to start from. How can we find them?

- From the previous knowledge of the experiment design
- Via the transfomed linear model
- Analysis of the expectation function f

2.3 nls-Function

In R: use function **nls** to obtain nonlinear least squares estimates:

```
nls(formula, data = parent.frame(), start, control = nls.control(),
    algorithm = "default", trace = FALSE, subset,
    weights, na.action, model = FALSE)
```
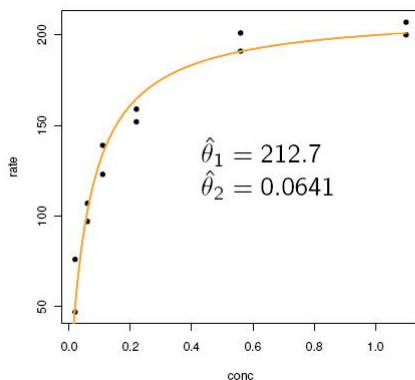

The fitted expectation function for the Puromycin data.

3 Tests and Inference Regions

3.1 Approximative Tests and Confidence Regions

Suppose, we have estimated $\hat{\boldsymbol{\theta}}$ with the LS method. Question: is this estimate consistent with the observations? Confidence regions or intervals contains all these values $\boldsymbol{\theta}$.

Main assumption: The LS estimate $\hat{\boldsymbol{\theta}}$ is asymptotically normal distributed:

$$\hat{\boldsymbol{\theta}} \overset{a}{\sim} N\left(\boldsymbol{\theta}, \frac{\boldsymbol{V}_\theta}{n}\right),$$

with asymptotic covariance matrix

$$\boldsymbol{V}_\theta = \sigma^2 \cdot \left(\boldsymbol{A}_\theta^T \cdot \boldsymbol{A}_\theta\right)^{-1}$$

Substitute $\boldsymbol{V}_\theta$ by its estimate:

$$\hat{\boldsymbol{V}}_\theta = \hat{\sigma}^2 \cdot \left(\boldsymbol{A}_{\hat{\theta}}^T \cdot \boldsymbol{A}_{\hat{\theta}}\right)^{-1} \text{ where } \hat{\sigma}^2 = \frac{1}{n-p}S(\hat{\boldsymbol{\theta}})$$

$$H_{0,j}: \quad \theta_j = 0 \text{ \textbf{against} } H_{A,j}: \quad \theta_j \neq 0$$

$$T_j = \frac{\hat{\theta}_j}{\sqrt{\hat{\sigma}^2(A^TA)_{jj}^{-1}}} \overset{H_{0,j}}{\sim} t_{n-p}$$

Confidence region:

$$\hat{\theta}_j \pm \sqrt{\hat{\sigma}^2(A^TA)_{jj}^{-1}} \cdot q_{1-\alpha/2}^{t_{n-p}}$$

3.2 Exact Tests and Confidence Regions

The quality of the aprroximative test depends strongly on the quality of the linear aprixidation of the expectation surface. We can avoid this issue and construct tests which do not use linearisation components.

To test the null-hypothesis:

$$H_0: \ \boldsymbol{\theta} = \boldsymbol{\theta}^*$$

we can compare the residual sums of squares $S(\boldsymbol{\theta}^*)$ with $S(\boldsymbol{\theta})$.

The test statistic is

$$T = \frac{n-p}{p} \cdot \frac{S(\boldsymbol{\theta}^*) - S(\hat{\boldsymbol{\theta}})}{S(\hat{\boldsymbol{\theta}})} \quad \overset{a}{\sim} \quad F_{p,n-p}.$$

This gives us the confidence region

$$\left\{\boldsymbol{\theta} \mid S(\boldsymbol{\theta}) \leq S(\hat{\boldsymbol{\theta}})\left(1 + \frac{p}{n-p} \cdot q\right)\right\}$$

where $q = q_{1-\alpha}^{F_{p,n-p}}$ is the $(1-\alpha)$-quantile of the $F(p, n-p)$ distribution.

This test is **not** based on the issue of the linearization => it is more precise than the approximative test!

For a single parameter:

$$H_0: \ \theta_j = \theta_j^*$$

$$\tilde{T}_k = (n-p) \cdot \frac{\tilde{S}_k(\boldsymbol{\theta}_k^*) - \tilde{S}(\hat{\boldsymbol{\theta}})}{S(\hat{\boldsymbol{\theta}})}$$

which is approx. $F_{1,n-p}$ - distributed.

We obtain the confidence interval by solving the equation $\tilde{T}_k = q_{1-\alpha/2}^{F_{1,n-p}}$ numerically for θ_k^*.

3.3   t-Test via F-Test

In the case of nonlinear regression the F-test is not equivalent to the f-test. But we can transform the the F-test into the t-test.

$$T_k(\theta_k^*) = \text{sign}\left(\hat{\theta}_k - \theta_k^*\right) \cdot (n-p)\frac{\sqrt{\tilde{S}_k(\theta_k^*) - S(\hat{\theta})}}{S(\hat{\theta})}$$

This statistic is approx. t_{n-p} - distributed.

3.4   Profile-t-Plot and Profile Traces

We introduce here some graphical techniques for assessing the severity of the nonlinearity in an estimation situation.

Look at the test statistic T_k as a function of θ_k, and call it **profile-t-function**:

$$T_k(\theta_k) = \text{sign}\left(\hat{\theta}_k - \theta_k\right) \cdot (n-p) \cdot \frac{\sqrt{\tilde{S}_k(\theta_k) - S(\hat{\theta})}}{S(\hat{\theta})}$$

The **profile-t-plot** provides exact likelihood intervals for individual parameters and, in addition, reveals how nonlinear the estimation sitiation is.
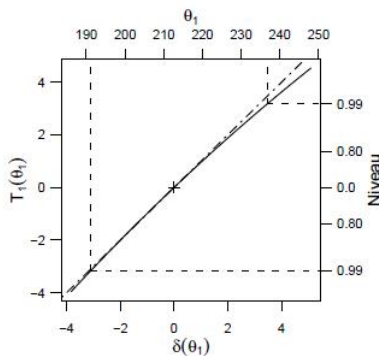
Usually we plot on the x-axis the standardised version of θ_k: $\delta_k(\theta_k) := \frac{\hat{\theta}_k - \theta_k^*}{se\langle(\hat{\theta}_k)\rangle}$ instead of θ_k.

Another tool is plotting the **likelihood profile traces**.

Here, we look at the estimated parameters $\hat{\theta}_j, j \neq k$, as a fuction $\hat{\theta}_j^{(k)}(\theta_k)$, while θ_k remains fixed. In addition, we may plot the likelihood contours:
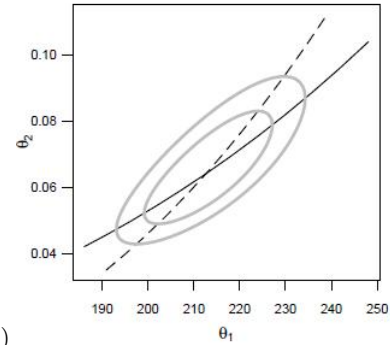
(Puromycin) Interpretation: traces are quite straight $\rightarrow$ linear approximation is good. Simular to the linear case: angle of intersection provides information about the correlation between the parameters.