Logistic Regression:

$Y_1, \ldots, Y_n$ are indep, $Y_i \sim \text{Bernoulli}(\pi_i)$

$E(Y_i) = \pi_i = P(Y_i = 1)$

assume

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta x_i \Rightarrow \pi_i = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}$$

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}, \quad 0 \leq \pi(x) < 1.$$

$\pi(x) = 0$ or 1 for some x, then this model is not appropriate

$$F(w) = \frac{e^w}{1+e^w} \quad \text{logistic } (0,1) \text{ dist}$$

$$\pi(x) = F(\alpha + \beta x)$$

$F(w)$ may be standard normal cdf — probit regression

Gumbel cdf — link fn is called log-log link

estimation

mle

$Y_i \sim \text{Bernoulli}(\pi_i)$

$\pi(x) = F(\alpha + \beta x)$

$F_i = F(\alpha + \beta x_i) = \pi(x_i)$

$$L(\alpha, \beta \mid \underline{y}) = \prod_{i=1}^{n} \pi(x_i)^{y_i} (1-\pi(x_i))^{1-y_i} = \prod_{i=1}^{n} F_i^{y_i}(1-F_i)^{1-y_i}$$

$$\log L = \sum_{i=1}^{n} \left\{ \log(1-F_i) + y_i \log\left(\frac{F_i}{1-F_i}\right) \right\}$$

let $\frac{dF(w)}{dw} = f(w)$ pdf, let $f_i = f(\alpha + \beta x_i)$

$$\frac{\partial}{\partial \alpha} \log(1-F_i) = -\frac{f_i}{1-F_i} = -\frac{F_i f_i}{F_i(1-F_i)}$$

$$\frac{\partial}{\partial \alpha} \log\left(\frac{F_i}{1-F_i}\right) = \frac{f_i}{F_i(1-F_i)}$$

$$\frac{\partial}{\partial \alpha} \log L = \sum_{i=1}^{n} (y_i - F_i) \frac{f_i}{F_i(1-F_i)} \quad — ① '$$

Ill$_y$

$$\frac{\partial}{\partial \beta} L = \sum_{i=1}^{n} (y_i - F_i) \frac{f_i}{F_i(1-F_i)} x_i \quad — ② '$$

For logistic regression with $F(w) = \frac{e^w}{1+e^w}$; $\frac{f_i}{F_i(1-F_i)} = 1$

①', ②' are somewhat simpler

①' $= 0$, ②' $= 0$ and solve for α and β. Solve numerically

Information matrix

$$I(\theta_1, \theta_2) = \begin{pmatrix} -\frac{\partial^2}{\partial \theta_1^2} \log L(\theta_1, \theta_2 | \underset{\sim}{y}) & -\frac{\partial^2}{\partial \theta_1 \partial \theta_2} \log L(\theta_1, \theta_2 | \underset{\sim}{y}) \\ -\frac{\partial^2}{\partial \theta_1 \partial \theta_2} \log L(\theta_1, \theta_2 | \underset{\sim}{y}) & -\frac{\partial^2}{\partial \theta_2^2} \log L(\theta_1, \theta_2 | \underset{\sim}{y}) \end{pmatrix}$$

predictor x
dataset $x_1 \rightarrow x_J$
n_j denote # of Bernoulli obs. at x_j
y_j^* # of successes in these n_j obs
$Y_j^* \sim Bin(n_j, \pi(x_j))$

For logistic regression

$$I(\alpha, \beta) = \begin{pmatrix} \sum_{j=1}^{J} n_j F_j(1-F_j) & \sum_{j=1}^{J} x_j n_j F_j(1-F_j) \\ \sum_{j=1}^{J} x_j n_j F_j(1-F_j) & \sum_{j=1}^{J} x_j^2 n_j F_j(1-F_j) \end{pmatrix}$$

Var. of mle $\hat{\alpha}$ and $\hat{\beta}$ are usually approximated using this matrix

$$I(\alpha, \beta)^{-1} \qquad I(\hat{\alpha}, \hat{\beta}) \qquad (I(\hat{\alpha}, \hat{\beta}))^{-1}.$$

$$= \begin{pmatrix} [se(\hat{\alpha})]^2 & * \\ * & (se(\hat{\beta}))^2 \end{pmatrix}$$

$100(1-\alpha)\%$ CI for β is $\hat{\beta} \pm z_{\frac{\alpha}{2}} se(\hat{\beta})$

$H_0 : \beta = 0$ Wald test stat $Z = \frac{\hat{\beta}}{se(\hat{\beta})} \approx N(0,1)$ if H_0 true and sample size large

H_0 is rejected if $|Z| \geq z_{\frac{\alpha}{2}}$

alter H_0 test using log LRT stat

$$-2 \log \lambda(y^*) = 2 [\log L(\hat{\alpha}, \hat{\beta} | y^*) - L(\hat{\alpha}_0 - 0 | y^*)]$$

$\hat{\alpha}_0$ is mle of α assuming $\beta = 0$

$\hat{\alpha}_0 = \sum_{i=1}^{n} y_i / n = \sum_{j=1}^{J} y_j^* / \sum_{j=1}^{J} n_j$ Under H_0 $-2 \log \lambda \sim \chi_1^2$ reject H_0 if $-2 \log \lambda \geq \chi^2_{1,\alpha}$