Discrete Choice Modeling

William Greene

Stern School of Business

New York University

6. NONLINEAR MODELS

Agenda

- Nonlinear Models
- Estimation Theory for Nonlinear Models
  - Estimators
  - Properties
  - M Estimation
    - Nonlinear Least Squares
    - Maximum Likelihood Estimation
  - GMM Estimation
    - Minimum Distance Estimation
    - Minimum Chi-square Estimation
- Computation – Nonlinear Optimization
  - Nonlinear Least Squares
  - Newton-like Algorithms; Gradient Methods
- (Background: JW, Chapters 12-14, Greene, Chapters 7,12-14)

What is the ‘Model?’

- Unconditional ‘characteristics’ of a population
- Conditional moments: $E[g(y)|x]$:
  median, mean, variance, quantile, correlations, probabilities…
- Conditional probabilities and densities
- Conditional means and regressions
- Fully parametric and semiparametric specifications
  - Parametric specification: Known up to parameter $\boldsymbol{\theta}$
  - Conditional means: $E[y|\mathbf{x}] = m(\mathbf{x}, \boldsymbol{\theta})$

What is a *Nonlinear* Model?

- Model: E[g(y)|**x**] = m(**x**,**θ**)
- Objective:
  - Learn about **θ** from **y**, **X**
  - Usually “estimate” **θ**
- Linear Model: Closed form; $\hat{\boldsymbol{\theta}}$ = h(**y**, **X**)
- Nonlinear Model
  - Not necessarily wrt m(**x**,**θ**). E.g., y = exp(**θ’x + ε**)
  - Wrt estimator: Implicitly defined. h(**y**, **X,** $\hat{\boldsymbol{\theta}}$)=0, E.g., E[y|**x**]= exp(**θ’x**)

What is an Estimator?

- Point and Interval

$$\hat{\theta} = f(\text{data} \,|\, \text{model})$$

$$I(\hat{\theta}) = \hat{\theta} \pm \text{sampling variability}$$

- Classical and Bayesian

$$\hat{\theta} = E[\theta \,|\, \text{data}, \text{prior } f(\theta)] = \text{expectation from posterior}$$

$$I(\hat{\theta}) = \text{ narrowest interval from posterior density}$$
$$\text{containing the specified probability (mass)}$$

Parameters

- Model parameters
- The parameter space: Estimators of ‘parameters’
- The true parameter(s)

$$\text{Example : } f(y_i \mid \mathbf{x}\boldsymbol{\beta}_i)\,\mathbf{x} = \frac{\exp(-y_i/\theta_i)}{\theta_i}, \; \theta_i = \exp(\;'\;_i)$$

$$\text{Model parameters : } \boldsymbol{\beta}$$

$$\text{Conditional Mean: } E(y_i \mid \mathbf{x}\boldsymbol{\beta}_i)\,\mathbf{x} = \theta_i = \exp(\;'\;_i)$$

The Conditional Mean Function

$m(\mathbf{x},\theta_0) = E[y \mid \mathbf{x}]$ for some $\boldsymbol{\theta}_0$ in $\boldsymbol{\Theta}$.

A property of the conditional mean:

$E_{y,x}(y - m(\mathbf{x},\theta))^2$ is minimized by $E[y \mid \mathbf{x}]$

(Proof, pp. 343-344, JW)

M Estimation

Classical estimation method

$$\hat{\theta} = \arg\min \frac{1}{n}\sum_{i=1}^{n} q(\mathbf{data}_i, \theta)$$

$$\text{Example : Nonlinear Least squares}$$

$$\hat{\theta} = \arg\min \frac{1}{n}\sum_{i=1}^{n} [y_i \text{-} E(y_i \,|\, \mathbf{x}_i, \theta)]^2$$

An Analogy Principle for M Estimation

The estimator $\hat{\theta}$ minimizes $\overline{q} = \frac{1}{n}\sum_{i=1}^{n} q(\text{data}_i, \theta)$

The true parameter θ_0 minimizes $q^* = E[q(\text{data}, \theta)]$

The weak law of large numbers:

$$\overline{q} = \frac{1}{n}\sum_{i=1}^{n} q(\text{data}_i, \theta) \xrightarrow{P} q^* = E[q(\text{data}, \theta)]$$

Estimation

$$\bar{q} = \frac{1}{n}\sum_{i=1}^{n} q(\text{data}_i, \theta) \xrightarrow{P} q^* = E[q(\text{data}, \theta)]$$

Estimator $\hat{\theta}$ minimizes $\bar{q}$

True parameter θ_0 minimizes q^*

$$\bar{q} \xrightarrow{P} q^*$$

Does this imply $\hat{\theta} \xrightarrow{P} \theta_0$?

Yes, if ...

Identification

Uniqueness :

If $\theta_1 \neq \theta_0$, then $m(x,\theta_1) \neq m(x,\theta_0)$ for some x

Examples in which this property is not met:

(1) (Multicollinearity)

(2) (Need for normalization) $E[y|x] = m(\beta' x/\sigma)$

(3) (Indeterminacy) $m(x,\theta) = \beta_1 + \beta_2 x + \beta_3 x^{\beta_4}$ when $\beta_3 = 0$

Continuity

$q(\text{data}_i, \theta)$ is

(a) Continuous in θ for all data_i and all θ

(b) Continuously differentiable. First derivatives are also continuous

(c) Twice differentiable. Second derivatives must be nonzero, though they need not be continuous functions of θ. (E.g. Linear LS)

Consistency

$$\bar{q} = \frac{1}{n}\sum_{i=1}^{n} q(\text{data}_i, \theta) \xrightarrow{P} q^* = E[q(\text{data}, \theta)]$$

Estimator $\hat{\theta}$ minimizes $\bar{q}$

True parameter θ_0 minimizes q^*

$$\bar{q} \xrightarrow{P} q^*$$

Does this imply $\hat{\theta} \xrightarrow{P} \theta_0$?

Yes. Consistency follows from identification and continuity with the other assumptions

Asymptotic Normality of M Estimators

First order conditions:

$$\frac{\partial (1/n)\Sigma_{i=1}^{N} q(\text{data}_i, \hat{\theta})}{\partial \hat{\theta}} = 0$$

$$= \frac{1}{n}\Sigma_{i=1}^{N} \frac{\partial q(\text{data}_i, \hat{\theta})}{\partial \hat{\theta}}$$

$$= \frac{1}{n}\Sigma_{i=1}^{N} \mathbf{g}(\text{data}_i, \hat{\theta}) = \overline{\mathbf{g}}(\text{data}, \hat{\theta})$$

For any $\hat{\theta}$, this is the mean of a random sample. We apply Lindberg-Feller CLT to assert the limiting normal distribution of $\sqrt{n}\ \overline{\mathbf{g}}(\text{data}, \hat{\theta})$.

Asymptotic Normality

A Taylor series expansion of the derivative

$$\bar{\mathbf{g}}(\text{data},\hat{\theta}) = \bar{\mathbf{g}}(\text{data},\theta_0) + \bar{\mathbf{H}}(\tilde{\theta})(\hat{\theta}-\theta_0) = 0$$

$$\bar{\mathbf{H}}(\tilde{\theta}) = \frac{1}{n}\Sigma_{i=1}^{n}\frac{\partial^2 q(\text{data}_i,\tilde{\theta})}{\partial\tilde{\theta}\partial\tilde{\theta}'}$$

$\tilde{\theta}$ = some point between $\hat{\theta}$ and θ_0

Then, $(\hat{\theta}-\theta_0) = [\bar{\mathbf{H}}(\tilde{\theta})]^{-1}\bar{\mathbf{g}}(\text{data},\theta_0)$ and

$$\sqrt{n}\ (\hat{\theta}-\theta_0) = [\bar{\mathbf{H}}(\tilde{\theta})]^{-1}\ \sqrt{n}\ \bar{\mathbf{g}}(\text{data},\theta_0)$$

Asymptotic Normality

$$\sqrt{n}\,(\hat{\theta}-\theta_0)=[\bar{\mathbf{H}}(\tilde{\theta})]^{-1}\,\sqrt{n}\,\bar{\mathbf{g}}(\text{data},\theta_0)$$

$[\bar{\mathbf{H}}(\tilde{\theta})]^{-1}$ converges to its expectation (a matrix)

$\sqrt{n}\,\bar{\mathbf{g}}(\text{data},\theta_0)$ converges to a normally distributed vector (Lindberg-Feller)

Implies limiting normal distribution of $\sqrt{n}\,(\hat{\theta}-\theta_0)$.

Limiting mean is 0.

Limiting variance to be obtained.

Asymptotic distribution obtained by the usual means.

Asymptotic Variance

$$\hat{\theta} \xrightarrow{a} \theta_0 + [\bar{\mathbf{H}}(\tilde{\theta})]^{-1} \ \bar{\mathbf{g}}(\text{data},\theta_0)$$

Asymptotically normal

Mean $= \theta_0$

$$\text{Asy.Var}[\hat{\theta}] = [\bar{\mathbf{H}}(\theta_0)]^{-1} \ \text{Var}[\bar{\mathbf{g}}(\text{data},\theta_0)] \ [\bar{\mathbf{H}}(\theta_0)]^{-1}$$

(A sandwich estimator, as usual)

What is $\text{Var}[\bar{\mathbf{g}}(\text{data},\theta_0)]$?

$$\frac{1}{n}E[\mathbf{g}(\text{data}_i,\theta_0)\mathbf{g}(\text{data}_i,\theta_0)']$$

Not known what it is, but it is easy to estimate.

$$\frac{1}{n}\times\left[\frac{1}{n}\Sigma_{i=1}^{n}\mathbf{g}(\text{data}_i,\hat{\theta})\mathbf{g}(\text{data}_i,\hat{\theta})'\right]$$

Estimating the Variance

$$\text{Asy.Var}[\hat{\theta}] = [\bar{\mathbf{H}}(\theta_0)]^{-1}\ \text{Var}[\bar{\mathbf{g}}(\text{data},\theta_0)]\ [\bar{\mathbf{H}}(\theta_0)]^{-1}$$

$$\text{Estimate }\ [\bar{\mathbf{H}}(\theta_0)]^{-1}\ \text{ with }\ \left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 m(\text{data}_i,\hat{\theta})}{\partial\hat{\theta}\partial\hat{\theta}'}\right]$$

$$\text{Estimate Var}[\bar{\mathbf{g}}(\text{data},\theta_0)]\ \text{ with }\ \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial m(\text{data}_i,\hat{\theta})}{\partial\hat{\theta}}\right)\left(\frac{\partial m(\text{data}_i,\hat{\theta})}{\partial\hat{\theta}'}\right)\right]$$

E.g., if this is linear least squares, $(1/2)\Sigma_{i=1}^{n}(y_i\text{-}x_i'\beta)^2$

$$m(\text{data}_i,\hat{\theta}) = (1/2)(y_i - x_i'b)^2$$

$$\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 m(\text{data}_i,\hat{\theta})}{\partial\hat{\theta}\partial\hat{\theta}'}\right] = (\mathbf{X}'\mathbf{X}/n)^{-1}$$

$$\frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial m(\text{data}_i,\hat{\theta})}{\partial\hat{\theta}}\right)\left(\frac{\partial m(\text{data}_i,\hat{\theta})}{\partial\hat{\theta}'}\right)\right] = (1/n^2)\Sigma_{i=1}^{N}e_i^2\mathbf{x}_i\mathbf{x}_i'$$

Nonlinear Least Squares

Gauss-Marquardt Algorithm

$$q_i = \text{ the conditional mean function}$$

$$= m(\mathbf{x}_i, \theta)$$

$$\mathbf{g}_i = \frac{\partial m(\mathbf{x}_i, \theta)}{\partial \theta} = \mathbf{x}_i^0 = \text{'pseudo} - \text{regressors'}$$

Algorithm - iteration

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} + [\mathbf{X^0{'}X^0}]^{-1}\mathbf{X^0{'}e^0}$$

Application - Income

German Health Care Usage Data, 7,293 Individuals, Varying Numbers of Periods
Variables in the file are

Data downloaded from Journal of Applied Econometrics Archive. This is an unbalanced panel with 7,293 individuals. They can be used for regression, count models, binary choice, ordered choice, and bivariate binary choice. This is a large data set. There are altogether 27,326 observations. The number of observations ranges from 1 to 7. (Frequencies are: 1=1525, 2=2158, 3=825, 4=926, 5=1051, 6=1000, 7=987). Note, the variable NUMOBS below tells how many observations there are for each person. This variable is repeated in each row of the data for the person. (Downloaded from the JAE Archive)

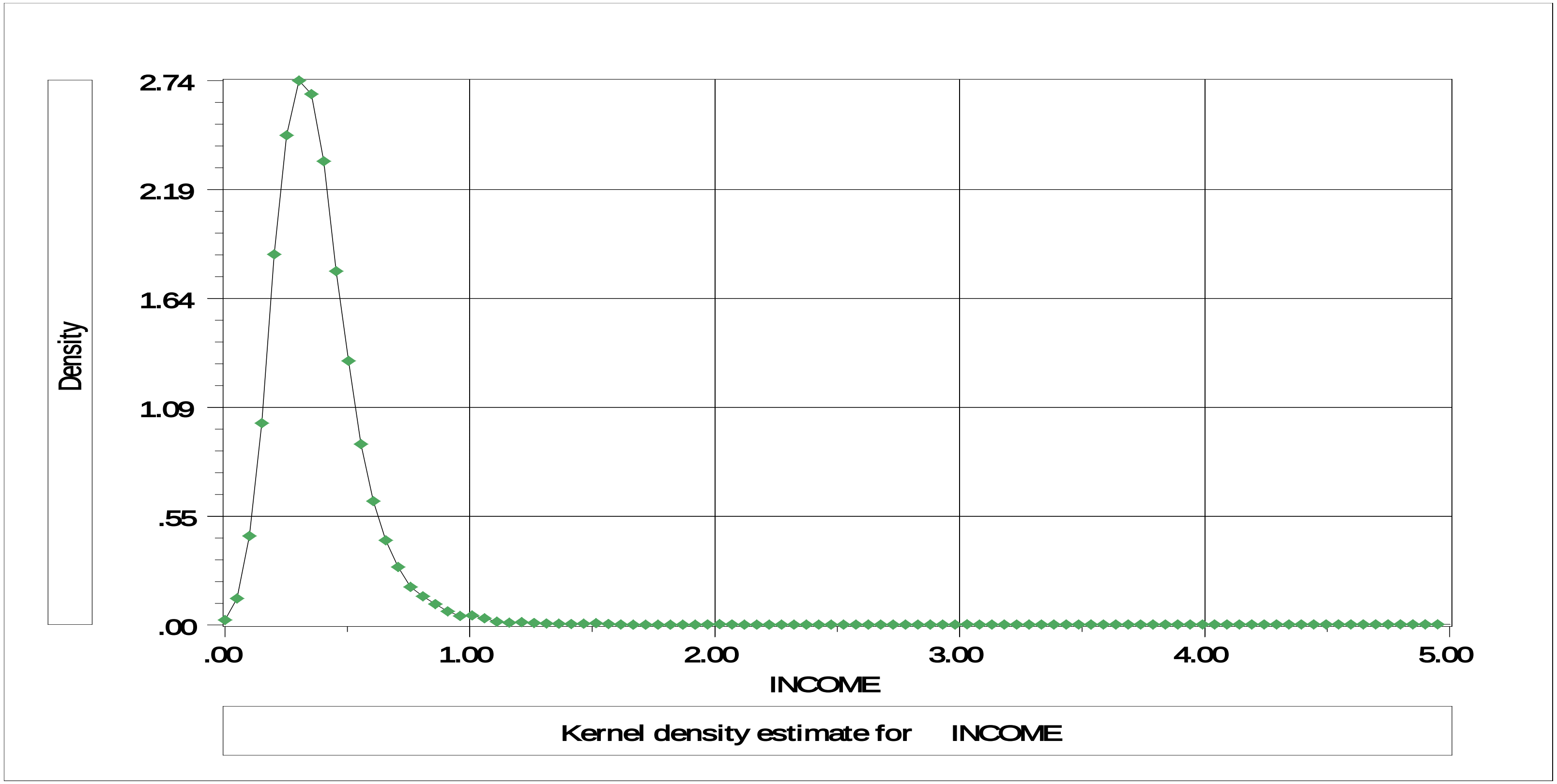
HHNINC = household nominal monthly net income in German marks / 10000.
(4 observations with income=0 were dropped)

HHKIDS = children under age 16 in the household = 1; otherwise = 0

EDUC = years of schooling

AGE = age in years

Income Data


Density
2.74
2.19
1.64
1.09
.55
.00
.00
1.00
2.00
3.00
4.00
5.00
INCOME
Kernel density estimate for INCOME

Exponential Model

$$f(\text{Income} \mid \text{Age}, \text{Educ}, \text{Married})$$

$$= \frac{1}{\theta_i} \exp\left(\frac{-\text{HHNINC}_i}{\theta_i} \right)$$

$$\theta_i = \exp(a_0 + a_1\text{Educ} + a_2\text{Married} + a_3\text{Age})$$

$$E[\text{HHNINC} \mid \text{Age}, \text{Educ}, \text{Married}] = \theta_i$$

Starting values for the iterations:

$$E[y_i \mid \text{nothing else}] = \exp(a_0)$$

$$\text{Start } a_0 = \log\overline{\text{HHNINC}},\ a_1 = a_2 = a_3 = 0$$

Conventional Variance Estimator

$$\frac{\Sigma_{i=1}^{n}[y_i - m(x_i,\hat{\theta})]^2}{n - \#\,\text{parameters}}(\mathbf{X}^{\mathbf{0}\prime}\mathbf{X}^{\mathbf{0}})^{-1}$$

Sometimes omitted.

Estimator for the M Estimator

$$q_i = (1/2)[y_i - \exp(\mathbf{x}_i'\beta)]^2 = (1/2)(y_i - \theta_i)^2$$

$$\mathbf{g}_i = -e_i\theta_i\mathbf{x}_i'$$

$$\mathbf{H}_i = y_i\theta_i\mathbf{x}_i\mathbf{x}_i'$$

Estimator is $[\Sigma_{i=1}^{N}\mathbf{H}_i]^{-1}[\Sigma_{i=1}^{N}\mathbf{g}_i\mathbf{g}_i'][\Sigma_{i=1}^{N}\mathbf{H}_i]^{-1}$

$$= [\Sigma_{i=1}^{N} - y_i\theta_i\mathbf{x}_i\mathbf{x}_i']^{-1}[\Sigma_{i=1}^{N}e_i^2\theta_i^2\mathbf{x}_i\mathbf{x}_i'][\Sigma_{i=1}^{N} - y_i\theta_i\mathbf{x}_i\mathbf{x}_i']^{-1}$$

This is the White estimator. See JW, p. 359.

Computing NLS

```
Reject; hhninc=0$
Calc  ; b0=log(xbr(hhninc))$
Names ; x=one,educ,married,age$
Nlsq  ; labels=a0,a1,a2,a3
      ; start=b0,0,0,0
      ; fcn=exp(a0’x)
      ; lhs=hhninc;output=3$
Create; thetai = exp(x'b)
      ; ei=hhninc-thetai
      ; gi2=(ei*thetai)^2
      ; hi=hhninc*thetai$
Matrix; varM = <x'[hi] x> * x'[gi2]x * <x'[hi] x> $
Matrix; stat(b,varM,x)$
```

Iterations ... Convergence

$$'\text{gradient}' = \mathbf{e}^{0\prime}\mathbf{X}^{0}(\mathbf{X}^{0\prime}\mathbf{X}^{0})^{-1}\mathbf{X}^{0\prime}\mathbf{e}^{0}$$

```
Begin NLSQ iterations. Linearized regression.
Iteration=  1; Sum of squares=  854.681775    ; Gradient=  90.0964694
Iteration=  2; Sum of squares=  766.073500    ; Gradient=  2.38006397
Iteration=  3; Sum of squares=  763.757721    ; Gradient=  .300030163E-02
Iteration=  4; Sum of squares=  763.755005    ; Gradient=  .307466962E-04
Iteration=  5; Sum of squares=  763.754978    ; Gradient=  .365064970E-06
Iteration=  6; Sum of squares=  763.754978    ; Gradient=  .433325697E-08
Iteration=  7; Sum of squares=  763.754978    ; Gradient=  .514374906E-10
Iteration=  8; Sum of squares=  763.754978    ; Gradient=  .610586853E-12
Iteration=  9; Sum of squares=  763.754978    ; Gradient=  .724960231E-14
Iteration= 10; Sum of squares=  763.754978    ; Gradient=  .860927011E-16
Iteration= 11; Sum of squares=  763.754978    ; Gradient=  .102139114E-17
Iteration= 12; Sum of squares=  763.754978    ; Gradient=  .118640949E-19
Iteration= 13; Sum of squares=  763.754978    ; Gradient=  .125019054E-21
Convergence achieved
```

NLS Estimates

```
+----------------------------------------------------+
| User Defined Optimization                          |
| Nonlinear   least squares regression               |
| LHS=HHNINC   Mean                 =   .3521352     |
|              Standard deviation   =   .1768699     |
| Residuals    Sum of squares       =   763.7550     |
+----------------------------------------------------+
+---------+--------------+----------------+--------+---------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] |
+---------+--------------+----------------+--------+---------+
 Conventional Estimates
 Constant     -1.89118955      .01879455  -100.624   .0000
 EDUC           .05471841      .00102649    53.306   .0000
 MARRIED        .23756387      .00765477    31.035   .0000
 AGE            .00081033      .00026344     3.076   .0021
+---------+--------------+----------------+--------+---------+
 Recomputed variances using sandwich estimator for M Estimation.
 B_1          -1.89118955      .01910054   -99.012   .0000
 B_2            .05471841      .00115059    47.557   .0000
 B_3            .23756387      .00842712    28.190   .0000
 B_4            .00081033      .00026137     3.100   .0019
```

Hypothesis Tests for M Estimation

Null hypothesis: $\mathbf{c}(\theta)=\mathbf{0}$ for some set of J functions

(1) continuous

(2) differentiable; $\dfrac{\partial\mathbf{c}(\theta)}{\partial\theta'}=\mathbf{R}(\theta)$, J × K Jacobian

(3) functionally independent: Rank $\mathbf{R}(\theta)$ = J

Wald: given $\hat\theta$, $\hat{\mathbf{V}}$=Est.Asy.Var[$\hat\theta$],

W=Wald distance

$$=[\mathbf{c}(\hat\theta)\text{-}\mathbf{c}(\theta)]'\{\mathbf{R}(\theta)\mathbf{V}(\theta)\mathbf{R}(\theta)'\}^{-1}[\mathbf{c}(\hat\theta)\text{-}\mathbf{c}(\theta)]'$$

$\rightarrow$ chi-squared[J]

Change in the Criterion Function

$$\bar{q}=\frac{1}{n}\sum_{i=1}^{n} q(\text{data}_i,\theta) \xrightarrow{P} q^*=E[q(\text{data},\theta)]$$

Estimator $\hat{\theta}$ minimizes $\bar{q}$

Estimator $\hat{\theta}^c$ minimizes $\bar{q}$ subject to restrictions $\mathbf{c}(\theta)=0$

$$\hat{\bar{q}}^c \ \geq \ \hat{\bar{q}}.$$

$$2n(\hat{\bar{q}}^c \ - \ \bar{q}\ ) \xrightarrow{D} \text{chi}-\text{squared}[J]$$

Score Test

LM Statistic

Derivative of the objective function

Score vector = $\dfrac{(1/n)\Sigma_{i=1}^{n}\partial q(\text{data}_i,\theta)}{\partial\theta} = \overline{\mathbf{g}}(\text{data},\theta)$

Without restrictions $\overline{\mathbf{g}}(\text{data},\hat{\theta}) = \mathbf{0}$

With null hypothesis, $\mathbf{c}(\hat{\theta})$ imposed

$\overline{\mathbf{g}}(\text{data},\hat{\theta}^c)$ generally not equal to $\mathbf{0}$. Is it close?

(Within sampling variability?)

Wald distance = $[\overline{\mathbf{g}}(\text{data},\hat{\theta}^c)]'\{\text{Var}[\overline{\mathbf{g}}(\text{data},\hat{\theta}^c)]\}^{-1}[\overline{\mathbf{g}}(\text{data},\hat{\theta}^c)]$

$\text{LM} \xrightarrow{D} \text{chi-squared}[J]$

Exponential Model

$$f(\text{Income} \mid \text{Age}, \text{Educ}, \text{Married})$$

$$= \frac{1}{\theta_i} \exp\left(\frac{-\text{HHNINC}_i}{\theta_i} \right)$$

$$\theta_i = \exp(a_0 + a_1 \text{Educ} + a_2 \text{Married} + a_3 \text{Age})$$

$$\text{Test } H_0\text{: } a_1 = a_2 = a_3 = 0$$

Wald Test

```
Matrix ; List ; R=[0,1,0,0 / 0,0,1,0 / 0,0,0,1]
        ; c=R*b ; Vc = R*Varb*R'
        ; Wald = c'<VC>c $
Matrix R        has  3 rows and  4 columns.
        .0000000D+00    1.00000    .0000000D+00  .0000000D+00
        .0000000D+00  .0000000D+00    1.00000    .0000000D+00
        .0000000D+00  .0000000D+00  .0000000D+00    1.00000
Matrix C        has  3 rows and  1 columns.
           .05472
           .23756
           .00081
Matrix VC       has  3 rows and  3 columns.
        .1053686D-05  .4530603D-06  .3649631D-07
        .4530603D-06  .5859546D-04 -.3565863D-06
        .3649631D-07 -.3565863D-06  .6940296D-07
Matrix WALD     has  1 rows and  1 columns.
         3627.17514
```

Change in Function

```
Calc ; M = sumsqdev $ (from unrestricted)
Calc ; b0 = log(xbr(hhninc)) $
Nlsq ; labels=a0,a1,a2,a3;start=b0,0,0,0
     ; fcn=exp(a0+a1*educ+a2*married+a3*age)
     ; fix=a1,a2,a3 ? Fixes at start values
     ; lhs=hhninc $
Calc ; M0 = sumsqdev $
```

Constrained Estimation

```
Nonlinear Estimation of Model Parameters
Method=BFGS  ; Maximum iterations=100
Start values:  -.10437D+01
1st derivs.       -.26609D-10
Parameters:    -.10437D+01
Itr  1 F=  .4273D+03 gtHg=  .2661D-10
                                * Converged
NOTE: Convergence in initial iterations is rarely
at a true function optimum. This may not be a
solution (especially if initial iterations stopped).
Exit from iterative procedure.    1 iterations completed.
```

Why did this occur? The starting value is the minimizer of the constrained function

Constrained Estimates

```
+----------------------------------------------------+
| User Defined Optimization                          |
| Nonlinear   least squares regression               |
| LHS=HHNINC   Mean                 =   .3521352     |
|              Standard deviation   =   .1768699     |
| Residuals    Sum of squares       =   854.6818     |
+----------------------------------------------------+
+---------+--------------+----------------+--------+---------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] |
+---------+--------------+----------------+--------+---------+
 A0           -1.04374019      .00303865  -343.488   .0000
 A1               .000000   ......(Fixed Parameter).......
 A2               .000000   ......(Fixed Parameter).......
 A3               .000000   ......(Fixed Parameter).......
```

```
--> calc ; m0=sumsqdev ; list ; df = 2*(m0 - m) $
    DF      =  181.854
```

LM Test

$$\text{Function: } q_i = (1/2)[y_i - \exp(a_0 + a_1\text{Educ}...)]^2$$

$$\text{Derivative: } \mathbf{g}_i = -e_i\theta_i\mathbf{x}_i$$

LM statistic

$$\text{LM}=(\Sigma_{i=1}^{n}\mathbf{g}_i)[\Sigma_{i=1}^{n}\mathbf{g}_i\mathbf{g}_i{}']^{-1}(\Sigma_{i=1}^{n}\mathbf{g}_i)$$

$$\text{All evaluated at } \hat{a}_0 = \log(\overline{y}), 0, 0, 0$$

LM Test

```
Name    ;x=one,educ,married,age$
Create  ;thetai=exp(x'b);ei=hhninc-thetai$
Create  ;gi=ei*thetai ; gi2 = gi*gi $
Matrix  ; list ; LM = 1'[gi]x * <x'[gi2]x> * x'[gi]1 $
```

```
Matrix LM       has  1 rows and  1 columns.
              1
       +--------------
      1| 1915.03286
```

Maximum Likelihood Estimation

- Fully parametric estimation
  - Density of y_i is fully specified
- The likelihood function = the joint density of the observed random variable.
- Example: density for the exponential model

$$f(y_i \mid \mathbf{x}_i)\boldsymbol{\beta} = \frac{1}{\theta_i}\exp\left(-\frac{y_i}{\theta_i}\right), \theta_i = \exp(\mathbf{x}_i' \ )$$

$$E[y_i \mid \mathbf{x}_i] = \theta_i, \ \ Var[y_i \mid \mathbf{x}_i] = \theta_i^2$$

NLS (M) estimator examined earlier

operated only on $E[y_i \mid \mathbf{x}_i] = \theta_i$.

The Likelihood Function

$$f(y_i \mid \mathbf{x}_i)\boldsymbol{\beta} = \frac{1}{\theta_i}\exp\left(-\frac{y_i}{\theta_i}\right), \theta_i = \exp(\mathbf{x}_i' \;)$$

$$\text{Likelihood} = f(y_1, \ldots, y_n \mid \mathbf{x}_1, \ldots, \mathbf{x}_n)$$

By independence,

$$L(\boldsymbol{\beta} \mid \textbf{data}) = \prod_{i=1}^{n} \frac{1}{\theta_i}\exp\left(-\frac{y_i}{\theta_i}\right), \theta_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$$

The MLE , $\hat{\boldsymbol{\beta}}_{\mathbf{MLE}}$, maximizes the likelihood function

Log Likelihood Function

$$f(y_i \mid \mathbf{x}_i)\boldsymbol{\beta} = \frac{1}{\theta_i}\exp\left(-\frac{y_i}{\theta_i}\right), \theta_i = \exp(\mathbf{x}_i' \ )$$

$$L(\boldsymbol{\beta} \mid \mathbf{data}) = \prod_{i=1}^{n} \frac{1}{\theta_i}\exp\left(-\frac{y_i}{\theta_i}\right), \theta_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$$

The MLE , $\hat{\boldsymbol{\beta}}_{\mathbf{MLE}}$, maximizes the likelihood function

$\log L(\boldsymbol{\beta} \mid \mathbf{data})$ is a monotonic function. Therefore

The MLE , $\hat{\boldsymbol{\beta}}_{\mathbf{MLE}}$, maximizes the log likelihood function

$$\log L(\boldsymbol{\beta} \mid \mathbf{data}) = \sum_{i=1}^{n} \text{-}\log\theta_i - \frac{y_i}{\theta_i}$$

Conditional and Unconditional Likelihood

Unconditional joint density $f(y_i, \mathbf{x}_i \mid \theta, \delta)$

$\theta =$ our parameters of interest

δ = parameters of the marginal density of $\mathbf{x}_i$

Unconditional likelihood function

$$L(\theta,\delta|y,\mathbf{X}) = \prod_{i=1}^{n} f(y_i, \mathbf{x}_i \mid \theta, \delta)$$

$$f(y_i, \mathbf{x}_i \mid \theta, \delta) = f(y_i \mid \mathbf{x}_i, \theta, \delta) g(\mathbf{x}_i \mid \theta, \delta)$$

$$L(\theta,\delta|y,\mathbf{X}) = \prod_{i=1}^{n} f(y_i \mid \mathbf{x}_i, \theta, \delta) g(\mathbf{x}_i \mid \theta, \delta)$$

Assuming the parameter space partitions

$$\log L(\theta,\delta|y,\mathbf{X}) = \sum_{i=1}^{n} \log f(y_i \mid \mathbf{x}_i, \theta) + \sum_{i=1}^{n} \log g(\mathbf{x}_i \mid \delta)$$

$=$ conditional log likelihood + marginal log likelihood

Concentrated Log Likelihood

$\hat{\theta}_{MLE}$ maximizes logL(θ|data)

Consider a partition, θ=(β,α) two parts.

Maximum occurs where $\dfrac{\partial \text{logL}}{\partial \begin{pmatrix} \beta \\ \alpha \end{pmatrix}} = 0$

Joint solution equates both derivatives to 0.

If ∂logL/$\partial\alpha$=0 admits an implicit solution for

α in terms of β, $\hat{\alpha}_{MLE} = \hat{\alpha}(\hat{\beta})$, then write

logL$_c$(β,α(β))=a function only of β.

Concentrated log likelihood can be maximized

for β, then the solution computed for α.

The solution must occur where $\hat{\alpha}_{MLE} = \hat{\alpha}(\hat{\beta})$, so restrict

the search to this subspace of the parameter space.

A Concentrated Log Likelihood

Fixed effects exponential regression: $\theta_{it} = \exp(\alpha_i + x_{it}'\beta)$

$$\log L = \sum_{i=1}^{n}\sum_{t=1}^{T}(-\log\theta_{it} - y_{it} / \theta_{it})$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T}(-(\alpha_i + x_{it}'\beta) - y_{it}\exp(-\alpha_i - x_{it}'\beta))$$

$$\frac{\partial \log L}{\partial \alpha_i} = \sum_{t=1}^{T} -1 - y_{it}\exp(-\alpha_i - x_{it}'\beta)(-1)$$

$$= -T + \sum_{t=1}^{T} y_{it}\exp(-\alpha_i - x_{it}'\beta)$$

$$= -T + \exp(-\alpha_i)\sum_{t=1}^{T} y_{it}\exp(-x_{it}'\beta) = 0$$

Solve this for $\alpha_i = \log\left(\frac{\Sigma_{t=1}^{T}\left[y_{it} / \exp(x_{it}'\beta)\right]}{T}\right) = \alpha_i(\beta)$

Concentrated log likelihood has $\theta_{it}^{c} = \left(\frac{\Sigma_{t=1}^{T} y_{it} / \exp(x_{it}'\beta)}{T}\right)\exp(x_{it}'\beta)$

ML and M Estimation

$$\log L(\theta) = \sum_{i=1}^{n} \log f(y_i \mid x_i, \theta)$$

$$\hat{\theta}_{MLE} = \arg\max \sum_{i=1}^{n} \log f(y_i \mid x_i, \theta)$$

$$= \arg\min \quad -\frac{1}{n}\sum_{i=1}^{n} \log f(y_i \mid x_i, \theta)$$

The MLE is an M estimator. We can use all of the previous results for M estimation.

Regularity Conditions

- Conditions for the MLE to be consistent, etc.
- Augment the continuity and identification conditions for M estimation
- Regularity:
  - Three times continuous differentiability of the log density
  - Finite third moments of log density
  - Conditions needed to obtain expected values of derivatives of log density are met.
- (See Greene (Chapter 14))

Consistency and Asymptotic Normality of the MLE

- Conditions are identical to those for M estimation
- Terms in proofs are log density and its derivatives
- Nothing new is needed.
  - Law of large numbers
  - Lindberg-Feller central limit applies to derivatives of the log likelihood.

Asymptotic Variance of the MLE

Based on results for M estimation

$$\text{Asy.Var}[\hat{\theta}_{\text{MLE}}]$$

$$=\{\text{-E}[\text{Hessian}]\}^{-1}\{\text{Var}[\text{first derivative}]\}\{\text{-E}[\text{Hessian}]\}^{-1}$$

$$=\left\{\text{-E}\left[\frac{\partial^2 \log L}{\partial\theta\partial\theta'}\right]\right\}^{-1}\text{Var}\left[\frac{\partial \log L}{\partial\theta}\right]\left\{\text{-E}\left[\frac{\partial^2 \log L}{\partial\theta\partial\theta'}\right]\right\}^{-1}$$

The Information Matrix Equality

Fundamental Result for MLE

The variance of the first derivative equals the negative of the expected second derivative.

$$\text{-E}\left[\frac{\partial^2 \log L}{\partial\theta\partial\theta'}\right] = \text{ The Information Matrix}$$

$$\text{Asy.Var}[\hat{\theta}_{\text{MLE}}]$$

$$= \left\{\text{-E}\left[\frac{\partial^2 \log L}{\partial\theta\partial\theta'}\right]\right\}^{-1}\left\{\text{-E}\left[\frac{\partial^2 \log L}{\partial\theta\partial\theta'}\right]\right\}\left\{\text{-E}\left[\frac{\partial^2 \log L}{\partial\theta\partial\theta'}\right]\right\}^{-1}$$

$$= \left\{\text{-E}\left[\frac{\partial^2 \log L}{\partial\theta\partial\theta'}\right]\right\}^{-1}$$

Three Variance Estimators

- Negative inverse of expected second derivatives matrix. (Usually not known)
- Negative inverse of actual second derivatives matrix.
- Inverse of variance of first derivatives

Asymptotic Efficiency

- M estimator based on the conditional mean is semiparametric. Not necessarily efficient.
- MLE is fully parametric. It is efficient among all consistent and asymptotically normal estimators when the density is as specified.
- This is the Cramer-Rao bound.
- Note the implied comparison to nonlinear least squares for the exponential regression model.

Invariance

Useful property of MLE

If $\lambda = g(\theta)$ is a continuous function of θ,

the MLE of λ is $g(\hat{\theta}_{MLE})$

E.g., in the exponential FE model, the

MLE of $\lambda_i = \exp(-\alpha_i)$ is $\exp(-\hat{\alpha}_{i,MLE})$

Application: Exponential Regression

```
+---------------------------------------------+
| Exponential (Loglinear) Regression Model    |
| Maximum Likelihood Estimates                |
| Dependent variable               HHNINC     |
| Number of observations            27322     |
| Iterations completed                 10     |
| Log likelihood function        1539.191     |
| Number of parameters                  4     |
| Restricted log likelihood      1195.070     |
| Chi squared                    688.2433     |
| Degrees of freedom                    3     |
+---------------------------------------------+
+---------+--------------+----------------+--------+---------+----------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
+---------+--------------+----------------+--------+---------+----------+
          Parameters in conditional mean function
 Constant     -1.82555590      .04219675   -43.263    .0000
 EDUC           .05545277      .00267224    20.751    .0000    11.3201838
 MARRIED        .23664845      .01460746    16.201    .0000     .75869263
 AGE           -.00087436      .00057331     1.525    .1272    43.5271942
 NLS Results with recomputed variances using results for M Estimation.
 B_1          -1.89118955      .01910054   -99.012    .0000
 B_2            .05471841      .00115059    47.557    .0000
 B_3            .23756387      .00842712    28.190    .0000
 B_4            .00081033      .00026137     3.100    .0019
```

Variance Estimators

$$\text{LogL} = \sum\nolimits_{i=1}^{n} -\log\theta_i - y_i\,/\,\theta_i,\ \ \theta_i = \exp(\mathbf{x\beta})$$

$$\mathbf{g} = \frac{\partial \log L}{\partial\beta} = \sum\nolimits_{i=1}^{n} -\mathbf{x_i} + (y_i\,/\,\theta_i)\mathbf{x_i} = \sum\nolimits_{i=1}^{n} [(y_i\,/\,\theta_i) - 1]\mathbf{x_i}$$

Note, $E[y_i \mid \mathbf{x_i}] = \theta_i$, so $E[\mathbf{g}] = \mathbf{0}$

$$\mathbf{H} = \frac{\partial^2 \log L}{\partial\beta\partial\beta'} = \sum\nolimits_{i=1}^{n} -(y_i\,/\,\theta_i)\mathbf{x_i}\mathbf{x_i}'$$

$E[\mathbf{H}] = \sum\nolimits_{i=1}^{n} -\mathbf{x_i}\mathbf{x_i}'$ = **-X'X** (known for this particular model)

Three Variance Estimators

Berndt-Hall-Hall-Hausman (BHHH)

$$\left[\sum_{i=1}^{n} \mathbf{g}_{\mathbf{i}}\mathbf{g}_{\mathbf{i}}'\right]^{-1} = \left[\sum_{i=1}^{n} [(y_i / \hat{\theta}_i) - 1]^2 \mathbf{x}_{\mathbf{i}}\mathbf{x}_{\mathbf{i}}'\right]^{-1}$$

Based on actual second derivatives

$$\left[-\sum_{i=1}^{n} \mathbf{H}_{\mathbf{i}}\right]^{-1} = \left[\sum_{i=1}^{n} (y_i / \hat{\theta}_i)\mathbf{x}_{\mathbf{i}}\mathbf{x}_{\mathbf{i}}'\right]^{-1}$$

Based on expected second derivatives

$$\left\{E\left[-\sum_{i=1}^{n} \mathbf{H}_{\mathbf{i}}\right]\right\}^{-1} = \left[\sum_{i=1}^{n} \mathbf{x}_{\mathbf{i}}\mathbf{x}'_{\mathbf{i}}\right]^{-1} = (\mathbf{X'X})^{-1}$$

Variance Estimators

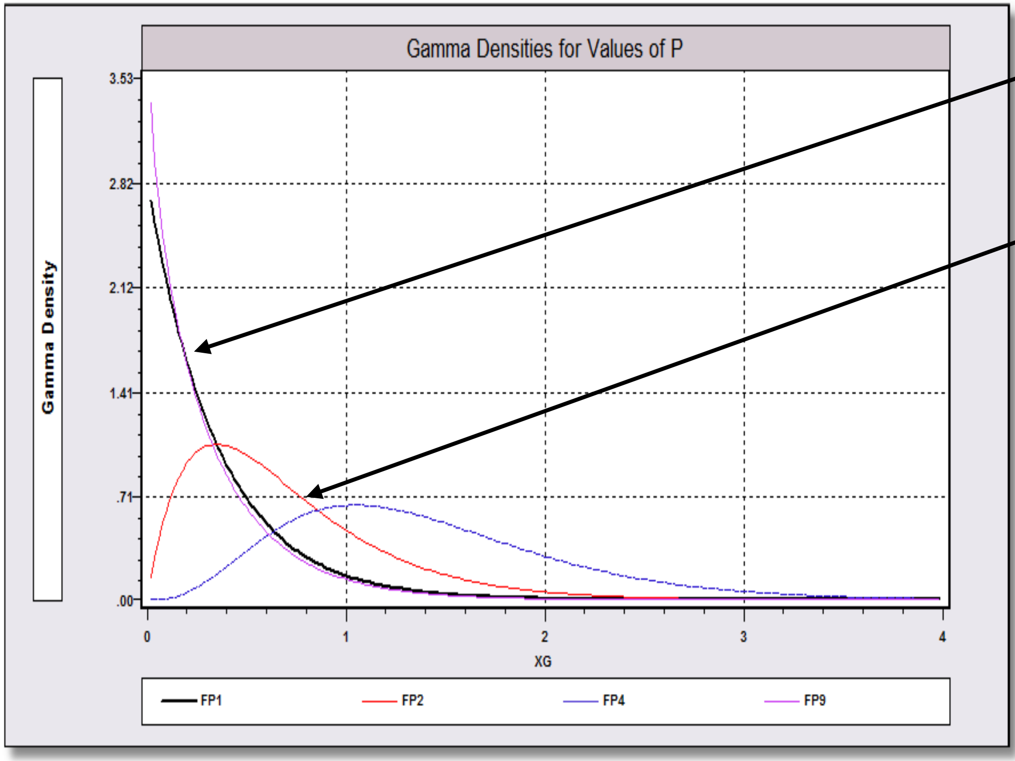
```
Loglinear ; Lhs=hhninc;Rhs=x ; Model = Exponential
create;thetai=exp(x'b);hi=hhninc/thetai;gi2=(hi-1)^2$
matr;he=<x'x> ; ha=<x'[hi]x> ; bhhh=<x'[gi2]x>$
matr;stat(b,ha);stat(b,he);stat(b,bhhh)$
+---------+--------------+----------------+--------+---------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] |
+---------+--------------+----------------+--------+---------+
 B_1          -1.82555590      .11129890   -16.402   .0000  ACTUAL
 B_2            .05545277      .00617308     8.983   .0000
 B_3            .23664845      .04609371     5.134   .0000
 B_4           -.00087436      .00164729     -.531   .5956
+---------+--------------+----------------+--------+---------+
 B_1          -1.82555590      .04237258   -43.083   .0000  EXPECTED
 B_2            .05545277      .00264541    20.962   .0000
 B_3            .23664845      .01442783    16.402   .0000
 B_4           -.00087436      .00055100    -1.587   .1125
+---------+--------------+----------------+--------+---------+
 B_1          -1.82555590      .05047329   -36.169   .0000  BHHH
 B_2            .05545277      .00326769    16.970   .0000
 B_3            .23664845      .01604572    14.748   .0000
 B_4           -.00087436      .00062011    -1.410   .1585
```

Hypothesis Tests

- Trinity of tests for nested hypotheses
  - Wald
  - Likelihood ratio
  - Lagrange multiplier
- All as defined for the M estimators

Example: Exponential vs. Gamma

$$\text{Gamma Distribution: } f(y_i \mid x_i, \theta, P) = \frac{\exp(-y_i / \theta_i) y_i^{P-1}}{\theta_i^P \; \Gamma(P)}$$



Exponential: P = 1

P > 1

Log Likelihood

$$\log L = \Sigma_{i=1}^{n}\left[-P\log\theta_i - \log\Gamma(P) - y_i/\theta_i + (P-1)\log y_i\right]$$

$$\Gamma(1) = 0! = 1$$

$$\log L = \Sigma_{i=1}^{n} -\log\theta_i - y_i/\theta_i$$
$$+ (P-1)\log y_i - (P-1)\log\theta_i - \log\Gamma(P)$$
$$= \Sigma_{i=1}^{n} -\log\theta_i - y_i/\theta_i + (P-1)\log(y_i/\theta_i) - \log\Gamma(P)$$
$$= \text{Exponential } \log L + \text{part due to } P \neq 1$$

Estimated Gamma Model

```
+---------------------------------------------+
| Gamma (Loglinear) Regression Model          |
| Dependent variable               HHNINC     |
| Number of observations            27322     |
| Iterations completed                 18     |
| Log likelihood function        14237.33     |
| Number of parameters                  5     |
| Restricted log likelihood      1195.070     |
| Chi squared                    26084.52     |
| Degrees of freedom                    4     |
| Prob[ChiSqd > value] =         .0000000     |
+---------------------------------------------+
+---------+--------------+----------------+--------+---------+----------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
+---------+--------------+----------------+--------+---------+----------+
          Parameters in conditional mean function
 Constant      3.45583194      .02043321   169.128   .0000
 EDUC          -.05545277      .00118268   -46.888   .0000    11.3201838
 MARRIED       -.23664845      .00646494   -36.605   .0000     .75869263
 AGE            .00087436      .00025374     3.446   .0006    43.5271942
          Scale parameter for gamma model
 P_scale       5.10528380      .04232988   120.607   .0000
```

Testing P = 1

- Wald: W = $(5.10528380-1)^2/.04232988^2$

  = 9405.7
- Likelihood Ratio:
  - logL|(P=1)= 1539.19
  - logL|P = 14237.33
  - LR = 2(14237.33 - 1539.19) = 25396.27
- **Lagrange Multiplier…**

Derivatives for the LM Test

$$\log L = \Sigma_{i=1}^{n} -\log\theta_i - y_i/\theta_i + (P-1)\log(y_i/\theta_i) - \log\Gamma(P)$$

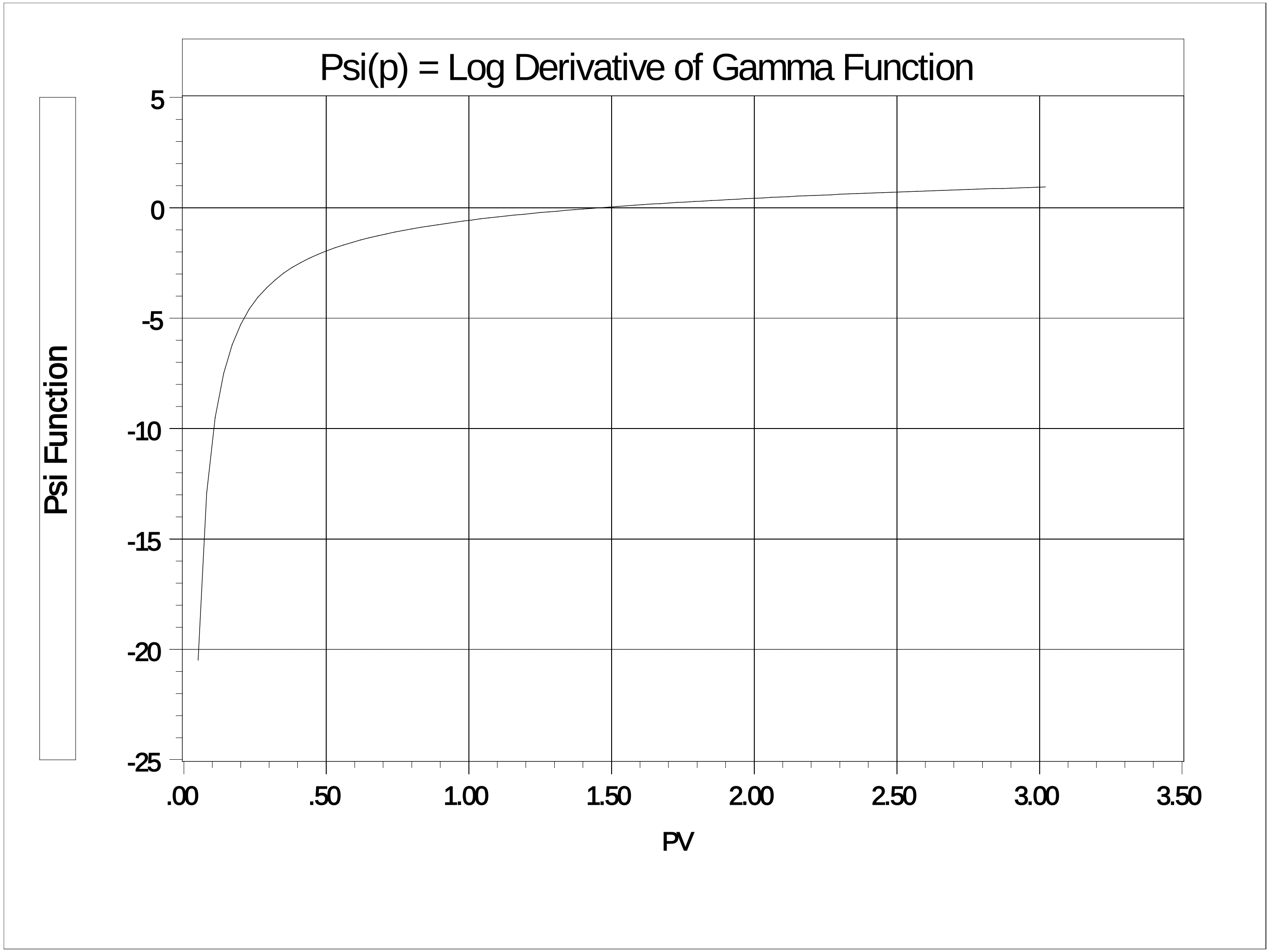
$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \Sigma_{i=1}^{n}(y_i/\theta_i - P)\mathbf{x}_i = \boxed{\Sigma_{i=1}^{n} g_{x,i}\mathbf{x}_i}$$

$$\frac{\partial \log L}{\partial P} = \Sigma_{i=1}^{n} \log(y_i/\theta_i) - \psi(P) = \boxed{\Sigma_{i=1}^{n} g_{P,i}}$$

$$\psi(1) = -.5772156649$$

For the LM test, we compute these at the exponential MLE and P = 1.

Psi Function = dlogΓ(P)/dP



Derivatives

$$\mathbf{g}_i = \begin{bmatrix} g_{x,i}(1) \\ g_{x,i}(\text{Educ}) \\ g_{x,i}(\text{Married}) \\ g_{x,i}(\text{Age}) \\ g_{P,i} \end{bmatrix}, \mathbf{g} = \sum_{i=1}^{27322} \mathbf{g}_i = \begin{bmatrix} .0000008237 \\ .00001586 \\ -.0000005033 \\ -.00001444 \\ -.00104 \end{bmatrix} \text{when P is unrestricted}$$

$$= \begin{bmatrix} .00006737 \\ .00092 \\ .000027881 \\ .0020 \\ 13007.80 \end{bmatrix} \text{ when P = 1 in the gamma model}$$

Score Test

Test the hypothesis that the derivative vector equals zero when evaluated for the larger model with the restricted coefficient vector.

Estimator of zero is $\bar{\mathbf{g}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i$

Statistic = chi squared = $\bar{\mathbf{g}}'[\text{Var }\bar{\mathbf{g}}]^{-1}\bar{\mathbf{g}}$

Use $\left(\frac{1}{n}\right)\left(\frac{1}{n}\right)\sum_{i=1}^{n}\mathbf{g}_i\mathbf{g}_i'$ (the n's will cancel).

chi squared = $\left(\sum_{i=1}^{n}\mathbf{g}_i\right)'\left[\sum_{i=1}^{n}\mathbf{g}_i\mathbf{g}_i'\right]^{-1}\left(\sum_{i=1}^{n}\mathbf{g}_i\right)$

Calculated LM Statistic

```
Create  ;thetai=exp(x’b) ; gi=(hhninc/thetai – 1)
Create  ;gpi=log(hhninc/thetai)-psi(1)$
Create  ;g1i=gi;g2i=gi*educ;g3i=gi*married;g4i=gi*age;g5i=gpi$
Namelist;ggi=g1i,g2i,g3i,g4i,g5i$
Matrix  ;list ; lm = 1'ggi * <ggi'ggi> * ggi'1 $
```

```
  Matrix LM
         1
  +--------------
1|  26596.92
```

```
? Use built-in procedure.
? LM is computed with actual Hessian instead of BHHH
logl;lhs=hhninc;rhs=one,educ,married,age;model=g;start=b,1;maxit=0$
```

```
| LM Stat. at start values       9602.409     |
```

Clustered Data and Partial Likelihood

Panel Data: $y_{it} \mid \mathbf{x}_{it}, t = 1,\ldots,T_i$

Some connection across observations within a group

Assume marginal density for $y_{it} \mid \mathbf{x}_{it} = f(y_{it} \mid \mathbf{x}_{it}, \theta)$

Joint density for individual i is

$$f(y_{i1},\ldots,y_{i,T_i} \mid \mathbf{X}_i) \neq \prod_{t=1}^{T_i} f(y_{it} \mid \mathbf{x}_{it}, \theta)$$

$$\text{"Pseudo} - \text{log Likelihood"} = \sum_{i=1}^{n} \log \prod_{t=1}^{T_i} f(y_{it} \mid \mathbf{x}_{it}, \theta)$$

$$= \sum_{i=1}^{n} \sum_{t=1}^{T} \log f(y_{it} \mid \mathbf{x}_{it}, \theta)$$

Just the pooled log likelihood, ignoring the panel aspect of the data.

Not the correct log likelihood. Does maximizing wrt θ work? Yes, if the marginal density is correctly specified.

Inference with ‘Clustering’

(1) Estimator is consistent

(2) Asymptotic Covariance matrix needs adjustment

Asy.Var[] = $[\text{Hessian}]^{-1}$Var[gradient]$[\text{Hessian}]^{-1}$

$$\mathbf{H} = \sum_{i=1}^{n}\sum_{t=1}^{T_i}\mathbf{H}_{it}$$

$\mathbf{g} = \sum_{i=1}^{n}\mathbf{g}_i$, where $\mathbf{g}_i = \sum_{t=1}^{T_i}\mathbf{g}_{it}$

Terms in $\mathbf{g}_i$ are not independent, so estimation of the variance cannot be done with $\sum_{i=1}^{n}\sum_{t=1}^{T_i}\mathbf{g}_{it}\mathbf{g}_{it}'$

But, terms across i are independent, so we estimate

Var[**g**] with $\sum_{i=1}^{n}\left(\sum_{t=1}^{T_i}\mathbf{g}_{it}\right)\left(\sum_{t=1}^{T_i}\mathbf{g}_{it}\right)'$

$$\text{Est.Var}[\hat{\theta}_{\mathbf{PMLE}}] = \left(\sum_{i=1}^{n}\sum_{t=1}^{T_i}\hat{\mathbf{H}}_{it}\right)^{-1}\left\{\sum_{i=1}^{n}\left(\sum_{t=1}^{T_i}\hat{\mathbf{g}}_{it}\right)\left(\sum_{t=1}^{T_i}\hat{\mathbf{g}}_{it}\right)'\right\}\left(\sum_{i=1}^{n}\sum_{t=1}^{T_i}\hat{\mathbf{H}}_{it}\right)^{-1}$$

(Generally inserts a correction term n/(n-1) before the middle term.)

Cluster Estimation

```
+---------------------------------------------+
| Exponential (Loglinear) Regression Model    |
| Maximum Likelihood Estimates                |
+---------------------------------------------+
+-----------------------------------------------------------------------+
| Covariance matrix for the model is adjusted for data clustering.      |
| Sample of  27322 observations contained   7293 clusters defined by    |
| variable ID       which identifies by a value a cluster ID.           |
+-----------------------------------------------------------------------+
+---------+--------------+----------------+--------+---------+----------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
+---------+--------------+----------------+--------+---------+----------+
          Parameters in conditional mean function
 Constant      1.82555590      .03215706    56.770   .0000
 EDUC          -.05545277      .00195517   -28.362   .0000    11.3201838
 MARRIED       -.23664845      .01338104   -17.685   .0000     .75869263
 AGE            .00087436      .00044694     1.956   .0504    43.5271942
+---------+--------------+----------------+--------+---------+----------+
          Uncorrected Standard Errors
 Constant      1.82555590      .04219675    43.263   .0000
 EDUC          -.05545277      .00267224   -20.751   .0000    11.3201838
 MARRIED       -.23664845      .01460746   -16.201   .0000     .75869263
 AGE            .00087436      .00057331     1.525   .1272    43.5271942
```

On Clustering

- The theory is very loose.
- That the marginals would be correctly specified while there remains ‘correlation’ across observations is ambiguous
- It seems to work pretty well in practice (anyway)
- BUT… It does not imply that one can safely just pool the observations in a panel and ignore unobserved common effects.
- The correction redeems the variance of an estimator, not the estimator, itself.

‘Robust’ Estimation

- If the model is misspecified in some way, then the information matrix equality does not hold.
- Assuming the estimator remains consistent, the appropriate asymptotic covariance matrix would be the ‘robust’ matrix, actually, the original one.

$$\text{Asy.Var}[\hat{\theta}_{\text{MLE}}] = [-\text{E}[\text{Hessian}]]^{-1}\,\text{Var}[\text{gradient}]\,[-\text{E}[\text{Hessian}]]^{-1}$$

(Software can be coerced into computing this by telling it that clusters all have one observation in them.)

Two Step Estimation and Murphy/Topel

Likelihood function defined over two parameter vectors

$$\log L = \sum_{i=1}^{n} \log f(y_i \mid x_i, z_i, \theta, \delta)$$

(1) Maximize the whole thing. (FIML)

(2) Typical Situation: Two steps

$$\text{E.g., } f(\text{HHNINC}|\text{educ,married,age,Ifkids}) = \frac{1}{\theta_i}\exp\left(-\frac{y_i}{\theta_i}\right),$$

$$\theta_i = \exp(\beta_0 + \beta_1 \text{Educ} + \beta_2 \text{Married} + \beta_3 \text{Age} + \beta_4 \Pr[\text{IfKids}])$$

$$\text{If}[\text{Kids} \mid \text{age, bluec}] = \text{Logistic Regression}$$

$$\Pr[\text{IfKids}] = \exp(\delta_0 + \delta_1 \text{Age} + \delta_2 \text{Bluec}) / [1 + \exp(\delta_0 + \delta_1 \text{Age} + \delta_2 \text{Bluec})]$$

(3) Two step strategy: Fit the stage one model (δ) by MLE

first, insert the results in $\log L(\theta, \hat{\delta})$ and estimate θ.

Two Step Estimation

(1) Does it work? Yes, with the usual identification conditions, continuity, etc. The first step estimator is assumed to be consistent and asymptotically normally distributed.

(2) The asymptotic covariance matrix at the second step that takes $\hat{\delta}$ as if it were known is too small.

(3) Repair to the covariance matrix by the Murphy Topel Result (1983, 2002)

Murphy-Topel - 1

$\log L_1(\delta)$ defines the first step estimator. Let

$\hat{\mathbf{V}}_1$ = Estimated asymptotic covariance matrix for $\hat{\delta}$

$$\mathbf{g}_{i,1} = \frac{\partial \log f_{i,1}(\ldots,\delta)}{\partial \delta}. \; (\hat{\mathbf{V}}_1 \text{ might } = [\Sigma_{i=1}^{n} \hat{\mathbf{g}}_{i,1} \hat{\mathbf{g}}_{i,1}']^{-1})$$

$\log L(\theta, \hat{\delta})$ defines the second step estimator using the estimated value of δ.

$\hat{\mathbf{V}}_2$ = Estimated asymptotic covariance matrix for $\hat{\theta} \mid \hat{\delta}$

$$\mathbf{g}_{i,2} = \frac{\partial \log f_{i,2}(\ldots,\theta,\hat{\delta})}{\partial \theta}. \; (\hat{\mathbf{V}}_2 \text{ might } = [\Sigma_{i=1}^{n} \hat{\mathbf{g}}_{i,2} \hat{\mathbf{g}}_{i,2}']^{-1})$$

$\hat{\mathbf{V}}_2$ is too small

Murphy-Topel - 2

$\hat{\mathbf{V}}_1 =$ Estimated asymptotic covariance matrix for $\hat{\delta}$

$\mathbf{g}_{i,1} = \partial \log f_{i,1}(\ldots,\delta) / \partial\delta$. ($\hat{\mathbf{V}}_1$ might = $[\Sigma_{i=1}^{n} \hat{\mathbf{g}}_{i,1}\hat{\mathbf{g}}'_{i,1}]^{-1}$)

$\hat{\mathbf{V}}_2 =$ Estimated asymptotic covariance matrix for $\hat{\theta} \,|\, \hat{\delta}$

$\mathbf{g}_{i,2} = \partial \log f_{i,2}(\ldots,\theta,\hat{\delta}) / \partial\theta$. ($\hat{\mathbf{V}}_2$ might = $[\Sigma_{i=1}^{n} \hat{\mathbf{g}}_{i,2}\hat{\mathbf{g}}'_{i,2}]^{-1}$)

$\mathbf{h}_{i,2} = \partial \log f_{i,2}(\ldots,\theta,\hat{\delta}) / \partial\hat{\delta}$

$\mathbf{C} \quad = \Sigma_{i=1}^{n} \hat{\mathbf{g}}_{i,2}\hat{\mathbf{h}}'_{i,2}$ (the off diagonal block in the Hessian)

$\mathbf{R} \quad = \Sigma_{i=1}^{n} \hat{\mathbf{g}}_{i,2}\hat{\mathbf{g}}_{i,1}$ (cross products of derivatives for two logL's)

M&T: Corrected $\hat{\mathbf{V}}_2 = \hat{\mathbf{V}}_2 + \hat{\mathbf{V}}_2[\mathbf{C}\hat{\mathbf{V}}_1\mathbf{C}' - \mathbf{C}\hat{\mathbf{V}}_1\mathbf{R}' - \mathbf{R}\hat{\mathbf{V}}_1\mathbf{C}']\hat{\mathbf{V}}_2$

Application of M&T

```
Names    ; z1=one,age,bluec$
Logit    ; lhs=hhkids ; rhs=z1 ; prob=prifkids $
Matrix   ; v1=varb$
Create   ; gi1=hhkids-prifkids$
Names    ; z2 = one,educ,married,age,prifkids
Loglinear; lhs=hhninc;rhs=z2;model=e$
Matrix   ; v2=varb$
Create   ; gi2=hhninc*exp(z2'b)-1$
Create   ; hi2=gi2*b(5)*prifkids*(1-prifkids)$
Create   ; varc=gi1*gi2 ; varr=gi1*hi2$
Matrix   ; c=z2'[varc]z1 ; r=z2'[varr]z1$
Matrix   ; q=c*v1*c'-c*v1*r'-r*v1*c’
         ; mt=v2+v2*q*v2;stat(b,mt)$
```

M&T Application

```
+---------------------------------------------+
| Multinomial Logit Model                     |
| Dependent variable                HHKIDS    |
+---------+--------------+----------------+--------+---------+----------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
+---------+--------------+----------------+--------+---------+----------+
          Characteristics in numerator of Prob[Y = 1]
 Constant      2.61232320       .05529365    47.245   .0000
 AGE           -.07036132       .00125773   -55.943   .0000    43.5271942
 BLUEC         -.02474434       .03052219     -.811   .4175     .24379621
+---------------------------------------------+
| Exponential (Loglinear) Regression Model    |
| Dependent variable                HHNINC    |
+---------+--------------+----------------+--------+---------+----------+
|Variable | Coefficient  | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
+---------+--------------+----------------+--------+---------+----------+
          Parameters in conditional mean function
 Constant     -3.79588863       .44440782    -8.541   .0000
 EDUC          -.05580594       .00267736   -20.844   .0000    11.3201838
 MARRIED       -.20232648       .01487166   -13.605   .0000     .75869263
 AGE            .08112565       .00633014    12.816   .0000    43.5271942
 PRIFKIDS      5.23741034       .41248916    12.697   .0000     .40271576
+---------+--------------+----------------+--------+---------+
 B_1          -3.79588863       .44425516    -8.544   .0000
 B_2           -.05580594       .00267540   -20.859   .0000
 B_3           -.20232648       .01486667   -13.609   .0000
 B_4            .08112565       .00632766    12.821   .0000
 B_5           5.23741034       .41229755    12.703   .0000
 Why so little change?  N = 27,000+.  No new variation.
```

GMM Estimation

$$\bar{\mathbf{g}}(\boldsymbol{\beta}) = \frac{1}{N}\Sigma_{i=1}^{N}\mathbf{m}_i(y_i,\mathbf{x}_i,\boldsymbol{\beta})$$

Asy.Var$[\bar{\mathbf{g}}(\boldsymbol{\beta})]$ estimated with

$$\mathbf{W} = \frac{1}{N}\left(\frac{1}{N}\Sigma_{i=1}^{N}\mathbf{m}_i(y_i,\mathbf{x}_i,\boldsymbol{\beta})\mathbf{m}_i(y_i,\mathbf{x}_i,\boldsymbol{\beta})'\right)$$

The GMM estimator of $\boldsymbol{\beta}$ then minimizes

$$q = \left(\frac{1}{N}\Sigma_{i=1}^{N}\mathbf{m}_i(y_i,\mathbf{x}_i,\boldsymbol{\beta})\right)'\mathbf{W}^{-1}\left(\frac{1}{N}\Sigma_{i=1}^{N}\mathbf{m}_i(y_i,\mathbf{x}_i,\boldsymbol{\beta})\right).$$

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}_{\mathbf{GMM}}] = [\mathbf{G'W^{-1}G}]^{-1},\ \mathbf{G} = \frac{\partial\bar{\mathbf{g}}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}'}$$

GMM Estimation-1

- GMM is broader than M estimation and ML estimation
- Both M and ML are GMM estimators.

$$\bar{\mathbf{g}}\boldsymbol{\beta}\ ) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f(y_i \mid x_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \text{ for MLE}$$

$$\bar{\mathbf{g}}\boldsymbol{\beta}\ ) = \frac{1}{n}\sum_{i=1}^{n} -e_i \frac{\partial E(y_i \mid x_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \text{ for NLSQ}$$

GMM Estimation - 2

Exactly identified GMM problems

When $\bar{\mathbf{g}}(\boldsymbol{\beta}) = \frac{1}{N}\Sigma_{i=1}^{N}\mathbf{m}_i(y_i, \mathbf{x}_i, \boldsymbol{\beta}) = \mathbf{0}$ is K equations in K unknown parameters (the exactly identified case), the weighting matrix in

$$q = \left(\frac{1}{N}\Sigma_{i=1}^{N}\mathbf{m}_i(y_i, \mathbf{x}_i, \boldsymbol{\beta})\right)'\mathbf{W}^{-1}\left(\frac{1}{N}\Sigma_{i=1}^{N}\mathbf{m}_i(y_i, \mathbf{x}_i, \boldsymbol{\beta})\right)$$

is irrelevant to the solution, since we can set exactly $\bar{\mathbf{g}}(\boldsymbol{\beta}) = \mathbf{0}$ so q = 0. And, the asymptotic covariance matrix (estimator) is the product of 3 square matrices and becomes $[\mathbf{G'W^{-1}G}]^{-1} = \mathbf{G^{-1}WG'^{-1}}$

Optimization - Algorithms

Maximize or minimize (optimize) a function $F(\theta)$

Algorithm = rule for searching for the optimizer

Iterative algorithm: $\theta^{(k+1)} = \theta^{(k)} + \text{Update}^{(k)}$

Derivative (gradient) based algorithm

$$\theta^{(k+1)} = \theta^{(k)} + \text{Update}(\mathbf{g}^{(k)})$$

Update is a function of the gradient.

Compare to 'derivative free' methods

(Discontinuous criterion functions)

Optimization

Algorithms

Iteration $\theta^{(k+1)} = \theta^{(k)} + \text{Update}^{(k)}$

General structure: $\theta^{(k+1)} = \theta^{(k)} + \lambda^{(k)}\mathbf{W}^{(k)}\mathbf{g}^{(k)}$

$\mathbf{g}^{(k)}$ = derivative vector, points to a better value than $\theta^{(k)}$

= direction vector

$\lambda^{(k)}$ = 'step size'

$\mathbf{W}^{(k)}$ = a weighting matrix

Algorithms are defined by the choices of $\lambda^{(k)}$ and $\mathbf{W}^{(k)}$

Algorithms

Steepest Ascent: $\lambda^{(k)} = \dfrac{-\mathbf{g}^{(\mathbf{k})\prime}\mathbf{g}^{(\mathbf{k})}}{\mathbf{g}^{(\mathbf{k})\prime}\mathbf{H}^{(\mathbf{k})}\mathbf{g}^{(\mathbf{k})}}$, $\mathbf{W}^{(k)} = \mathbf{I}$

$\mathbf{g}^{(\mathbf{k})}$ = first derivative vector

$\mathbf{H}^{(\mathbf{k})}$ = second derivatives matrix

Newton's Method: $\lambda^{(k)} = -1$, $\mathbf{W}^{(k)} = [\mathbf{H}^{(\mathbf{k})}]^{-1}$

(Sometimes called Newton-Raphson.

Method of Scoring: $\lambda^{(k)} = -1$, $\mathbf{W}^{(k)} = [E[\mathbf{H}^{(\mathbf{k})}]]^{-1}$

(Scoring uses the expected Hessian. Usually inferior to Newton's method. Takes more iterations.)

BHHH Method (for MLE): $\lambda^{(k)} = -1$, $\mathbf{W}^{(k)} = [\Sigma_{i=1}^{n}\mathbf{g}_i^{(\mathbf{k})}\mathbf{g}_i^{(\mathbf{k})}{}']^{-1}$

Line Search Methods

Squeezing: Essentially trial and error

$\lambda^{(k)} = 1, 1/2, 1/4, 1/8, ...$

Until the function improves

Golden Section: Interpolate between $\lambda^{(k)}$ and $\lambda^{(k-1)}$

Others : Many different methods have been suggested

Quasi-Newton Methods

How to construct the weighting matrix:

Variable metric methods:

$$\mathbf{W}^{(k)} = \mathbf{W}^{(k-1)} + \mathbf{E}^{(k-1)},\ \mathbf{W}^{(1)} = \mathbf{I}$$

Rank one updates: $\mathbf{W}^{(k)} = \mathbf{W}^{(k-1)} + \mathbf{a}^{(k-1)}\mathbf{a}^{(k-1)\prime}$

(Davidon Fletcher Powell)

There are rank two updates (Broyden) and higher.

Stopping Rule

When to stop iterating: 'Convergence'

(1) Derivatives are small? Not good.

Maximizer of F() is the same as that of .0000001F(), but the derivatives are small right away.

(2) Small absolute change in parameters from one iteration to the next? Problematic because it is a function of the stepsize which may be small.

(3) Commonly accepted 'scale free' measure

$$\Delta = \mathbf{g}^{(k)\prime}[\mathbf{H}^{(k)}]^{-1}\mathbf{g}^{(k)}$$

For Example

```
Nonlinear Estimation of Model Parameters
Method=BFGS  ; Maximum iterations=  4
Convergence criteria:gtHg   .1000D-05 chg.F   .0000D+00 max|dB|   .0000D+00
Start values:  -.10437D+01   .00000D+00   .00000D+00   .00000D+00   .10000D+01
1st derivs.    -.23934D+05  -.26990D+06  -.18037D+05  -.10419D+07   .44370D+05
Parameters:    -.10437D+01   .00000D+00   .00000D+00   .00000D+00   .10000D+01
Itr  1 F=  .3190D+05 gtHg=  .1078D+07 chg.F=  .3190D+05 max|db|=  .1042D+13
Try =  0 F=  .3190D+05 Step=  .0000D+00 Slope= -.1078D+07
Try =  1 F=  .4118D+06 Step=  .1000D+00 Slope=  .2632D+08
Try =  2 F=  .5425D+04 Step=  .5214D-01 Slope=  .8389D+06
Try =  3 F=  .1683D+04 Step=  .4039D-01 Slope= -.1039D+06
1st derivs.    -.45100D+04  -.45909D+05  -.18517D+04  -.95703D+05  -.53142D+04
Parameters:    -.10428D+01   .10116D-01   .67604D-03   .39052D-01   .99834D+00
Itr  2 F=  .1683D+04 gtHg=  .1064D+06 chg.F=  .3022D+05 max|db|=  .4538D+07
Try =  0 F=  .1683D+04 Step=  .0000D+00 Slope= -.1064D+06
Try =  1 F=  .1006D+06 Step=  .4039D-01 Slope=  .7546D+07
Try =  2 F=  .1839D+04 Step=  .4702D-02 Slope=  .1847D+06
Try =  3 F=  .1582D+04 Step=  .1855D-02 Slope=  .7570D+02
...
1st derivs.    -.32179D-05  -.29845D-04  -.28288D-05  -.16951D-03   .73923D-06
Itr 20 F=  .1424D+05 gtHg=  .1389D-07 chg.F=  .1155D-08 max|db|=  .1706D-08
                        * Converged
Normal exit from iterations. Exit status=0.
Function=  .31904974915D+05, at entry, -.14237328819D+05 at exit
```