
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 $y_1, \ldots, y_n$ is

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

If the π_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

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

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

$$D = 2\left\{\ell(\tilde{\pi}_1, \ldots, \tilde{\pi}_n) - \ell(\widehat{\pi}_1, \ldots, \widehat{\pi}_n)\right\} = -2\ell(\widehat{\beta}),$$

where we have set

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

We could now note that the logistic regression model is a canonical exponential family model with minimal sufficient statistic $X^{\mathrm{T}}y$, and therefore the maximum likelihood estimators and all derived quantities, including $\ell(\widehat{\beta})$ 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

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

from which we see that $y^{\mathrm{T}}X$, or equivalently $X^{\mathrm{T}}y$, is sufficient for β and that

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

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

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

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

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

say. Hence

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

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

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

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

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 β, with logarithm

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

after a little algebra.

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

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

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

We saw in (a) that $\ell(\beta) < 0$, and here we see that $\ell(t\gamma) \to 0$ when $t \to \infty$, so the MLE is given by $\lim_{t \to \infty} t\gamma$. Since we cannot have $\gamma = 0$ (otherwise $x_j^{\mathrm{T}} \gamma = 0$ 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 $\widehat{\beta}$, 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 ± 36 or so.