Chapter 13
Introduction to Nonlinear Regression (非線性迴歸) and Neural Networks (類神經網路)

許湘伶

Applied Linear Regression Models (Kutner, Nachtsheim, Neter, Li)
Examples of nonlinear regression model:

- growth from birth to maturity in human: nonlinear in nature
- dose-response relationships

Purpose:

- obtain estimates of parameters
- Neural network model: data mining applications
- logistic regression models: Chap. 14
the general linear regression model (6.7):

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \]

A polynomial regression model in one or more predictor variables is linear in the parameters. Ex:

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2} + \beta_4 X_{i2}^2 + \beta_5 X_{i1}X_{i2} + \varepsilon_i \]

\[ \log_{10} Y_i = \beta_0 + \beta_1 \sqrt{X_{i1}} + \beta_2 \exp(X_{i2}) + \varepsilon_i \quad (\text{transformed}) \]
In general, we can state a linear regression model as:

$$Y_i = f(X_i, \beta) + \varepsilon_i = X_i' \beta + \varepsilon_i$$

where $X_i = [1 \ X_{i1} \cdots X_{i,p-1}]'$

Nonlinear regression models

- the same basic form as that in (13.4):

$$Y_i = f(X_i, \gamma) + \varepsilon_i$$

- $f(X_i, \gamma)$: mean response given by the nonlinear response function $f(X, \gamma)$
- $\varepsilon_i$: error term; assumed to have $E\{\varepsilon_i\} = 0$, constant variance and to be uncorrelated
- $\gamma$: parameter vector
exponential regression model: in growth (生長) studies; concentration (濃度)

\[ Y_i = \gamma_0 \exp(\gamma_1 X_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{indep.}}{\sim} N(0, \sigma^2) \]
logistic regression models (邏輯式迴歸模型): in population studies (人口研究);

\[ Y_i = \frac{\gamma_0}{1 + \gamma_1 \exp(\gamma_2 X_i)} + \varepsilon_i, \quad \varepsilon_i \text{ indep. } \sim N(0, \sigma^2) \]
Logistic regression models (邏輯式迴歸模型):

- The response variable is qualitative (0,1): purchase a new car
- The error terms are not normally distributed with constant variance (Chap. 14)
General Form of Nonlinear Regression Models:

- The error terms $\varepsilon_i$ are often assumed to be independent normal random variables with constant variance.
- **Important difference:** $\{\beta_i\}$ is not necessarily directly related to $\{X_i\}$ in the model
  - **linear regression:** $(p - 1) X$ variables $\Rightarrow p$ regression coefficients
    \[ Y_i = \sum_{j=0}^{p-1} \beta_j X_{ji} + \varepsilon_i \]
  - **nonlinear regression:**
    \[ Y_i = \gamma_0 \exp(\gamma_1 X_i) + \varepsilon_i \quad (p = 2, q = 1) \]
Nonlinear regression model

- \( q \): \#\{X variables\}
- \( p \): \#\{regression parameters\}
- \( X_i \): the observations on the X variables \textit{without the initial element} 1

The general form of a nonlinear regression model:

\[
Y_i = f(X_i, \gamma) + \varepsilon_i
\]
intrinsically linear(實質線性的) response functions: nonlinear response functions can be linearized by a transformation

Ex:

\[
f(X, \gamma) = \gamma_0[\exp(\gamma_1 X)]
\]

\[
\Rightarrow g(X, \gamma) = \log_e f(X, \gamma) = \beta_0 + \beta_1 X
\]

\[
\beta_0 = \log \gamma_0, \quad \beta_1 = \gamma_1
\]

Just because a nonlinear response function is intrinsically linear does not necessarily imply that linear regression is appropriate. (∵ the error term in the linearized model will no longer be normal with constant variance)
Estimation of regression parameters:

1. least squares method
2. maximum likelihood method

Also as in linear regression, both of these methods of estimation yield the same parameter estimates when the error terms in (13.12) are independent normal with constant variance.

It is usually not possible to find analytical expression for LSE and MLE for nonlinear regression models.

numerical search procedures must be used: require intensive computations
The concepts of LSE for linear regression also extend directly to nonlinear regression models.

The least squares criterion:

\[ Q = \sum_{i=1}^{n} [Y_i - f(X_i, \gamma)]^2 \]

\( Q \) must be minimized with respect to \( \gamma_0, \ldots, \gamma_{p-1} \).

A difference from linear regression is that the solution of the normal equations usually requires an iterative numerical search procedure because analytical solutions generally cannot be found.
13.2 Least Squares Estimation in Nonlinear Regression

Solution of Normal Equations

\[ Y_i = f(X_i, \gamma) + \varepsilon_i \]
\[ \Rightarrow Q = \sum_{i=1}^{n} [Y_i - f(X_i, \gamma)]^2 \]

\[ \Rightarrow g = \arg \min_{\gamma} Q \quad (g : \text{the vector of the LSE } g_k) \]

(partial derivative of \( Q \) with respect to \( \gamma_k \))

\[ \Rightarrow \frac{\partial Q}{\partial \gamma_k} = \sum_{i=1}^{n} -2[Y_i - f(X_i, \gamma)] \left[ \frac{\partial f(X_i, \gamma)}{\partial \gamma_k} \right] \bigg|_{\gamma_k = g_k} \overset{\text{set}}{=} 0 \]
Solution of Normal Equations (cont.)

- The $p$ normal equations:

$$
\sum_{i=1}^{n} Y_i \left[ \frac{\partial f(X_i, \gamma)}{\partial \gamma_k} \right]_{\gamma=g} - \sum_{i=1}^{n} f(X_i, g) \left[ \frac{\partial f(X_i, \gamma)}{\partial \gamma_k} \right]_{\gamma=g} = 0, \\
\quad k = 0, 1, \ldots, p - 1
$$

- $g$: the vector of the least squares estimates $g_k$

$$
g_{p \times 1} = [g_0, \ldots, g_{p-1}]'
$$

- nonlinear in the parameter estimates $g_k$
- numerical search procedure are required
- multiple solution may be possible
Example: Severely injured patients

- **Y**: Prognostic (預後的) Index
- **X**: Days Hospitalized (住院治療)
- Related earlier studies: the relationship between **Y** and **X** is exponential

\[ Y_i = \gamma_0 \exp(\gamma_1 X_i) + \varepsilon_i \]
13.2 Least Squares Estimation in Nonlinear Regression

Solution of Normal Equations (cont.)

\[
\frac{\partial f(X_i, \gamma)}{\partial \gamma_0} = \exp(\gamma_1 X_i)
\]

\[
\frac{\partial f(X_i, \gamma)}{\partial \gamma_1} = \gamma_0 X_i \exp(\gamma_1 X_i)
\]

\[
\Rightarrow \left. \frac{\partial Q}{\partial \gamma_k} \right|_{g} = 0
\]

\[
\Rightarrow \sum Y_i \exp(g_1 X_i) - g_0 \sum \exp(2g_1 X_i) = 0
\]

\[
\sum Y_i X_i \exp(g_1 X_i) - g_0 \sum X_i \exp(2g_1 X_i) = 0
\]

No closed-form solution exists for \( g = (g_0, g_1)^T \)
Gauss-Newton Method (高斯-牛顿法)

- linearization method:
  1. use a Taylor series expansion to approximate the nonlinear regression model

\[
(f(x) = f(a) + \sum_{n=1}^{\infty} \frac{f^{(n)}}{n!}(x - a)^n)
\]

2. employ OLS to estimate the parameters
13.2 Least Squares Estimation in Nonlinear Regression

Gauss-Newton Method (高斯-牛頓法) (cont.)

圖片來源：Wiki: Taylor’s theorem
https://en.wikipedia.org/wiki/Taylor’s_theorem
Gauss-Newton Method (高斯-牛顿法) (cont.)

Gauss-Newton Method:

1. **initial parameters** $\gamma_0, \ldots, \gamma_{p-1}$: $g_0^{(0)}, \ldots, g_{p-1}^{(0)}$

2. approximate the mean responses $f(X_i, \gamma)$ in the Taylor series expansion around $g_k^{(0)}$:

   $$f(X_i, \gamma) \approx f(X_i, g^{(0)}) + \sum_{k=0}^{p-1} \left[ \frac{\partial f(X_i, \gamma)}{\partial \gamma_k} \right]_{\gamma=g^{(0)}} \gamma_k - g_k^{(0)}$$

3. obtain **revised estimated regression coefficients** $g_k^{(1)}$: (later)

   $$g_k^{(1)} = g_k^{(0)} + b_k^{(0)}$$
Gauss-Newton Method (高斯-牛頓法) (cont.)

\[ f(x_i, \gamma) \approx f(x_i, g^{(0)}) + \sum_{k=0}^{p-1} \left[ \frac{\partial f(x_i, \gamma)}{\partial \gamma_k} \right] (\gamma_k - g^{(0)}_k) \]

\[ = f_i^{(0)} + \sum_{k=0}^{p-1} D_{ik}^{(0)} \beta^{(0)}_k \]

\[ \Rightarrow Y_i \approx f_i^{(0)} + \sum_{k=0}^{p-1} D_{ik}^{(0)} \beta^{(0)}_k + \varepsilon_i \]

\[ (Y_i^{(0)} = Y_i - f_i^{(0)}) \quad Y_i^{(0)} \approx \sum_{k=0}^{p-1} D_{ik}^{(0)} \beta^{(0)}_k + \varepsilon_i \quad \text{(no intercept)} \quad (13.24) \]

The purpose of fitting the linear regression model approximation (13.24) is therefore to estimate \( \beta^{(0)}_k \) and use these estimates to adjust the initial starting estimates of the regression parameters.
Matrix Form: \( Y_i^{(0)} \approx \sum_{k=0}^{p-1} D_{ik}^{(0)} \beta_k^{(0)} + \varepsilon_i \)

\[
Y^{(0)} \approx D^{(0)} \beta^{(0)} + \varepsilon
\]  (13.25)

where:

(13.25a) \[
Y^{(0)} = \begin{bmatrix}
Y_1 - f_1^{(0)} \\
\vdots \\
Y_n - f_n^{(0)}
\end{bmatrix}
\]

(13.25b) \[
D^{(0)} = \begin{bmatrix}
D_{10}^{(0)} & \cdots & D_{1,p-1}^{(0)} \\
\vdots & \ddots & \vdots \\
D_{n0}^{(0)} & \cdots & D_{n,p-1}^{(0)}
\end{bmatrix}
\]

(13.25c) \[
\beta^{(0)} = \begin{bmatrix}
\beta_0^{(0)} \\
\vdots \\
\beta_{p-1}^{(0)}
\end{bmatrix}
\]

(13.25d) \[
\varepsilon = \begin{bmatrix}
\varepsilon_1 \\
\vdots \\
\varepsilon_n
\end{bmatrix}
\]
Gauss-Newton Method (高斯-牛顿法) (cont.)

- the $D$ matrix of partial derivative play the role of the $X$ matrix (without a column of 1s for the intercept)
- Estimate $\beta^{(0)}$ by OLS:
  \[ b^{(0)} = (D^{(0)\prime} D^{(0)})^{-1} D^{(0)\prime} Y^{(0)} \]
- obtain revised estimated regression coefficients $g_k^{(1)}$:
  \[ g_k^{(1)} = g_k^{(0)} + b_k^{(0)} \]
Gauss-Newton Method (高斯-牛顿法) (cont.)

- Evaluated for $g^{(0)}$ by $SSE^{(0)}$:

$$SSE^{(0)} = \sum_{i=1}^{n} [Y_i - f(X_i, g^{(0)})]^2 = \sum_{i=1}^{n} [Y_i - f_i^{(0)}]^2$$

- After the end of the first iteration:

$$SSE^{(1)} = \sum_{i=1}^{n} [Y_i - f(X_i, g^{(1)})]^2 = \sum_{i=1}^{n} [Y_i - f_i^{(1)}]^2$$

- If the Gauss-Newton method is working effectively in the first iteration, $SSE^{(1)}$ should be smaller than $SSE^{(0)}$. (∵ $g^{(1)}$ should be better estimates)
The Gauss-Newton method repeats the procedure with $g^{(1)}$ now used for the new starting values.

Until $g^{(s+1)} - g^{(s)}$ and/or $SSE^{(s+1)} - SSE^{(s)}$ become negligible.

The Gauss-Newton method works effectively in many nonlinear regression applications. (Sometimes may require numerous iterations before converging.)
Gauss-Newton Method (cont.)

Example: Severely injured patients

- **Initial**: Transformed $Y$ 
  \[ \log \gamma_0 \exp(\gamma_1 X) = \log \gamma_0 + \gamma_1 X \]

\[ Y_i' = \beta_0 + \beta_1 X_i + \varepsilon_i \]

OLS \( \Rightarrow \) \( b_0 = 0.40371, \quad b_1 = -0.03797 \)

\[ g_0^{(0)} = \exp(b_0) = 56.6646, \quad g_1^{(0)} = b_1 = -0.03797 \]

(a) Estimates of Parameters and Least Squares Criterion Measure

<table>
<thead>
<tr>
<th>Iteration</th>
<th>( g_0 )</th>
<th>( g_1 )</th>
<th>SSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>56.6646</td>
<td>-.03797</td>
<td>56.0869</td>
</tr>
<tr>
<td>1</td>
<td>58.5578</td>
<td>-.03953</td>
<td>49.4638</td>
</tr>
<tr>
<td>2</td>
<td>58.6065</td>
<td>-.03959</td>
<td>49.4593</td>
</tr>
<tr>
<td>3</td>
<td>58.6065</td>
<td>-.03959</td>
<td>49.4593</td>
</tr>
</tbody>
</table>
13.2 Least Squares Estimation in Nonlinear Regression

**Gauss-Newton Method (高斯-牛顿法) (cont.)**

(b) Final Least Squares Estimates

<table>
<thead>
<tr>
<th>$k$</th>
<th>$g_k$</th>
<th>$s{g_k}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>58.6065</td>
<td>1.472</td>
</tr>
<tr>
<td>1</td>
<td>-0.03959</td>
<td>0.00171</td>
</tr>
</tbody>
</table>

\[
MSE = \frac{49.4593}{13} = 3.80456
\]

(c) Estimated Approximate Variance-Covariance Matrix of Estimated Regression Coefficients

\[
s^2[g] = MSE(D'D)^{-1} = 3.80456 \begin{bmatrix}
5.696E-1 & -4.682E-4 \\
-4.682E-4 & 7.697E-7
\end{bmatrix}
\]

\[
= \begin{bmatrix}
2.1672 & -1.781E-3 \\
-1.781E-3 & 2.928E-6
\end{bmatrix}
\]

\[
\hat{Y} = (58.6065) \exp(-0.03959X)
\]
The choice of initial starting values:

- **a poor choice** may result in slow convergence, convergence to a local minimum, or even divergence
- **Good starting values**: result in faster convergence, will lead to a solution that is the global minimum rather than a local minimum

A variety of methods for obtaining starting values:

1. related earlier studies
2. select $p$ representative observations $\Rightarrow$ solve for $p$ parameters, then used as the starting values
3. do a grid search in the parameter space $\Rightarrow$ using as the starting values that $g$ for which $Q$ is smallest
Some properties that exist for linear regression least squares do not hold for nonlinear regression least squares.

Ex: $\sum e_i \neq 0$; $SSR + SSE \neq SSTO$; $R^2$ is not a meaningful descriptive statistic for nonlinear regression.

Two other direct search procedures:

1. The method of **steepest descent searches**

2. The **Marquardt algorithm**: seeks to utilize the best feature of the Gauss-Newton method and the method of steepest descent.

   a middle ground (折衷; 妥協) between these two method
Model building and diagnostics

- The model-building process for nonlinear regression models often differs somewhat from that for linear regression models.
- Validation of the selected nonlinear regression model can be performed in the same fashion as for linear regression models.
- Use of diagnostics tools to examine the appropriateness of a fitted model plays an important role in the process of building a nonlinear regression model.
When replicate observations are available and the sample size is reasonably large, the appropriateness of a nonlinear regression function can be tested formally by means of the lack of fit test for linear regression models. (the test is an approximate one)

Plots: $e_i$ vs. $t_i$, $\hat{Y}_i$, $X_{ik}$ can be helpful in diagnosing departures from the assumed model

Unequal variances $\Rightarrow$ WLS; transformations
Inferences

- Inferences about the regression parameters in nonlinear regression are usually based on large-sample theory.
- When $n$ is large, LSE and MLE for nonlinear regression models with normal error terms are approximately normally distributed and almost unbiased and have almost minimum variance.
- Estimate of Error Term Variance:

$$MSE = \frac{SSE}{n - p} = \frac{\sum (Y_i - f(X_i, g))^2}{n - p}$$

(Not unbiased estimator of $\sigma^2$ but the bias is small when $n$ is large)
Large-Sample Theory

When the error terms $\varepsilon_i$ are independent $N(0, \sigma^2)$ and the sample size $n$ is reasonably large, the sampling distribution of $\mathbf{g}$ is approximately normal. The expected value of the mean vector is approximately:

$$E\{\mathbf{g}\} \approx \gamma$$

The approximate variance-covariance matrix of the regression coefficients is estimated by:

$$s^2\{\mathbf{g}\} = MSE(D'D)^{-1}$$
Inferences (cont.)

- **No simple rule exists** that tells us when it is appropriate to use the large-sample inference methods and when it is not appropriate.

- **However, a number of guidelines have been developed** that are helpful in assessing the appropriateness of using the large-sample inference procedures in a given application.

- **When the diagnostics suggest that large-sample inference procedures are not appropriate** in a particular instance, remedial measures should be explored.
  - reparameterize the nonlinear regression model
  - bootstrap estimated of precision and confidence intervals instead of the large-sample inferences
Interval Estimation

- Large-sample theorem: approximate result for a single $\gamma_k$

$$\frac{g_k - \gamma_k}{s\{ g_k \}} \sim t(n - p), \quad k = 0, 1, \ldots, p - 1$$

$$\Rightarrow g_k \pm t(1 - \alpha/2; n - p)s\{ g_k \}$$

- Several $\gamma_k$: $m$ parameters to be estimated with approximate family confidence coefficient $1 - \alpha \Rightarrow$ the Bonferroni confidence limits:

$$g_k \pm Bs\{ g_k \} \quad B = t(1 - \alpha/2m; n - p)$$
Test Concerning a single $\gamma_k$

- A large-sample test:

$$H_0 : \gamma_k = \gamma_k^0 \text{ vs. } H_a : \gamma_k \neq \gamma_k^0$$

$$t^* = \frac{g_k - \gamma_k^0}{s\{g_k\}}$$

- If $|t^*| \leq t(1 - \alpha/2; n - p)$, conclude $H_0$
- If $|t^*| > t(1 - \alpha/2; n - p)$, conclude $H_a$

Test concerning several $\gamma_k$

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div MSE(F)$$

$$\text{approx} \sim F(df_R - df_F, df_F) \text{ when } H_0 \text{ holds}$$