Non-Linear Models - Further Statistical Methods

Sarah Filippi
University of Oxford

Hilary Term 2015

1With grateful acknowledgements to Professor Ripley for his notes from 2012–13.

Outline

In one sense, this is a very large topic: everything that is not linear models. But it is usually taken to mean some specific sets of departures from linear models—and not always the same sets, depending on the author.

1 Non-linear least square and some specific non-linear models
2 Basis expansion and regularisation
3 Kernel smoothing methods
4 Additive and general additive linear models

References

A thorough treatment of non-linear regression is given in:

• Bates and Watts (1988). *Nonlinear Regression Analysis and Its Applications*
• Seber and Wild (1989). *Nonlinear Regression*
• Gallant (1987). *Nonlinear Statistical Models*
• Ratkowsky (1983). *Nonlinear Regression Modeling: A Unified Practical Approach*
• Ross (1990). *Nonlinear Estimation*
• Venables and Ripley (2002). *Modern Applied Statistics with S*

In the preparation of this course, I have also used

• Clark (2013). *Generalized additive models*
Linear models

There are several specific aspects to linear models as used in regression:

(a) There is a linear predictor which is a linear function of the predictors $x_1, \ldots, x_p$, and perhaps one of these predictors is constant. So conventionally we write

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Note that this is linear in the parameters: the predictors might be non-linear functions (such as polynomials or splines) of the independent variables.

(b) The linear predictor models the mean of the distribution of the dependent variable.

(c) Fitting is by least squares, which is maximum likelihood estimation (for $\beta$) for iid Normal errors.

You have seen linear predictors used in other contexts:
- logistic regression and log-linear models,
- accelerated-life and proportional hazard survival models,
- hierarchical models.

These differ in points (b) and (c): the linear predictor models some other aspect of the distribution (the log-mean, probabilities on a logistic scale \ldots), and fitting is by maximum likelihood or something similar (maximum partial likelihood, REML).

A lot of the special properties of linear models come from the combination of a linear predictor and fitting by least-squares.

The log-likelihood is concave (indeed quadratic) in the parameters $\beta$. This has several consequences:

- There is a unique local maximum to the likelihood, which is the global maximum.
- The least-squares/maximum likelihood solution can be found analytically.
- The asymptotic theory for MLEs is exact, since the log-likelihood is exactly quadratic.

Note that if we need to estimate other parameters (e.g. the error variance $\sigma^2$) the inference task may not be quite as simple.

This section of the course is about what may happen if we relax some of these assumptions:

- fitting by something other than least squares
- or predicting by a non-linear function of the parameters.
In general, numerical optimization methods have to be used to estimate the model parameters. These have some common characteristics:

- They find a local maximum/minimum of the fit criterion. There may well be multiple local optima.
- They need a starting point, and that often needs to be a set of parameters for which the model fits quite well, to ensure that an appropriate local optimum is found.
- The only distribution theory for the parameter estimates is asymptotic, but simulation-based methods can be used.
- They usually need complete identifiability.

### 1. Non-linear least-square and some specific non-linear models

Subject-specific theory may suggest that the response varies non-linearly with the independent variables. Perhaps the simplest example is the exponential growth or decay with time $t$:

$$E[Y] = a \exp \pm b t$$

One could argue that this has a linear predictor $\log(\log(a) \pm b t)$, but that does not predict the mean (it might predict the log-mean of a lognormal distribution).

Note that the parameter $b$ has an interpretation as a rate of growth/decay, and it is typical of specific non-linear models that they are specified with some interpretable parameters.

### Non linear regression

The general form of a nonlinear regression model is

$$y = E(Y | X = x) + \epsilon = m(x, \theta) + \epsilon$$

where

- $m(x, \theta)$ is called the *kernel mean function*
- an independent error term $\epsilon$ with an unknown variance $\sigma^2$ to be estimated.
Parameter estimation for non linear regression

Given a set of observation \( \{x_i, y_i\}_{i=1}^{N} \), the parameter \( \theta \) can be estimated by minimising the residual sum of squares

\[
S(\theta) = \sum_{i=1}^{N} (y_i - m(x_i, \theta))^2
\]

• Contrary to the linear least-square problem, there is no general formula of the solution.
• Use an iterative procedure starting with an initial guess for the parameters.
• The R function \texttt{nls} enables to estimate the parameters of such models: it requires the model definition, the dataset and initial values for all the model parameters.

Data points on weight loss with a fitted exponential decay:

Example 1: Weight loss

Data frame \texttt{wtloss} in package \texttt{MASS} gives the weight, in kilograms, of an obese patient at 52 time points over an 8 month period of a weight rehabilitation programme.

An exponential decay to a non-zero asymptote looks a suitable model

\[
\text{weight} = b_0 + b_1 2^{-\text{Days}/\text{th}} + \epsilon
\]

and the curve was fitted by

\[
\text{nls}(\text{Weight} ~ b_0 + b_1 2^{-\text{Days}/\text{th}}, \text{data = wtloss, start = list(b0 = 90, b1 = 95, th = 120)})
\]

The starting values were chosen based on their interpretation: the initial weight is \( b_0 + b_1 \), the target weight is \( b_0 \) and the ‘half-life’ is \( \text{th} \).

The model summary is

Formula: Weight ~ b0 + b1 * 2^(-Days/th)

Parameters:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| b0       | 81.374     | 2.269   | 35.86    | <2e-16 |
| b1       | 102.684    | 2.083   | 49.30    | <2e-16 |
| th       | 141.911    | 5.295   | 26.80    | <2e-16 |

Residual standard error: 0.8949 on 49 degrees of freedom

Number of iterations to convergence: 3
Achieved convergence tolerance: 4.39e-06
Self-starting non-linear regression

For some functions including exponential decays, starting values for non-linear regressions can be calculated by an automatic procedure.

Example: a logistic growth model

\[ y = \frac{\theta_1}{1 + \exp[-(\theta_2 + \theta_3 x)]} + \epsilon \]

A procedure to obtain initial parameter values from a set of observations \( \{x_i, y_i\}_{i=1}^N \) is:

1. Rewrite the model

\[ \log \left[ \frac{y / \theta_1}{1 - y / \theta_1} \right] \approx \theta_2 + \theta_3 x \]

2. Since \( \theta_1 \) is the asymptote, a starting value for \( \theta_1 \) is some value larger than any value in the data

3. Starting values for \( \theta_2 \) and \( \theta_3 \) can be estimated by fitting a linear regression using the functions \texttt{lm} and \texttt{logit}

Profile likelihood to assess standard error estimates

- The usual standard errors produced by nonlinear routines is based on the t-statistics (Wald statistics).
- For their validity, they depend on the objective function being quadratic around the solution.
- To check this, we can use profile methods.
- If the quadratic approximation is valid, the plot of

\[ \tau(\alpha) = \text{sign}(\alpha - \hat{\alpha}) \sqrt{S(\alpha) - S(\hat{\alpha})} \]

against \( \alpha \) is a straight line; where \( s \) is the standard error of the estimate.
- A confidence interval for \( \alpha \) is \( \{\alpha, -t \leq \tau(\alpha) < t\} \) where \( t \) is an appropriate percentage point from the \( t_{n-p} \) distribution.

Non-linear models are often associated with families of solutions to simple differential equations: a prominent field of application is the pharmaceutical industry, to PK (pharmacokinetics) and PD (pharmacodynamics) models.

The R package \texttt{stats} contains several specific non-linear models for which self-starting (SS) procedures are supplied:

- functions \texttt{SSasymp}, \texttt{SSasympOff} and \texttt{SSasympOrig} fit various forms of exponential decay
- \texttt{SSbiexp} fits the sum of two exponentials
- \texttt{SSweibull} fits a Weibull curve of the form \( a - b \exp(c x^\rho) \) (exponential growth/decay in the \( \rho \)th power of the explanatory variable)

We can use an SS model for the Weight loss example by

\[
\text{nls(Weight} ~ \text{SSasymp(Days, target, start, lograte), data = wtloss)}
\]
Example 2: Stormer viscosity data

The Stormer viscometer measures the viscosity of a fluid by measuring the time taken for an inner cylinder in the mechanism to perform a fixed number of revolutions in response to an actuating weight. The viscometer is calibrated by measuring the time taken with varying weights while the mechanism is suspended in fluids of accurately known viscosity. The dataset `stormer` in package `MASS` comes from such a calibration, and theoretical considerations suggest a non-linear relationship between time \( T \), weight \( w \) and viscosity \( v \) of the form

\[
T = \frac{\beta_1 v}{w - \beta_2} + \epsilon
\]

where \( \beta_1 \) and \( \beta_2 \) are unknown parameters to be estimated. This model is partially linear: given \( \beta_2 \) we can fit \( \beta_1 \) by linear regression.

Williams (1959) suggested finding initial values of the parameters by fitting the rewritten model

\[
wT = \beta_1 v + \beta_2 T + (w - \beta_2)\epsilon
\]

by least squares.

Doing so and then fitting the non-linear model gives

```r
> fm0 <- lm(Wt*Time ~ Viscosity + Time - 1, data = stormer)
> b0 <- coef(fm0); names(b0) <- c("b1", "b2"); b0
  b1  b2
28.876 2.8437
>
> storm.fm <- nls(Time ~ b1*Viscosity/(Wt-b2), data = stormer, start = b0)
> storm.fm

b1  b2
29.401 2.218

residual sum-of-squares: 825.1
```

Confidence region for the regression parameters

Denote by \((\hat{\beta}_1, \hat{\beta}_2)\) the parameters minimising the residual sum of squares

\[
S(\beta_1, \beta_2) = \sum_{i=1}^{N} \left( T_i - \frac{\beta_1 v_i}{w_i - \beta_2} \right)^2
\]

If \( \beta_1, \beta_2 \) are true parameters, the extra sum of squares

\[
F(\beta_1, \beta_2) = \frac{(S(\beta_1, \beta_2) - S(\hat{\beta}_1, \hat{\beta}_2))/2}{S(\hat{\beta}_1, \hat{\beta}_2)/(N - 2)}
\]

is approximately distributed as \( F_{p, N-p} \) where \( p = 2 \).
Example 3: Enzyme kinetics

The classic Michaelis–Menten model for enzyme kinetics has two parameters $K_m, V_{max}$ and relates the reaction rate $v$ to the concentration of substrate $S$ by

$$v = \frac{V_{max} S}{K_m + S}$$

Variations with a competitive inhibition

$$v = \frac{V_{max} S}{K_m(1 + I/K_i) + S}$$

and non-competitive inhibition

$$v = \frac{V_{max} S}{(S + K_m)(1 + I/K_i)}$$

where $I$ is the concentration of an inhibitor.

These come from the solutions of differential equations. We consider two datasets from CRAN package `nlstools`.

First from an experiment without an inhibitor:

```r
data(vmkm, package = "nlstools") # 22 obs
plot(vmkm, xlim=c(0, 2), ylim=c(0, 0.7)) # to include origin
fm <- nls(v ~ SSmicmen(S, Vm, Km), data = vmkm)
summary(fm)
Parameters:
   Estimate Std. Error t value Pr(>|t|)
Vm  1.5460    0.1454 10.635 1.11e-09
Km  2.5861    0.3704  6.981 8.94e-07
Residual standard error: 0.02572 on 20 degrees of freedom

Spred <- seq(0, 2, length = 100)
lines(Spred, predict(fm, data.frame(S=Spred)))
```

The second example has 72 observations with an inhibitor:

```r
data(vmkmki, package = "nlstools")
summary(vmkmki)
plot(v ~ S, data = vmkmki)
coplot(v ~ S | I, data = vmkmki)
```
We start with a simple model to find starting values:
\[
\text{nls}(v \sim \text{SSmicmen}(S, Vm, Km), \text{data} = \text{vmkmki})
\]
and then start the inhibitor models with a large value of \( k_i \) (small inhibitor effect).
\[
\text{fmc} <- \text{nls}(v \sim \frac{S}{S + Km * (1 + I/Ki)} * Vm, \\
\text{list}(Vm=16, Km=21, Ki=1000), \text{data} = \text{vmkmki})
\]
\[
\text{fmnc} <- \text{nls}(v \sim \frac{S}{((S + Km) * (1 + I/Ki))} * Vm, \\
\text{list}(Vm=16, Km=21, Ki=1000), \text{data} = \text{vmkmki})
\]
We now have three models, with RSS 406.5 (70 df), 177.3 and 55.0 (69 df).

The first model is nested within each of the others, so we could use an (approximate) \( F \) test:
\[
> \text{anova(fm0, fmc)}
\]
Analysis of Variance Table
Model 1: \( v \sim \text{SSmicmen}(S, Vm, Km) \)
Model 2: \( v \sim \frac{S}{(S + Km) * (1 + I/Ki)} * Vm \)
\[
\begin{array}{|c|c|c|c|c|c|c|}
\hline
\text{Res.Df} & \text{Res.Sum Sq} & \text{Df} & \text{Sum Sq} & \text{F value} & \text{Pr(>F)} \\
\hline
1 & 70 & 406.47 & & & \\
2 & 69 & 177.25 & 1 & 229.21 & 89.227 & 4.633e-14 \\
\hline
\end{array}
\]
\[
> \text{anova(fm0, fmnc)}
\]
Analysis of Variance Table
Model 1: \( v \sim \text{SSmicmen}(S, Vm, Km) \)
Model 2: \( v \sim \frac{S}{((S + Km) * (1 + I/Ki))} * Vm \)
\[
\begin{array}{|c|c|c|c|c|c|c|}
\hline
\text{Res.Df} & \text{Res.Sum Sq} & \text{Df} & \text{Sum Sq} & \text{F value} & \text{Pr(>F)} \\
\hline
1 & 70 & 406.47 & & & \\
2 & 69 & 54.97 & 1 & 351.5 & 441.24 < 2.2e-16 *** \\
\hline
\end{array}
\]
but there is no need of a formal test: the non-competitive inhibition model is clearly the best.

- We can do most of the regression diagnostics we would do for a linear model for mildly non-linear fits such as these.
- The one thing that is harder (at best) is drop-one-observation diagnostics such as studentized residuals and Cook’s distance.
- We can also use simulation-based methods to explore the sampling distribution of the parameters or other quantities of interest, including bootstrapping and the jackknife.
- There are some functions to do so in package nlstools.