Nonlinear Regression

KNNL – Chapter 13

Linear models

Recall the linear regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

- Linear models are at best *approximations* in most applications
- There are situations in which it is known theoretically that the relationship is *not* linear!
- Examples: growth follows exponential model, dose response models follow *S*-shaped curves

Intrinsically Linear models: Nonlinear Relations wrt *X*, but linear wrt β's

1) Polynomial Models: $E\{Y_i\} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2$

$$\frac{\partial E\{Y_i\}}{\partial X_i} = \frac{\partial}{\partial X_i}\left[\beta_0 + \beta_1 X_i + \beta_2 X_i^2\right] = 0 + \beta_1 + 2\beta_2 X_i = h(X_i)$$

$$\frac{\partial E\{Y_i\}}{\partial \beta_0} = 1 \qquad \frac{\partial E\{Y_i\}}{\partial \beta_1} = X_i \qquad \frac{\partial E\{Y_i\}}{\partial \beta_2} = X_i^2 \quad \text{None are functions of } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

2) Transformed Variable Models: $E\left\{\sqrt{Y_i}\right\} = \beta_0 + \beta_1 \ln(X_{i1}) + \beta_2 \left(\frac{1}{X_{i2}}\right)$

$$\frac{\partial E\left\{\sqrt{Y_i}\right\}}{\partial X_{i1}} = \beta_1 \left(\frac{1}{X_{i1}}\right) = h_1(X_{i1}) \qquad \frac{\partial E\left\{\sqrt{Y_i}\right\}}{\partial X_{i2}} = -\beta_2 \left(\frac{1}{X_{i2}^2}\right) = h_2(X_{i2})$$

$$\frac{\partial E\{Y_i\}}{\partial \beta_0} = 1 \qquad \frac{\partial E\{Y_i\}}{\partial \beta_1} = \ln(X_{i1}) \qquad \frac{\partial E\{Y_i\}}{\partial \beta_2} = \left(\frac{1}{X_{i2}}\right) \quad \text{None are functions of } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

In each case: $E\{Y_i\} = f(\mathbf{X_i}, \boldsymbol{\beta}) = \mathbf{X_i'}\boldsymbol{\beta}$

Case 1: $\mathbf{X_i'} = \begin{bmatrix} 1 & X_i & X_i^2 \end{bmatrix}$ Case 2: $\mathbf{X_i'} = \begin{bmatrix} 1 & \ln(X_{i1}) & \dfrac{1}{X_{i2}} \end{bmatrix}$

Nonlinear Regression Models

Nonlinear Regression models often use $\boldsymbol{\gamma}$ as vector of coefficients to distinguish from linear models:

Exponential Regression Models (Often used for modeling growth, where rate of growth changes):

$$E\{Y_i\} = \gamma_0 \exp(\gamma_1 X_i) \quad \Rightarrow \quad \frac{\partial E\{Y_i\}}{\partial \gamma_0} = \exp(\gamma_1 X_i) \qquad \frac{\partial E\{Y_i\}}{\partial \gamma_1} = \gamma_0 X_i \exp(\gamma_1 X_i) \qquad \text{functions of } \boldsymbol{\gamma}$$

$$f(\mathbf{X_i}, \boldsymbol{\gamma}) = \gamma_0 \exp(\gamma_1 X_i) \neq \mathbf{X_i'}\boldsymbol{\gamma}$$

More general exponential model (with errors independent and $N(0, \sigma^2)$):

$$Y_i = \gamma_0 + \gamma_1 \exp(\gamma_2 X_i) + \varepsilon_i \quad \text{Typically, } \gamma_0 > 0, \gamma_1 < 0, \gamma_2 < 0$$

$\Rightarrow$ Intercept: $E(Y_i \mid X_i = 0) = \gamma_0 + \gamma_1(1) = \gamma_0 + \gamma_1$

$\Rightarrow$ Asymptote: $E(Y_i \mid X_i \to \infty) = \gamma_0 + \gamma_1(1) = \gamma_0$

$\Rightarrow$ "Half-way" Point: $E\left(Y_i \mid X_i = \frac{0.693}{|\gamma_2|}\right) = \gamma_0 + \gamma_1 \exp\left(\gamma_2\left(\frac{0.693}{|\gamma_2|}\right)\right) = \gamma_0 + \gamma_1 \exp(-0.693) = \gamma_0 + \left(\frac{\gamma_1}{2}\right)$

Example:
A Nonlinear Regression Model

$$Y = \gamma_0 + \frac{\gamma_1 x}{\gamma_2 + x} + \varepsilon$$

This is a simple Maximum Effect (E_{max}) model,

which is a special case of the so-called Hill Model:

- x is usually a dose of some chemical (treatment);
- γ_0 = Mean Response at Dose 0 (or untreated patients);
- γ_1 = Maximal Effect = $\lim_{x\to\infty} \frac{\gamma_1 x}{\gamma_2 + x}$
- $\gamma_0 + \gamma_1$ = Maximum Mean Response
- γ_2 = Dose providing 50% of maximal effect (ED_{50}) = $\frac{\gamma_1 \gamma_2}{\gamma_2 + \gamma_2}$

Nonlinear Least Squares

$$f(\mathbf{X_i},\boldsymbol{\gamma}) = f_i(\boldsymbol{\gamma}) = f(\gamma_0,\gamma_1,\gamma_2) = \gamma_0 + \frac{\gamma_1 X_i}{\gamma_2 + X_i}$$

$$\frac{\partial f(\mathbf{X_i},\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}'} = F_i(\boldsymbol{\gamma}) = \frac{\partial f_i(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}'} = \begin{bmatrix} 1 & \dfrac{X_i}{\gamma_2 + X_i} & \dfrac{-\gamma_1 X_i}{(\gamma_2 + X_i)^2} \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbf{f}(\boldsymbol{\gamma}) = \begin{bmatrix} f_1(\gamma) \\ \vdots \\ f_n(\gamma) \end{bmatrix} = \begin{bmatrix} \gamma_0 + \dfrac{\gamma_1 X_1}{\gamma_2 + X_1} \\ \vdots \\ \gamma_0 + \dfrac{\gamma_1 X_n}{\gamma_2 + X_n} \end{bmatrix}$$

$$\mathbf{F}(\boldsymbol{\gamma}) = \begin{bmatrix} F_1(\gamma) \\ \vdots \\ F_n(\gamma) \end{bmatrix} = \begin{bmatrix} 1 & \dfrac{X_1}{\gamma_2 + X_1} & \dfrac{-\gamma_1 X_1}{(\gamma_2 + X_n)^2} \\ \vdots & \vdots & \vdots \\ 1 & \dfrac{X_n}{\gamma_2 + X_n} & \dfrac{-\gamma_1 X_n}{(\gamma_2 + X_n)^2} \end{bmatrix}$$

F acts like the **X** matrix in linear regression (but depends on parameters)

Nonlinear Least Squares

Goal: Choose $\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2$ that minimize error sum of squares:

$$Q = SSE(\boldsymbol{\gamma}) = \sum_{i=1}^{n}\left(Y_i - \left[\gamma_0 + \frac{\gamma_1 X_i}{\gamma_2 + X_i}\right]\right)^2 =$$

$$= \left(\mathbf{Y} - \mathbf{f}(\boldsymbol{\gamma})\right)'\left(\mathbf{Y} - \mathbf{f}(\boldsymbol{\gamma})\right)$$

$$\frac{\partial Q}{\partial \gamma_j} = -2\sum_{i=1}^{n}\left(Y_i - \left[\gamma_0 + \frac{\gamma_1 X_i}{\gamma_2 + X_i}\right]\right)F_i(\gamma_j) \quad j = 0,1,2$$

$$\frac{\partial Q}{\partial \boldsymbol{\gamma}'} = -2\left[\mathbf{Y} - \mathbf{f}(\boldsymbol{\gamma})\right]^{\mathrm{T}} \mathbf{F}(\boldsymbol{\gamma}) \overset{set}{=} \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$$

Numerical Search (Gauss-Newton Method)

1) Start with initial guesses $\gamma_0^{(0)}, \gamma_1^{(0)}, \gamma_2^{(0)}$;
2) Calculate $\boldsymbol{F}^{(0)}$;
3) Approximate the model with a linear model and obtain the slopes:

$$\boldsymbol{b}^{(0)} = (\boldsymbol{F}^{(0)\prime} \boldsymbol{F}^{(0)})^{-1} \boldsymbol{F}^{(0)\prime} \boldsymbol{Y}^{(0)}$$

4) Get revised estimates $\gamma_k^{(1)} = \gamma_k^{(0)} + b_k^{(0)}, k = 1, 2, 3$
5) Go back to step 2);
6) Continue until convergence.

Example – Iteration History (Tolerance = .0001)

iteration	g0	g1	g2	SSE	Delta(g)
0	5.0000	28.0000	160.0	13541.6	
1	6.2379	28.5863	140.9	12945.5	365.5745418
2	6.1771	28.1281	133.7	12942.9	52.82513814
3	6.1507	27.9163	129.9	12942.2	14.44158887
4	6.1361	27.7967	127.8	12942.0	4.506448063
5	6.1277	27.7272	126.5	12941.9	1.510150161
6	6.1227	27.6861	125.8	12941.9	0.526393989
7	6.1197	27.6615	125.4	12941.9	0.187692352
8	6.1180	27.6467	125.1	12941.9	0.067822683
9	6.1169	27.6377	125.0	12941.9	0.024703325
10	6.1162	27.6323	124.9	12941.9	0.009040833
11	6.1158	27.6291	124.8	12941.9	0.003318268
12	6.1156	27.6271	124.8	12941.9	0.001220029
13	6.1155	27.6259	124.8	12941.9	0.000449042
14	6.1154	27.6251	124.7	12941.9	0.000165379
15	6.1153	27.6247	124.7	12941.9	6.09317E-05

$$\hat{Y} = 6.12 + \frac{27.62X}{124.7 + X}$$

Estimated Variance-Covariance Matrix

$$s^2\left\{\hat{\boldsymbol{\gamma}}\right\} = s^2\left(\hat{\mathbf{F}}'\hat{\mathbf{F}}\right)^{\mathbf{-1}}$$

$$s^2 = \frac{\left(\mathbf{Y}-\hat{\mathbf{f}}\right)^{\mathbf{T}}\left(\mathbf{Y}-\hat{\mathbf{f}}\right)}{n-p}$$

$$s\left\{\hat{\gamma}_i\right\} = s\sqrt{\left(\hat{\mathbf{F}}'\hat{\mathbf{F}}\right)^{\mathbf{-1}}_{(i+1,i+1)}}$$

Note: KNNL uses **g** for $\hat{\boldsymbol{\gamma}}$ and **D** for $\hat{\mathbf{F}}$

Variance Estimates/Confidence Intervals

$$s^2 = \frac{\sum_{i=1}^{163}\left(Y_i - f_i\left(\hat{\boldsymbol{\gamma}}\right)\right)^2}{163-3} = 80.89$$

$$\mathbf{s}^2\left\{\hat{\boldsymbol{\gamma}}\right\} = s^2\left(\hat{\mathbf{F}}'\hat{\mathbf{F}}\right)^{-1} = \begin{bmatrix} 1.1594 & -0.7219 & 15.609 \\ -0.7219 & 12.081 & 130.14 \\ 15.609 & 130.14 & 2238.76 \end{bmatrix}$$

Parameter	Estimate	Std. Error	95% CI
γ_0	6.12	1.08	(3.96 , 8.28)
γ_1	27.62	3.48	(20.66 , 34.58)
γ_2	124.7	47.31	(30.08 , 219.32)

Notes on Nonlinear Least Squares

Large-Sample Theory:

When $\varepsilon_i \sim N\left(0,\sigma^2\right)$ independent , for large n: $\hat{\boldsymbol{\gamma}}$ is approximately normal

$E\left\{\hat{\boldsymbol{\gamma}}\right\} \approx \boldsymbol{\gamma}$ Approximate $\sigma^2\left\{\hat{\boldsymbol{\gamma}}\right\}$ estimated by $\mathbf{s^2}\left\{\hat{\boldsymbol{\gamma}}\right\} = MSE\left(\hat{\mathbf{F}}'\hat{\mathbf{F}}\right)^{\mathbf{-1}}$

$\Rightarrow \quad \hat{\boldsymbol{\gamma}} \sim N\left(\boldsymbol{\gamma}, \sigma^2\left(\mathbf{F'F}\right)^{\mathbf{-1}}\right)$ (Approximately)

For small samples:

- When errors are normal, independent, with constant variance, we can often use the t-distribution for tests and confidence intervals (software packages do this implicitly)
- When the extent of nonlinearity is extreme, or normality assumptions do not hold, should use bootstrap to estimate standard errors of regression coefficients

Neural Network Modeling

In a neural network model the i^{th} outcome Y_i is modeled as a nonlinear function of *m derived* predictors H_{i0}, …, $H_{i,m\text{-}1}$:

$$Y_i = g_Y(\beta_0 H_{i0} + \beta_1 H_{i1} + \ldots + \beta_{m\text{-}1} H_{i,m\text{-}1}) + \varepsilon_i$$

Each H_{ij} is a nonlinear function of the original predictors:

$$H_{ij} = g_j(\boldsymbol{X}_i' \boldsymbol{\alpha}_j)$$

Typical choice for the *activation* functions g is the logistic function:

$$g(z) = \frac{1}{1 + e^{-z}}$$

Neural Network Terminology

Statistics	Machine Learning
coefficient	weight
predictor	input
response	output
parameter estimation	training (learning)
steepest descent	back propagation
intercept	bias term
derived predictor	hidden node
penalty function	weight decay