Nonlinear Statistical Models, Chapter 1, Univariate Nonlinear Regression

by

A. Ronald Gallant
Department of Economics
University of North Carolina
Chapel Hill NC 27599-3305 USA

References

Gallant, A. Ronald (1987) *Nonlinear Statistical Models,* Wiley, New York.

Gallant, A. Ronald (1992) *Nonlinear Regression Asymptotics,* Manuscript, Department of Economics, University of North Carolina.

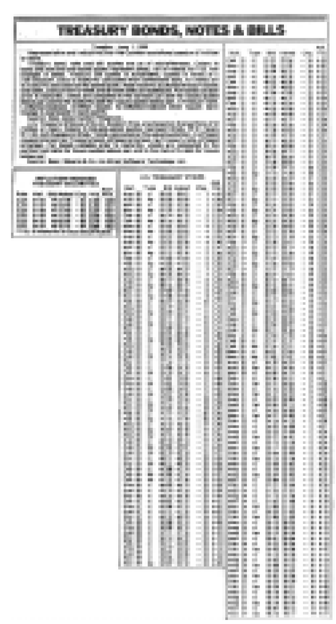
Gallant, A. Ronald (1997) *Introduction to Econometric Theory,* Princeton University Press, Princeton NJ.

Fletcher, R. (1987) *Practical Methods of Optimization, Second Edition,* Wiley, New York

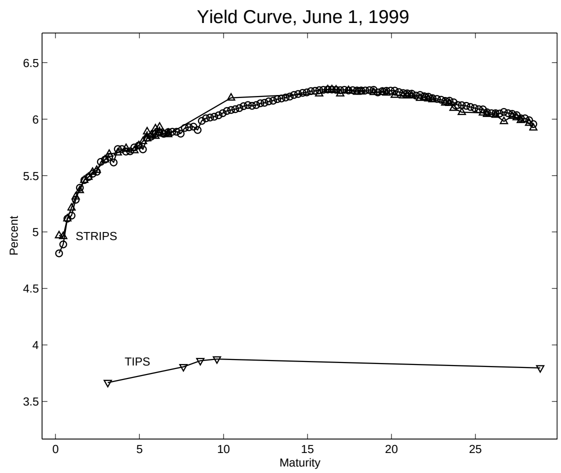
Topics

- Examples & Least Squares Estimates
- Notation & Taylor's Theorem
- Statistical Properties
- Computations
- Hypothesis Tests
- Confidence Intervals

Bond Prices



Source: Wall Street Journal, June 2, 1999
Files: strips99.dat, tips99.dat



Shown are continuously compounded yields. The return for TIPS is computed by adding the coupon rate to the continuously compounded return on the principal.

Strips marked with a triangle are principal strips; strips marked with a circle are coupon interest strips. There is no conceptual difference between them. Lines merely connect points.

Consumption Based Asset Pricing (1)

Given random income $\{w_t\}$, price level $\{p_t\}$, and securities that sell at price $\{S_{jt}\}$ at time t, have payoff $\{S_{j,t+m_j}\}$ at time $t+m_j$, and cannot be sold in the interim, the consumer's problem is to choose consumption $\{c_t\}$ and portfolio $\{q_{jt}\}$ to maximize

$$\mathcal{E}_0\left(\sum_{t=0}^{\infty}\delta^s\frac{c_t^{1-\gamma}}{1-\gamma}\right)$$

subject to

$$p_tc_t+\sum_{j=1}^{J}q_{jt}S_{jt}\le w_t+\sum_{j=1}^{J}q_{j,t-m_j}S_{jt}$$

where $0<\delta<1$ and $0\le\gamma$. The solution to this problem must satisfy the Euler equation

$$S_t=\mathcal{E}_t\left\{\left[\delta^{m_j}\left(\frac{c_{t+m_j}}{c_t}\right)^{-\gamma}\frac{p_t}{p_{t+m_j}}\right]S_{t+m_j}\right\}.$$

Reference: Soderlind, Paul, and Lars Svensson (1999) "New Techniques to Extract Market Expectations from Financial Instruments," *Journal of Monetary Economics* 40, 383–429.

Consumption Based Asset Pricing (2)

Putting $T=t+m$, the term

$$D(t,T)=\delta^{T-t}\left(\frac{c_T}{c_t}\right)^{-\gamma}\frac{p_t}{p_T}$$

is called, variously, the stochastic discount factor, the pricing kernel, or state price density. In logs,

$$\log D(t,T)=m\log\delta-\gamma(\log c_T-\log c_t)-(\log p_T-\log p_t)$$

Assume that log consumption follows a drifting random walk with normally distributed increments

$$\log c_{s+1}-\log c_s\sim N(\mu_c,\sigma_c^2),$$

that the price level follows a trending autoregression with normal errors

$$\log p_{s+1}-g(s+1)\sim N\left\{\rho\left[\log p_s-g(s)\right],\sigma_p^2\right\},$$

and that consumption and inflation are independent. Then, conditional on c_t and p_t,

$$\log c_T\sim N\Big(m\mu_c+\log c_t,\,m\sigma_c^2\Big)$$

$$\log p_T-\log p_t\sim N\left\{g(t+m)-g(t)+(\rho^m-1)\left[\log p_t-g(t)\right],\sigma_p^2\sum_{j=0}^{m-1}\rho^{2j}\right\}$$

Choice of $g(s)$

We shall choose $g(s)$ in

$$\log p_T-\log p_t\sim N\left\{g(t+m)-g(t)+(\rho^m-1)\left[\log p_t-g(t)\right],\sigma_p^2\sum_{j=0}^{m-1}\rho^{2j}\right\}$$

to satisfy the differential equation

$$dg(t)=\left\{t\rho^{t-1}[g(t)-a-bt]+\rho^tb-b\right\}dt$$

which integrates to

$$g(t+m)-g(t)-(\rho^m-1)\,g(t)=-(\rho^m-1)\,(a+bm)$$

to give

$$\log p_T-\log p_t\sim N\left\{(\rho^m-1)\left[\log p_t-a-bm\right],\sigma_p^2\left(\frac{1-\rho^{2m}}{2-2\rho^2}\right)\right\}.$$

There is no particular merit to this choice other than it fits the data much better than many other more obvious choices.

U.S. Treasury Strips

A discount bond that pays

$$S_T = \$1$$

at time $T = t + m$ will have price

$$S_t = \mathcal{E}_t D(t,T) = \mathcal{E}_t \exp[\log D(t,T)],$$

which, from the formula for the moment generating function of the normal, is

$$S_t = \delta^m \exp\left[-m\left(\gamma\mu_c - \frac{\gamma^2\sigma_c^2}{2}\right) + (1-\rho^m)(\log p_t - a - bm) + \sigma_p^2\left(\frac{1-\rho^{2m}}{2-2\rho^2}\right)\right]$$

This derivation has assumed that the time increment is one year and that m is an integer. Although we could derive the formula on a daily basis, keep an exact count of days within a month, and account for leap years, we shall not. Rather, we shall merely apply this formula with fractional m.

U.S. Treasury Inflation Protected Bonds (1)

The value of the principal payment

$$B_T = \$\left(\frac{p_T}{p_t}\right)P_t$$

of an inflation indexed bond that has accrued principal P_t at time t and matures at time $T = t + m$ is

$$B_t = P_t\,\delta^m \exp\left[-m\left(\gamma\mu_c - \frac{\gamma^2\sigma_c^2}{2}\right)\right]$$

The value of the stream of semi-annual coupon payments

$$C_{T_j} = \$\,\frac{r}{2}P_t\left(\frac{p_{T_j}}{p_t}\right) \qquad j = 1,\ldots,J$$

is

$$C_t = \frac{r}{2}P_t\sum_{t=1}^{J}\delta^{m_j}\exp\left[-m_j\left(\gamma\mu_c - \frac{\gamma^2\sigma_c}{2}\right)\right].$$

where $J = \lceil 2m \rceil$, and

$$T_j = T - \frac{1}{2}(j-1) \qquad m_j = m - \frac{1}{2}(j-1)$$

The bond price at time t is the sum

$$S_t = B_t + C_t.$$

U.S. Treasury Inflation Protected Bonds (2)

For TIPS, we shall compute the payoff as

$$R_T = P_t \exp(rm).$$

With this assumptions, a continuously compounded yield on a TIPS is

$$\begin{aligned} y &= \frac{\log R_T - \log(\text{Asked})}{m} \\ &= r + \frac{\log P_t - \log(\text{Asked})}{m}. \end{aligned}$$

The only reason for making this assumption is allow us to display a yield for TIPS on graphics. It does not affect any computations.

Bond Prices

To simplify notation, rewrite

$$\begin{aligned} S_t &= \delta^m \exp\left[-m\left(\gamma\mu_c - \frac{\gamma^2\sigma_c^2}{2}\right) + (1-\rho^m)(\log p_t - a - bm) + \sigma_p^2\left(\frac{1-\rho^{2m}}{2-2\rho^2}\right)\right] \\ B_t &= P_t\,\delta^m \exp\left[-m\left(\gamma\mu_c - \frac{\gamma^2\sigma_c^2}{2}\right)\right] \\ C_t &= \frac{r}{2}P_t\sum_{t=1}^{J}\delta^{m_j}\exp\left[-m_j\left(\mu_c - \frac{\gamma^2\sigma_c}{2}\right)\right]. \end{aligned}$$

as

$$\begin{aligned} S_t &= \theta_1^m \exp\left[(1-\theta_2^m)(\theta_3 + \theta_4 m) + \theta_5(1-\theta_2^{2m})\right] \\ B_t &= P_t\,\theta_1^m \\ C_t &= \frac{rP_t}{2}\sum_{t=1}^{J}\theta_1^{m_j}. \end{aligned}$$

where

$$\begin{aligned} \theta_1 &= \delta\,\exp\left[-\gamma\mu_c + \frac{\gamma^2\sigma_c^2}{2}\right] \\ \theta_2 &= \rho \\ \theta_3 &= \log p_t - a \\ \theta_4 &= -b \\ \theta_5 &= \frac{\sigma_p^2}{2-2\rho^2}. \end{aligned}$$

Nonlinear Regression Model

$$y_i = \begin{cases} -\frac{1}{m}\log(\text{Asked}) & \text{strip} \\ \frac{1}{m}[\log R_T - \log(\text{Asked})] & \text{tip} \end{cases}$$

$$f(x_i, \theta) = \begin{cases} -\log\theta_1 - (1-\theta_2^m)\left(\frac{\theta_3}{m} + \theta_4\right) - \frac{\theta_5}{m}\left(1-\theta_2^{2m}\right) & \text{strip} \\ \frac{1}{m}\left[\log R_T - \log\left(P_t\,\theta_1^m + \frac{rP_t}{2}\sum_{t=1}^{J}\theta_1^{m_j}\right)\right] & \text{tip} \end{cases}$$

$$x_i = \begin{cases} (m, 1, 0, 0) & \text{strip } (x_{i3} = 0) \\ (m, P_t, \frac{rP_t}{2}, 1) & \text{tip } (x_{i3} = 0) \end{cases}$$

$$i = 1, \ldots, n = 179$$

SAS code (data preparation)

```
data strips;
  infile 'strips99.dat';
  input mm yy src $ bid0 bid1 ask0 ask1 chg yld;
  if (yy = 99) then yy=-1;
  type = 0;
  /* June 1 trade date, June 3 settlement date */
  mat = 1.0 + yy + (mm-6.0)/12.0 + 13.0/365.25;
  prn = 1.0; cpn = 0.0;
  ask = (ask0 + ask1/32)/100.0;
  J = ceil(2.0*mat);
  y = -log(ask)/mat;
  pmt = 1;
  keep mat ask prn pmt cpn J y type;

data tips;
  infile 'tips99.dat';
  input r mm yy bid0 bid1 ask1 chg yld prn;
  if (yy = 99) then yy=-1;
  r = r/100.0;
  type = 1;
  /* June 1 trade date, June 3 settlement date */
  mat = 1.0 + yy + (mm-6.0)/12.0 + 13.0/365.25;
  prn = prn/1000.0;
  ask = prn*(bid0 + ask1/32)/100.0;
  J = ceil(2.0*mat);
  cpn = (r/2.0)*prn;
  pmt = prn*exp(r*mat);
  y = log(pmt)/mat - log(ask)/mat;
  keep mat ask prn pmt cpn J y type;

data bonds;
  set strips tips;
```


SAS code (nonlinear regression)

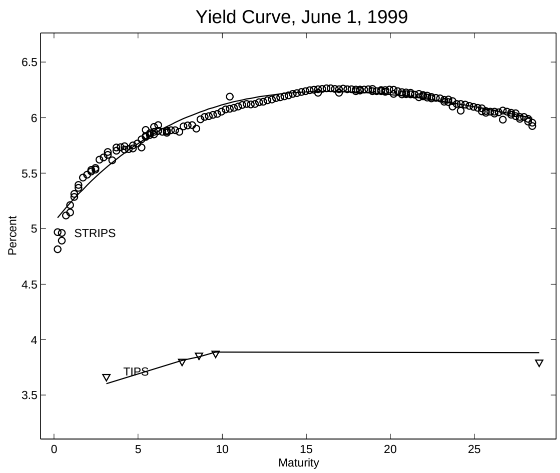
```
proc nlin data=bonds method=gauss iter=400 convergence=1.0e-5;
  parms t1=0.96 t2=0.9 t3=0.01 t4=0.01 t5=0.01;
  if (type = 1 ) then
  do
    B = prn*(t1**mat); dBwt1 = mat*B/t1;
    C = 0.0;           dCwt1 = 0.0;
    do jj=1 to J;
      matj = mat-(jj-1.0)/2.0;
      Cj = cpn*(t1**matj);
      C = C + Cj;
      dCwt1 = dCwt1 + matj*Cj/t1;
    end;
    f = log(pmt)/mat - log(B+C)/mat;
    dfwt1 = -(dBwt1+dCwt1)/((B+C)*mat);
    dfwt2 = 0; dfwt3 = 0; dfwt4 = 0; dfwt5 = 0;
  end;
  else
  do
    tmp1 = t2**mat; tmp2 = tmp1**2;
    f = - log(t1) - (1.0-tmp1)*(t3/mat+t4) - (t5/mat)*(1.0-tmp2);
    dfwt1 = - (1.0/t1);
    dfwt2 = (mat*tmp1/t2)*(t3/mat+t4)+(t5/mat)*(2.0*mat*tmp2)/t2;
    dfwt3 = - (1.0-tmp1)*(1.0/mat);
    dfwt4 = - (1.0-tmp1);
    dfwt5 = - (1.0/mat)*(1.0-tmp2);
  end;
  model y = f;
  der.t1=dfwt1; der.t2=dfwt2; der.t3=dfwt3;
  der.t4=dfwt4; der.t5=dfwt5;
  output out = fit p = yhat;

data _null_;
  set fit;
  file "fit.dat";
  put mat 10.5 y 10.5 yhat 10.5 type 4.0;
```


SAS output

```
Non-Linear Least Squares Summary Statistics     Dependent Variable Y

  Source                DF Sum of Squares     Mean Square

  Regression             5  0.62926865067   0.12585373013
  Residual             174  0.00006009685   0.00000034538
  Uncorrected Total    179  0.62932874752

  (Corrected Total)    178  0.00397610036


  Parameter     Estimate    Asymptotic             Asymptotic 95 %
                            Std. Error         Confidence Interval
                                               Lower         Upper
  T1         0.960443236 0.00030435051  0.9598425357  0.9610439365
  T2         0.955169625 0.00492005543  0.9454588504  0.9648804005
  T3        -2.940159096 0.80538700259 -4.5297615176 -1.3505566750
  T4         0.015765485 0.00772758056  0.0005134622  0.0310175074
  T5         1.358617510 0.38900892850  0.5908257119  2.1264093075
```

The implied real rate is

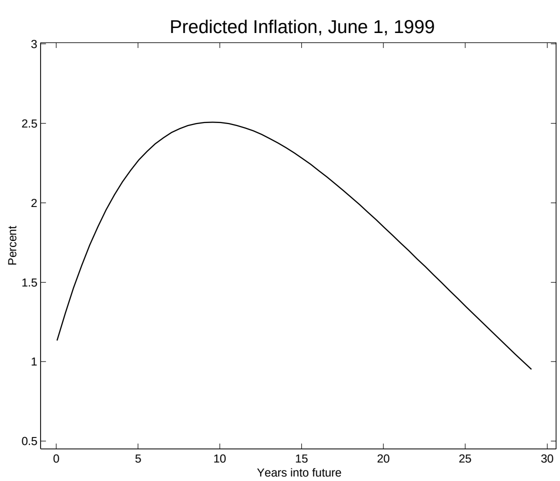
$$-100\log\hat{\theta}_1 = 4.04\%$$



Solid line is the nonlinear least squares fit. Yields are continuously compounded yields. The return for TIPS is computed by adding the coupon rate to the continuously compounded return on the principal.

<u>Matlab code</u> (predicted inflation)

```
T1 = 0.960443236; T2 = 0.955169625; T3 = -2.940159096;
T4 = 0.015765485; T5 = 1.358617510;

mat =  0.05:.5:29.55;   n = length(mat);

%logEd is the logarithm of expected deflator, given p_t. If p_t = 1,
%logEd is the logarithm of expected 1/P_T given p_t, which is

logEd = (1.0 - T2.^mat).*(T3 + T4.*mat) + T5.*(1.0 - T2.^(2.0*mat));

%inflation is defined as the change in -log((EP_t/p_t));

inflat = -100.0*(logEd(2:n) - logEd(1:n-1))./(mat(2:n) - mat(1:n-1));

left = min(mat) - 1.0;
rite = max(mat) + 1.0;
bot = min(inflat) - .5;
top = max(inflat) + .5;

figure(1);
plot(mat(1:n-1),inflat,'-','LineWidth',1.0);
axis([left rite bot top]);
title( '\fontsize{16} Predicted Inflation, June 1, 1999');
xlabel('Years into future');
ylabel('Percent');

print -r300 -deps2 bonds03.ps;
```

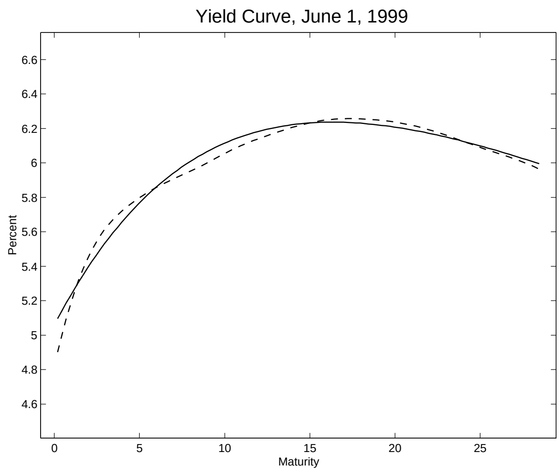
Inflation is defined here in terms of the price deflator. Plotted is the change in the conditional expectation of the deflator given past prices expressed as a percentage, which is

$$-100\left(\frac{d}{d\tau}\right)\log\left[\mathcal{E}\left(\frac{p_t}{p_{t+\tau}}\,\middle|\,p_t\right)\right],$$

against years into the future τ.

<u>Matlab code</u> (local quadratic regression)

```
 .
 .
 .

n = length(y);
X = [ones(n,1),m,m.^2];
h = 2;

for i=1:n
   w = normpdf(m(i),m,h);
   WX = [w.*X(:,1),w.*X(:,2),w.*X(:,3)];
   beta = inv(WX'*X)*WX'*y;
   yhat(i) = beta(1) + beta(2)*m(i) + beta(3)*m(i)*m(i);
end

 .
 .
 .
```



Solid line is the nonlinear least squares fit. Dashed line is the local quadratic fit. Yields are continuously compounded yields.

Statistical Model

$$y_t = f(x_t, \theta) + e_t \quad t = 1, 2, \ldots, n$$

- y_t the dependent variable, univariate, observed
- x_t the explanatory variables, k-variate, observed
- θ model parameters, p-variate, unknown (to be estimated)
- e_t the error, univariate, unobserved (because θ is unknown) $\mathcal{E}(e_t) = 0$, $\mathrm{Var}(e_t) = \sigma^2$, iid

Least Squares Estimator

$$\hat{\theta}_n = \underset{\theta \in \Theta}{\mathrm{argmin}}\ \mathrm{SSE}(\theta)$$

$$\mathrm{SSE}(\theta) = \sum_{t=1}^{n} [y_t - f(x_t, \theta)]^2$$

Example 1, Chapt. 1, NLSM

$$f(x, \theta) = \theta_1 x_1 + \theta_2 x_2 + \theta_4 e^{\theta_3 x_3}$$

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \quad k = 3$$

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{pmatrix} \quad p = 4$$

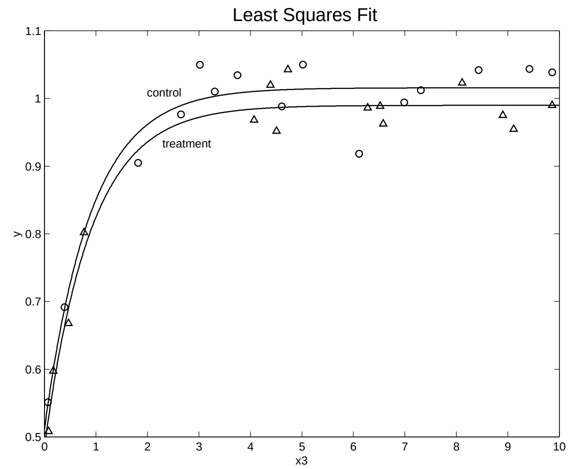
Table 1. Data Values for Example 1.

t	y	x_1	x_2	x_3
1	0.98610	1	1	6.28
2	1.03848	0	1	9.86
3	0.95482	1	1	9.11
4	1.04184	0	1	8.43
5	1.02324	1	1	8.11
6	0.90475	0	1	1.82
7	0.96263	1	1	6.58
8	1.05026	0	1	5.02
9	0.98861	1	1	6.52
10	1.03437	0	1	3.75
11	0.98982	1	1	9.86
12	1.01214	0	1	7.31
13	0.66768	1	1	0.47
14	0.55107	0	1	0.07
15	0.96822	1	1	4.07
16	0.98823	0	1	4.61
17	0.59759	1	1	0.17
18	0.99418	0	1	6.99
19	1.01962	1	1	4.39
20	0.69163	0	1	0.39
21	1.04255	1	1	4.73
22	1.04343	0	1	9.42
23	0.97526	1	1	8.90
24	1.04969	0	1	3.02
25	0.80219	1	1	0.77
26	1.01046	0	1	3.31
27	0.95196	1	1	4.51
28	0.97658	0	1	2.65
29	0.50811	1	1	0.08
30	0.91840	0	1	6.11

Source: Gallant (1976)
File: amstat.dat

SAS code (nonlinear regression)

```
data amstat;
  infile "amstat.dat";
  input t y x1 x2 x3;

proc nlin data=amstat method=gauss iter=50 convergence=1.0e-5;
  parms t1=-0.048660 t2=1.038835 t3=-0.737919 t4=-0.513623;
  model y=t1*x1+t2*x2+t4*exp(t3*x3);
  der.t1=x1; der.t2=x2; der.t3=t4*x3*exp(t3*x3);
  der.t4=exp(t3*x3);
  output out = fit p = yhat;

data _null_;
  set fit;
  file "fit.dat";
  put t 5.0 y 10.5 x1 5.0 x2 5.0 x3 10.5 yhat 10.5;
```

Solid lines are the nonlinear least squares fit. Circles indicate control observations ($x_1 = 0$); triangle indicate treatment observations ($x_1 = 1$);

Example 1 (continued)

The inputs correspond to a one way "treatment-control" design that uses experimental material whose age ($= x_3$) affects the response exponentially. That is, the first observation

$$x_1 = (1, 1, 6.28)'$$

represents experimental material with attained age $x_3 = 6.28$ months that was (randomly) allocated to the treatment group and has expected response

$$f(x_1, \theta^o) = \theta_1^o + \theta_2^o + \theta_4^o e^{6.28\theta_3^o}.$$

Similarly, the second observation

$$x_1 = (0, 1, 9.86)'$$

represents an allocation of material with attained age $x_3 = 9.86$ to the control group, with expected response

$$f(x_2, \theta^o) = \theta_2^o + \theta_4^o e^{9.86\theta_3^o}.$$

and so on. The parameter θ_1^o is the treatment effect. The data of Table 1 are simulated.

Topics

- Examples & Least Squares Estimates
- Notation & Taylor's Theorem
- Statistical Properties
- Computations
- Hypothesis Tests
- Confidence Intervals

Vector Notation(1)

The nonlinear regression equations

$$y_t = f(x_t, \theta^o) + e_t \quad t = 1, 2, \ldots, n$$

may be written in a convenient vector form

$$y = f(\theta^o) + e$$

by adopting conventions analogous to those employed in linear regression; namely

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$f(\theta) = \begin{pmatrix} f(x_1, \theta) \\ f(x_2, \theta) \\ \vdots \\ f(x_n, \theta) \end{pmatrix}$$

$$e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

Vector Notation(2)

The sum of squared deviations

$$\text{SSE}(\theta) = \sum_{t=1}^{n} [y_t - f(x_t, \theta)]^2$$

of the observed y_t from the predicted value $f(x_t, \theta)$ corresponding to a trial value of the parameter θ becomes

$$\text{SSE}(\theta) = [y - f(\theta)]'[y - f(\theta)] = \|y - f(\theta)\|^2$$

in this vector notation.

For Example 1,

$$y_t = \theta_1 x_{1t} + \theta_2 x_{2t} + \theta_4 e^{\theta_3 x_{3t}} + e_t \quad t = 1, \ldots, 30$$

these vectors are

$$y = \begin{pmatrix} 0.98610 \\ 1.03848 \\ \vdots \\ 0.50811 \\ 0.91840 \end{pmatrix}$$

$$f(\theta) = \begin{pmatrix} \theta_1 + \theta_2 + \theta_4 e^{\theta_3 6.20} \\ \theta_2 + \theta_4 e^{\theta_3 9.86} \\ \vdots \\ \theta_1 + \theta_2 + \theta_4 e^{\theta_3 0.08} \\ \theta_2 + \theta_4 e^{\theta_3 6.11} \end{pmatrix}$$

$$e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_{29} \\ e_{30} \end{pmatrix}$$

Linear Pseudo-Model

The estimators employed in nonlinear regression can be characterized as linear and quadratic forms in the vector e which are similar to those that appear in linear regression. Let

$$F(\theta) = \frac{\partial}{\partial \theta'} f(\theta);$$

i.e., $F(\theta)$ is the matrix with typical element $(\partial/\partial\theta_j) f(x_t, \theta)$, where t is the row index and j is the column index. The matrix $F(\theta^o)$ plays the same role as the design matrix X in the linear regression

$$\text{"}y\text{"} = X\beta + e.$$

The appropriate analogy is obtained by setting

$$\text{"}y\text{"} = y - f(\theta^o) + F(\theta^o)\theta^o$$

and

$$X = F(\theta^o).$$

We shall write F for the matrix $F(\theta)$ when it is evaluated at $\theta = \theta^o$, i.e.,

$$F = F(\theta^o).$$

For Example 1,

$$y_t = \theta_1 x_{1t} + \theta_2 x_{2t} + \theta_4 e^{\theta_3 x_{3t}} + e_t \quad t = 1, \ldots, 30$$

$$F(\theta) = \begin{pmatrix} 1 & 1 & 6.28\,\theta_4 e^{6.28\theta_3} & e^{6.28\theta_3} \\ 0 & 1 & 9.86\,\theta_4 e^{9.86\theta_3} & e^{9.86\theta_3} \\ 1 & 1 & 9.11\,\theta_4 e^{9.11\theta_3} & e^{9.11\theta_3} \\ 0 & 1 & 8.43\,\theta_4 e^{8.43\theta_3} & e^{8.43\theta_3} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0.08\,\theta_4 e^{0.08\theta_3} & e^{0.08\theta_3} \\ 0 & 1 & 6.11\,\theta_4 e^{6.11\theta_3} & e^{6.11\theta_3} \end{pmatrix}$$

which is of order 30×4.

Gradients, Jacobians, and Hessians(1)

Suppose that $s(\theta)$ is a real valued function of a p-dimensional argument θ. The notation $(\partial/\partial\theta)s(\theta)$ denotes the **gradient** of $s(\theta)$:

$$\frac{\partial}{\partial\theta}s(\theta) = \begin{pmatrix} \frac{\partial}{\partial\theta_1}s(\theta) \\ \frac{\partial}{\partial\theta_2}s(\theta) \\ \vdots \\ \frac{\partial}{\partial\theta_p}s(\theta) \end{pmatrix}$$

a p by 1 (column) vector with typical element $(\partial/\partial\theta_i)s(\theta)$. Its transpose is denoted by

$$\frac{\partial}{\partial\theta'}s(\theta) = \left(\frac{\partial}{\partial\theta_1}s(\theta), \ \frac{\partial}{\partial\theta_2}s(\theta), \ \ldots, \ \frac{\partial}{\partial\theta_p}s(\theta) \right).$$

Gradients, Jacobians, and Hessians(2)

Suppose that all second order derivatives of $s(\theta)$ exist. They can be arranged in a p by p matrix, known as the **Hessian** matrix of the function $s(\theta)$,

$$\frac{\partial^2}{\partial\theta\partial\theta'}s(\theta) = \begin{pmatrix} \frac{\partial^2}{\partial\theta_1\partial\theta_1}s(\theta) & \frac{\partial^2}{\partial\theta_1\partial\theta_2}s(\theta) & \cdots & \frac{\partial^2}{\partial\theta_1\partial\theta_p}s(\theta) \\ \frac{\partial^2}{\partial\theta_2\partial\theta_1}s(\theta) & \frac{\partial^2}{\partial\theta_2\partial\theta_2}s(\theta) & \cdots & \frac{\partial^2}{\partial\theta_2\partial\theta_p}s(\theta) \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2}{\partial\theta_p\partial\theta_1}s(\theta) & \frac{\partial^2}{\partial\theta_p\partial\theta_2}s(\theta) & \cdots & \frac{\partial^2}{\partial\theta_p\partial\theta_p}s(\theta) \end{pmatrix}$$

If the second order derivatives of $s(\theta)$ are continuous in θ, then the Hessian matrix is symmetric (Young's Theorem).

Gradients, Jacobians, and Hessians(3)

Let $f(\theta)$ be an n by 1 (column) vector valued function of a p-dimensional argument θ. The **Jacobian** of

$$f(\theta) = \begin{pmatrix} f_1(\theta) \\ f_2(\theta) \\ \vdots \\ f_n(\theta) \end{pmatrix}$$

is the n by p matrix

$$\frac{\partial}{\partial\theta'}f(\theta) = \begin{pmatrix} \frac{\partial}{\partial\theta_1}f_1(\theta) & \frac{\partial}{\partial\theta_2}f_1(\theta) & \cdots & \frac{\partial}{\partial\theta_p}f_1(\theta) \\ \frac{\partial}{\partial\theta_1}f_2(\theta) & \frac{\partial}{\partial\theta_2}f_2(\theta) & \cdots & \frac{\partial}{\partial\theta_p}f_2(\theta) \\ \vdots & \vdots & & \vdots \\ \frac{\partial}{\partial\theta_1}f_n(\theta) & \frac{\partial}{\partial\theta_2}f_n(\theta) & \cdots & \frac{\partial}{\partial\theta_p}f_n(\theta) \end{pmatrix}.$$

Gradients, Jacobians, and Hessians(4)

Let $h'(\theta)$ be a 1 by n (row) vector valued function

$$h'(\theta) = \begin{pmatrix} h_1(\theta), & h_2(\theta), & \ldots, & h_n(\theta) \end{pmatrix}.$$

Then its "gradient" is

$$\frac{\partial}{\partial \theta} h'(\theta) = \begin{pmatrix} \frac{\partial}{\partial \theta_1} h_1(\theta) & \frac{\partial}{\partial \theta_1} h_2(\theta) & \cdots & \frac{\partial}{\partial \theta_1} h_n(\theta) \\ \frac{\partial}{\partial \theta_2} h_1(\theta) & \frac{\partial}{\partial \theta_2} h_2(\theta) & \cdots & \frac{\partial}{\partial \theta_2} h_n(\theta) \\ \vdots & \vdots & & \vdots \\ \frac{\partial}{\partial \theta_p} h_1(\theta) & \frac{\partial}{\partial \theta_p} h_2(\theta) & \cdots & \frac{\partial}{\partial \theta_p} h_n(\theta) \end{pmatrix}.$$

The following rule governs transposition

$$\left(\frac{\partial}{\partial \theta'} f(\theta) \right)' = \frac{\partial}{\partial \theta} f'(\theta).$$

Gradients, Jacobians, and Hessians(5)

The Hessian matrix of $s(\theta)$ can be obtained by successive differentiation variously as

$$\begin{aligned} \frac{\partial^2}{\partial \theta \partial \theta'} s(\theta) &= \frac{\partial}{\partial \theta} \left(\frac{\partial}{\partial \theta'} s(\theta) \right) \\ &= \frac{\partial}{\partial \theta} \left(\frac{\partial}{\partial \theta} s(\theta) \right)' \\ &= \frac{\partial}{\partial \theta'} \left(\frac{\partial}{\partial \theta} s(\theta) \right) \qquad \text{(if symmetric)} \\ &= \frac{\partial}{\partial \theta'} \left(\frac{\partial}{\partial \theta'} s(\theta) \right)' \qquad \text{(if symmetric).} \end{aligned}$$

Gradients, Jacobians, and Hessians(6)

Product Rule: If $f(\theta)$ and $h'(\theta)$ are as above, then

$$\frac{\partial}{\partial \theta'} h'(\theta) f(\theta) = h'(\theta) \frac{\partial}{\partial \theta'} f(\theta) + f'(\theta) \frac{\partial}{\partial \theta'} h(\theta)$$

Gradients, Jacobians, and Hessians(7)

Chain Rule: Let $g(\rho)$ be a p by 1 (column) vector valued function of an r-dimensional argument ρ, and let $f(\theta)$ be as above, then

$$\frac{\partial}{\partial \rho'} f[g(\rho)] = \frac{\partial}{\partial \theta'} f[g(\rho)] \frac{\partial}{\partial \rho'} g(\rho)$$

or, perhaps better,

$$\frac{\partial}{\partial \rho'} f[g(\rho)] = \left[\frac{\partial}{\partial \theta'} f(\theta) \right]_{\theta = g(\rho)} \frac{\partial}{\partial \rho'} g(\rho).$$

The Jacobian of a composition is the product of the Jacobians.

Application

$$F(\theta) = \frac{\partial}{\partial \theta'} f(\theta)$$

$$\text{SSE}(\theta) = [y - f(\theta)]'[y - f(\theta)]$$

$$\begin{aligned} \frac{\partial}{\partial \theta'}\text{SSE}(\theta) &= [y - f(\theta)]'[-F(\theta)] \quad \text{product rule} \\ &\quad + [y - f(\theta)]'[-F(\theta)] \\ &= -2[y - f(\theta)]'F(\theta) \end{aligned}$$

$$\frac{\partial}{\partial \theta}\text{SSE}(\theta) = -2F'(\theta)[y - f(\theta)] \quad \text{transpose}$$

First Order Conditions

If $\hat{\theta}$ minimizes SSE(θ), then

$$\frac{\partial}{\partial \theta}\text{SSE}(\hat{\theta}) = 0$$

so that

$$\frac{\partial}{\partial \theta}\text{SSE}(\hat{\theta}) = -2F'(\theta)[y - f(\hat{\theta}) = 0$$

or

$$\hat{F}'\hat{e} = 0.$$

Residuals are orthogonal to the columns of $\hat{F}$.

Taylor's Theorem

(mean value form of the remainder)

First order:

$$s(\theta) = s(\theta^*) + \frac{\partial}{\partial \theta'} s(\bar{\theta})(\theta - \theta^*)$$

Second order:

$$\begin{aligned} s(\theta) &= s(\theta^*) + \frac{\partial}{\partial \theta'} s(\theta^*)(\theta - \theta^*) \\ &\quad + \frac{1}{2}(\theta - \theta^*)' \left[\frac{\partial^2}{\partial \theta \partial \theta'} s(\bar{\theta}) \right] (\theta - \theta^*) \end{aligned}$$

where

$$\bar{\theta} = \lambda\theta^* + (1 - \lambda)\theta \quad 0 \le \lambda \le 1$$

Topics

- Examples & Least Squares Estimates
- Notation & Taylor's Theorem
- Statistical Properties
- Computations
- Hypothesis Tests
- Confidence Intervals

Setup

$y_t = f(x_t, \theta^o) + e_t \qquad t = 1, \ldots, n$

e_t iid. $P(e)$

$\mathcal{E} e_t = \int_{\mathcal{E}} e \, dP(e) = 0$

$\mathrm{Var}(e_t) = \int_{\mathcal{E}} e^2 \, dP(e) = \sigma^2$

θ^o in Θ, a closed and bounded subset of $\Re^p$

$\hat{\theta}_n = \underset{\theta \epsilon \Theta}{\mathrm{argmin}} s_n(\theta)$

$s_n(\theta) = (1/n) \sum_{t=1}^n [y_t - f(x_t, \theta)]^2$

Almost Sure Convergence (1)

Because

$$s_n(\theta) = \frac{1}{n} \sum_{t=1}^{n} [y_t - f(x_t, \theta)]^2$$

for $n = 1, 2, \ldots$ is a sequence of functions, it can possess a limit for each fixed θ in the ordinary calculus sense:

$$\lim_{n\to\infty} s_n(\theta) = s^*(\theta)$$

Holding $\{x_t\}$ fixed, for some sequences of errors $\{e_t\}$ the limit will exist, for others it will not.

Almost sure convergence in this context is defined as follows:

$$P\left[\{e_t\} : \lim_{n\to\infty} s_n(\theta) \neq s^*(\theta)\right] = 0$$

That is, the probability of getting a sequence of errors for which convergence fails is zero.

If $\{x_t\}$ is a random sequence rather than a fixed sequence, then $P(\cdot)$ is interpreted as be conditional distribution of $\{e_t\}$ given $\{x_t\}$.

Almost Sure Convergence (2)

Almost sure convergence is the standard calculus notion of convergence and is subject to all the standard manipulative rules

For instance,

$\lim_{n\to\infty} a_n(\theta) = a^*(\theta)$

$\lim_{n\to\infty} b_n(\theta) = b^*(\theta)$

$\lim_{n\to\infty} c_n = c^*$

implies

$\lim_{n\to\infty} a_n(\theta) + b_n(\theta) + c_n = a^*(\theta) + b^*(\theta) + c^*$

$\lim_{n\to\infty} a_n(\theta)/c_n = a^*(\theta)/c^*$ if $c^* \neq 0$

etc.

Almost Sure Convergence (3)

Especially important are the rules regarding continuity.

If G[(a,b,c)] is continuous with respect to some norm $\|(a, b, c)\|$ then

$$\lim_{n\to\infty} \|(a_n, b_n, c_n) - (a^*, b^*, c^*)\| = 0$$

implies

$$\lim_{n\to\infty} G[(a_n, b_n, c_n)] = G[(a^*, b^*, c^*)].$$

Important Example

If attention is restricted to continuous functions $s(\theta)$ that are defined on a closed and bounded set Θ, then the argmin function is continuous with respect to uniform convergence. Therefore,

$$\lim_{n\to\infty} \max_{\theta\in\Theta} |s_n(\theta) - s^*(\theta)| = 0$$

implies that

$$\lim_{n\to\infty} \underset{\theta\in\Theta}{\operatorname{argmin}} s_n(\theta) = \underset{\theta\in\Theta}{\operatorname{argmin}} s^*(\theta)$$

An assumption such as $s^*(\theta)$ has a unique minimum is necessary in addition to make sure that the argmin function is well defined when applied to $s^*(\theta)$. It is possible to get by with less, but for our applications, a unique minimum is a reasonable assumption.

Proof

Let

$$\theta^o = \underset{\theta\in\Theta}{\operatorname{argmin}} s^*(\theta)$$

$$\hat{\theta}_n = \underset{\theta\in\Theta}{\operatorname{argmin}} s_n(\theta)$$

If Θ is closed and bounded then every subsequence $\{\hat{\theta}_{n_m}\}$ of $\{\hat{\theta}_n\}$ has a convergent subsubsequence $\{\hat{\theta}_{n_{m_j}}\}$ with limit point

$$\lim_{j\to\infty} \hat{\theta}_{n_{m_j}} = \theta^{\#}$$

Now

$$s_{n_{m_j}}(\hat{\theta}_{n_{m_j}}) \le s_{n_{m_j}}(\theta^o)$$

and uniform convergence taken together imply

$$s^*(\theta^{\#}) \le s^*(\theta^o)$$

Uniqueness of θ^o implies $\theta^{\#} = \theta^o$. Thus, every limit point of $\{\hat{\theta}_n\}$ is θ^o.

Consequence

Applying these ideas to the least squares estimator

$$\hat{\theta}_n = \underset{\theta\in\Theta}{\operatorname{argmin}} s_n(\theta)$$

where

$$s_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} [y_t - f(x_t,\theta)]^2.$$

We now know that to prove consistency of the nonlinear least squares estimator we must (1) show that the residual sum of squares function has a uniform limit, (2) show that the limit function has a unique minimum, and (3) compute this minimum.

Strong Law of Large Numbers for $\{e_t\}$

"Sample averages converge to population averages."

That is,

$$\lim_{n\to\infty} \left| \frac{1}{n}\sum_{t=1}^{n} g(e_t) - \int g(e)\,dP(e) \right| = 0$$

for any $g(e)$ for which $\int |g(e)|\,dP(e) < \infty$.

Stability Condition on $\{x_t\}$

For some fixed sequences the statement

> "Sample averages converge to population averages."

can also be true. Chaotic data, data obtained by replicating a fixed set of points, and a sequence obtained by sampling a distribution exhibit this behavior:

For some μ, called the design measure,

$$\lim_{n\to\infty}\left|\frac{1}{n}\sum_{t=1}^{n} g(x_t) - \int g(x)\,d\mu(x)\right| = 0$$

for any $g(x)$ for which $\int |g(x)|\,d\mu P(x) < \infty$.

This stability condition is referred to as "$\{x_t\}$ is a Cesaro sum generator" in the text.

Uniform SLLN for the Joint Process $\{(x_t, e_t)\}$

If $\{e_t\}$ is iid and $\{x_t\}$ is a Cesaro sum generator, then

$$\lim_{n\to\infty}\max_{\theta\in\Theta}\left|\frac{1}{n}\sum_{t=1}^{n} g(e_t, x_t, \theta) - \int\int g(e, x, \theta)\,dP(e)d\mu(x)\right| = 0$$

for continuous functions $g(e, x, \theta)$ for which

$$\int\int \max_{\theta\in\Theta}|g(e, x, \theta)|\,dP(e)d\mu(x) < \infty.$$

Consistency (1)

We can now establish consistency.

We now know that if $\{e_t\}$ is iid, $\{x_t\}$ is a Cesaro sum generator, and

$$s_n(\theta) = \frac{1}{n}\sum_{t=1}^{n}[y_t - f(x_t, \theta)]^2,$$

then

$$\lim_{n\to\infty}\max_{\theta\in\Theta}\left|s_n(\theta) - \int\int[e + f(x, \theta^o) - f(x, \theta)]^2\,dP(e)d\mu(x)\right| = 0$$

This is the uniform convergence we need. The consequence is that the least square estimator

$$\hat{\theta}_n = \underset{\theta\in\Theta}{\text{argmin}}\, s_n(\theta)$$

will converge to whatever minimizes

$$s^*(\theta) = \int\int[e + f(x, \theta^o) - f(x, \theta)]^2\,dP(e)d\mu(x).$$

Consistency (2)

$$\begin{aligned}
s^*(\theta) &= \int\int[e + f(x, \theta^o) - f(x, \theta)]^2\,dP(e)d\mu(x) \\
&= \int\int e^2\,dP(e)d\mu(x) \\
&\quad + 2\int\int e[f(x, \theta^o) - f(x, \theta)]\,dP(e)d\mu(x) \\
&\quad + \int\int[f(x, \theta^o) - f(x, \theta)]^2\,dP(e)d\mu(x) \\
&= \int e^2\,dP(e) \\
&\quad + 2\int e\,dP(e)\int[f(x, \theta^o) - f(x, \theta)]\,d\mu(x) \\
&\quad + \int[f(x, \theta^o) - f(x, \theta)]^2\,d\mu(x) \\
&= \sigma^2 + \int[f(x, \theta^o) - f(x, \theta)]^2\,d\mu(x)
\end{aligned}$$

Consistency (3)

The least square estimator

$$\hat{\theta}_n = \underset{\theta\in\Theta}{\operatorname{argmin}} s_n(\theta)$$

will converge to whatever minimizes

$$s^*(\theta) = \sigma^2 + \int [f(x,\theta^o) - f(x,\theta)]^2 \, d\mu(x).$$

The true value of the parameter θ^o is certainly a minimum. If it is also a unique minimum then

$$\lim_{n\to\infty} \hat{\theta}_n = \theta^o.$$

The condition that $s^*(\theta)$ have a unique minimum is the identification condition for nonlinear least squares.

Consistency (4)

Consider Example 1

$$y_t = \theta_1 x_{1t} + \theta_2 x_{2t} + \theta_4 e^{\theta_3 x_{3t}} + e_t,$$

with data

t	y	x_1	x_2	x_3
1	0.98610	1	1	6.28
2	1.03848	0	1	9.86
3	0.95482	1	1	9.11
4	1.04184	0	1	8.43
5	1.02324	1	1	8.11
6	0.90475	0	1	1.82
⋮				
29	0.50811	1	1	0.08
30	0.91840	0	1	6.11

On pages 19–24 of the text, the design measure $\mu(x)$ is derived, $s^*(\theta)$ is computed, and the conclusion is that

$$s^*(\theta) = 0,\ \theta_3^o \neq 0,\ \theta_4^o \neq 0 \ \Rightarrow\ \theta = \theta^o.$$

As you will see, this is a lot of trouble to work out. Few would bother to do so. Most just rely on a common sense inspection of the model and on the optimization algorithm used to compute $\hat{\theta}_n$ to detect problems.

For instance, it is easy to see that if $\theta_4^o = 0$, then it will be impossible to determine what θ_3^o is. Similarly, if $\theta_3^o = 0$, then it is easy to see that one can estimate the sum $\theta_2^o + \theta_4^o$ but not θ_2^o and θ_4^o individually.

Asymptotic Normality (1)

First Order Conditions

$$\frac{\partial}{\partial\theta} s_n(\theta) = 0$$

Taylor's Expansion of FOC

$$\left[\frac{\partial^2}{\partial\theta\partial\theta'} s_n(\bar{\theta}_n)\right] \sqrt{n}(\hat{\theta}_n - \theta^o) = -\sqrt{n}\frac{\partial}{\partial\theta} s_n(\theta^o)$$

where $\bar{\theta}_n$ is on the line segment joining θ^o to $\hat{\theta}n$. Because $\bar{\theta}_n$ must therefore be closer to θ^o than $\hat{\theta}n$ is and $\lim_{n\to\infty}\hat{\theta}n = \theta^o$, we have $\lim_{n\to\infty}\bar{\theta}_n = \theta^o$ as well.

The second order expansion is not strictly correct: Each row of $\frac{\partial^2}{\partial\theta\partial\theta'}s(\bar{\theta})$ should have its own $\bar{\theta}_i$, $i = 1,\ldots,p$. This leads to cluttered notation, so it will just be understood in the transparancies. The text, *Nonlinear Statistical Models*, handles this detail correctly.

Asymptotics of RHS

$$-\sqrt{n}\frac{\partial}{\partial\theta} s_n(\theta^o) = \frac{2}{\sqrt{n}} \sum_{t=1}^n \frac{\partial}{\partial\theta} f(x_t,\theta^o)\, e_t$$

Mean: $\mathcal{E}\left[-\sqrt{n}\frac{\partial}{\partial\theta}s_n(\theta^o)\right] = 0$

Variance:

$$\begin{aligned}
\mathcal{I}_n &= \text{Var}\left[-\sqrt{n}\frac{\partial}{\partial\theta}s_n(\theta^o)\right] \\
&= \frac{4\sigma^2}{n}\sum_{t=1}^n \left[\frac{\partial}{\partial\theta}f(x_t,\theta^o)\right]\left[\frac{\partial}{\partial\theta}f(x_t,\theta^o)\right]' \\
&= \frac{4\sigma^2}{n}F'F
\end{aligned}$$

Limiting Variance:

$$\begin{aligned}
\mathcal{I} &= \lim_{n\to\infty}\mathcal{I}_n \\
&= 4\sigma^2\int\left[\frac{\partial}{\partial\theta}f(x_t,\theta^o)\right]\left[\frac{\partial}{\partial\theta}f(x_t,\theta^o)\right]' d\mu(x) \\
&= 4\sigma^2 Q
\end{aligned}$$

Central Limit Theorem:

$$-\sqrt{n}\frac{\partial}{\partial\theta}s_n(\theta^o) \xrightarrow{\mathcal{L}} N_p(0,\mathcal{I})$$

Asymptotics of LHS

$$
\begin{aligned}
\mathcal{J}_n &= \left[\frac{\partial^2}{\partial\theta\partial\theta'} s_n(\bar{\theta}_n)\right] \\
&= \frac{2}{n}\sum_{t=1}^{n}\left[\frac{\partial}{\partial\theta} f(x_t,\bar{\theta}_n)\right]\left[\frac{\partial}{\partial\theta} f(x_t,\bar{\theta}_n)\right]' \\
&\quad + \frac{2}{n}\sum_{t=1}^{n} e_t \left[\frac{\partial^2}{\partial\theta\partial\theta'} f(x_t,\bar{\theta}_n)\right]
\end{aligned}
$$

A consequence of the uniform strong law of large numbers is that a joint limit can be computed as an iterated limit; i.e.

$$\lim_{n\to\infty}\max_{\theta\in\Theta}|g_n(\theta)-g(\theta)| = 0 \ \ \& \ \lim_{n\to\infty}\bar{\theta}_n = \theta^o \quad \Rightarrow \quad \lim_{n\to\infty} g_n(\bar{\theta}_n) = g(\theta^o)$$

Therefore:

$$
\begin{aligned}
\mathcal{J} &= \lim_{n\to\infty}\mathcal{J}_n \\
&= 2\int\left[\frac{\partial}{\partial\theta} f(x_t,\theta^o)\right]\left[\frac{\partial}{\partial\theta} f(x_t,\theta^o)\right]' d\mu(x) \\
&\quad + 2\int e\,dP(e)\int\frac{\partial^2}{\partial\theta\partial\theta'} f(x_t,\theta^o)\,d\mu(x) \\
&= 2Q
\end{aligned}
$$

LHS & RHS Combined

Slutsky's Theorem:

$$\mathcal{J}_n\sqrt{n}(\hat{\theta}_n - \theta^o) = -\sqrt{n}\frac{\partial}{\partial\theta}s_n(\theta^o)$$

$$-\sqrt{n}s_n(\theta^o) \xrightarrow{\mathcal{L}} N_p(0,\mathcal{I})$$

$$\mathcal{J} = \lim_{n\to\infty}\mathcal{J}_n$$

imply

$$\sqrt{n}(\hat{\theta}_n - \theta^o) \xrightarrow{\mathcal{L}} N_p(0,\mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1}).$$

Because $\mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-1} = (2Q)^{-1}(4\sigma^2 Q)(2Q)^{-1} = \sigma^2 Q^{-1}$, we have

$$\sqrt{n}(\hat{\theta}_n - \theta^o) \xrightarrow{\mathcal{L}} N_p(0,\sigma^2 Q^{-1})$$

Further, $\sigma^2 Q^{-1}$ can be estimated consistently by $\hat{V} = \text{SSE}(\hat{\theta}_n)(\hat{F}'\hat{F})^{-1}$. Why?

Topics

- Examples & Least Squares Estimates
- Notation & Taylor's Theorem
- Statistical Properties
- Computations
- Hypothesis Tests
- Confidence Intervals

Computations

The best reference for nonlinear optimization is

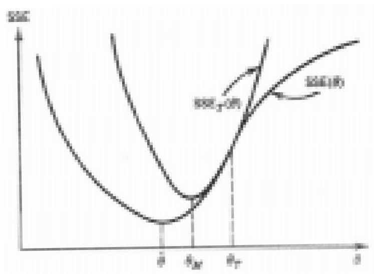
Fletcher, R. (1987) *Practical Methods of Optimization, Second Edition,* Wiley, New York

an honorable mention is

Gill, Philip E., Walter Murray, and Margaret H. Wright (1981) *Practical Optimization,* Academic Press, New York

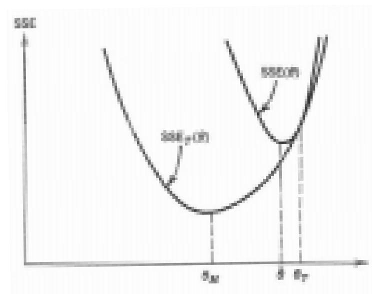
The best routine available is NPSOL by Murray, Gill, and Wright which is available from the Office of Technology Licensing, Stanford University, and is in the NaG Library.

Computations (an adequate approximation)



The idea is to obtain a quadratic approximation $\text{SSE}_{T_0}(\theta)$ that is tangent to the residual sum of squares surface $\text{SSE}(\theta)$ at a trial value of the parameter θ_{T_0}, as shown for the case $p = 1$ above. The minimum θ_{M_0} of the approximating quadratic is an approximation to $\hat{\theta}$. The process is iterated, putting $\theta_{T_{i+1}} = \theta_{M_i}$, until the sequence θ_{T_i} appears to have converged. The limit is accepted as $\hat{\theta}$.

Computations (an inadequate approximation)



Sometimes the minimum θ_{M_i} of the approximating quadratic overshoots $\hat{\theta}$, as shown above. But also as shown, all points on the line joining θ_{M_i} and θ_{T_i}

$$\theta = \theta_{T_i} + \lambda(\theta_{M_i} - \theta_{T_i}) \quad 0 < \lambda \le \lambda^*$$

for λ^* small enough will lead to an improvement. The idea is to try to find a λ_i with

$$\text{SSE}[\theta_{T_i} + \lambda_i(\theta_{M_i} - \theta_{T_i})] < \text{SSE}(\theta_{T_i})$$

and put $\theta_{T_{i+1}} = \theta_{T_i} + \lambda_i(\theta_{M_i} - \theta_{T_i})$.

Quadratic Approximations (Gauss-Newton)

$$\text{SSE}(\theta) = \|y - f(\theta)\|^2$$

$$\text{SSE}_T(\theta) = \|y - f(\theta_T) + F(\theta_T)(\theta - \theta_T)\|^2$$

$$(\theta_M - \theta_T) = [F'(\theta_T)F(\theta_T)]^{-1}F'(\theta_T)[y - f(\theta_T)]$$

or

$$\theta_M = \theta_T + D_T$$

where

$$D_T = [F'(\theta_T)F(\theta_T)]^{-1}F'(\theta_T)[y - f(\theta_T)],$$

which is called the Gauss-Newton downhill direction.

Quadratic Approximations (Newton)

$$\begin{aligned}\text{SSE}_T(\theta) &= \text{SSE}(\theta_T) + \left[\frac{\partial}{\partial\theta'}\text{SSE}(\theta_T)\right](\theta - \theta_T) \\ &+ \frac{1}{2}(\theta - \theta_T)'\left[\frac{\partial^2}{\partial\theta\partial\theta'}\text{SSE}(\theta_T)\right](\theta - \theta_T)\end{aligned}$$

The minimum is

$$\theta_M = \theta_T + D_T$$

where

$$\begin{aligned}D_T &= -\left[\frac{\partial^2}{\partial\theta\partial\theta'}\text{SSE}(\theta_T)\right]^{-1}\frac{\partial}{\partial\theta'}\text{SSE}(\theta_T) \\ &= \left[F'(\theta_T)F(\theta_T) - \sum_{t=1}^{n}\tilde{e}_t\frac{\partial^2}{\partial\theta\partial\theta'}f(x_t,\theta_T)\right]^{-1}F'(\theta_T)\tilde{e}\end{aligned}$$

where

$$\tilde{e} = y - f(\theta_T)$$

which is actually the Gauss-Newton downhill direction with a correction term added to the matrix that gets inverted.

Quadratic Approximations (Steepest Descent)

$$D_T = F'(\theta_T)[y - f(\theta_T)],$$

Quadratic Approximations (Marquardt)

$$D_T = [F'(\theta_T)F(\theta_T) + \delta S]^{-1}F'(\theta_T)[y - f(\theta_T)],$$

where S is $F'(\theta_T)F(\theta_T)$ with all off diagonal elements put to zero.

Which is Better?

Gauss-Newton

$$D_T = [F'(\theta_T)F(\theta_T)]^{-1}F'(\theta_T)[y - f(\theta_T)],$$

or Newton

$$D_T = \left[F'(\theta_T)F(\theta_T) - \sum_{t=1}^{n}\tilde{e}_t \frac{\partial^2}{\partial\theta\partial\theta'} f(x_t, \theta_T)\right]^{-1} F'(\theta_T)\tilde{e}$$

or steepest descent, or Marquardt, or something else?

In my opinion, one should just use Gauss-Newton because the matrix $F'(\theta_T)F(\theta_T)$ is always positive semi-definite and one only has to compute first derivatives.

Numerical analysts argue that something that is fast far from the solution like Gauss-Newton or steepest descent should be used initially and then one should switch over to the Newton method to speed convergence towards the end.

Line Search

There are numerous suggestions in the literature. The two most commonly used are chopping and quadratic interpolation.

Chopping: Accept the first λ in a decreasing sequence such as $1, \frac{1}{2}, \frac{1}{4}, \ldots$ for which

$$\text{SSE}[\theta_T + \lambda(\theta_M - \theta_T)] < \text{SSE}(\theta_T)$$

Quadratic Interpolation: Fit a quadratic in λ to the three points

x-axis	y-axis
$\lambda = 0$	$\text{SSE}(\theta_T)$
$\lambda = \frac{1}{2}$	$\text{SSE}\left[\theta_T + \frac{1}{2}(\theta_M - \theta_T)\right]$
$\lambda = 1$	$\text{SSE}(\theta_M)$

Put λ to the minimum of the quadratic.

Line Search (pitfalls)

Chopping: There may be no λ in the sequence $1, \frac{1}{2}, \frac{1}{4}, \ldots$ that leads to improvement because $\theta_T + \lambda(\theta_M - \theta_T)$ gets to within machine precision of θ_T. If this happens, either announce convergence or announce failure, your choice. Often convergence of the minimization algorithm can be accelerated by starting the chopping sequence off with some $\lambda > 1$.

Quadratic Interpolation: The minimizer λ_M of the quadratic in λ does not necessarily satisfy

$$\text{SSE}[\theta_T + \lambda_M(\theta_M - \theta_T)] < \text{SSE}(\theta_T)$$

Always check this condition before taking the step. If the condition is not satisfied, start chopping as above, but starting from λ_M.

The Modified Gauss-Newton Algorithm

0. Choose a starting value θ_0. Compute

$$D_0 = [F'(\theta_0)F(\theta_0)]^{-1}F'(\theta_0)[y - f(\theta_0)],$$

Find λ_0 between 0 and 1 such that

$$\text{SSE}[\theta_0 + \lambda_0 D_0] < \text{SSE}(\theta_0)$$

1. Put $\theta_1 = \theta_0 + \lambda_0 D_0$. Compute

$$D_1 = [F'(\theta_1)F(\theta_1)]^{-1}F'(\theta_1)[y - f(\theta_1)],$$

Find λ_1 between 0 and 1 such that

$$\text{SSE}[\theta_1 + \lambda_1 D_1] < \text{SSE}(\theta_1)$$

2. Put $\theta_2 = \theta_1 + \lambda_1 D_1$.

-
-
-

Some Comments

The modified Gauss-Newton method is due to H. O. Hartley (1961), "The modified Gauss-Newton method for the fitting of nonlinear regression functions by least squares," *Technometrics* 3, 269–280.

Modified means line searched.

Algorithms like this that use an approximation to the Hessian are called quasi Newton by numerical analysts. The most popular general quasi Newton algorithm (not just for least squares problems) uses rank one numerically updated Hessians and is called Broyden-Fletcher-Goldfarb-Shanno (BFGS).

The Newton algorithm with line search is the same as above but with the Newton downhill direction D_i substituted.

Marquardt requires that δ decrease to zero as iterations continue.

Stopping Rules

Stop when

$$\|\theta_i - \theta_{i+1}\| < \epsilon(\|\theta_i\| + \tau) \quad \text{for } i = 1, \ldots, p$$

and, simultaneously,

$$\|\text{SSE}(\theta_i) - \text{SSE}(\theta_{i+1})\| < \epsilon(\|\text{SSE}(\theta_i)\| + \tau)$$

where $\tau > 0$ and $\epsilon > 0$ are preset tolerances. A standard choice is $\epsilon = 10^{-3}$ and $\epsilon = 10^{-5}$.

Some authors would also check whether

$$\|D_i\| < \epsilon$$

simultaneously with the above before stopping.

Starting Values

Homily:

> A plot of $f(x_t, \theta_0)$ against t must resemble a plot of y_t against t.

One Method:

> A perfect fit to representative values.

Consider Example 1

$$y_t = \theta_1 x_{1t} + \theta_2 x_{2t} + \theta_4 e^{\theta_3 x_{3t}} + e_t,$$

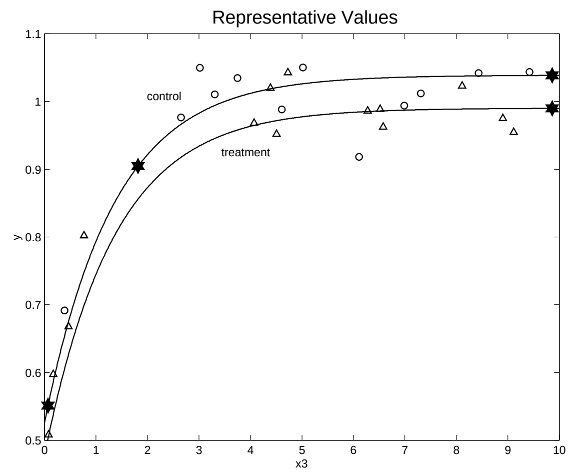
with data

t	y	x_1	x_2	x_3
1	0.98610	1	1	6.28
2	1.03848	0	1	9.86
3	0.95482	1	1	9.11
4	1.04184	0	1	8.43
5	1.02324	1	1	8.11
6	0.90475	0	1	1.82
7	0.96263	1	1	6.58
8	1.05026	0	1	5.02
9	0.98861	1	1	6.52
10	1.03437	0	1	3.75
11	0.98982	1	1	9.86
12	1.01214	0	1	7.31
13	0.66768	1	1	0.47
14	0.55107	0	1	0.07
⋮				

Solve

$$\begin{aligned} t = 14: \quad 0.55107 &= \theta_2 + \theta_4 e^{\theta_3 0.07} \\ t = 6: \quad 0.90475 &= \theta_2 + \theta_4 e^{\theta_3 1.82} \\ t = 2: \quad 1.03848 &= \theta_2 + \theta_4 e^{\theta_3 9.86} \\ t = 11: \quad 0.98982 &= \theta_1 + \theta_2 + \theta_4 e^{\theta_3 9.86} \end{aligned}$$

to get starting values.



Plotted is

$$y = \theta_1 x_1 + \theta_2 x_2 + \theta_4 e^{\theta_3 x_3}$$

against x_3 with

$$\theta = (-0.048660, 1.0038835, -0.737919, -0.513623)$$

which is a perfect fit to the data marked with a star.

Fit to Representative Values (SAS code)

```
data amstat;
  infile "amstat.dat";
  input t y x1 x2 x3;

data work;
  set amstat;
  if t=2 or t=6 or t=11 or t=14 then output;
  delete;

proc nlin data=work method=gauss iter=50 convergence=1.0e-5;
  parms t1=0 t2=0 t3=-1 t4=-1;
  model y=t1*x1+t2*x2+t4*exp(t3*x3);
  der.t1=x1; der.t2=x2; der.t3=t4*x3*exp(t3*x3); der.t4=exp(t3*x3);
```


Fit to Representative Values (SAS output)

```
            Non-Linear Least Squares Iterative Phase
          Dependent Variable Y   Method: Gauss-Newton
       Iter        T1              T2        Sum of Squares
                   T3              T4

         0          0               0       5.397072
                   -1.000000       -1.000000

         1         -0.048660        1.038596       0.000447
                   -0.826742       -0.510747

         2         -0.048660        1.038769   0.0000039585
                   -0.729756       -0.513288

         3         -0.048660        1.038834  1.8362822E-10
                   -0.737864       -0.513620

         4         -0.048660        1.038835  3.3698672E-19
                   -0.737919       -0.513623

         5         -0.048660        1.038835  8.6281662E-32
                   -0.737919       -0.513623

NOTE: Convergence criterion met.
```

Fit to Data (SAS code)

```
data amstat;
  infile "amstat.dat";
  input t y x1 x2 x3;

proc model data=amstat;
  y=t1*x1+t2*x2+t4*exp(t3*x3);
  parms t1=-0.048660 t2=1.038835 t3=-0.737919 t4=-0.513623;
  fit y / ols converge=1.0e-8 maxiter=50 method=gauss covb;
```


Fit to Data (SAS output, page 1)

```
                         MODEL Procedure
                          OLS Estimation

            Nonlinear OLS Summary of Residual Errors

                DF    DF
Equation Model Error       SSE       MSE R-Square Adj R-Sq

Y            4    26    0.0305  0.001173   0.9576   0.9527


                Nonlinear OLS Parameter Estimates

                               Approx.       'T'   Approx.
 Parameter    Estimate      Std Err     Ratio  Prob>|T|

 T1          -0.025890      0.01262     -2.05    0.0505
 T2           1.015680    0.0099379    102.20    0.0001
 T4          -0.504903      0.02566    -19.68    0.0001
 T3          -1.115697      0.16354     -6.82    0.0001


   Number of Observations       Statistics for System
   Used                 30       Objective    0.001017
   Missing               0       Objective*N    0.0305
```


Fit to Data (SAS output, page 2)

```
                   Covariance of Estimates

CovB          T1          T2          T4          T3

T1      0.000159   -0.000079   -0.000044   -0.000177
T2     -0.000079   0.0000988   -1.851E-6    0.000607
T4     -0.000044   -1.851E-6    0.000658    0.002356
T3     -0.000177    0.000607    0.002356      0.0267
```

This matrix is

$$s^2 \left(\hat{F}'\hat{F}\right)^{-1}$$

where

$$s^2 = \frac{\text{SSE}(\hat{\theta}_n)}{n-p}$$

.

Difficult Cases

In difficult cases, with numerous local minima, such as neural nets, flexible form demand systems, sums of exponentials, etc. a reasonable strategy is the following:

Get one reasonable start value θ. Let generate a random point δ with distance from zero $\|\delta\| = 10^{-8}$ and iterate the Gauss-Newton method 15 times starting from $\theta + \delta$. Do this 100 times, saving the final value θ_{15} and corresponding SSE(θ_{15}). Do this again for $\|\delta\| = 10^{-7}$, for $\|\delta\| = 10^{-6}$, ..., for $\|\delta\| = 10^{-1}$. Of the 800 values thus produced, iterate the best 50 to convergence. Select the best of these as the answer $\hat{\theta}$.

Topics

- Examples & Least Squares Estimates
- Notation & Taylor's Theorem
- Statistical Properties
- Computations
- Hypothesis Tests
- Confidence Intervals

Tests of Hypotheses

$$y_t = f(x_t, \theta^o) + e_t \quad t = 1, \ldots, n$$

$$h : \Theta \to \Re^q$$

$$\text{H} : h(\theta^o) = 0 \text{ against A} : h(\theta^o) \neq 0$$

Notation:

$$H(\theta) = \frac{\partial}{\partial \theta'} h(\theta) = \begin{pmatrix} \frac{\partial}{\partial \theta_1} h_1(\theta) & \cdots & \frac{\partial}{\partial \theta_p} h_1(\theta) \\ \vdots & & \vdots \\ \frac{\partial}{\partial \theta_1} h_q(\theta) & \cdots & \frac{\partial}{\partial \theta_p} h_q(\theta) \end{pmatrix}$$

Example:

$$y_t = \theta_1 x_{1t} + \theta_2 x_{2t} + \theta_4 e^{\theta_3 x_{3t}} + e_t$$

$$\text{H} : \theta_3 \theta_4 e^{\theta_3} - \frac{1}{5} = 0$$

$$H(\theta) = \left(0, 0, \theta_4 (1 + \theta_3) e^{\theta_3}, e^{\theta_3}\right)$$

$$p = 4, \; q = 1$$

Wald Test(1)

Recall:

$$\sqrt{n}(\hat{\theta}_n - \theta^o) \xrightarrow{\mathcal{L}} N_p\left(0, \sigma^2 Q^{-1}\right)$$

$$\frac{1}{n} \hat{F}\hat{F} \to Q$$

$$\frac{1}{n} \text{SSE}(\hat{\theta}) \to \sigma^2$$

Taylor's Theorem:

$$\sqrt{n}\left[h(\hat{\theta}_n) - h(\theta^o)\right] = H(\bar{\theta}_n)\sqrt{n}(\hat{\theta}_n - \theta^o)$$

Slutsky's Theorem implies

$$\sqrt{n}\left[h(\hat{\theta}_n) - h(\theta^o)\right] \xrightarrow{\mathcal{L}} N_q\left(0, \sigma^2 H Q^{-1} H'\right)$$

Therefore: If H : $h(\theta^o) = 0$ is true, then

$$W = \frac{n h'(\hat{\theta}_n)\left[H\left(\frac{1}{n}\hat{F}\hat{F}\right)^{-1} H'\right]^{-1} h(\hat{\theta}_n)}{\frac{1}{n} SSE(\hat{\theta})} \xrightarrow{\mathcal{L}} \mathcal{X}_q$$

Wald Test(2)

The statistic

$$W = \frac{n h'(\hat{\theta}_n)\left[H(\hat{F}\hat{F})^{-1} H'\right]^{-1} h(\hat{\theta}_n)}{SSE(\hat{\theta}_n)}$$

is called the Wald test statistic, after Abraham Wald. It is to be compared to the quantiles of the chi squared distribution on q degrees of freedom. One rejects for large W.

Often one computes

$$W = \frac{h'(\hat{\theta}_n)\left[H(\hat{F}\hat{F})^{-1} H'\right]^{-1} h(\hat{\theta}_n)}{q s^2}$$

instead and compares to the quantiles of the F-distribution with q numerator degrees of freedom and $n-p$ denominator degrees of freedom because this agrees with the formulas used in linear models and gives more accurate answers in small samples.

Consider Example 1:

$$y_t = \theta_1 x_{1t} + \theta_2 x_{2t} + \theta_4 e^{\theta_3 x_{3t}} + e_t$$

$$\mathsf{H} : \theta_3 \theta_4 e^{\theta_3} - \frac{1}{5} = 0$$

$$H(\theta) = \left(0, 0, \theta_4(1+\theta_3)e^{\theta_3}, e^{\theta_3}\right)$$

Computations:

$$\hat{h} = h(\hat{\theta}_n) = (-1.1157)(-0.50490)e^{-1.1157} - \frac{1}{5} = -0.0154$$

$$\hat{H} = H(\hat{\theta}_n) = (0,\ 0,\ 0.019142,\ -0.365599)$$

$$\hat{H}(\hat{F}\hat{F})^{-1}\hat{H}' = 0.055256$$

$$s^2 = 0.00117291$$

$$W = \frac{(-0.0154)(0.055256)^{-1}(-0.0154)}{(1)(0.00117291)} = 3.66$$

$$F(0.95, 1, 26) = 4.22$$

Wald Test (SAS code)

```
data amstat;
  infile "amstat.dat";
  input t y x1 x2 x3;

proc model data=amstat;
  y=t1*x1+t2*x2+t4*exp(t3*x3);
  parms t1=-0.048660 t2=1.038835 t3=-0.737919 t4=-0.513623;
  fit y / ols converge=1.0e-8 maxiter=50 method=gauss covb;
  test t3*t4*exp(t3)-0.20 = 0 ,/ wald;
```

Wald Test (SAS output)

```
                     Test Results

  Test       Type       Statistic      Prob.     Label

  Test0      Wald            3.66     0.0556
  T3*T4*EXP(T3)-0.20 =
```


Constrained and Unconstrained Estimates

$$y_t = f(x_t, \theta^o) + e_t \quad t = 1, \ldots, n$$

$$\mathsf{H} : h(\theta^o) = 0 \text{ against } \mathsf{A} : h(\theta^o) \neq 0$$

Unconstrained Estimate:

$$\hat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmin}}\ \mathsf{SSE}(\theta)$$

Constrained Estimate:

$$\tilde{\theta}_n = \underset{h(\theta)=0}{\operatorname{argmin}}\ \mathsf{SSE}(\theta)$$

Likelihood Ratio Test(1)

The statistic

$$L = \frac{n\left[\mathsf{SSE}(\tilde{\theta}_n) - \mathsf{SSE}(\hat{\theta}_n)\right]}{\mathsf{SSE}(\hat{\theta}_n)}$$

is, after some algebra, the likelihood ratio test statistic for H : $h(\theta^o) = 0$ against $A : h(\theta^o) \neq 0$ under the assumption that the errors $\{e_t\}$ are normally distributed. It is to be compared to the quantiles of the chi squared distribution on q degrees of freedom. One rejects for large L.

Often one computes

$$L = \frac{\left[\mathsf{SSE}(\tilde{\theta}_n) - \mathsf{SSE}(\hat{\theta}_n)\right]/q}{\mathsf{SSE}(\hat{\theta}_n)/(n-p)}$$

instead and compares to the quantiles of the F-distribution with q numerator degrees of freedom and $n-p$ denominator degrees of freedom because this agrees with the formulas used in linear models and gives more accurate answers in small samples.

Likelihood Ratio Test(2)

The derivation of the asymptotic distribution of the likelihood ratio test is not difficult but it is time consuming and therefore will be omitted.

What takes time is in getting the asymptotic distribution of the constrained estimator $\tilde{\theta}_n$. The rest of the derivation is a straightforward application of Taylor's Theorem.

Each of the following references contains the derivation. The second is recommended and can be downloaded from the course web page.

Gallant, A. Ronald (1987) *Nonlinear Statistical Models,* Wiley, New York.

Gallant, A. Ronald (1992) *Nonlinear Regression Asymptotics,* Manuscript, Department of Economics, University of North Carolina.

Gallant, A. Ronald (1997) *Introduction to Econometric Theory,* Princeton University Press, Princeton NJ.

Computing $\tilde{\theta}_n$ (1)

$\tilde{\theta}$ minimizes SSE(θ)
subject to $h(\theta) = 0$

Direct Approach:

Use software such as NPSOL from the Office of Technical Licensing, Stanford University.

Indirect Approach:

Rewrite the hypothesis as a functional dependence

$$\text{H} : h(\theta^o) = 0 \text{ against A} : h(\theta^o) \neq 0$$

$\Leftrightarrow$

H : $\theta^o = g(\rho)$ for some ρ against A : $\theta^o \neq g(\rho)$ for any ρ

$h(\theta) \in \Re^q$, $\Theta \in \Re^p$, $\rho \in \Re^r$, $p = r + q$

Computing $\tilde{\theta}_n$ (2)

Once the hypothesis is written as a functional dependence, fit the model

$$y_t = f[x_t, g(\rho)] + e_t$$

to get the unconstrained estimate $\hat{\rho}_n$ and then put

$$\tilde{\theta}_n = g(\hat{\rho}_n)$$

Example:

$$y_t = \theta_1 x_{1t} + \theta_2 x_{2t} + \theta_4 e^{\theta_3 x_{3t}} + e_t$$

$$\text{H} : \theta_3 \theta_4 e^{\theta_3} - \frac{1}{5} = 0$$

$\Leftrightarrow$

$$\text{H} : \theta_4 = 5(\theta_3 e^{\theta_3})^{-1}$$

That is, θ_1, θ_2, and θ_3 are free parameters and the value of θ_4 is implied by the parametric restriction $h(\theta) = 0$. This can be expressed as the functional dependence

$$(\theta_1, \theta_2, \theta_3, \theta_4) = g(\rho_1, \rho_2, \rho_3)$$

where

$$\begin{aligned} \theta_1 &= \rho_1 \\ \theta_2 &= \rho_2 \\ \theta_3 &= \rho_3 \\ \theta_4 &= 5(\rho_3 e^{\rho_3})^{-1} \end{aligned}$$

Constrained Estimation (SAS code)

```
data amstat;
  infile "amstat.dat";
  input t y x1 x2 x3;

proc model data=amstat;
  t1=r1; t2=r2; t3=r3; t4=1.0/(5.0*r3*exp(r3));
  y=t1*x1+t2*x2+t4*exp(t3*x3);
  parms r1=-0.048660 r2=1.038835 r3=-0.737919;
  fit y / ols converge=1.0e-5 maxiter=150 method=gauss;
```

Constrained Estimation (SAS output)

```
                         MODEL Procedure
                          OLS Estimation

           Nonlinear OLS Summary of Residual Errors

               DF    DF
   Equation Model Error       SSE       MSE R-Square Adj R-Sq

   Y            3    27    0.0349  0.001294   0.9514   0.9478


              Nonlinear OLS Parameter Estimates

                                Approx.       'T'   Approx.
    Parameter    Estimate       Std Err     Ratio  Prob>|T|

    R1          -0.023019       0.01315     -1.75    0.0915
    R2           1.019656       0.01010    100.98    0.0001
    R3          -1.160403       0.16300     -7.12    0.0001
```


Consider Example 1:

$$y_t = \theta_1 x_{1t} + \theta_2 x_{2t} + \theta_4 e^{\theta_3 x_{3t}} + e_t$$

$$\text{H} : \theta_3 \theta_4 e^{\theta_3} - \frac{1}{5} = 0$$

$$H(\theta) = \left(0, 0, \theta_4 (1 + \theta_3) e^{\theta_3}, e^{\theta_3}\right)$$

Computations:

$$\text{SSE}(\tilde{\theta}) = 0.3493$$

$$\text{SSE}(\hat{\theta}) = 0.3049$$

$$L = \frac{(0.3493 - 0.3049)/1}{(0.3049)/26} = 3.78$$

$$F(0.95, 1, 26) = 4.22$$

Likelihood Ratio Test (SAS code)

```
data amstat;
  infile "amstat.dat";
  input t y x1 x2 x3;

proc model data=amstat;
  y=t1*x1+t2*x2+t4*exp(t3*x3);
  parms t1=-0.048660 t2=1.038835 t3=-0.737919 t4=-0.513623;
  fit y / ols converge=1.0e-8 maxiter=50 method=gauss;
  test t3*t4*exp(t3)-0.20 = 0 ,/ lr;
```

Likelihood Ratio Test (SAS output)

```
                          Test Results

   Test       Type       Statistic      Prob.     Label

   Test0      L.R.            3.78     0.0518
   T3*T4*EXP(T3)-0.20 =
```


Lagrange Multiplier Test (1)
aka Efficient Score Test
aka nR^2 Test

Intuition:

$$R = \frac{n\tilde{D}'(\tilde{F}'\tilde{F})\tilde{D}}{\text{SSE}(\tilde{\theta})}$$

where

$$\tilde{D} = (\tilde{F}'\tilde{F})^{-1}\tilde{F}'\tilde{e} \quad \tilde{e} = y - f(\tilde{\theta})$$

is the Gauss-Newton step from $\tilde{\theta}$ to $\hat{\theta}$. That is, one expects that

$$\hat{\theta} \doteq \tilde{\theta} + \tilde{D}.$$

Thus, if the constraint $h(\theta) = 0$ is markedly false, then one expects that $\tilde{D} \doteq \hat{\theta} - \tilde{\theta}$ will be large and that R will therefore be large. Conversely, if $h(\theta) = 0$ is approximately true, then $\tilde{D}$ and therefore R should be small.

Lagrange Multiplier Test (2)

$$R = \frac{n\tilde{D}'(\tilde{F}'\tilde{F})\tilde{D}}{\text{SSE}(\tilde{\theta})} = \frac{(n/4)\tilde{\lambda}'\tilde{H}(\tilde{F}'\tilde{F})^{-1}\tilde{H}'\tilde{\lambda}}{\text{SSE}(\tilde{\theta})}$$

The Lagrangian for the constrained optimization problem is

$$\mathcal{L}(\theta,\lambda) = \text{SSE}(\theta) + \lambda' h(\theta)$$

with first order condition

$$0 = -2\tilde{e}'\tilde{F} + \tilde{\lambda}'\tilde{H}$$

so that

$$\tilde{D} = (\tilde{F}'\tilde{F})^{-1}\tilde{F}'\tilde{e} = (1/2)\tilde{H}'\tilde{\lambda}$$

The shadow price of the constraint $h(\theta) = 0$ in SSE units is λ. When the constraint is severely binding, one expects that λ and hence R will be large.

Lagrange Multiplier Test (3)

The statistic

$$R = \frac{n\tilde{D}'(\tilde{F}'\tilde{F})\tilde{D}}{\text{SSE}(\tilde{\theta})}$$

is to be compared to the quantiles of the chi squared distribution on q degrees of freedom. One rejects for large R. To make degrees of freedom corrections, compare to

$$d = \frac{nF}{(n-p)/q + F}$$

where F is the quantile of the F-distribution with q numerator degrees of freedom and $n-p$ denominator degrees of freedom.

A difficulty with the Lagrange multiplier test is the division by $\text{SSE}(\tilde{\theta})$ instead of $\text{SSE}(\hat{\theta})$; if the hypothesis is false then the former is larger than the latter, which reduces power.

Lagrange Multiplier Test (4)

$$R = \frac{n\tilde{D}'(\tilde{F}'\tilde{F})\tilde{D}}{\text{SSE}(\tilde{\theta})}$$

where

$$\tilde{D} = (\tilde{F}'\tilde{F})^{-1}\tilde{F}'\tilde{e}$$

Computation:

Regress $\tilde{e} = y - f(\tilde{\theta})$ on $\tilde{F} = \frac{\partial}{\partial\theta}f(\tilde{\theta})$ with no intercept term in the regression. Then

$$\text{SSE}(\tilde{\theta}) = \text{uncorrected sum of squares}$$

$$\tilde{D}'(\tilde{F}'\tilde{F})\tilde{D} = \text{regression sum of squares}$$

$$R = n \times \text{uncorrected } R^2 \text{ statistic}$$

Lagrange Multiplier Computations (SAS code)

```
data amstat;
  infile "amstat.dat";
  input t y x1 x2 x3;

data work;
  set amstat;
  r1 = -0.023019; r2 = 1.019656; r3 = -1.160403;
  t1=r1; t2=r2; t3=r3; t4=1.0/(5.0*r3*exp(r3));
  e = y - t1*x1 - t2*x2 - t4*exp(t3*x3);
  f1=x1; f2=x2; f3=t4*x3*exp(t3*x3); f4=exp(t3*x3);

proc reg data=work;
  model e = f1 f2 f3 f4 / noint;
```

Lagrange Multiplier Computations (SAS output)

```
                         Analysis of Variance

                            Sum of         Mean
Source          DF         Squares       Square      F Value   Prob>F

Model            4         0.00444      0.00111        0.946   0.4531
Error           26         0.03049      0.00117
U Total         30         0.03493

    Root MSE       0.03425     R-square       0.1271
    Dep Mean       0.00000     Adj R-sq      -0.0072
    C.V.      7641577.2219

                         Parameter Estimates

                         Parameter      Standard    T for H0:
     Variable  DF         Estimate         Error   Parameter=0

     F1         1        -0.002857    0.01261060        -0.227
     F2         1        -0.003987    0.00982955        -0.406
     F3         1         0.043420    0.15679176         0.277
     F4         1         0.045355    0.02612781         1.736
```


Consider Example 1:

$$y_t = \theta_1 x_{1t} + \theta_2 x_{2t} + \theta_4 e^{\theta_3 x_{3t}} + e_t$$

$$\mathrm{H} : \theta_3 \theta_4 e^{\theta_3} - \frac{1}{5} = 0$$

Computations:

$$\mathrm{SSE}(\tilde{\theta}) = 0.3493$$

$$\tilde{D}'(\tilde{F}'\tilde{F})\tilde{D} = 0.00444$$

$$R = \frac{(30)(0.0444)}{(0.3493)/26} = 3.81$$

$$R = (30)(0.1271) = 3.81$$

$$\mathcal{X}^2(0.95, 1) = 3.841 \quad d = 4.19$$

Lagrange Multiplier Test (SAS code)

```
data amstat;
  infile "amstat.dat";
  input t y x1 x2 x3;

proc model data=amstat;
  y=t1*x1+t2*x2+t4*exp(t3*x3);
  parms t1=-0.048660 t2=1.038835 t3=-0.737919 t4=-0.513623;
  fit y / ols converge=1.0e-8 maxiter=50 method=gauss;
  test t3*t4*exp(t3)-0.20 = 0 ,/ lm;
```

Lagrange Multiplier Test (SAS output)

```
                          Test Results

    Test       Type       Statistic      Prob.     Label

    Test0      L.M.            3.81     0.0509
    T3*T4*EXP(T3)-0.20 =
```


Lagrange Multiplier Test (5)

As for the likelihood ratio test, the derivation of the asymptotic distribution of the Lagrange multiplier test is not difficult but it is time consuming and therefore will be omitted.

Each of the following references contains the derivation. The second is recommended and can be downloaded from the course web page.

Gallant, A. Ronald (1987) *Nonlinear Statistical Models,* Wiley, New York.

Gallant, A. Ronald (1992) *Nonlinear Regression Asymptotics,* Manuscript, Department of Economics, University of North Carolina.

Gallant, A. Ronald (1997) *Introduction to Econometric Theory,* Princeton University Press, Princeton NJ.

Lack of Invariance of the Wald Test (1)

Me: I want to test the hypothesis that the half life in the following exponential model is 2 hours. My parameters are Cl and V, which is the standard parameterization in pharmacokinetic applications. The value of D_0 is known.

$y_t = \frac{D_0}{V} e^{-\frac{Cl}{V}t} + e_t$

Half life: $\frac{V}{Cl} \log 2$

H: $\frac{V}{Cl} = \frac{2}{\log 2}$

You: You use the standard parameterization of the model in the statistical literature:

$y_t = D_0 \theta_1 e^{-\theta_2 t} + e_t$

Half life: $\frac{\log 2}{\theta_2}$

H: $\theta_2 = \frac{\log 2}{2}$

Both of us are using the same model and testing the same hypothesis. With the same data, one would expect that we should both get the same result. But if we use the Wald test, one of us might accept and the other reject.

The relation between the models has the form $\theta = g(\rho)$; that is,

$$\theta = (\theta_1, \theta_2) = (\frac{1}{V}, \frac{Cl}{V}) = g(Cl, V) = g(\rho)$$

Lack of Invariance of the Wald Test (2)

Here is why this happens:

Me: $y = f(\theta) + e$

H: $\theta = \theta^*$

$W = (\hat{\theta} - \theta^*)'(\hat{F}'\hat{F})(\hat{\theta} - \theta^*)/(ps^2)$

You: $y = f[g(\rho)] + e$

H: $\rho = \rho^*$ where $g(\rho^*) = \theta^*$

$W = (\hat{\rho} - \rho^*)'(\hat{G}'\hat{F}'\hat{F}\hat{G})(\hat{\rho} - \rho^*)/(ps^2)$

The two statistics would be the same if

$$\hat{\theta} - \theta^* = \hat{G}(\hat{\rho} - \rho^*)$$

but this is not the case in general. The difference is the second order term in a Taylor's expansion:

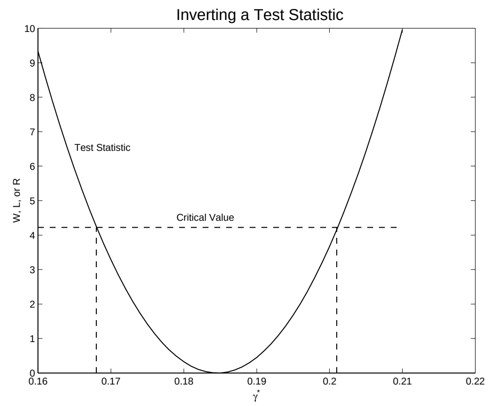
$$(\hat{\theta} - \theta^*) - \hat{G}(\hat{\rho} - \rho^*) = \frac{1}{2}(\hat{\rho} - \rho^*)' \frac{\partial^2}{\partial\rho\partial\rho'} g(\bar{\rho})(\hat{\rho} - \rho^*)$$

- Wald Test
  - Advantages: Can be computed from $\hat{\theta}$ only, which is useful if $f(x, \theta)$ is linear and $h(\theta)$ is not.
  - Disadvantages: Asymptotics are inaccurate. Not invariant to reparametrization.
- Likelihood Ratio Test
  - Advantages: Asymptotics are very accurate. Invariant to reparametrization. Better power than the Lagrange multiplier test.
  - Disadvantages: Requires two optimizations.
- Lagrange Multiplier Test
  - Advantages: Asymptotics are accurate. Invariant to reparametrization. Can be computed from $\tilde{\theta}$ only, which is useful if $f[x, g(\rho)]$ is linear.
  - Disadvantages: Spurious acceptance because $\tilde{D} = 0$ at every local minimum, local maximum, and saddle point. Power is not as good as the likelihood ratio test.

Topics

- Examples & Least Squares Estimates
- Notation & Taylor's Theorem
- Statistical Properties
- Computations
- Hypothesis Tests
- Confidence Intervals

Confidence Intervals



To set a confidence interval on a nonlinear function $\gamma(\theta)$, invert one of the three tests. That is, let

$$h(\theta) = \gamma(\theta) - \gamma^*$$

and put in the interval all γ^* for which

$$\text{H} : h(\theta) = 0$$

is accepted.

Wald Test

The Wald test accepts when

$$\frac{|\gamma(\hat{\theta}) - \gamma^*|}{s\left[\hat{H}(\hat{F}'\hat{F})^{-1}\hat{H}'\right]^{\frac{1}{2}}} \leq t_{\alpha/2}$$

The points that satisfy the inequality are

$$\gamma(\hat{\theta}) \pm t_{\alpha/2}\, s\left[\hat{H}(\hat{F}'\hat{F})^{-1}\hat{H}'\right]^{\frac{1}{2}}$$

Consider Example 1:

$$y_t = \theta_1 x_{1t} + \theta_2 x_{2t} + \theta_4 e^{\theta_3 x_{3t}} + e_t$$

$$\gamma(\theta) = \theta_3 \theta_4 e^{\theta_3}$$

Wald Test (SAS code)

```
proc model data=amstat;
  y=t1*x1+t2*x2+t4*exp(t3*x3);
  parms t1=-0.048660 t2=1.038835 t3=-0.737919 t4=-0.513623;
  fit y / ols converge=1.0e-8 maxiter=50 method=gauss;
  estimate "GrowthRate" t3*t4*exp(t3) ,/ covb;
```

Wald Test (SAS output)

```
                       Approx.      'T'  Approx.
Term         Estimate   Std Err    Ratio Prob>|T| Label

GrowthRate   0.184592  0.008050    22.93   0.0001 T3*T4*EXP(T3)
```

Computations:

$$\begin{aligned} &\gamma(\hat{\theta}) \pm t_{\alpha/2}\, s\left[\hat{H}(\hat{F}'\hat{F})^{-1}\hat{H}'\right]^{\frac{1}{2}} \\ &= 0.1846 \pm (2.054)(0.00805) \\ &= [0.168, 0.201] \end{aligned}$$

Likelihood Ratio Test

The likelihood ratio test accepts when

$$\frac{\text{SSE}(\tilde{\theta}_{\gamma^*}) - \text{SSE}(\hat{\theta})}{\text{SSE}(\tilde{\theta}_{\gamma^*})/(n-p)} \leq F_{\alpha/2}$$

where

$$\tilde{\theta}_{\gamma^*} = \underset{\gamma(\theta)=\gamma^*}{\text{argmin}}\, \text{SSE}(\theta)$$

Consider Example 1:

$$y_t = \theta_1 x_{1t} + \theta_2 x_{2t} + \theta_4 e^{\theta_3 x_{3t}} + e_t$$

$$\gamma(\theta) = \theta_3 \theta_4 e^{\theta_3}$$

Likelihood Ratio Test (SAS code)

```
proc model data=amstat;
  y=t1*x1+t2*x2+t4*exp(t3*x3);
  parms t1=-0.048660 t2=1.038835 t3=-0.737919 t4=-0.513623;
  fit y / ols converge=1.0e-8 maxiter=50 method=gauss;
  test t3*t4*exp(t3)=0.166 ,/ lr;
  test t3*t4*exp(t3)=0.167 ,/ lr;
  test t3*t4*exp(t3)=0.168 ,/ lr;
  test t3*t4*exp(t3)=0.200 ,/ lr;
  test t3*t4*exp(t3)=0.201 ,/ lr;
  test t3*t4*exp(t3)=0.202 ,/ lr;
```

Likelihood Ratio Test (SAS output)

```
                          Test Results

    Test       Type       Statistic      Prob.     Label

    Test0      L.R.            4.62     0.0316
    T3*T4*EXP(T3)=0.166
    Test1      L.R.            4.18     0.0408
    T3*T4*EXP(T3)=0.167
    Test2      L.R.            3.76     0.0523
    T3*T4*EXP(T3)=0.168
    Test3      L.R.            3.78     0.0518
    T3*T4*EXP(T3)=0.200
    Test4      L.R.            4.29     0.0382
    T3*T4*EXP(T3)=0.201
    Test5      L.R.            4.84     0.0278
    T3*T4*EXP(T3)=0.202
```


Likelihood Ratio Test (Matlab code)

```
A = [1 0.166 0.166^2 ; 1 0.167 0.167^2 ; 1 0.168 0.168^2];
y = [4.62 ; 4.18 ; 3.76];
b = inv(A)*y;
root_l = (- b(2) - sqrt(b(2)^2 - 4*b(3)*(b(1)-4.22)))/(2*b(3))

A = [1 0.200 0.200^2 ; 1 0.201 0.201^2 ; 1 0.202 0.202^2];
y = [3.78 ; 4.29 ; 4.84];
b = inv(A)*y;
root_r = (- b(2) + sqrt(b(2)^2 - 4*b(3)*(b(1)-4.22)))/(2*b(3))
```

Likelihood Ratio Test (Matlab output)

```
root_l =

    0.1669

root_r =

    0.2009
```

Confidence Interval:

[0.167, 0.201]

Lagrange Multiplier Test

The Lagrange multiplier test accepts when

$$\frac{n\tilde{D}'\tilde{F}'\tilde{F}\tilde{D}}{\mathsf{SSE}(\tilde{\theta}_{\gamma^*})} \leq d_{\alpha/2} \doteq \mathcal{X}^2_{\alpha/2}$$

where

$$\tilde{\theta}_{\gamma^*} = \underset{\gamma(\theta)=\gamma^*}{\text{argmin}} \; \mathsf{SSE}(\theta)$$

Consider Example 1:

$$y_t = \theta_1 x_{1t} + \theta_2 x_{2t} + \theta_4 e^{\theta_3 x_{3t}} + e_t$$

$$\gamma(\theta) = \theta_3 \theta_4 e^{\theta_3}$$

Lagrange Multiplier Test (SAS code)

```
proc model data=amstat;
  y=t1*x1+t2*x2+t4*exp(t3*x3);
  parms t1=-0.048660 t2=1.038835 t3=-0.737919 t4=-0.513623;
  fit y / ols converge=1.0e-8 maxiter=50 method=gauss;
  test t3*t4*exp(t3)=0.166 ,/ lm;
  test t3*t4*exp(t3)=0.167 ,/ lm;
  test t3*t4*exp(t3)=0.168 ,/ lm;
  test t3*t4*exp(t3)=0.200 ,/ lm;
  test t3*t4*exp(t3)=0.201 ,/ lm;
  test t3*t4*exp(t3)=0.202 ,/ lm;
```

Lagrange Multiplier Test (SAS output)

```
                          Test Results

    Test       Type       Statistic      Prob.     Label

    Test0      L.M.            4.60     0.0320
    T3*T4*EXP(T3)=0.166
    Test1      L.M.            4.23     0.0398
    T3*T4*EXP(T3)=0.167
    Test2      L.M.            3.85     0.0497
    T3*T4*EXP(T3)=0.168
    Test3      L.M.            3.81     0.0509
    T3*T4*EXP(T3)=0.200
    Test4      L.M.            4.25     0.0391
    T3*T4*EXP(T3)=0.201
    Test5      L.M.            4.71     0.0300
    T3*T4*EXP(T3)=0.202
```

Lagrange Multiplier Test (Matlab code)

```
y = [4.60 ; 4.23 ; 3.85];
b = inv(A)*y;
root_l = (- b(2) - sqrt(b(2)^2 - 4*b(3)*(b(1)-4.19)))/(2*b(3))

A = [1 0.200 0.200^2 ; 1 0.201 0.201^2 ; 1 0.202 0.202^2];
y = [3.81 ; 4.25 ; 4.71];
b = inv(A)*y;
root_r = (- b(2) + sqrt(b(2)^2 - 4*b(3)*(b(1)-4.19)))/(2*b(3))
```

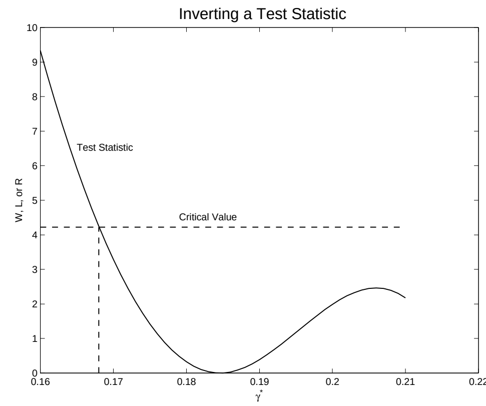
Lagrange Multiplier Test (Matlab output)

```
root_l =

    0.1671

root_r =

    0.2009
```

Confidence Interval:

$$[0.167, 0.201]$$

Confidence Interval Problems



Neither the likelihood ratio test statistic nor the Lagrange multiplier test statistic are guaranteed to plot above their critical values. This can result in open ended confidence intervals as shown above. Models with exponential terms in them sometimes exhibit this behavior. Also, the test statistic can oscillate about its critical value resulting in confidence sets that are a union of disjoint intervals. This can happen with spline models where the join point is estimated. The Wald test does not have these problems and always produces a confidence interval that is symmetric about the estimate of the parametric function.

Topics

- Examples & Least Squares Estimates
- Notation & Taylor's Theorem
- Statistical Properties
- Computations
- Hypothesis Tests
- Confidence Intervals