

**Solution 1**

- (a) The density for a binary observation can be written as  $\pi_j^{y_j}(1 - \pi_j)^{1-y_j}$ , so the log likelihood for independent binary data  $y_1, \dots, y_n$  is

$$\ell(\pi_1, \dots, \pi_n) = \sum_{j=1}^n y_j \log \pi_j + (1 - y_j) \log(1 - \pi_j).$$

If the  $\pi_j$  are unconnected, then it is easy to check that the maximising probabilities are  $\tilde{\pi}_j = y_j$ , so  $\pi_j^{y_j}(1 - \pi_j)^{1-y_j} = 1$  (noting that  $0^0 = 1$ ), and therefore

$$\ell(\tilde{\pi}_1, \dots, \tilde{\pi}_n) = 0$$

is the highest possible value of the log likelihood function. Therefore the deviance for a model in which  $\pi_j(\beta) = \exp(x_j^T \beta) / \{1 + \exp(x_j^T \beta)\}$  is

$$D = 2 \{ \ell(\tilde{\pi}_1, \dots, \tilde{\pi}_n) - \ell(\hat{\pi}_1, \dots, \hat{\pi}_n) \} = -2\ell(\hat{\beta}),$$

where we have set

$$\ell(\hat{\beta}) = \ell(\hat{\pi}_1, \dots, \hat{\pi}_n) = \sum_{j=1}^n y_j \log \pi_j(\hat{\beta}) + (1 - y_j) \log \{1 - \pi_j(\hat{\beta})\}.$$

We could now note that the logistic regression model is a canonical exponential family model with minimal sufficient statistic  $X^T y$ , and therefore the maximum likelihood estimators and all derived quantities, including  $\ell(\hat{\beta})$  and therefore the deviance, are functions of this. Hence the deviance is a function only of the fitted model, not of the individual observations, and thus cannot measure model fit.

In more detail, we write

$$\begin{aligned} \sum_{j=1}^n y_j \log \pi_j + (1 - y_j) \log(1 - \pi_j) &= \sum_{j=1}^n y_j \beta x_j - y_j \log \{1 + \exp(x_j^T \beta)\} - (1 - y_j) \log \{1 + \exp(x_j^T \beta)\} \\ &= y^T X \beta + \sum_{j=1}^n \log \{1 + \exp(x_j^T \beta)\}, \end{aligned}$$

from which we see that  $y^T X$ , or equivalently  $X^T y$ , is sufficient for  $\beta$  and that

$$\frac{\partial \ell(\beta)}{\partial \beta} = X^T y - \sum_{j=1}^n x_j^T \frac{e^{x_j^T \beta}}{1 + e^{x_j^T \beta}} = X^T y - X^T \pi(\beta),$$

and that (after a little work) and with  $W = \text{diag}\{\pi_1(1 - \pi_1), \dots, \pi_n(1 - \pi_n)\}$ ,

$$-\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} = X^T W X,$$

which is positive definite when  $X$  has rank  $p$  and all the  $\pi_j$  satisfy  $0 < \pi_j < 1$ . If so, the maximum likelihood estimator is unique and satisfies

$$X^T y = X^T \pi(\hat{\beta}) = X^T \hat{\pi},$$

say. Hence

$$\ell(\hat{\beta}) = \ell\{\pi_1(\hat{\beta}), \dots, \pi_n(\hat{\beta})\} = \sum_{j=1}^n y_j \log \pi_j(\hat{\beta}) + (1 - y_j) \log\{1 - \pi_j(\hat{\beta})\},$$

and thus  $D = -2\ell(\hat{\beta})$  depends only on  $\hat{\pi}_1, \dots, \hat{\pi}_n$ , where  $\hat{\pi}_j = \pi_j(\hat{\beta})$ . It is therefore useless as a measure of fit.

- (b) In this case  $\hat{\pi} = \bar{y} = n^{-1} \sum_{j=1}^n y_j$ . As the data are binary,  $y_j^2 = y_j$  for all  $j$ , and Pearson's statistic

$$P = \sum_{i=1}^n \frac{(y_j - \bar{y})^2}{\bar{y}(1 - \bar{y})} = \frac{1}{\bar{y}(1 - \bar{y})} \left( \sum_{i=1}^n y_j - 2\bar{y} \sum_{i=1}^n y_j + n\bar{y}^2 \right) = \frac{n\bar{y} - n\bar{y}^2}{\bar{y}(1 - \bar{y})} = n$$

is clearly also uninformative about model fit.

## Solution 2

- (a) The likelihood  $L(\beta)$  for discrete responses such as these is a product of probabilities, so  $L(\beta) < 1$  for all  $\beta$ , with logarithm

$$\ell(\beta) = \sum_{j=1}^n y_j \log P(Y_j = 1) + (1 - y_j) \log P(Y_j = 0) = \sum_{j=1}^n y_j x_j^T \beta - \log \left( 1 + e^{x_j^T \beta} \right)$$

after a little algebra.

- (b) The log likelihood can be re-expressed as

$$\begin{aligned} \ell(t\gamma) &= \sum_{j: x_j^T \gamma > 0} \left\{ t x_j^T \gamma - \log \left( 1 + e^{t x_j^T \gamma} \right) \right\} - \sum_{j: x_j^T \gamma < 0} \log \left( 1 + e^{t x_j^T \gamma} \right) \\ &= - \sum_{j: x_j^T \gamma > 0} \log \left( 1 + e^{-t x_j^T \gamma} \right) - \sum_{j: x_j^T \gamma < 0} \log \left( 1 + e^{t x_j^T \gamma} \right). \end{aligned}$$

Both sums here are positive and both tend monotonically down to zero as  $t \rightarrow \infty$ , because  $e^{-t x_j^T \gamma} \rightarrow 0$  when  $x_j^T \gamma > 0$  and  $e^{t x_j^T \gamma} \rightarrow 0$  when  $x_j^T \gamma < 0$ ; recall that none of the  $\gamma^T x_j$  equal zero.

We saw in (a) that  $\ell(\beta) < 0$ , and here we see that  $\ell(t\gamma) \rightarrow 0$  when  $t \rightarrow \infty$ , so the MLE is given by  $\lim_{t \rightarrow \infty} t\gamma$ . Since we cannot have  $\gamma = 0$  (otherwise  $x_j^T \gamma = 0$  for all  $j$ ), some element of the MLE must equal  $\pm\infty$ . This corresponds to a perfect fit of the model to the data (i.e., the fitted probability for every  $y_j = 1$  is 1, and the fitted probability for every  $y_j = 0$  is 0).

- (c) In panel A the 0s and 1s are not separated, so maximum likelihood estimation should work OK.

In panel B there is total separation of the 0s and 1s (i.e., there is a line that separates them perfectly), so the problem found in (c) will arise. In B the log likelihood has a maximum at infinity, leading to divergence of some component of  $\hat{\beta}$ , which we would expect to lead to 'large estimates' of at least one parameter when the iterations stop. In R, for example, the estimates are often  $\pm 36$  or so.