

Binary Response Regression

Treating the x_i as fixed and the y_i 's as random we have the probability function of each y_i is:

$$\begin{aligned}
 & [y_i; \beta_0, \beta_1, x_i] \\
 &= \begin{cases} F(\beta_0 + \beta_1 x_i) & y_i = 1 \\ 1 - F(\beta_0 + \beta_1 x_i) & y_i = 0 \end{cases} \\
 &= F(\beta_0 + \beta_1 x_i)^{y_i} \\
 &\quad \times \{1 - F(\beta_0 + \beta_1 x_i)\}^{1-y_i}, \quad y_i = 0, 1.
 \end{aligned}$$

for $1 \leq i \leq n$.

In the BPD/birthweight example let

$$y_i = \text{BPD}_i$$

$$x_i = \text{birthweight}_i$$

$$1 \leq i \leq n \quad (n = 223)$$

The likelihood of (β_0, β_1) is then

$$\begin{aligned}
 \mathcal{L}(\beta_0, \beta_1) &= \prod_{i=1}^n F(\beta_0 + \beta_1 x_i)^{y_i} \\
 &\quad \times \{1 - F(\beta_0 + \beta_1 x_i)\}^{1-y_i}
 \end{aligned}$$

The log-likelihood of (β_0, β_1) is

$$\begin{aligned}\ell(\beta_0, \beta_1) &= \sum_{i=1}^n [y_i \ln F(\beta_0 + \beta_1 x_i) \\ &\quad + (1 - y_i) \ln \{1 - F(\beta_0 + \beta_1 x_i)\}] \\ &= \sum_{i=1}^n [y_i \ln \left\{ \frac{F(\beta_0 + \beta_1 x_i)}{1 - F(\beta_0 + \beta_1 x_i)} \right\} \\ &\quad + \ln \{1 - F(\beta_0 + \beta_1 x_i)\}] \\ &= \sum_{i=1}^n [y_i \text{logit}\{F(\beta_0 + \beta_1 x_i)\} \\ &\quad + \ln \{1 - F(\beta_0 + \beta_1 x_i)\}] \\ &\quad \text{where } \text{logit}(u) = \ln \left(\frac{u}{1 - u} \right), \quad 0 < u < 1.\end{aligned}$$

Question:

How might we choose $F : \mathbb{R} \rightarrow (0, 1)$?

Any **cumulative distribution function** has this property.

e.g.

$$F(x) = P(Z \leq x) = \Phi(x), \quad Z \sim N(0, 1).$$

The choice $F = \Phi$ leads to what is known as **probit regression**.

Probit regression log-likelihood

$$\ell(\beta_0, \beta_1) = \sum_{i=1}^n [y_i \text{logit}\{\Phi(\beta_0 + \beta_1 x_i)\} \\ + \ln \{1 - \Phi(\beta_0 + \beta_1 x_i)\}]$$

Question: Can we make $\ell(\beta_0, \beta_1)$ less complicated?

Answer:

How about choosing F so that

$$y_i \text{logit}\{F(\beta_0 + \beta_1 x_i)\} = y_i(\beta_0 + \beta_1 x_i) ?$$

Wanted:

F such that

$$\text{logit}\{F(u)\} = u$$

i.e.

$$F(u) = \text{logit}^{-1}(u).$$

$$y = \ln\left(\frac{1}{1/x - 1}\right)$$

$$e^y = \frac{1}{1/x - 1}$$

$$e^{-y} = 1/x - 1$$

$$1/x = 1 + e^{-y}$$

$$x = \frac{1}{1 + e^{-y}} = \frac{e^y}{1 + e^y}$$

i.e.

Class Exercise

Find the explicit form of logit^{-1} where

$$y = \text{logit}(x) = \ln\left(\frac{x}{1-x}\right).$$

Answer

$$\text{logit}^{-1}(u) = \frac{e^u}{1 + e^u}.$$

\Rightarrow Take

$$F(x) = \frac{e^x}{1 + e^x}.$$

\Rightarrow

$$\ell(\beta_0, \beta_1) = \sum_{i=1}^n \{y_i(\beta_0 + \beta_1 x_i) - \ln(1 + e^{\beta_0 + \beta_1 x_i})\}.$$

Note

$$F(x) = \frac{e^x}{1+e^x}$$

is the cumulative distribution function of the random variable X having density

$$f_X(x) = F'(x) = \frac{e^x}{(1+e^x)^2},$$

known as the **logistic** density.

Hence, this choice of F for binary responses is known as

logistic regression.

An immediate benefit of the simple form of $\ell(\beta_0, \beta_1)$ is

Fisher information calculation.

$$\frac{\partial}{\partial \beta_0} \ell(\beta_0, \beta_1) = \sum_{i=1}^n \left(y_i - \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right).$$

$$\frac{\partial^2}{\partial \beta_0^2} \ell(\beta_0, \beta_1) = - \sum_{i=1}^n \frac{e^{\beta_0 + \beta_1 x_i}}{(1 + e^{\beta_0 + \beta_1 x_i})^2}.$$

\implies

$$\begin{aligned} -E \left\{ \frac{\partial^2}{\partial \beta_0^2} \ell(\beta_0, \beta_1) \right\} &= \sum_{i=1}^n \frac{e^{\beta_0 + \beta_1 x_i}}{(1 + e^{\beta_0 + \beta_1 x_i})^2} \\ &= (1, 1) \text{ entry of } I_n(\beta_0, \beta_1) \end{aligned}$$

We will return to Fisher information later.

First, address the problem of finding

$$(\hat{\beta}_0, \hat{\beta}_1) = \text{maximisers of } \ell(\beta_0, \beta_1).$$

Streamlined matrix notation:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

Also let

$$b(x) = \ln(1 + e^x).$$

⇒ the log-likelihood is

$$\ell(\boldsymbol{\beta}) = \mathbf{y}^T \mathbf{X} \boldsymbol{\beta} - \mathbf{1}^T b(\mathbf{X} \boldsymbol{\beta}).$$

Maximum occurs when

$$D \ell(\boldsymbol{\beta}) = \mathbf{0}.$$

Newton-Raphson iteration is:

$$\boldsymbol{\beta}_{i+1} = \boldsymbol{\beta}_i - \{\mathbf{H} \ell(\boldsymbol{\beta}_i)\}^{-1} D \ell(\boldsymbol{\beta}_i).$$

So we need $D \ell(\boldsymbol{\beta})$ and $\mathbf{H} \ell(\boldsymbol{\beta}) \dots$

$$\begin{aligned} d \ell(\boldsymbol{\beta}) &= d \{ \mathbf{y}^T \mathbf{X} \boldsymbol{\beta} - \mathbf{1}^T b(\mathbf{X} \boldsymbol{\beta}) \} \\ &= \mathbf{y}^T \mathbf{X} d \boldsymbol{\beta} - \mathbf{1}^T \text{diag}\{b'(\mathbf{X} \boldsymbol{\beta})\} \mathbf{X} d \boldsymbol{\beta} \\ &= \{ \mathbf{y} - b'(\mathbf{X} \boldsymbol{\beta}) \}^T \mathbf{X} d \boldsymbol{\beta} \end{aligned}$$

⇒ By the First Identification Theorem, and using $b'(u) = e^u / (1 + e^u)$,

$$D \ell(\boldsymbol{\beta}) = \left(\mathbf{y} - \frac{e^{\mathbf{X} \boldsymbol{\beta}}}{1 + e^{\mathbf{X} \boldsymbol{\beta}}} \right)^T \mathbf{X}$$

Similarly,

$$\mathbf{H} \ell(\boldsymbol{\beta}) = -\mathbf{X}^T \text{diag}\{b''(\mathbf{X} \boldsymbol{\beta})\} \mathbf{X} = -\mathbf{X}^T \mathbf{W}_\beta \mathbf{X}$$

where

$$\mathbf{W}_\beta = \text{diag}\{b''(\mathbf{X}\beta)\} = \text{diag}\left\{\frac{e^{\mathbf{X}\beta}}{(1 + e^{\mathbf{X}\beta})^2}\right\}.$$

and is called the
adjusted response variable.

The Newton-Raphson update is

$$\begin{aligned}\beta_{i+1} &= \beta_i + (\mathbf{X}^T \mathbf{W}_{\beta_i} \mathbf{X})^{-1} \mathbf{X}^T \left(\mathbf{y} - \frac{e^{\mathbf{X}\beta_i}}{1 + e^{\mathbf{X}\beta_i}} \right) \\ &= \beta_i + (\mathbf{X}^T \mathbf{W}_{\beta_i} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}_{\beta_i} \mathbf{y}_{\beta_i}^{\text{adj}}\end{aligned}$$

$$\text{where } \mathbf{y}_\beta^{\text{adj}} = \mathbf{W}_\beta^{-1} \left(\mathbf{y} - \frac{e^{\mathbf{X}\beta}}{1 + e^{\mathbf{X}\beta}} \right)$$

Logistic Regression Fitting Algorithm:

Step 0: Get initial value β .

Step 1: $\mathbf{W}_\beta \leftarrow \text{diag}\left\{\frac{e^{\mathbf{X}\beta}}{(1+e^{\mathbf{X}\beta})^2}\right\}$.

Step 2: $\mathbf{y}_\beta^{\text{adj}} \leftarrow \mathbf{W}_\beta^{-1} \left(\mathbf{y} - \frac{e^{\mathbf{X}\beta}}{1+e^{\mathbf{X}\beta}} \right)$.

Step 3: $\beta \leftarrow \beta + (\mathbf{X}^T \mathbf{W}_\beta \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}_\beta \mathbf{y}_\beta^{\text{adj}}$.

Iterate Steps 1–3 until convergence.

This algorithm is called
iteratively reweighted least squares.

$\ell(\boldsymbol{\beta})$ is well-behaved so convergence to unique global maximum often occurs quickly.

(see Assignment 4, Question 2).

Fisher Information for Logistic Regression

$$\begin{aligned} I_n(\boldsymbol{\beta}) &= -E[\mathbf{H}_{\boldsymbol{\beta}}\ell(\boldsymbol{\beta})] \\ &= \mathbf{X}^T \text{diag} \left\{ \frac{e^{\mathbf{X}\boldsymbol{\beta}}}{(1 + e^{\mathbf{X}\boldsymbol{\beta}})^2} \right\} \mathbf{X} \end{aligned}$$

$$\Rightarrow \widehat{\text{se}}(\widehat{\beta}_1) = \sqrt{(2,2) \text{ entry of } \left[\mathbf{X}^T \text{diag} \left\{ \frac{e^{\mathbf{X}\widehat{\boldsymbol{\beta}}}}{(1 + e^{\mathbf{X}\widehat{\boldsymbol{\beta}}})^2} \right\} \mathbf{X} \right]^{-1}}$$

Computer Demonstration

We will now demonstrate these results in **R**.

Extensions

1. Several predictors;
higher degree polynomials.
2. Other GLMs (e.g. count data).

Extension 1 involves simply making

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{bmatrix}$$

more general.

For example:

$$\text{logit}\{P(\text{BPD}_i = 1)\} = \beta_0 + \beta_1 \text{birthweight}_i + \beta_2 \text{birthweight}_i^2$$

set

$$\mathbf{X} = \begin{bmatrix} 1 & \text{birthweight}_1 & \text{birthweight}_1^2 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & \text{birthweight}_n & \text{birthweight}_n^2 \end{bmatrix}$$

or

$$\text{logit}\{P(\text{BPD}_i = 1)\} = \beta_0 + \beta_1 \text{birthweight}_i + \beta_2 \text{MumSmoker}_i$$

where

MumSmoker_i = indicator of Mum of i th baby is a smoker

$$\mathbf{X} = \begin{bmatrix} 1 & \text{birthweight}_1 & \text{MumSmoker}_1 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & \text{birthweight}_n & \text{MumSmoker}_n \end{bmatrix}$$

Extension 2 often simply involve replacing

$$b(x) = \ln(1 + e^x)$$

by other b functions.

e.g. for count data and Poisson model set

$$b(x) = e^x$$

(see Question 1 of Assignment 4)

