Review of Newton-Raphson Iteration (Newton's Method)

<u>Goal</u>: Maximize $\ln L(\theta)$ using iterative scheme

θ - scalar.

<u>Idea</u>: 2nd order TS approximation of $\ln L(\theta)$ about

arbitrary point θ_1 :

doesn't depend on θ !

$$\ln L(\theta) = \ln L(\theta_1) + \frac{\partial \ln L(\theta_1)}{\partial \theta}(\theta-\theta_1) + \frac{1}{2}\frac{\partial^2 \ln L(\theta_1)(\theta-\theta_1)^2}{\partial \theta^2} + \text{error}$$

Now maximize 2nd order TSE

$$\max_{\theta} \quad \ln L(\theta_1) + \frac{\partial \ln L(\theta_1)}{\partial \theta}(\theta-\theta_1) + \frac{1}{2}\frac{\partial^2 \ln L(\theta_1)}{\partial \theta^2}(\theta-\theta_1)^2$$

F.O.C's

$$\frac{d \ln L(\theta_1)}{d\theta} + \frac{d^2 \ln L(\theta_1)}{\partial \theta^2}(\hat{\theta}_2 - \theta_1) = 0$$

$$\Rightarrow \quad \hat{\theta}_2 = \theta_1 - \left[\frac{d^2 \ln L(\theta_1)}{d\theta^2}\right]^{-1}\frac{d \ln L(\theta_1)}{d\theta}$$

or

$$\hat{\theta}_2 = \theta_1 - H(\theta_1)^{-1} * S(\theta_1)$$

The iterative scheme is then

$$\hat{\theta}_{n+1} = \hat{\theta}_n - H(\hat{\theta}_n)^{-1} \cdot S(\hat{\theta}_n)$$

and iteration stops when

$$S(\hat{\theta}_n) \approx 0$$

and

$$\hat{\theta}_{n+1} \approx \hat{\theta}_n .$$

<u>Remarks</u>

(i) In the N-R scheme it can be shown that convergence occurs in one step (iteration) if $\ln L(\theta)$ is quadratic in θ

(ii) The N-R iteration scheme computes an estimate of the information matrix, $I(\theta) = -E[H(\theta)]$, as a by product of the algorithm. That is, at convergence, $\hat{\theta}_n = \hat{\theta}_{MLE}$ and

$$-H(\hat{\theta}_n)^{-1} = I(\hat{\theta}_{MLE})^{-1} = \hat{\mathrm{var}}(\hat{\theta}_{ML}$$

(iii) In the context of MLE, the N-R algorithm can be simplified using the result

$$I(\theta) = -E\left[\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right] = E\left[S(\theta)^2\right]$$

$$= E\left[\left(\frac{\partial \ln L(\theta)}{\partial \theta}\right)^2\right]$$

(a) Method of Scoring

replace $H(\hat{\theta}_n)$ with $E[H(\hat{\theta}_n)]$

if $E[H(\hat{\theta}_n)]$ can be computed easily.

Then $-E[H(\hat{\theta}_n)]^{-1} = \hat{var}(\hat{\theta}_{MLE})$.

(b) BHHH method

replace $H(\hat{\theta}_n) = \dfrac{\partial^2 \ln L(\hat{\theta}_n)}{\partial \theta} = \sum_{i=1}^{n} \dfrac{\partial^2 \ln l}{\partial \theta}$

with $-\sum_{i=1}^{n} \left(\dfrac{\partial \ln L(\hat{\theta}_n)}{\partial \theta}\right)^2$

The advantage here is that we only need 1st derivatives

Also $\left(\sum_{i=1}^{n} \left(\dfrac{\partial \ln L(\hat{\theta}_n)}{\partial \theta}\right)^2\right)^{-1} = I(\hat{\theta}_{MLE})^{-1}$

$= \hat{var}(\hat{\theta}_{MLE})$

Practical considerations

(1) Need to choose starting values for the iteration

Choose θ_1

(2) Compute analytical or numerical derivatives for

$S(\theta)$ and $H(\theta)$

(i) Analytical derivatives give faster iterations

(ii) Numerical derivatives less accurate (in pr) but generally very good

(iii) Analytic derivatives are more trouble to compute — requires programming

MATLAB, GAUSS have procedures to compute numerical derivatives.

Application to Logit & Probit Estimation

$$y_i^* = x_i'\beta + \epsilon_i$$

$\epsilon_i \sim$ iid with CDF F & pdf f

$$\left.\begin{aligned} y_i &= 1 \text{ if } y_i^* > 0 \\ &= 0 \text{ if } y_i^* \le 0 \end{aligned}\right\} \Pr(Y_i = 1 \mid x_i) = F(x_i'\beta)$$

Logit: $\epsilon_i \sim$ logistic, $F(x) = \Lambda(x)$, $f(x) = \lambda(x)$

Probit: $\epsilon_i \sim \cancel{\text{Prob}}\ N(0,1)$, $F(x) = \Phi(x)$, $f(x) = \varphi(x)$

$$\left(\text{Recall}, \ \Lambda(x) = \frac{e^x}{e^x + 1}\right)$$

Given an iid sample

$$L(\beta \mid y, x) = \prod_{i=1}^{n} F(x_i'\beta)^{y_i} (1 - F(x_i'\beta))^{1-y_i}$$

$$\ln L(\beta \mid y, x) = \sum_{i=1}^{n} \left\{ y_i \ln F(x_i'\beta) + (1-y_i)\ln(1 - F(x_i'\beta)) \right.$$

F.O.C's are

$$\frac{\partial \ln L(\beta \mid y, x)}{\partial \beta} = \sum_{i=1}^{n} \left\{ y_i \frac{f(x_i'\beta)}{F(x_i'\beta)} \cdot x_i + \frac{(1-y_i)(-f(x_i'\beta))\, x_i}{1 - F(x_i'\beta)} \right.$$

1.e.

$$S(\hat{\beta}_{MLE}) = \sum_{i=1}^{\hat{}} \left\{ y_i \frac{f(x_i'\hat{\beta}_{MLE})}{F(x_i'\hat{\beta}_{MLE})} - (1-y_i)\frac{f(x_i'\hat{\beta}_{MLE})}{1-F(x_i'\hat{\beta}_{MLE})} \right\} x_i = 0$$

In the logit & probit models the F.O.C's are K nonlinear equations are can only be solved numerically. Newton's method gives

$$\hat{\beta}_{n+1} = \hat{\beta}_n - H(\hat{\beta}_n)^{-1} S(\hat{\beta}_n)$$

where

$$H(\hat{\beta}_n) = \frac{\partial^2 \ln L(\hat{\beta}_n)}{\partial\beta\,\partial\beta'}$$

$$S(\hat{\beta}_n) = \frac{\partial \ln L(\hat{\beta}_n)}{\partial\beta}$$

Note: Starting values can be computed from the linear probability model

Remarks

(1) In the logit model, it can be shown that

$$H(\beta) = -\sum_1^{\hat{}} \Lambda(x_i'\beta)(1-\Lambda(x_i'\beta))\,\underset{\sim}{x_i}\,\underset{\sim}{x_i}'$$

which is independent of y_i and is always negative definite. Hence $\ln L(\beta)$ is globally concave for the logit

Also, it is easy to show that

$$S(\beta) = \sum_{i}^{n} (y_i - \Lambda(x_i'\beta))x_i$$

Hence, NR iteration can be conducted with "analytic derivatives" and will converge very quickly.

(2) In the probit model, it can be shown that

$$S(\beta) = \sum_{i=1}^{n} m_i x_i$$

where

$$m_i = \frac{q_i \,\varphi(q_i x_i'\beta)}{\Phi(q_i x_i'\beta)},$$

$$q_i = 2y_i - 1,$$

and

$$H(\beta) = -\sum_{i=1}^{n} m_i(m_i + x_i'\beta)x_i x_i'.$$

~~Further~~ Hence NR iteration can be conducted with analytic derivatives ~~and~~. Also, it can be shown that $H(\beta)$ is negative definite for all β $\Rightarrow$ $\ln L(\beta)$ is

For both logit e Probit

$$\hat{Var}(\hat{\beta}_{MLE}) = -H(\hat{\beta}_{MLE})^{-1}$$

can be computed analytically.

[give example of Logit e Probit Estimation]

Interpreting the Model

1. Marginal Effects

$$\frac{\partial \Pr(Y_i = 1 \mid \underset{\sim}{x}_i)}{\partial x_{ki}} = \frac{\partial F(\underset{\sim}{x}_i' \underset{\sim}{\beta})}{\partial (\underset{\sim}{x}_i' \underset{\sim}{\beta})} \cdot \frac{\partial (\underset{\sim}{x}_i' \underset{\sim}{\beta})}{\partial x_{ki}}$$

$$= f(\underset{\sim}{x}_i' \underset{\sim}{\beta}) \cdot \beta_k$$

Depends on (1) value of x_i

(2) value of β

(3) value of β_k

<u>Note</u>: Since $f(\cdot) > 0 \Rightarrow$ sign of β_k indicates sign of marginal effect.

Estimated marginal effect

$$\frac{\partial \widehat{\Pr}(Y_i = 1 \mid x_i)}{\partial x_{ki}} = f(\underset{\sim}{x}_i' \underset{\sim}{\hat{\beta}}_{MLE}) \cdot \hat{\beta}_{MLE_k}$$

<u>Note</u>: A std. error for this marginal effect can be computed using the delta method

– see Greene pg. 885

Measuring Goodness of Fit

(1) McFadden's Likelihood Ratio Index

((Reduces to usual R^2 in linear regression model!))

$$LRI = 1 - \frac{\ln L(\hat{\beta}_{MLE})}{\ln L(\text{all slopes} = \text{zero})}$$

$LRI \approx 0$ if model with all slopes $= 0$

fits the data as well as the model

with estimated slopes

LRI approaches 1 the better the fit.

(2) Prediction Table

2×2 Table of hits & misses

of a prediction rule

Classify $\hat{y}_i = 1$ if $\hat{Pr}(y_i = 1 | x_i) >$ cutoff probabilit