The Bernoulli Distribution

$$f(y;\pi) = \pi^{y}(1-\pi)^{1-y}, \quad y=0,1$$

Result 1: $E(y) = \pi$

Proof:

$$E(y) = P(y=0)\cdot 0 + P(y=1)\cdot 1$$

$$= (1-\pi)\cdot 0 + \pi \cdot 1 = \pi$$

Result 2: $Var(y) = \pi(1-\pi)$

Proof:

First note that $Var(y) = E(y^2) - [E(y)]^2$

for all random variables y.

Let us first focus on getting $E(y^2)$.

$$E(y^2) = P(y=0)\cdot 0^2 + P(y=1)\cdot 1^2$$

$$= (1-\pi)\cdot 0^2 + \pi \cdot 1^2 = \pi$$

Therefore,

$$Var(y) = E(y^2) - [E(y)]^2$$

$$= \pi - \pi^2 = \pi(1-\pi)$$

Now let us get the likelihood function of a random sample $y_1, y_2, \cdots, y_n$ drawn from the Bernoulli distribution.

Result 3 : The likelihood function for a rample sample drawn from the Bernoulli distribution is

$$L(\pi; y) = \prod_{i=1}^{n} f(y_i; \pi)$$
$$= \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i}$$

Proof:

$$L(\pi; y) = \prod_{i=1}^{n} f(y_i; \pi)$$
$$= \pi^{y_1}(1-\pi)^{1-y_1} \cdot \pi^{y_2}(1-\pi)^{1-y_2}$$
$$\cdots \pi^{y_n}(1-\pi)^{1-y_n}$$
$$= \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i}$$

Result 4 : The log likelihood function is

$$\log L(\pi; y) = \sum_{i=1}^{n} \left[(1-y_i)\log(1-\pi) + y_i \log(\pi)\right]$$

Proof:

$$\log L(\pi; y) = \log[\pi^{y_1}(1-\pi)^{1-y_1}] +$$
$$\log[\pi^{y_2}(1-\pi)^{1-y_2}] + \cdots + \log[\pi^{y_n}(1-\pi)^{1-y_n}]$$

$$= \sum_{i=1}^{n} \left\{ \log\left[\pi^{y_i} (1-\pi)^{1-y_i} \right] \right\}$$

$$= \sum_{i=1}^{n} \left[y_i \log(\pi) + (1-y_i) \log(1-\pi) \right].$$

Result 5: The Score for the log Likelihood function is

$$s(\pi; y) = \frac{d}{d\pi} \log L(\pi; y)$$

$$= \sum_{i=1}^{n} \frac{y_i - \pi}{\pi(1-\pi)}$$

Proof:

$$s(\pi; y) = \frac{d}{d\pi}\left[\sum_{i=1}^{n} \left[y_i \log(\pi) + (1-y_i) \log(1-\pi) \right] \right]$$

$$= \sum_{i=1}^{n} \left[\frac{d}{d\pi} y_i \log(\pi) + \frac{d}{d\pi} (1-y_i) \log(1-\pi) \right]$$

$$= \sum_{i=1}^{n} \left[y_i \cdot \frac{1}{\pi} + (1-y_i) \frac{1}{(1-\pi)} (-1) \right]$$

$$= \sum_{i=1}^{n} \left[\frac{y_i(1-\pi) + (1-y_i)(-1)\pi}{\pi(1-\pi)} \right]$$

$$= \sum_{i=1}^{n} \left[\frac{y_i - y_i\pi - \pi + y_i\pi}{\pi(1-\pi)} \right]$$

$$= \sum_{i=1}^{n} \left[\frac{y_i - \pi}{\pi(1-\pi)} \right].$$

Result 6: The Hessian of the Log Likelihood function is

$$H(\pi; y) = \frac{d\, s(\pi; y)}{d\pi}$$

$$= \sum_{i=1}^{n} \left[\frac{-y_i}{\pi^2} - \frac{(1-y_i)}{(1-\pi)^2} \right]$$

Proof:

$$H(\pi; y) = \frac{d}{d\pi} \left[\sum_{i=1}^{n} \left(\frac{y_i - \pi}{\pi(1-\pi)} \right) \right]$$

Let us focus on deriving

$$\frac{d}{d\pi} \left(\frac{y_i - \pi}{\pi(1-\pi)} \right) = y_i \frac{d}{d\pi} \pi^{-1}(1-\pi)^{-1} - \frac{d}{d\pi}(1-\pi)^{-1}$$

$$= y_i \left(-\frac{1}{\pi^2(1-\pi)} + \frac{1}{\pi(1-\pi)^2} \right) - \frac{1}{(1-\pi)^2}$$

$$= \frac{-y_i}{\pi^2} - \frac{(1-y_i)}{(1-\pi)^2}$$

Therefore

$$H(\pi; y) = \sum_{i=1}^{n} \left[\frac{-y_i}{\pi^2} - \frac{(1-y_i)}{(1-\pi)^2} \right]$$

Result 7: The Maximum Likelihood Estimator of π is $\hat{\pi} = \sum_{i=1}^{n} y_i / n = \bar{y}$

Proof:

Set the score to zero and solve for $\hat{\pi}$ as in

$$s(\hat{\pi}; y) = \sum_{i=1}^{n} \left(\frac{y_i - \hat{\pi}}{\hat{\pi}(1-\hat{\pi})} \right) = 0$$

$$\sum_{i=1}^{n} \frac{y_i}{\hat{\pi}(1-\hat{\pi})} - \frac{n\hat{\pi}}{\hat{\pi}(1-\hat{\pi})} = 0$$

$$\frac{1}{\hat{\pi}(1-\hat{\pi})} \left[\sum_{i=1}^{n} y_i - n\hat{\pi} \right] = 0$$

$$\Rightarrow \quad \hat{\pi} = \sum_{i=1}^{n} y_i / n = \bar{y}$$

(Note $\bar{y}$ is assumed here to be an interior solution. That is $\bar{y} \neq 0$ or $\neq 1$.)

Result 8: (Second Order Condition) $\hat{\pi} = \bar{y}$ provides a maximum to the Log Likelihood function.

Proof:

For this to be so the Hessian (second derivative of the Log Likelihood function) must be negative when evaluated at the maximum likelihood estimate $\hat{\pi} = \bar{y}$ (assuming an interior solution where $\bar{y} \neq 0$ or $\bar{y} \neq 1$).

$$H(\hat{\pi}; y) = \sum_{i=1}^{n} \left[\frac{-y_i}{\hat{\pi}^2} - \frac{(1-y_i)}{(1-\hat{\pi})^2} \right]$$

$$= -\frac{n\bar{y}}{\hat{\pi}^2} - \frac{(n-n\bar{y})}{(1-\hat{\pi})^2}$$

$$= -\frac{n\bar{y}}{\bar{y}^2} - \frac{(n-n\bar{y})}{(1-\bar{y})^2}$$

$$= -\frac{n}{\bar{y}} - \frac{n}{(1-\bar{y})}$$

$$= \frac{-n(1-\bar{y}) - n\bar{y}}{\bar{y}(1-\bar{y})}$$

$$= -\frac{n}{\bar{y}(1-\bar{y})} < 0$$

for interior solutions ($\bar{y} \neq 0$ or $\bar{y} \neq 1$). Therefore, the maximum likelihood estimate of π does maximize the likelihood function.

The Probit and Logit Models

They are conditional probability versions of the Bernoulli Distribution.

What we want to do now is make our Bernoulli distribution dependent on some exogenous variables $X_1, X_2, \cdots, X_k$.

The Likelihood Function then becomes a conditional probability model as in

$$\Pi_i = F(\underset{\sim}{X}_i' \underset{\sim}{\beta})$$

where $F(\underset{\sim}{X}_i' \underset{\sim}{\beta})$ denotes a cumulative

probability model. The two popular CDF's include the standard normal CDF

$$F(\underset{\sim}{X}_i' \underset{\sim}{\beta}) = \int_{-\infty}^{\underset{\sim}{X}_i' \underset{\sim}{\beta}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v^2}{2}\right) dv$$

and the standard logistic CDF

$$F(\underset{\sim}{X}_i' \underset{\sim}{\beta}) = \frac{\exp(\underset{\sim}{X}' \underset{\sim}{\beta})}{1+\exp(\underset{\sim}{X}' \underset{\sim}{\beta})} = \int_{-\infty}^{\underset{\sim}{X}_i' \underset{\sim}{\beta}} \frac{\exp(v)}{(1+\exp(v))^2} dv$$

The Likelihood Function for $y_1, y_2, \cdots, y_n$

is then

$$L(\underset{\sim}{X}, \underset{\sim}{\beta}; y) = \prod_{i=1}^{n} F(\underset{\sim}{X}_i' \underset{\sim}{\beta})^{y_i} (1 - F(\underset{\sim}{X}_i' \underset{\sim}{\beta}))^{1-y_i}$$

with Log Likelihood Function

$$\log[L(X, \underset{\sim}{\beta}; y)]$$

$$= \sum_{i=1}^{n} \left\{ y_i \log[F(\underset{\sim}{X}_i' \underset{\sim}{\beta})] + (1-y_i) \log[1 - F(\underset{\sim}{X}_i' \underset{\sim}{\beta})] \right\}$$

But from Maximum Likelihood Theory we know that

$$\underset{\sim}{\hat{\beta}} \overset{asy}{\sim} N\left(\underset{\sim}{\beta}, -E\left[\frac{\partial^2 \log L}{\partial \underset{\sim}{\beta} \partial \underset{\sim}{\beta}'}\right]^{-1}_{\big| \underset{\sim}{\beta} = \underset{\sim}{\hat{\beta}}}\right) = N(\underset{\sim}{\beta}, I(\underset{\sim}{\hat{\beta}})^{-1})$$

or, asymptotically

$$\underset{\sim}{\hat{\beta}} \overset{asy}{\sim} N\left(\underset{\sim}{\beta}, -\left[\frac{\partial^2 \log L}{\partial \underset{\sim}{\beta} \partial \underset{\sim}{\beta}'}\right]^{-1}_{\big| \underset{\sim}{\hat{\beta}} = \underset{\sim}{\beta}}\right) = N(\underset{\sim}{\beta}, -H(\underset{\sim}{\hat{\beta}})^{-1})$$

Numerical Optimization

1) The Newton-Raphson iterative method can solve for $\hat{\beta}$ by the formula

$$\hat{\beta}_{i+1} = \hat{\beta}_i - \left[\frac{\partial^2 \log L}{\partial \beta \partial \beta'}\right]^{-1}_{\beta=\hat{\beta}} \cdot \left[\frac{\partial \log L}{\partial \beta}\right]_{\beta=\hat{\beta}}$$

2) The Method of Scoring iterative method can solve for $\hat{\beta}$ by the formula

$$\hat{\beta}_{i+1} = \hat{\beta}_i - \left[E\left(\frac{\partial^2 \log L}{\partial \beta \partial \beta'}\right)_{\beta=\hat{\beta}}\right]^{-1} \cdot \left[\frac{\partial \log L}{\partial \beta}\right]_{\beta=\hat{\beta}}$$

where

$$E\left[\frac{\partial^2 \log L}{\partial \beta \partial \beta'}\right] = -\sum_{i=1}^{n} \frac{f(X_i'\beta)}{F(X_i'\beta)[1-F(X_i'\beta)} X_i X_i'$$

For the Probit Model we have the Score Vector

$$\frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{n}\left[y_i \frac{f(X_i'\beta)}{F(X_i'\beta)} - (1-y_i)\frac{f(X_i'\beta)}{1-F(X_i'\beta)}\right] X_i$$

and the Hessian matrix

$$\frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\sum_{i=1}^{n} f(\mathbf{x}_i'\boldsymbol{\beta}) \Bigg[y_i \frac{f(\mathbf{x}_i'\boldsymbol{\beta}) + (\mathbf{x}_i'\boldsymbol{\beta}) F(\mathbf{x}_i'\boldsymbol{\beta})}{[F(\mathbf{x}_i'\boldsymbol{\beta})]^2}$$

$$+ (1-y_i) \frac{f(\mathbf{x}_i'\boldsymbol{\beta}) - (\mathbf{x}_i'\boldsymbol{\beta})[1-F(\mathbf{x}_i'\boldsymbol{\beta})]}{[1-F(\mathbf{x}_i'\boldsymbol{\beta})]^2} \Bigg] \mathbf{x}_i \mathbf{x}_i'$$

For the Logit Model we have the Score vector

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} y_i \frac{1}{(1+\exp(\mathbf{x}_i'\boldsymbol{\beta}))} \mathbf{x}_i$$

$$- \sum_{i=1}^{n} (1-y_i) \frac{1}{(1+\exp(\mathbf{x}_i'\boldsymbol{\beta}))} \mathbf{x}_i$$

$$= \sum_{i=1}^{n} [y_i F(-\mathbf{x}_i'\boldsymbol{\beta}) - (1-y_i) F(\mathbf{x}_i'\boldsymbol{\beta})] \mathbf{x}_i$$

and Hessian Matrix

$$\frac{\partial^2 \log L}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = -\sum_{i=1}^{n} \frac{\exp(-\mathbf{x}_i'\boldsymbol{\beta})}{[1+\exp(-\mathbf{x}_i'\boldsymbol{\beta})]^2} \mathbf{x}_i \mathbf{x}_i'$$

$$= -\sum_{i=1}^{n} f(\mathbf{x}_i'\boldsymbol{\beta}) \mathbf{x}_i \mathbf{x}_i'$$

For the Probit and Logit Models it can be shown that the Hessian, $H = \frac{\partial^2 \log L}{\partial \underset{\sim}{\beta} \partial \underset{\sim}{\beta}'}$

is negative definite for all values of $\underset{\sim}{\beta}$.

Consequently, the Newton-Raphson iteration is gauranteed to converge (regardless of starting point) to a unique optimum $\hat{\underset{\sim}{\beta}}$

eventually despite the fact that the Normal equations associated with Log Likelihood function are not analytically solvable!

Two Alternatives for Getting the Variance-Covariance Matrix of the ML Estimates, $\hat{\underset{\sim}{\beta}}$

1) $\text{asy } \widehat{Var}(\hat{\underset{\sim}{\beta}}) = \left[-\hat{H} \right]^{-1}$

2) $\text{asy } \widehat{Var}(\hat{\underset{\sim}{\beta}}) = \left[-E(H) \Big|_{H=\hat{H}} \right]^{-1}$

assuming that $E(H)$ can be evaluated. Fortunately, for the Probit and Logit models, it can.

For more discussion see Ch. 4 in W & B.

In the case that one is confident that $\pi_i = \exp(x_i' \underset{\sim}{\beta})$ is true but is not confident in the choice of likelihood function there is the option of computing the

Huber/White Quasi-Maximum Variance-Covariance matrix as

$$\text{asy}\,\widehat{\text{Var}}_{QML}(\underset{\sim}{\hat{\beta}}) = \hat{H}^{-1} \underset{\sim}{\hat{g}}\underset{\sim}{\hat{g}}' \hat{H}^{-1}$$

where $\underset{\sim}{\hat{g}}$ = the score vector of all of the observations as in the previously derived

$$\left.\frac{\partial \log L}{\partial \underset{\sim}{\beta}}\right|_{\underset{\sim}{\beta} = \underset{\sim}{\hat{\beta}}} = \underset{\sim}{\hat{g}}.$$