Notes on logistic regression

Assume there are independent random variables \( \{Y_1, ..., Y_n\} \), where \( Y_i \sim \text{Binomial}(m_i, \pi_i) \).

The log-likelihood function becomes

\[
l(\pi, y) = \sum_{i=1}^{n} [y_i \log \left( \frac{\pi_i}{1 - \pi_i} \right) + m_i \log(1 - \pi_i)]
\]

Define the logit

\[g(\pi_i) = \log \left( \frac{\pi_i}{1 - \pi_i} \right) = \eta_i = \sum_{j=1}^{k} x_{ij} \beta_j,\]

where \( i = 1, ..., n \) and \( k \) predictors usually include the constant term (intercept). Consider the first derivative of the log-likelihood function with respect to \( \beta_j \)

\[
\frac{\partial l}{\partial \beta_j} = \sum_i \left( \frac{\partial l}{\partial \pi_i} \right) \left( \frac{\partial \pi_i}{\partial \beta_j} \right) = \sum_i \left[ \frac{y_i - m_i \pi_i}{\pi_i(1 - \pi_i)} \right] \frac{\partial \pi_i}{\partial \eta_i} x_{ij} = \sum_i (y_i - m_i \pi_i) x_{ij}
\]

The vectorized form will be

\[
\frac{\partial l}{\partial \beta} = X^T (Y - \mu),
\]

where \( X \in \mathbb{R}^{k \times n} \), \( Y \in \mathbb{R}^n \), \( \mu \in \mathbb{R}^n \) with elements being \( m_i \pi_i \).

Computing

A bit digression for Newton-Raphson. Suppose we search for a root for \( F(x) = 0 \) and start with \( (x_0, F_0) = F(x_0) \). The next point for evaluation will be

\[
x_1 = x_0 + \left( \frac{\partial F}{\partial x} \right)^{-1} \big|_{x_0} (0 - F_0) = x_0 + \left( \frac{\partial x}{\partial F} \right)^{-1} \big|_{x_0} (0 - F_0).
\]

Then, repeat the process using \( x_1 \) as the new starting point. Stop until there is no longer material difference between two steps. In our case, we search roots for Eqn (1), which resembles the normal equation from a least-squares regression. Direct calculation involving \( \beta \) is quite complicated. Instead, we consider a two-stage procedure.

At first, we regard \( Y \) as the response, \( \mu \) as a function of \( \eta \). Suppose current estimate for \( (\beta, \eta, \mu) \) are \( (\hat{\beta}_0, \hat{\eta}_0, \hat{\mu}_0) \). Define the adjusted response

\[
z_0 = \hat{\eta}_0 + (y - \hat{\mu}_0) \left( \frac{dn}{d\mu} \right) (0).
\]

Note: in general \( Y - \hat{\mu} \neq 0 \) unless a saturated model.

At the second stage, we regression \( z_0 \) on covariate \( X \) with weight matrix \( W \), where the “variance” of \( z_0 \) is about

\[
W^{-1} = \left( \frac{dn}{d\mu} \right)^2 V = \text{diag} \{m_i \pi_i (1 - \pi_i)^2\} \text{diag} \{m_i \pi_i (1 - \pi_i)\} = \text{diag} \left\{ \frac{1}{m_i \pi_i (1 - \pi_i)} \right\}
\]

evaluated at \( (\hat{\beta}_0, \hat{\eta}_0, \hat{\mu}_0) \). That is

\[
X^T W X \beta = X^T W Z \Rightarrow \hat{\beta} = (X^T W X)^{-1} X^T W Z.
\]

Based on the new \( \hat{\beta} \), update \( (\eta, \mu) \). Repeat until converge. See Chapter 2 and 4 in Generalized Linear Model, McCullagh & Nelder, 1989, 2nd Ed. Chapman.
Properties of MLE

The $r$th element of the Fisher Information matrix $I(\beta)$ for $\beta$ is

$$ -E(\frac{\partial^2 l}{\partial \beta_r \partial \beta_s}) = E(\frac{\partial l}{\partial \beta_r} \frac{\partial l}{\partial \beta_s}) $$

$$ = \sum_i E[(\frac{\partial l}{\partial \pi_i})^2] \frac{\partial \pi_i}{\partial \beta_r} \frac{\partial \pi_i}{\partial \beta_s} $$

$$ = \sum_i \frac{m_i}{\pi_i(1-\pi_i)} \frac{\partial \pi_i}{\partial \beta_r} \frac{\partial \pi_i}{\partial \beta_s} $$

$$ = \sum_i \frac{m_i}{\pi_i(1-\pi_i)} \left( \frac{d\pi_i}{d\eta} \right)^2 x_{ir} x_{is} $$

$$ = \sum_i m_i \pi_i(1-\pi_i) x_{ir} x_{is} $$

Therefore, $I(\beta) = X^T W X$, where $W = \text{diag}\{m_i \pi_i(1-\pi_i)\}$.

Chapter 7 (Statistical Inference, Casella & Berger, 1990, Duxbury) provides a good summary of results about MLE. The MLE in this case is consistent and asymptotically efficient, i.e. achieves the Cramer-Rao Lower Bound as $n \to \infty$.

$$ E[\hat{\beta} - \beta] = O(n^{-1}) $$

$$ \text{Cov}(\hat{\beta}) = (X^T W X)^{-1} \{1 + O(n^{-1})\} $$

Deviance

For the fitted value,

$$ l(\hat{\pi}, Y) = \sum_i \{y_i \log(\hat{\pi}_i) + (m_i - y_i) \log(1 - \hat{\pi}_i)\} $$

The maximum achievable (saturated) $\hat{\pi}_i = y_i / m_i$. Therefore, the deviance

$$ D(Y, \hat{\pi}) = 2l(\hat{\pi}, Y) - 2l(\hat{\pi}, Y) $$

$$ = 2 \sum_i \{y_i \log(y_i / \hat{\mu}_i) + (m_i - y_i) \log(\frac{m_i - y_i}{m_i - \hat{\mu}_i})\} $$

Score Test

Let $U(\beta) = \frac{\partial l}{\partial \beta}$ and let $I(\beta)$ be Fisher Information matrix. Let $\hat{\beta}_0$ is the MLE under the $H_0$. The test statistic is

$$ U(\hat{\beta}_0)^T I(\hat{\beta}_0)^{-1} U(\hat{\beta}_0) \sim \chi_k^2 $$

where $k$ is the number of constraints under the $H_0$.

Exercise: show for a two-by-two contingency table, the score test is same as the Pearson Chi-square test.
Suppose we have a \((a, 1), (b, 0), (c, 0), (d, 0))\) contingency table, where \(b(1, 0)\) means the frequency count for \((y = 1, x = 0)\) is \(b\). Let \(n = a + b + c + d\).

For Pearson Chi-square test, the test statistic is the sum of four item with the first being

\[
\frac{[a - (a + c)(a + b)/n]^2}{(a + c)(a + b)/n} = \frac{na - (a + c)(a + b)/n}{n(a + c)(a + b)} = \frac{[ad - bc]a}{n(a + c)(a + b)} = \frac{(b + d)(c + d)(ad - bc)^2}{n(a + c)(a + b)(b + d)(c + d)}
\]

Notice that

\((c + d)(b + d) + (a + c)(b + d) + (a + b)(b + d) + (a + b)(a + c) = n(b + d) + (a + b)n = n^2\).

Therefore, the \(X^2 = n(ad - bc)^2/[(a + b)(a + c)(b + d)(c + d)]\).

Follow the Score Test formula on page 17,

\[
\frac{\sum x_i(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{\sum x_iy_i - \sum x_i\bar{y}}{\sqrt{\sum (x_i - \bar{x})^2 - n\bar{x}^2}} = \frac{[(a + b)/n][(c + d)/n][(a + c) - (a + c)^2/n]}{na - (a + b)(a + c) = \sqrt{(a + b)(c + d)/n}[n(a + c) - (a + c)^2]} = \frac{ad - bc}{\sqrt{(a + b)(a + c)(b + d)(c + d)/n}}
\]

In addition, we consider the hypothesis test about equality of two population proportions, where \(\hat{p}_1 = a/(a + c)\), \(\hat{p}_2 = b/(b + d)\), and \(\hat{p} = (a + b)/n\). Therefore, the \(z\)-test statistic becomes

\[
\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})(1/n_1 + 1/n_2)}} = \frac{a/(a + c) - b/(b + d)}{\sqrt{(a + b)(c + d)/[1/(a + c) + 1/(b + d)]}/n^2} = \frac{a(b + d) - b(a + c)}{\sqrt{(a + b)(c + d)/(a + c)(b + d)(b + d) + (a + c)/n^2}} = \frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)/n}}
\]