James B. McDonald
Brigham Young University
12/2012

V. Nonlinear Models

1. Introduction
2. Estimation
    a. Nonlinear least squares
    b. Maximum likelihood
    c. Estimation subject to constraints
    d. Other methods of estimation
        (1) Method of moments
        (2) Generalized Method of Moments
        (3) Methods for grouped data
        (4) Extremum Estimators (QMLE)
3. Computational Algorithms
4. Statistical Inference
    a. Individual parameters
    b. More general tests
        (1) Wald
        (2) LM
        (3) LR
5. Some Applications
    a. Production functions
    b. Models for income distribution
    c. A precautionary note
    d. Binary or limited dependent variables
    e. Ordered or categorical data
    f. Poisson regression
    g. Censored regression
    h. Truncated and grouped regression
    i. Interval regression
    j. Quantile regression
    k. Hazard functions
    j. GMM
    k. Bootstrapping
6. Some exercises

James B. McDonald
Brigham Young University
12/2012

V. Nonlinear Models

1. Introduction

Many nonlinear models can be transformed into a form such that linear techniques can be employed in estimating the unknown parameters. One such example is the Cobb-Douglas production function

$$Y_t = AL_t^{\beta_1}K_t^{\beta_2}\varepsilon_t .$$

Taking the logarithm of each side yields the equivalent expression

$$\ln(Y_t) = \ln A + \beta_1 \ln(L_t) + \beta_2 \ln(K_t) + \ln\varepsilon_t.$$

which is linear in the parameters β_2 ε β_3 .

The second representation can be easily estimated using methods associated with linear models. If the (ln ε_t)'s are distributed normally, with zero mean, constant variance and independent of each other, then an application of least squares will yield maximum likelihood **estimators.** Recall that $\ln(\varepsilon_t) \sim N[0, \sigma^2]$ implies that $\varepsilon_t \sim LN[\mu = 0, \sigma^2]$[1].

Many applied problems are associated with nonlinear models which can't be so easily estimated. For example, the constant elasticity of substitution production function (CES) is defined by

$$Y_t = A(\delta L_t^{\rho} + (1-\delta_t)K_t^{\rho})^{M/\rho}\ \varepsilon_t$$

This is obviously a nonlinear relationship. Taking the logarithms of both sides yields

$$\ln Y_t = \ln A + (M/\rho) \ln(\delta L_t^{\rho} + (1-\delta)K_t^{\rho}) + \ln\varepsilon_t$$

which is nonlinear in the parameters and is of a form which can't be directly estimated using linear techniques.

---

[1] $E(\varepsilon_t) = e^{\mu+.5\sigma^2}$

The CES production function is an example of a nonlinear model which is of the general form (1.1) $Y_t = f(X_t; \beta) + \varepsilon_t$

where Y_t denotes an endogenous variable, X_t represents a Kx1 vector of exogenous variables and ε_t is a random disturbance. f( ) represents a known function with unknown parameter values, denoted by β. Given N observations for Y_t and the vector X_t, we can also write

(1.1)' $Y=f(X;\beta) + \varepsilon$

where $Y=(Y_1,....,Y_N)'$, $X=(X_1,....,X_N)'$, and $\varepsilon=(\varepsilon_1,...,\varepsilon_N)'$. The vector function $f(X;\beta)$ is then defined to be $(f(X_1;\beta), \ldots, f(X_N;\beta))'$. Also assume that β has K* elements, $\beta=(\beta_1, \ldots, \beta_{K*})'$. The number of unknown parameters (K*) need not equal the number of independent variables (K), e.g., the C.E.S. function production has four unknown parameters (A, δ, M, ρ) and two independent variables (K, L). For notational convenience, we will assume that the number of parameters is K. The reader is referred to the excellent survey paper, "Nonlinear Regression Models" by T. Amemiya in Handbook of Econometrics. Section 2 will discuss methods of estimating models which are nonlinear in the parameters. Section 3 will summarize algorithms used in obtaining estimators. Methods of statistical inference using the Wald, likelihood ratio and Lagrangian multiplier test will be presented in Section 4, and applications will be considered in the Section 5.

2. Estimation

The estimation problem is basically the problem of obtaining estimators of the vector of parameters β in

(2.1) $Y=f(X;\beta) + \varepsilon$

from the sample data in Y and X. Many estimation techniques are available, but two of the most commonly used are nonlinear least squares and maximum likelihood estimation., although GMM, extremum, kernel, and other methods are discussed in the chapter on Estimation

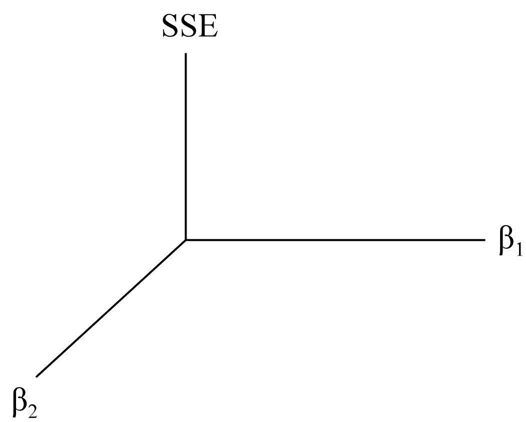
Frameworks. Each of these methods generally involves solving a nonlinear system of equations. As previously noted, some computational methods of solving these equations will be surveyed in section three.

a. Nonlinear Least Squares Estimators (NLS):

The sum of squared errors associated with (2.1) is given by

$$
\begin{aligned}
(2.2)\qquad SSE(\beta) &= \sum_t \varepsilon_t^2 \\
&= \varepsilon'\varepsilon \\
&= (Y-f(X;\beta))'\,(Y-f(X;\beta)).
\end{aligned}
$$

The least squares estimator of β, β_{NLS}, satisfies the necessary conditions:



$$(2.3)\qquad \frac{dSSE(\beta)}{d\beta} = 0$$

These equations are typically nonlinear in the β_i's.

If the errors are independently and identically distributed with constant variance and a zero mean, $\varepsilon \sim N[0, \sigma^2 I]$, and the function f( ) satisfies certain regularity conditions, then the least squares estimator β_{NLS} will be consistent and asymptotically normal. σ^2 can be estimated by

$$(2.4)\qquad \hat{\sigma}^2_{NLS} = s^2_{NLS} = SSE\left(\hat{\beta}_{NLS}\right)/\left(N-K\right)$$

and, the variance matrix of β_{NLS} can be estimated by

$$(2.5)\qquad Var\left(\hat{\beta}_{NLS}\right) = \hat{\sigma}^2_{NLS}\left[Z\left(\hat{\beta}_{NLS}\right)'Z\left(\hat{\beta}_{NLS}\right)\right]^{-1}$$

where $Z(\beta) = \left(\frac{\partial f(X;\beta)}{\partial \beta'}\right)_{NxK}$

$$= \begin{bmatrix} \frac{\partial f(X_1;\beta)}{\partial \beta_1} & \cdots & \frac{\partial f(X_1;\beta)}{\partial \beta_K} \\ \vdots & & \\ \frac{\partial f(X_N;\beta)}{\partial \beta_1} & \cdots & \frac{\partial f(X_N;\beta)}{\partial \beta_K} \end{bmatrix}$$

evaluated at β_{NLS}.

The **sandwich estimator** can also be used to provide an estimator of the variance.

In the linear model $Y = X\beta + \varepsilon$

$$Z(\beta) = \frac{\partial (X\beta)}{\partial \beta'} = X$$

and we obtain the regular least squares result

$$\beta_{NLS} = (X'X)^{-1}X'Y \overset{a}{\sim} N(\beta, \sigma^2(X'X)^{-1}).$$

Consider the nature of the matrix $Z(\beta)$ corresponding to the CES production function given on page 2.

Nonlinear least squares can be easily performed using popular software packages.

A simple example of using nonlinear least squares to estimate $y = \beta_1 + \beta_2 / x + \varepsilon$

Obviously OLS could be used with a transformed variable.

STATA commands

nl(y=eqn), initial(coef value coef value)   generic form

nl (y={b1}+{b2}/x), initial (b0 1 b1 3) with initial starting values or

nl (y={b1}+{b2}/x), nolog   suppresses output and doesn't specify initial starting values

nl (y={b1}+{b2}/x), initial (b0 1 b1 3) robust   reports robust standard errors

As another example, the CES production function just discussed could be estimated using the commands:

STATA:

```
gen ly=log(y)
nl (ly={g}+({M}/{rho})*log({d}*L^{rho}+(1-{d})*k^{rho}) ), initial (g 1 M 1 rho .5 d .5)
```

It is easy! Try it you will like it. To summarize, the "*nl*" option in STATA selects the parameter estimates to minimize the sum of squared errors.

b. Maximum Likelihood Estimation

In order to obtain MLE of β in (1.1), the distribution of the vector

$$\varepsilon = Y - f(X;\beta)$$

must be given. Let this density be denoted by $g(\varepsilon;\theta)$ wher θ denotes the parameters associated with the density function g( ), e.g., σ^2 for the normal. The likelihood function L( ) and log likelihood functions, respectively, are defined by

$$\text{(2.6a)} \quad L(\beta,\psi) = g(\varepsilon;\theta)$$

$$= g(Y - f(X;\beta);\theta)$$

$$\text{(2.6b)} \quad \ell(\beta,\psi) = \ln g(Y - f(X;\beta);\theta).$$

If the random disturbances in the model are identically and independently (homoskedastic and not autocorrelated) distributed, then the log likelihood function (2.6b) can be written as a sum

$$\text{(2.7)} \quad \ell = \sum_{t=1}^{N} \ln[g(Y_t - f(X_t;\beta);\theta]$$

The maximum likelihood estimators of θ = (β,ψ) are defined by the solution of the equations

$$\text{(2.7a,b)} \quad \frac{\partial \ell}{\partial \beta} = 0$$

$$\frac{\partial \ell}{\partial \psi} = 0 \qquad .$$

These equations are typically nonlinear in the parameters, and the solution again requires nonlinear optimization procedures. Under quite general regularity conditions:

(R.1) the range of the random variable Y_t is independent of the vector of parameters $\theta = (\beta,\psi)$;

(R.2) the density g( ) possesses derivatives of at least third order with respect to θ and these derivatives are bounded,

$$\frac{|\partial^3 L|}{|\partial\theta_i^3|} < M_i(y)$$

$E(M_i\,(Y)) < K$ constant; then

the maximum likelihood estimators are asymptotically normal

$$\text{(2.9)} \qquad \tilde{\theta}_{ML} \overset{a}{\sim} N\left[\theta = \begin{pmatrix}\beta \\ \psi\end{pmatrix};\ -\left(E\frac{\partial^2 \ell}{\partial\theta\partial\theta'}\right)^{-1}\right]$$

or

$$\text{(2.10)} \qquad \sqrt{N}(\tilde{\theta}_{ML} - \theta) \overset{a}{\sim} N\left[0;\ N\left(-E\frac{\partial^2 \ell}{\partial\theta\partial\theta'}\right)^{-1}\right]$$

Note that the asymptotic variance covariance matrix is the *Cramer-Rao* lower bound.

One of the most common assumptions for the distribution of the random disturbances is that

$$\text{(2.11)} \qquad \varepsilon \sim N[0,\Sigma]$$

in which case the log likelihood function (2.6b) is given by

$$\text{(2.12)} \qquad \ell = \ln L = -(Y-f(X;\beta))'\Sigma^{-1}(Y-f(X;\beta))/2$$

$$-1/2\ \ln|\Sigma| - \frac{N}{2} \qquad \ln(2\pi).$$

If the normal errors are also independently and identically distributed, i.e.,

$$\varepsilon \sim N[0,\Sigma = \sigma^2\ I],$$

the corresponding loglikelihood function is given by

$$(2.13) \qquad \ell = \ln L = -\frac{1}{2\sigma^2}\sum_{t=1}^{N}(Y_t - f(X_t;\beta))^2 - \frac{N}{2}(\ln\sigma^2 + \ln 2\Pi)$$

$$= -\frac{SSE(\beta)}{2\sigma^2} - \frac{N}{2}(\ln\sigma^2 + \ln 2\Pi)$$

$$= -\frac{(Y - f(X;\beta))'(Y - f(X;\beta))}{2\sigma^2} - \frac{N}{2}(\ln\sigma^2 + \ln 2\Pi)$$

The maximum likelihood estimators in this case are defined by

$$\frac{\partial\ell}{\partial\beta} = -\frac{1}{2\sigma^2}\frac{\partial SSE(\beta)}{\partial\beta} = \left(\frac{1}{\sigma^2}\right)\left(\frac{\partial f}{\partial\beta}\right)(f(X;\beta) - Y) = \frac{\partial f(\varepsilon)}{\partial\beta}\left(\frac{\varepsilon}{\sigma^2}\right)$$

$$\frac{\partial\ell}{\partial\sigma^2} = \frac{SSE(\beta)}{2\sigma^4} - \frac{N}{2\sigma^2}$$

$$= \frac{\varepsilon'\varepsilon}{2\sigma^4} - \frac{N}{2\sigma^2} \quad .$$

Setting the derivatives to zero and solving implies

$$\hat{\beta}_{NLS} = \tilde{\beta}_{ML}$$

$$\tilde{\sigma}_{ML} = SSE(\beta_{ML})/N.$$

Substituting SSE(β)/N into (2.13) for σ^2 yields what is called the *concentrated likelihood function*

$$(2.13)' \qquad \ell = -\frac{N}{2}\left(1 + \ln(2\pi) + \ln\left(\frac{SSE}{N}\right)\right) \quad .$$

Recall from our previous discussion of MLE estimators that the asymptotic variance matrix (Cramer Rao) can also be obtained from

$$\left[-E\left(\frac{d^2\ell}{d\theta^2}\right)\right]^{-1} = \left[E\left[\frac{d\ell}{d\theta}\ \frac{d\ell}{d\theta'}\right]\right]^{-1} \quad .$$

Given the previous results, the second expression for the variance covariance matrix, i.e.,

$$\text{(2.14)} \qquad Var(\beta_{ML}) = \left[E\frac{dl}{d\beta}\ \frac{dl}{d\beta'}\right]^{-1}$$

$$= \left[E\ \frac{df}{d\beta}\left(\frac{\varepsilon}{\sigma^2}\right)\frac{\varepsilon'}{\sigma^2}\left(\frac{df}{d\beta}\right)'\right]^{-1}$$

$$= \left[\frac{1}{\sigma^4}\frac{df}{d\beta}E(\varepsilon\varepsilon')\left(\frac{df}{d\beta}\right)'\right]^{-1}$$

$$= \sigma^2\left[\frac{df}{d\beta}\left(\frac{df}{d\beta}\right)'\right]^{-1} \qquad .$$

Note that (2.5) corresponds to the results for nonlinear least squares, as it should. Thus we see that least squares and MLE will be equivalent for the case of normally distributed residuals which are uncorrelated and homoskedastic. In summary, NLS and MLE will be equivalent if $\varepsilon \sim N[0, \sigma^2 I]$.

Exercise:

Consider the case of the generalized error distribution in which the density of the random disturbances is given by

$$\mathrm{GED}(\varepsilon_t; \psi=(p,s)) = \frac{p e^{-(|\varepsilon_t|^p/s^p)}}{2\Gamma(1/p)}$$

$$= \frac{p e^{-(|Y_t - f(X_t;\beta)|^p/\sigma^p)}}{2\sigma\Gamma(1/p)}$$

(a) Write out an expression for the log likelihood function.

(b) Interpret the sum in the log likelihood function corresponding to

(1) $p = 1$

(2) $p = 2$

(c) What relationship does this density function have to the normal density? Hint: $\Gamma(1/2) = \sqrt{\pi}$

(d) How could MLE of $\beta = (\beta_1, \ldots, \beta_K)'$, σ and "p" be obtained (conceptually)?

The **STATA commands** for obtaining estimates of the unknown regression and distributional parameters with the GED distribution are given as follows:

```
clear
cap prog drop GED
program define GED
version 1.0
args lnf xb s p
quietly replace `lnf'=ln(abs(`p'))-((abs($ML_y1-`xb')/`s')^(abs(`p')))-ln(2*`s')-lngamma((abs(1/`p')))
end

clear
infile y x using g:\american.dat


ml model lf GED (y=x) (s:) (p:), technique(dfp)
ml search
ml maximize
```

For comparative purposes, the commands for MLE with a normal pdf are given by (STATA 10)

STATA 10

STATA commands for a normal pdf in a linear regression model

```
cap prog drop normalpdf
program define normalpdf
version 1.0
args lnf mu sigma (lnf, mu, and sigma are the parameters of estimation)
quietly replace `lnf'=ln(normalden($ML_y1, `mu', `sigma'))
end
ml model lf normalpdf (y=x's) (sigma: ), technique(dfp)
ml search
ml maximize, difficult
```

The fifth problem in the homework section involves an application of MLE using STATA with a GED pdf.

c. Estimation subject to constraints

Assume that we believe that the parameters θ satisfy the constraint

(2.15) $q(\theta)=0$

where $q(\theta)$ can denote an rx1 vector of univariate real valued functions if multiple constraints exist (r constraints or restrictions). For example, if the CES production function is characterized by constant returns to scale, it can be shown that

$$M = 1$$

or

$$M - 1 = 0.$$

There are several approaches to estimation subject to constraints on the parameters. One is to reduce the number of parameters by substitution. Another approach provides some necessary groundwork for testing hypotheses of the form

$$q(\theta)= 0$$

where $q(\theta)$ is differentiable and will be briefly discussed.

Assume that the objective function to be maximized is $H(\theta)$, e.g., if nonlinear least squares estimation is used then $H(\theta)= - SSE(\beta)$; if MLE is used then

$$H(\theta)=\ell(\beta,\sigma^2).$$

A second estimation procedure which takes account of constraints on parameters is based on Lagrangian multipliers.

Define

$$\mathcal{L}(\theta,\lambda)=H(\theta) + \lambda' q(\theta)$$

where $\lambda'=(\lambda_1, \ldots, \lambda_r)$ and $q(\theta)=(q_1(\theta), \ldots, q_r(\theta))'$ correspond to r constraints imposed on the parameters. λ_i is generally referred to as the i^{th} Lagrangian multiplier which corresponds to the i constraint $q_i(\theta)=0$.

The necessary conditions associated with this constrained optimization problem,

(2.16) $$\max_{\theta} H(\theta)$$

subject to $q(\theta)=0$,

are given by

(2.17) $$\frac{\partial \mathcal{L}(\theta,\lambda)}{\partial \theta} = 0$$

$$\frac{\partial \mathcal{L}(\theta,\lambda)}{\partial \lambda} = q(\theta) = 0 .$$

The Lagrangian multipliers in this formulation can be used to test the validity of the constraints.

d. Other methods of estimation (generally require nonlinear optimization procedures–covered in another section of the notes in more detail)

(1) Method of moments. Sample moments of the data are equated to the theoretical moments which are functions of the parameters. These are equated and the resulting estimators are usually consistent, but not necessarily efficient.

(2) Generalized Method of Moments. See the discussion in the Regression Section which describes estimation where more moment restrictions are considered than the number of parameters.

(3) Methods for grouped data. Data is in the form of cell frequencies. Predicted frequencies depend upon unknown parameters. Estimation procedures are based on "matching" observed

and predicted frequencies. Possible criteria: Least Squares, MLE, minimum or modified minimum chi-square.

(4) Extremum estimators. As discussed earlier, extremum estimators are defined by the equation

$$\tilde{\theta} = \arg\max_{\theta} H(\theta) \quad .$$

Where under appropriate regularity conditions,

$$\tilde{\theta} \sim^{a} N\left[\theta; C = B^{-1}AB^{-1}\right]$$

with $B = \left(E\frac{d^2 H}{d\theta d\theta'} \right)$ and $\mathrm{A} = \left(E\frac{\mathrm{dH}}{\mathrm{d}\theta}\frac{\mathrm{dH}}{\mathrm{d}\theta'} \right)$.

This estimator includes the nonlinear least squares and maximum likelihood estimators as special cases. Additionally if MLE is performed with a questionable specification for the pdf, the corresponding estimator is frequently referred to as a Quasi Maximum Likelihood estimator (QMLE).

3. Computation

The derivation of either nonlinear least squares or maximum likelihood estimators requires the solution of a system of nonlinear equations. For example if the estimation problem has been written as a maximization problem

$$(3.1) \qquad \max_{\theta} H(\theta), \qquad H(\theta) = \begin{cases} -SSE(\beta) \text{ for L.S.} \\ \ell(\beta,\sigma^2) \text{ for M.L.E} \\ -g(\beta)'Var^{-1}(g(\beta))g(\beta) \end{cases}$$

estimation requires solving the equations:

$$(3.2) \qquad H_\theta(\theta) = \frac{dH(\theta)}{d\theta} = 0 \qquad .$$

Before discussing numerical methods of computing solutions to (3.2) it should be mentioned that

(1) Round off error can accumulate.

(2) The objective function may have a "plateau" or several local optima.

(3) Many techniques are based upon gradient techniques.

a. Generalized Gradient Methods

The procedure for maximizing (3.1) generally involves selecting an <u>initial guess for θ</u>, and then determining whether this corresponds to a local maximum. If not, determine a <u>direction "δ"</u> in which to adjust and <u>how far (step size) "t" to move</u> in that direction. This procedure is continued until the desired convergence is achieved.

The gradient vector $\left(\frac{dH}{d\theta}\right)$

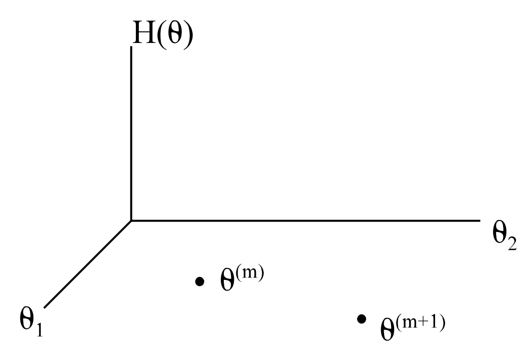
indicates the direction of movement or adjustment in the θ's which will result in the "most rapid" increase in $H(\theta)$. Some algorithms modify the direction of movement by multiplying the gradient vector by a matrix P to yield $P\left(\frac{dH}{d\theta}\right)$.

Let "t" denote a scale factor which determines the distance moved. Thus, an iterative search procedure can be implemented as follows:

(3.3) $$\boxed{\theta^{(n+1)} = \theta^{(n)} + t_n\left\{P_n\left(\frac{dH(\theta^{(n)})}{d\theta}\right)\right\}}$$ or

(3.4) $$\boxed{\begin{aligned}\Delta\theta &= t_n\left\{P_n\left(\frac{dH(\theta^{(n)})}{d\theta}\right)\right\}\\ &= t_n\delta_n\end{aligned}}$$.

This can be visualized as in the following figure



<u>What type of matrices P_n will lead to an improvement, i.e. $H(\theta^{(n+1)}) > H(\theta^{(n)})$?</u>

Consider the Taylor Series expansion of $H(\theta)$ about $\theta^{(n)}$

(3.5) $$H(\theta^{(n+1)}) \doteq H(\theta^{(n)}) + \frac{dH(\theta^{(n)})}{d\theta'}\Delta\theta^{(n)}$$

$$+ \; 1/2\Delta\theta^{(n)'}\frac{dH^2(\theta^{(n)})}{d\theta d\theta'}\Delta\theta^{(n)} \qquad .$$

Substituting (3.4) into (3.5) yields

$$(3.5)' \quad H(\theta^{(n+1)}) \; = \; H(\theta^{(n)}) \; + \; \frac{dH(\theta^{(n)})}{d\theta'} \; t_n P_n \frac{dH}{d\theta}$$

$$+ \; \frac{t_n^2}{2} \; \frac{dH(\theta^{(n)})}{d\theta'} P_n' \frac{d^2H(\theta^{(n)})}{d\theta d\theta'} \; P_n \frac{dH(\theta^{(n)})}{d\theta}$$

or

$$(3.6) \qquad H(\theta^{(n+1)}) \; - \; H(\theta^{(n)})$$

$$\doteq \; \frac{dH(\theta^{(n)})}{d\theta'}\left\{ t_n P_n \; + \; \frac{t_n^2}{2} P_n' \frac{[d^2H(\theta^{(n)})]}{d\theta d\theta'} P_n \right\}\frac{dH(\theta^{(n)})}{d\theta}$$

The iterative adjustment will correspond to an increase in $H(\theta)$, i.e.,

$$H(\theta^{(n+1)}) > H(\theta^{(n)})$$

if the bracketed expression is positive definite.

<u>Special cases</u>

(1.) <u>Simple Gradient</u>. (P_n=I).

$$\theta^{(n+1)} \; = \; \theta^{(n)} \; + \; t_n P_n \left(\frac{dH}{d\theta} \right)$$

$$= \; \theta^{(n)} \; + \; \left(\frac{dH}{d\theta} \right), \; P_n \; = \; I, \; t_n \; = \; 1$$

This method will converge, but may do so very slowly.

(2.) Newton Raphson $\left(P_N \; = \; \left(\frac{d^2H}{d\theta d\theta'} \right)^{-1} \right)$ ____________________

$$\theta^{(n+1)} \; = \; \theta^{(n)} \; + \; t_n P_n \left(\frac{dH}{d\theta} \right) \qquad .$$

The Newton Raphson works very well near an optimum. If we are a long way from an optimum, Newton-Raphson need not work well and may even send us in the wrong direction. If the original objective function H(θ) is quadratic, the Newton-Raphson algorithm will yield the solution in one step. The motivation for this selection of P_n is that (3.5)' is maximized with respect to $\theta^{(n+1)}$ to yield

$$P_n = \left(\frac{d^2H}{d\theta d\theta'} \right)^{-1} .$$

(3.) Fletcher-Powell. Similar to Newton-Raphson except numerical derivatives are used to evaluate $d^2H/d\theta d\theta'$ in the previous equation.

(4.) The Berndt, Hall, Hall, Hausman (BHHH) algorithm corresponds to using.

$$P_n = \left[\frac{dH}{d\theta} \frac{dH}{d\theta'} \right]^{-1}$$

(5) other variations exist and many are characterized by different selections for P_n and t_n.

b. Direct Search and Grid Search Techniques.

See Goldfeld and Quandt--Nonlinear Methods in Econometrics, 1972 Amsterdam: North Holland.

4. Statistical Inference

The exact distribution of parameter estimators in nonlinear models is generally unknown. However, approximating asymptotic distributions can be determined in some important leading cases. There is no generally valid rule of thumb relating sample size for a desired accuracy of the asymptotic distributions. Consequently, the only justification for statistical inference in nonlinear models must be based on asymptotic considerations. One approach to estimating the unknown distribution is using what is called the bootstrap which will be briefly discussed in another section.

The asymptotic distribution of θ_{MLE} is

$$N(\theta, \Sigma_{\theta_{MLE}})$$

where the asymptotic variance of θ_{MLE} is given by

$$\Sigma_{\theta_{MLE}} = -\left(E\frac{d^2\ell}{d\theta d\theta'}\right)^{-1} \quad \left(E\frac{d\ell}{d\theta}\frac{d\ell}{d\theta'}\right)^{-1} \qquad =$$

the Cramer-Rao matrix, and also by

$$\sigma^2_{MLE}\left[\frac{\partial f(X;\beta_{MLE})}{\partial\beta}\ \frac{\partial f(X;\beta_{MLE})}{\partial\beta'}\right]^{-1}$$

for the case of $\varepsilon \sim N[0;\sigma^2 I]$.

a. Confidence Intervals (asymptotic) for Individual θ_is.

(4.1) $\quad H_o\text{: } \theta_i = \theta_i^0$

Confidence intervals for θ_i based on the asymptotic distribution, are given by

(4.2) $\quad \theta_{ML,i} \pm Z_{\alpha/2}\, \sigma_{MLE,i}$

where $Z_{\alpha/2}$= appropriate critical level for a standard normal,

$\sigma^2_{MLE,i}$ = MLE of asymptotic variance of $\theta_{MLE,i}$

= i^{th} diagonal element of $\Sigma_{\theta_{MLE}}$

b. <u>More General Tests</u>, e.g.,

(4.3) $$\boxed{H_o:\ q(\theta) = 0,}$$

against the alternative H_a: $q(\theta) \neq 0$. $q(\theta)$ is a (rx1) vector of r- functional constraints on the individual parameters θ, $q(\theta)$ must be continuously differentiable and $(\partial q/\partial\theta')$ must be of full rank, r. If we want to test the hypothesis that the parameters in the CES production function satisfied the constraints

$$M = 1 \text{ and}$$

$$\delta = 1/2,$$

the constraints could be written in the form (4.3) as

$$q(\theta) = \begin{pmatrix} M & - & 1 \\ \delta & - & 1/2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} .$$

The Wald (W) test, Lagrangian Multiplier (LM) or Rao test and Likelihood ratio (LR) test can be used to test hypotheses of the form (4.3). They differ in ease of application, but share the same asymptotic distribution.

(1) <u>The Wald test</u>.

To perform a Wald test, the *unconstrained* (without the constraints imposed) model is estimated and the test statistic is defined by

(4.4) $$\boxed{\begin{aligned} W &= q(\theta_{ML})'(\mathrm{Var}(q(\theta_{ML}))^{-1}q(\theta_{ML}) \\ W &= q(\theta_{ML})'\left(\frac{\partial q}{\partial\theta'}\ \Sigma_{\theta_{ML}}\frac{\partial q}{\partial\theta}\right)^{-1} q(\theta_{ML}) \end{aligned}}$$

(evaluated at θ_{ML}). The asymptotic distribution of W is a chi-square with degrees of freedom equal to r (the number of independent constraints on the θ_i's),

$$\boxed{W \overset{a}{\sim} \chi^2\ (r)} .$$

Note: (a) The construction of the Wald test only requires the estimation of the unconstrained model.

(b) Equivalence of the two forms of the Wald test.

Given that

$$\theta_{ML} \overset{a}{\sim} N(\theta, \Sigma_{\theta_{ML}}),$$

it follows that $q(\theta_{MLE})$ has the following asymptotic distribution,

$$q(\theta_{ML}) \overset{a}{\sim} N\left[q(\theta) = 0,\ \Sigma_{q(\theta_{ML})} = \frac{dq}{d\theta'}\ \Sigma_{\theta_{ML}} \frac{dq}{d\theta}\right]$$

Demonstration(this methodology is very important for PhD bound students):

Consider the Taylor series expansion of $q(\theta_{ML})$ about the true value of θ:

$$q(\theta_{ML}) \doteq q(\theta) + \frac{dq}{d\theta'} (\theta_{ML} - \theta)$$

Consequently,

Asympt mean $q(\theta_{ML}) = q(\theta) = 0$

Asympt var $q(\theta_{ML}) = \dfrac{dq}{d\theta'} \Sigma_{\theta_{ML}} \dfrac{dq}{d\theta}$

Therefore,

$$q(\theta_{ML}) \overset{a}{\sim} N\left[q(\theta) = 0,\ \Sigma_{q(\theta_{ML})} = \frac{dq}{d\theta'} \Sigma_{\theta_{ML}} \frac{dq}{d\theta}\right].$$

(c) It follows that

$$\left(q\left(\theta_{MLE}\right) - 0\right)'\left(\frac{dq}{d\theta'} \Sigma_{\theta_{MLE}} \frac{dq}{d\theta}\right)^{-1} \left(q\left(\theta_{MLE}\right) - 0\right)$$

is asymptotically distributed as a $\chi^2(r)$.

(2) Lagrangian Multiplier (LM) or Rao Test.

The test statistic is defined by

$$(4.5)\quad \boxed{\begin{aligned} LM &= \left(\frac{\partial \ell}{\partial \theta}\right)' \left(\mathrm{Var}\left(\frac{\partial \ell}{\partial \theta}\right)\right)^{-1} \left(\frac{\partial \ell}{\partial \theta}\right) \\ LM &= \left(\frac{\partial \ell}{\partial \theta}\right)' \Sigma_{\theta_R} \left(\frac{\partial \ell}{\partial \theta}\right) \end{aligned}}$$

where the statistic $\frac{\partial \ell}{\partial \theta}$ is evaluated at the restricted maximum likelihood estimator (θ_R - the solution to 2.16). The motivation for the form and distribution of LM follows from the results established I.B Appendix B (Cramer-Rao Inequality):

The relevant results are

- $$E\left(\frac{d\ell}{d\theta}\right) = 0$$
- $$\mathrm{Var}\left(\frac{d\ell}{d\theta}\right) = -E\left(\frac{d^2\ell}{d\theta d\theta'}\right) = (\mathrm{Var}(\theta_R))^{-1} = E\left(\frac{\partial \ell}{\partial \theta}\frac{\partial \ell}{\partial \theta'}\right)$$

for the case where $q(\theta) = 0$. Therefore the quadratic form in (4.5) has an asymptotic chi-square distribution,

$$\boxed{LM \overset{a}{\sim} \chi^2(r)}$$

Note: (1)The evaluation of LM using (4.5) requires ML estimates of θ subject to the restriction, (θ_R), and also the associated ($\partial\ell/\partial\theta$).

(2) Unconstrained estimates of θ are not required.

(2) An alternative form for (4.5) can be expressed in terms of λ_{ML}, the vector of Lagrangian multipliers associated with $q(\theta) = 0$. See (2.16) and (2.17). Recall that the restricted MLE of θ, θ_R, satisfies

$$\max_{\ell} \ell(\theta)$$
$$\text{s.t. } q(\theta) = 0.$$

The solution can be structured in terms of

$$\ell(\theta) + \lambda' q(\theta).$$

Maximizing this with respect to θ yields the condition

$$\frac{\partial \ell(\theta)}{\partial \theta} + \frac{\partial q}{\partial \theta} \lambda_{ML} = 0$$

$$\lambda_{ML}' \frac{\partial q}{\partial \theta'} = -\frac{\partial l(\theta)}{\partial \theta'} .$$

Therefore,

(4.5)' $$\boxed{LM = \lambda_{ML}' \frac{\partial q}{\partial \theta'} \Sigma_{\theta_R} \frac{\partial q}{\partial \theta} \lambda_{ML}}$$

(4.5) and (4.5)' can be further simplified if the residuals are $N[0,\sigma^2 I]$. In this case

$$\frac{-\partial \ell}{\partial \beta'} = (Y - f(X,\beta_R))' \frac{df}{d\beta'} / \sigma_R^2$$

from (2.13)

Therefore,

$$LM = \left(\frac{(Y - f(X;\beta_R))'}{\sigma_R^2} \right) \frac{df}{d\beta'} \left[\sigma_R^2 \left(\frac{df}{d\beta} \frac{df}{d\beta'} \right)^{-1} \right]$$

$$\cdot \frac{df}{d\beta} \frac{(Y - f(X;\beta_R))}{\sigma_R^2}$$

or

(4.5)" $$\boxed{LM = \frac{(Y - f(X;\beta_R))' \frac{df}{d\beta'} \left(\left(\frac{df}{d\beta} \frac{df}{d\beta'} \right) \right)^{-1} \frac{df}{d\beta} (Y - f(X;\beta_R)}{SSE(\beta_R)}}$$

This form doesn't require estimates of λ_{ML}.

(3) The Likelihood Ratio Test.

This test involves obtaining MLE estimates of $\theta = (\beta',\psi')'$ with the constraints imposed and without the constraints imposed, denoted

(β_R,ψ_R) and (β_{ML}, ψ_{ML}).

The test statistic is defined by

$$(4.6) \qquad \boxed{\begin{aligned} LR &= 2[\ell(\beta_{ML},\psi_{ML}) - \ell(\beta_R, \psi_{ML}] \\ &= 2(\ell - \ell*) \end{aligned}}$$

and is asymptotically distributed as a chi square with the same degrees of freedom as for the Wald and LM test statistics. For the case in which

$$\varepsilon \sim N(0, \sigma^2 I),$$

the reader can use (2.13) to demonstrate that

$$(4.6)' \qquad LR = \frac{[SSE(\beta_R) - SSE(\beta_{ML})]}{\sigma^2} \qquad \text{or}$$

$$(4.6)'' \qquad \boxed{LR = (N)\left(\ln\left(\frac{SSE_R}{SSE}\right)\right) = N[\ln SSE_R - \ln SSE]}$$

depending on whether σ^2 is known or not. Consider the relationship between (4.6) and the familiar Chow test used to test hypotheses of the form $q(\beta) = 0$ for linear regression models.

It can also be shown that for the special case of errors

$$\varepsilon \sim N[0,\sigma^2 I]$$

in the linear regression model, i.e.,

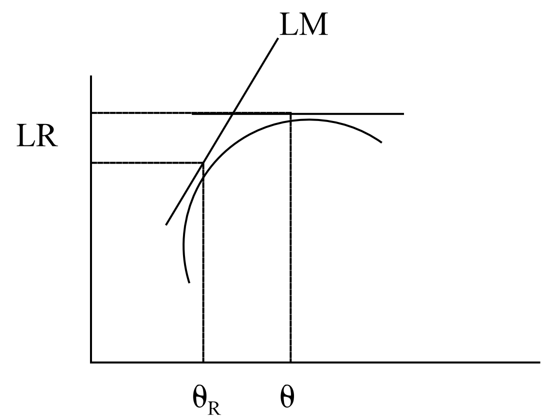
$$\begin{aligned} Y &= f(X;,\beta) + \varepsilon \\ &= X\beta + \varepsilon, \end{aligned}$$

the W, LM and LR tests for testing overall explanatory power H_0: $\beta_2 = \beta_3 = ... = \beta_K = 0$ can be expressed in terms of R^2 as

$$\boxed{\begin{aligned} W &= NR^2/(1-R^2) \\ LM &= NR^2 \\ LR &= -N\ln(1-R^2). \end{aligned}}$$

We also note in this case that W ≥ LR ≥ LM. A proof of this result uses the inequality X ≥ $\log_e(1 + X) \geq X/(1 + X)$.

The relationship between the LR, W and LM test can be easily visualized as follows:



The Wald, LR, and LM tests are appropriate for <u>nested</u> hypotheses, where one model includes another as a special case. Several approaches to testing <u>nonnested</u> hypotheses have been developed. One approach to testing nonnested hypotheses is to nest both hypotheses (models) as special cases of a more general model.

5. Some Applications

This section includes a brief discussion of a number of nonlinear econometric models including (a) the variable elasticity of substitution production function, (b) models for the distribution of income, (c) a precautionary note pertaining to computations associated with nonlinear models, (d) qualitative response models, (e) ordered or categorical data, (f) Poisson regression, (g) censored, (g) truncated or grouped regression models, and (h) quantitile regression.

a. Variable Elasticity of Substitution (VES) Production Function

The VES production function is defined by

(5.1) $$Y_t = A[(\rho - 1) K_t + L_t]^{\alpha\delta\rho} K_t^{\alpha(1-\delta\rho)} \varepsilon_t$$

$$= A[\rho-1 + L_t/K_t]^{\alpha\delta\rho} K_t^{\alpha} \varepsilon_t.$$

Taking the logarithm of both sides of (5.1) yields

(5.1)' $$\ln Y_t = \ln A + \alpha\delta\rho \ln[(\rho-1)K_t + L_t] + \alpha(1-\delta\rho)\ln K_t + \ln\varepsilon_t$$

$$= \ln A + \alpha \ln K_t + \alpha\delta\rho\ln[(\rho-1) + L_t/K_t] + \ln\varepsilon_t.$$

The elasticity of substitution of this production function is given by

(5.2) $$\sigma = \frac{\%\Delta(K/L)}{\%\Delta(\omega_L/\omega_K)}$$

$$= 1 + \left(\frac{\rho - 1}{1 - \delta\rho}\right) \frac{K}{L} .$$

The parameter α presents the returns to scale.

The VES production function includes the Cobb-Douglas production as a special case, i.e., if ρ=1 then (5.1) simplifies to

(5.3) $$Y_t = A L_t^{\alpha\delta} K_t^{\alpha-\alpha\delta}$$

$$= A L_t^{\alpha\delta} K_t^{\alpha(1-\delta)} \varepsilon_t$$

Note that the corresponding elasticity of substitution for the Cobb-Douglas production function is one. See (5.2) with ρ=1.

Estimation:

If the $\ln\varepsilon_t$ are identically and independently distributed as $N(0,\sigma^2)$, then nonlinear least squares estimators are the same as maximum likelihood estimators. The sum of squared errors is given by

$$(5.4) \quad SSE(A,\alpha,\delta,\rho) = \sum_{t=1}^{N} (\ln Y_t - \ln A - \alpha\delta\rho\ln[(\rho-1)K_t + L_t] - \alpha\,(1-\delta\rho)\ln K_t)^2$$

and the log likelihood function is given by (see equation (2.13))

$$(5.5) \quad \ell(A,\alpha,\delta,\rho,\sigma^2) = -SSE(\ )/2\sigma^2 - \frac{N}{2}(\ln\sigma^2 + \ln 2\pi).$$

The parameter σ^2 can be estimated by

$$(5.6) \quad \hat{\sigma}^2_{NLS} = SSE(\ )/N-4$$

$$\text{or} \quad \hat{\sigma}^2_{MLE} == SSE(\ )/N$$

where the least squares estimates of A, α, δ, ρ, are used in evaluating SSE. Fortunately econometric packages not only calculate the parameter estimates, but also estimate standard errors. The asymptotic variance covariance matrix of A_{NLS}, α_{NLS}, δ_{NLS}, ρ_{NLS} (same as maximum likelihood estimators) is given by

$$(5.7) \quad \sigma^2_{NLS}\left(\frac{\partial f}{\partial \beta}\ \frac{\partial f}{\partial \beta'}\right)^{-1}$$

$$\text{where } f(A,\alpha,\delta,\rho) = \begin{pmatrix} \ln A + \alpha\ln K_1 + \alpha\delta\rho\ln[(\rho-1) + L_1/K_1] \\ \vdots \\ \ln A + \alpha\ln K_N + \alpha\delta\rho\ln[(\rho-1) + L_N/K_N] \end{pmatrix}$$

(5.8)

$$\frac{\partial f}{\partial \beta} = \begin{pmatrix} \frac{1}{A} & \cdots & \frac{1}{A} \\ \delta\rho\ln[\rho-1 + L_1/K_1] + \ln K_1 & \cdots & \delta\rho\ln[\rho-1 + L_N/K_N] + \ln K_N \\ \alpha\rho\ln[\rho-1 + L_1/K_1] & \cdots & \alpha\rho\ln[\rho-1 + L_N/K_N] \\ \dfrac{\alpha\delta\rho}{\rho-1 + \frac{L_1}{K_1} + \alpha\delta\ln[\rho-1 + \frac{L_1}{K_1}]} & \cdots & \dfrac{\alpha\delta\rho}{\rho-1 + \frac{L_N}{K_N} + \ln[\rho-1 + \frac{L_N}{K_N}]} \end{pmatrix}$$

The columns correspond to observations and the rows correspond to derivatives of f( ) with respect to the parameters A, α, δ and ρ, respectively. The derivatives are evaluated at the non linear least squares estimates. The Cramer-Rao matrix could also be used. Several optimization algorithms use numerical approximations to (5.8). Numerical approximations of this matrix are standard output from many software packages.

Testing Hypotheses

In order to test hypotheses about individual parameters, the asymptotic standard errors in (5.7) can be used. For example, in order to test whether (5.1) is characterized by constant returns to scale, we test the hypothesis

(5.9) H_o: $\alpha = 1$.

The corresponding test statistic is given by

(5.10) $$\left(\frac{\alpha_{NLS} - 1}{\hat{\sigma}_{\alpha_{NLS}}}\right) \overset{a}{\sim} N(0,1)$$

where

$\hat{\sigma}_{NLS}$ = square root of the element in the 2^{nd} row and 2^{nd} column of (5.7).

In order to test whether (5.1) can be modeled by a Cobb-Douglas production function with constant returns to scale, we test the hypothesis

(5.11) H_o: $\rho=1$ and $\alpha=1$,

and the Wald test can be used, or the Lagrange or Likelihood ratio test could be applied.

Define

$$(5.12) \qquad q(\theta) = \begin{pmatrix} \rho - 1 \\ \alpha - 1 \end{pmatrix} .$$

The Wald test of $q(\beta) = 0$ is constructed as

$$(5.13) \qquad W = (q(\theta_{ML}))'(Var(q(\theta_{ML})))^{-1}\, q(\theta_{ML})$$

$$= (\rho_{ML}-1,\ \alpha_{ML}-1)\begin{bmatrix} \sigma_{\rho\rho} & \sigma_{\rho\alpha} \\ \sigma_{\alpha\rho} & \sigma_{\alpha\alpha} \end{bmatrix}^{-1}\begin{pmatrix} \rho_{ML}-1 \\ \alpha_{ML}-1 \end{pmatrix} \underset{\sim}{a} \chi^2(2).$$

Note that this test statistic only requires that the unconstrained model (5.1) be estimated.

In this case $Var(q(\theta_{ML}))$ is easily constructed from the simple form of the hypotheses being tested. More generally, note that

$$(5.14) \qquad Var(q(\theta_{ML})) = \left(\frac{\partial q}{\partial \theta'}\right)\Sigma_{\theta_{ML}}\left(\frac{\partial q}{\partial \theta}\right)$$

$$= \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}\begin{bmatrix} \sigma_{AA} & \sigma_{A\alpha} & \sigma_{A\delta} & \sigma_{A\rho} \\ \sigma_{\alpha A} & \sigma_{\alpha\alpha} & \sigma_{\alpha\delta} & \sigma_{\alpha\rho} \\ \sigma_{\delta A} & \sigma_{\delta\alpha} & \sigma_{\delta\delta} & \sigma_{\delta\rho} \\ \sigma_{\rho A} & \sigma_{\rho\alpha} & \sigma_{\rho\delta} & \sigma_{\rho\rho} \end{bmatrix}\begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{\rho\rho} & \sigma_{\rho\alpha} \\ \sigma_{\alpha\rho} & \sigma_{\alpha\alpha} \end{bmatrix}$$

where

$$(5.15) \qquad \frac{\partial q}{\partial \theta'} = \begin{bmatrix} \frac{\partial(\rho-1)}{\partial A} & \frac{\partial(\rho-1)}{\partial \alpha} & \frac{\partial(\rho-1)}{\partial \delta} & \frac{\partial(\rho-1)}{\partial \rho} \\ \frac{\partial(\alpha-1)}{\partial A} & \frac{\partial(\alpha-1)}{\partial \alpha} & \frac{\partial(\alpha-1)}{\partial \delta} & \frac{\partial(\alpha-1)}{\partial \rho} \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} .$$

The Lagrangian and Likelihood Ratio tests require that the constrained model be estimated. The constrained model associated with (5.11) is given by

$$(5.16) \qquad Y_t = AL_t^{\delta} K_t^{1-\delta} \varepsilon_t$$

or

(5.16)' $\ln Y_t = \ln A + \delta \ln L_t + (1-\delta)\ln K_t + \ln \varepsilon_t$

or

(5.16)" $\ln(Y_t/K_t) = \ln A + \delta \ln(L_t/K_t) + \ln \varepsilon_t$.

Equation (4.5)" can be used to evaluate <u>LM</u>.

$$(5.17) \qquad \frac{\partial f}{\partial \beta} = \begin{pmatrix} \frac{1}{A} & \cdots & \frac{1}{A} \\ \ln\left(\frac{L_1}{K_1}\right) & \cdots & \ln\left(\frac{L_N}{K_N}\right) \end{pmatrix}$$

$$(5.18) \qquad \frac{\partial f}{\partial \beta}\frac{\partial f}{\partial \beta'} = \begin{pmatrix} \frac{N}{A} & \sum_t \ln(L_t/K_t)/A \\ \sum_t \ln(L_t/K_t)/A & \sum_t \ln^2(L_N/K_N) \end{pmatrix}$$

(5.19) $(Y-f(x;\beta_R))' = (\ln Y_1 - \ln A - \delta \ln L_1 - (1-\delta)\ln K_1, \ldots \ldots, \ln Y_N - \ln A - \delta \ln L_N - (1-\delta)\ln K_N)$.

$SSE(\beta_R)$ is the sum of squared errors associated with the constrained model, (5.16)'. Substituting these expressions into (4.5)" yields the LM test statistic which has a $\chi^2(2)$ asymptotic distribution.

(4.6) and (4.6)" are easily implemented forms for the <u>likelihood ratio</u> test statistic. This form only requires that the sum of squared errors for the constrained model (5.16)" and the sum of squared errors for the unconstrained model (5.1)' be obtained. This test statistic is also asymptotically distributed as $\chi^2(2)$. If normality is not assumed, then equation (4.6) can be used.

b. Models for the Distribution of Income

Many of the models for the distribution of income are special cases of the *generalized beta of the second kind* (GB2) and *generalized beta of the first kind* (GB1).

$$\text{(5.20a)} \quad GB2(y;a,b,p,q) = \frac{|a|y^{(ap-1)}}{b^{ap}B(p,q)(1 + (y/b)^a)^{p+q}}$$

$$\text{(5.20b)} \quad GB1(y;a,b,p,q) = \frac{|a|y^{(ap-1)}(1 - (y/b)^a)^{q-1}}{b^{ap}B(p,q)}$$

These include the *generalized gamma* (GG)

$$\text{(5.21)} \quad GG(y;a,\beta,p) = \frac{|a|y^{ap-1}e^{-(y/\beta)^a}}{\beta^{ap}\Gamma(p)}$$

which corresponds to the limit of (5.20a-b) as $q \to \infty$, $b = \beta\, q^{1/a}$;

the *gamma* (G)

$$\text{(5.22)} \quad G(y;\beta,p) = GG(y;a,1,\beta,p) = \frac{y^{p-1}e^{-y/\beta}}{\beta^p\Gamma(p)};$$

the *Singh Maddala* (SM)

$$\text{(5.23)} \quad SM(y;a,b,q) = GB2(y;a,b,p=1,q) = \frac{|a|y^{a-1}}{b^{a}B(1,q)(1+(y/b)^a)^{q+1}};$$

the Fisk (F)

$$\text{(5.24)} \quad FISK(y;a,b) = GB2(y;a,b,p=q=1) = \frac{|a|y^{a-1}}{b^{a}(1+(y/b)^a)^2};$$

the Weibull (W)

$$\text{(5.25)} \quad W(y;a,\beta) = GG(y;a,\beta,p=1) = \frac{|a|y^{a-1}e^{-(y/\beta)^a}}{\beta^a}$$

and the Lognormal (LN)

$$(5.26)\quad LN(y;\mu,\sigma^2) = \frac{e^{\frac{-1}{2\sigma^2}(\ln y - \mu)^2}}{y\sqrt{2\pi}\,\sqrt{\sigma^2}}$$

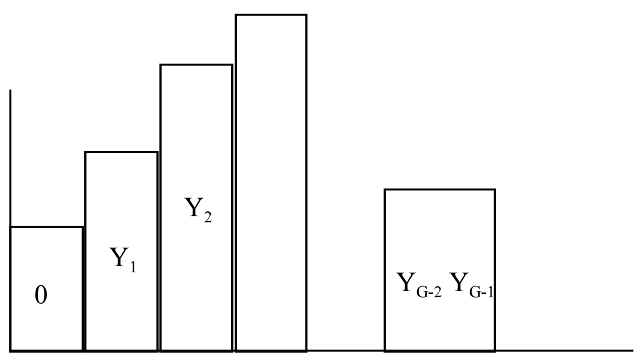
which is the limit of (5.21) as $a \to 0$, $b^a = \sigma^2 a^2$, $p = (a\mu + 1)/b^a$.

These distributions can be fit to individual observations using maximum likelihood procedures in the usual manner by maximizing

$$\ell = \sum_{t=1}^{n} \ln f(y_t;\theta)$$

with respect to the underlying parameters θ, where $f(y;\theta)$ denotes the assumed distribution.

If the data are in a grouped format, with G groups, n_i observations will be reported in the i^{th} group, (Y_{i-1}, Y_i), for i = 1, 2, ... G and alternative procedures must be employed.



The corresponding histogram might appear as above where $0 < Y_1 < Y_2 < \ldots < Y_{G-1} < Y_G = \infty$ partitions all possible incomes into g groups. Further, let $F(Y;\theta)$ denote the cumulative distribution function

$$F(Y;\theta) = \int_{-\infty}^{Y} f(y;\theta)dy.$$

The observed relative frequencies in the i^{th} interval are given by n_i/n where $(n = \sum_{i=1}^{G} n_i)$

and the predicted probabilities for the i^{th} interval is given by

$$p_i(\theta) = \int_{Y_{i-1}}^{Y_i} f(y;\theta)dy = F(Y_i;\theta) - F(Y_{i-1};\theta).$$

$p_i(\theta)$ is a fu

Estimators of the parameters θ can be obtained for grouped data by solving any of the following problems:

$$\min_{\theta} \left\{ SSE = \sum_{i=1}^{G} \left(\frac{n_i}{n} - p_i(\theta) \right)^2 \right\}$$

$$\min_{\theta} \left\{ SAE = \sum_{i=1}^{G} \left| \frac{n_i}{n} - p_i(\theta) \right| \right\}$$

$$\min_{\theta} \left\{ \chi^2 = \sum_{i=1}^{G} \left(\frac{n_i}{n} - p_i(\theta) \right)^2 / p_i(\theta) \right\}$$

$$\max_{\theta} \left\{ \ell = n! \prod_{i=1}^{G} (p_i(\theta))^{n_i} / n_i! \right\}.$$

The first two estimators might be thought of as least squares and least absolute deviations estimators. The third is a minimum chi-square estimator and the fourth is a maximum likelihood estimator. The third and fourth estimators are asymptotically efficient.

An example of fitting statistical distributions to the following grouped data from Census Population Reports for 1980 family income by maximizing the multinomial loglikelihood function is reported in table 1. The corresponding 1980 data are given by

(thousands)	(%)
2.5	2.1
5.0	4.1
7.5	6.2
10.0	6.5
12.5	7.3
15.0	6.9
20.0	14.0
25.0	13.7
35.0	19.8
50.0	12.8
∞	6.7
ENDPOINTS	(n = 40,000)

TABLE 1. ESTIMATED DISTRIBUTION FUNCTIONS
1980 FAMILY INCOME

	GB1	GB2	B1	GG	B2	SM	LN	G	W	F	E
a	1.4008	2.5373	1.0000	1.4008	1.0000	1.6971	μ=2.9372	1.0000	1.6057	2.2768	1.0000
b(β)	273102.	40.7667	163.757	(21.8145)	3535660.	87.6981	σ=0.7797	(11.0473)	(26.3368)	19.7450	(24.5944)
p	1.2454	0.6117	1.9173	1.2454	2.1555	1.0000		2.1557	1.0000	1.0000	1.0000
q	549517.	2.1329	11.3828		320081.	8.3679				1.0000	
Mean*	23.644	23.931	23.604	23.646	23.810	23.730	25.564	23.815	23.065	27.749	24.59
Gini**	.353	.359	.353	.353	.363	.355	.419	.363	.351	.439	.50
SAE	.0004	0.0002	0.0006	0.0004	0.0008	0.0003	0.0070	0.0008	0.0005	0.0053	0.02
SSE	.0545	0.0385	0.609	0.0545	0.0775	0.0495	0.2326	0.0775	0.0561	0.2038	0.41
χ^2	143.0	84.3	188.4	143.0	353.5	114.2	4620.7	353.6	183.6	2438.9	9107
$-\ell$	129.1	99.1	151.6	129.1	228.3	114.5	1859.4	228.3	150.0	1219.5	5108

*Census Estimate: 23.974
**Census Estimate: .365

Exercises:

(1) Compare the relative fits of the various distributions. Which distributions provide the best fits for 2, 3 and 4 parameter models.

(2) Indicate how a statistical test can be performed to check whether the SM model is consistent with the generalized beta of the second kind, H_0: GB2 = SM. Hint: consider the likelihood ratio test. Wald and LM tests could also be used, but would require more information than is reported in the table. A comparison of the sum of absolute errors (SAE), sum of squared errors, χ^2 value and log likelihood estimates reveals considerable difference in "goodness of fit" between the various models.

c. A Precautionary Note: Nonlinear Optimization

Numerous estimation problems can arise in nonlinear estimation which may yield questionable results. An indirect check of the validity of the parameter estimates obtained from a nonlinear optimization routine is provided by comparing estimated population characteristics such as the mean with independently obtained results where available. The following example provides an example of the importance of this.

Thurow's widely cited [AER, 1970] paper provides an example of estimation problems. The underlying distribution of income was assumed to be modeled by a beta density function of the first kind (B1).

$$B1(y;b,p,q) = \frac{y^{p-1}(b-y)^{q-1}}{B(p,q)b^{p+q}} \qquad 0 < y < b, \quad p,q > 0$$

Thurow assumed that the maximum income (b) was equal to \$15,000 and obtained separate estimates of p and q for the distribution of income (1959 dollars) of families and unrelated individuals for whites and nonwhites for the period 1949-1966. Income characteristics associated with the estimated parameter values for (p,q,b) are inferred and their relationship with hypothesized explanatory variables considered. Thurow's results raise questions as to whether economic growth is associated with a more egalitarian distribution as well as suggesting that inflation may lead to a more equal distribution of income for whites. The accuracy of the estimated (p,q)'s is a critical element in the validity of the analysis of the estimated relationship between the hypothesized explanatory variables and the distribution of income. Thurow's estimates of (p,q) were not reported in his paper, but were provided on request and are given in Table 2. The mean and Gini coefficient associated with the beta function are given by

$$E(Y) = \frac{bp}{p+q}.$$

$$G = \frac{\Gamma(p+q)\Gamma(p+1/2)\Gamma(q+1/2)}{\Gamma(p+q+1/2)\Gamma(p+1)\Gamma(q)\Gamma(1/2)},$$

(McDonald [1984]). The mean income level and Gini coefficients implied by Thurow's estimates can be readily obtained by substituting parameter estimates into these equations. . These corresponding estimates are reported in Table 2. Independent estimates obtained in various census publications are also given in this table to provide a useful comparison.

An analysis of the entries in this table suggests that the distribution of income for whites is more egalitarian with a higher mean than for nonwhites. This qualitative result is consistent with Thurow's estimates; however, other implications of these two sets of estimates are not. For example, all of the associated estimated density functions are either "∪" shaped ($p < 1$, $q < 1$) or "L" shaped ($p < 1$, $q > 1$) rather than "∩" shaped. The magnitude and intertemporal behavior (reductions in excess of 30 percent) of the associated Gini coefficients implied by the estimated parameters (p,q) for the period under consideration are inconsistent with the census estimates and provide additional evidence of an estimation problem. The agreement between the implied and census estimates of the mean is much closer than for the Gini coefficients.

Thus there is relatively close agreement between the two estimates of mean income, but very poor agreement between the measures of inequality. The estimation procedure appears to have roughly preserved the mean characteristic, but implicitly modeled intra and/or intergroup variation incorrectly. The results could also have been partially due to the conjunction of the nature of the income groups and treatment of the maximum income.

The previous section included examples in which estimated distributions (B1) of the form considered by Thurow provide relatively accurate estimates of population characteristics.

THURÓW'S ESTIMATES OF p,q[a]

	Whites				Nonwhites			
Year	p	q	mean[b]	Gini[b]	p	q	mean[b]	Gini[b]

1949	.258	.666	4.188	(4.052)	.615	(.404)	.160	.930	2.202	(2.044)	.752	(.443)
1950	.279	.687	4.332	(4.294)	.600	(.407)	.172	.930	2.341	(2.245)	.738	(.438)
1951	.269	.625	4.513	(4.441)	.596	(.387)	.182	.908	2.505	(2.334)	.725	(.433)
1952	.298	.649	4.720	(4.579)	.576	(.398)	.205	.921	2.731	(2.518)	.702	(.407)
1953	.327	.667	4.935	(4.798)	.556	(.395)	.228	.920	2.979	(2.638)	.679	(.428)
1954	.334	.697	4.859	(4.741)	.557	(.401)	.217	.929	2.840	(2.579)	.691	(.456)
1955	.368	.718	5.083	(5.034)	.536	(.397)	.225	.913	2.966	(2.679)	.681	(.431)
1956	.411	.731	5.398	(5.311)	.511	(.391)	.249	.978	3.044	(2.875)	.666	(.427)
1957	.406	.728	5.370	(5.229)	.513	(.385)	.269	1.025	3.118	(2.860)	.652	(.435)
1958	.411	.750	5.310	(5.291)	.514	(.388)	.276	1.075	3.064	(2.893)	.651	(.448)
1959	.460	.765	5.633	(5.571)	.488	(.396)	.286	1.051	3.209	(2.977)	.641	(.452)
1960	.504	.815	5.732	(5.646)	.473	(.398)	.330	1.061	3.559	(3.276)	.608	(.459)
1961	.622	.979	5.828	(5.817)	.443	(.408)	.346	1.173	3.417	(3.268)	.607	(.462)
1962	.663	.971	6.086	(5.987)	.426	(.395)	.338	1.107	3.509	(3.278)	.607	(.443)
1963	.712	.985	6.294	(6.167)	.411	(.396)	.356	1.073	3.737	(3.513)	.591	(.440)
1964	.785	1.017	6.534	(6.333)	.391	(.400)	.406	1.095	4.057	(3.788)	.562	(.444)
1965	.842	1.029	6.750	(6.552)	.376	(.393)	.452	1.124	4.302	(3.859)	.538	(.427)
1966	.955	1.044	7.166	(6.912)	.348	(.390)	.514	1.104	4.765	(4.192)	.504	(.426)

[a]Thurow did not estimate the parameter b, but rather assumed it to be 15 ($15,000) and included any higher incomes in the group with an upper bound of $15,000.

[b]The mean and Gini coefficients were evaluated using the given equations. The numbers in parentheses are the corresponding census estimates reported in current populations reports (P60). The nominal figures for mean income were adjusted by the CPI to obtain the figures in 1959 dollars.

d. Models with binary dependent variables or limited dependent variables

(1) Introduction

Consider models in which one might want to explain:

(a) when there will be a default on a loan ($Y = 1$) or no default ($Y = 0$);

(b) whether a tax return has been filed by someone who has misrepresented their financial position ($Y = 1$) or accurately reflects their financial situation ($Y = 0$);

(c) The market share of a firm ($0 \leq Y \leq 1$)

These are known as limited dependent variable problems. Amemiya (1981) has an excellent survey paper on these models in the Journal of Economic Literature.

In each case the dependent variable (Y) in the function

$$Y = f(X;\beta) + \varepsilon$$

is constrained in value.

Numerous approaches have been adopted to solve this problem and these include regression analysis, linear probability models, discriminant analysis and limited dependent models.

(2) Linear Probability Model (LPM)

Let $Y_t = X_t\beta + \varepsilon_t$

where

$Y_t = 1$ if first option chosen
$\quad = 0$

X_t vector of values of attributes (independent variables)

β corresponding vector of unknown coefficients

ε_t independently distributed random variable with zero mean

The specification of the model implies:

$$E(Y_t) = X_t\beta.$$

Now let $P_t = Prob(Y_t = 1)$

$$Q_t = 1 - P_t = Prob(Y_t = 0)$$

So that

$$E(Y_t) = 1 \bullet Prob(Y_t = 1) + 0\ Prob(Y_t = 0)$$

$$= 1 \bullet P_t + 0 \bullet Q_t$$

$$= P_t .= X_t\beta.$$

Thus the regression equation can be interpreted as describing, given the vector X_t, the probability that the first choice is made. The vector β measures the effect on the probability of choosing the first alternative corresponding to unit changes in explanatory variables, $\beta_i = \frac{\partial Y_t}{\partial X_{ti}}$. If we estimate the equation using OLS, we may obtain estimates of β. However, there is some question about the applicability of OLS in this model. To explore this issue, note the following:

$$\varepsilon_t = Y_t - X_t\beta$$

$$E(\varepsilon_t) = P_t(1 - X_t\beta) + (1 - P_t)(-X_t\beta) = P_t - X_t\beta$$

$$E(\varepsilon_t) = 0 \text{ implies } P_t = X_t\beta \text{ and } (1 - P_t) = 1 - X_t\beta.$$

To find the variance of the error term ε_t, we evaluate

$$Var(\varepsilon_t) = E(\varepsilon_t^2) = (1 - X_t\beta)^2 P_t + (-X_t\beta)^2(1 - P_t)$$

$$= (1 - X_t\beta)^2(X_t\beta) + (X_t\beta)^2(1 - X_t\beta)$$

$$= (1 - X_t\beta)(X_t\beta)$$

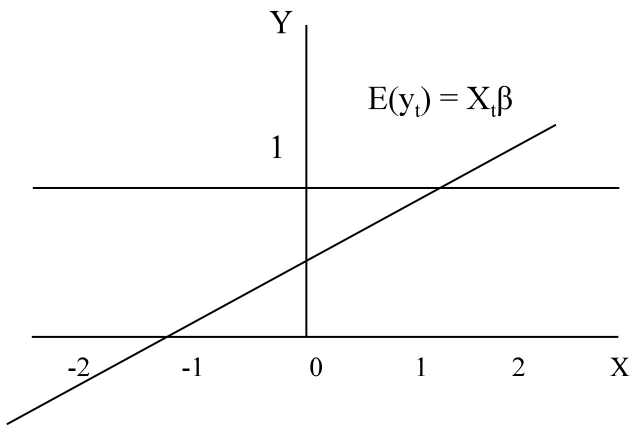
$$= P_t(1 - P_t)$$

$$= E(Y_t)(1 - E(Y_t)) = (X_t\beta)(1 - X_t\beta),$$

which clearly shows the error is heteroskedastic. One possible solution is to use weighted least squares.

Another problem with the LPM model is that of prediction:

Note that with the linear probability model there is a chance that predicted values for Y_t may lie outside the interval [0, 1].

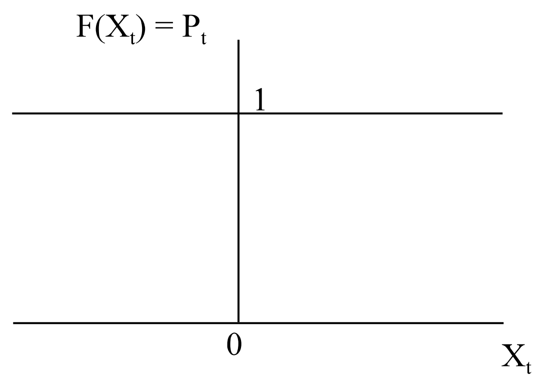


One possible solution is to set all predictions greater than 1 equal to 1 and all predictions less than 0 equal to zero. However, this approach presents a problem in running weighted least squares.

(3) Limited Dependent Variable Models

(a) Introduction

Multiple discriminant analysis (MDA) (Multivariate Statistics classes) is still another approach to problems of this type. MDA is closely related to the LPM model. Another possibility for binary or limited dependent variables is to use constrained estimation. Since Y_t or a transformation is constrained to the interval (0,1), functional forms $F(x_t)$ which are constrained to the interval (0,1) can be selected. This quite naturally suggests the use of cumulative distributions for $F(X_t)$.



This possibility admits many alternative models:

$$P_t = \Pr(Y_t = 1 | X_t) = F(X_t\beta;\theta) = \int_{-\infty}^{X_t\beta} f(s;\theta)ds$$

where $f(s;\theta)$ denotes a "well behaved" density function with distributional parameters θ. $X_t\beta$ are often referred to as the "scores." $F(x_t\beta;\theta)$ is then in the form of a cumulative distribution function. Two models which have been widely used are the standard normal and logistic models:

	$f(s;\theta)$	$F(z) = \int_{-\infty}^{z} f(s;\theta)ds$	
Normal	$\frac{e^{-s^2/2}}{\sqrt{2\pi}}$	$\int_{-\infty}^{z} \frac{e^{-s^2/2}}{\sqrt{2\pi}}$	Probit
Logistic	$\frac{e^{-s}}{(1 + e^{-s})^2}$	$\frac{1}{1 + e^{-z}}$	Logit

These two distributions are only two of many which could have been used; however, this literature has been dominated by the use of the probit (based on the normal) and logit (based on the log-logistic) models.

(b) Interpretation of Coefficients

Using Leibnitz's Rule (also the chain rule in this case)

$$\frac{\partial P_t}{\partial X_{ti}} = \frac{\partial F(X_t\beta)}{\partial X_{ti}} = \frac{\partial F(X_t\beta)}{\partial(X_t\beta)}\frac{\partial(X_t\beta)}{\partial X_{ti}} = f(X_t\beta)\beta_i$$

Thus the β_i 's, alone, do not yield the change in P_t corresponding to a change in X_{ti}, it is necessary to estimate $f(X_t\beta)$,the pdf evaluated at the score. However, the relative values of the beta's estimate the relative impact of changes in the exogenous variables on P_t.

(c) Estimation

The estimation of limited dependent models depends upon the model or density selected and the nature of the data: (i) $Y_t = 0$ or 1 or (ii) $0 < Y_t < 1$.

(i) $Y_t = 0$ or 1.

If we have a <u>micro data based on discrete choices</u>, then Y = 0 or 1.

The likelihood function in this case is given by

$$L(\beta;\theta:\ Y_t) = \prod_{t=1}^{n} P_t^{Y_t}(1 - P_t)^{1-Y_t}$$

$$= \prod_{t=1}^{n} F(X_t\beta;\theta)^{Y_t}(1 - F(X_t\beta;\theta))^{1-Y_t}$$

and the log likelihood function is

$$\boxed{\ell(\beta,\theta) = \sum_{t=1}^{n}\left\{Y_t \ln F(X_t\beta;\theta) + (1 - Y_t)\ln(1 - F(X_t\beta;\theta)\right\}.}$$

which is optimized over the parameters β and θ to obtain maximum likelihood estimators. This procedure can be quite involved if the expression for the cumulative distribution is complicated. Recall that

$$F(X_t\beta;\theta) = \Pr(s \le X_t\beta) = \int_{-\infty}^{X_t\beta} f(s;\theta)ds$$

where θ denotes unknown distributional parameters.

The STATA commands for fitting the probit and logit models are as follows:

```
probit y x's
logit y x's
```

Using the command **dprobit** reports the marginal effects. **dlogit** is not functional in version 10 of STATA; however, the ***margins*** command available in Stata 11 and 12 facilitates the evaluation of the marginal impact of changes in the independent variables on the dependent variable in a many econometric models: ***margins, dydx(*) atmeans*** or ***margins, dydx(selected x's) atmeans***

The command
estat class, cutoff(#) reports the prediction matrix corresponding to the selected threshold value.

Qualitative response models and heteroskedasticity

There is a literature on solutions to treating heteroskedasticity in qualitative response models (Green 6th ed. , 788-790). The STATA command for one model of heteroskedasticity in the probit model is
hetprob y x's, het (all or selected X;s).
This command uses the normal cdf where the σ parameter is expressed as a function of the explanatory variables $\sigma = e^{x\delta}$. If tl δ vector = 0, then $\sigma = 1$ which is the regular probit model. Thus, the heteroskedasticity specification includes the regular probit model as a special case. The likelihood ratio test statistic can be used to check for statistically significant differences.

Qualitative response models and specification error

If the true qualitative response model corresponds to a pdf other than the normal or loglogistics, a nonlinear index (rather than $(X_t\beta)$), or involves heteroskedasticity, then the probit/logit estimators will be inconsistent.

Alternatives to the normal or log-logistics include using more general pdf's such as the EGB2, SGT, or IHS. Klein-Spadey (1993, An efficient semi-parametric estimator for binary response models, Econometrica, 387-421) propose the use of a kernel estimator in the log-likelihood function. The Klein-Spadey methodology can be implemented in Stata by downloading and installing a subroutine with the command, ***findit ST0144***.

Qualitative response models and endogenous regressors

If the right hand side variables (X) includes any endogenous regressors, you will probably want to consider using the "ivprobit" estimator. The form for this command is

ivprobit depvar X1 (Y1=X2 or Y1=X1 X2) where X2 denotes the instruments

(ii) Bounded data $0 < Y_t < 1$.

If we have a discrete choice model with grouped data or a model with the dependent variable strictly between 0 and 1, alternative estimation techniques are available. For the discrete choice problem, the MLE approach just considered can be adopted.

One approach is to use

$$\hat{p}_t = \frac{v_t}{m_t}$$

v_t = number choosing the first response in the t^{th} group

m_t = number in the t^{th} group

$$F^{-1}(\hat{P}_t) = X_t\beta \text{ or } F^{-1}(Y_t) = X_t\beta \text{ .}$$

If F is known, then regression techniques can be employed to estimate the vector β.

Recall that the probit model is based upon the normal cumulative distribution function and

$$F(X_t\beta) = \int_{-\infty}^{X_t\beta} \frac{s^{-s^2/2}ds}{\sqrt{2\pi}}.$$

The logit model is based upon the logistic distribution function

$$F(X_t\beta) = \frac{1}{1 + e^{-X_t\beta - \varepsilon_t}}.$$

The probit model involves rather complicated estimation and there is no compelling reason that the normal should be used. The logit approximates the probit model, but has thicker tails. The logit model is particularly well suited for grouped data or other situations in which $0 < Y_t = F(x_t B) < 1$.

Note that

$$F^{-1}(Y_t) = \ln\left(\frac{Y_t}{1 - Y_t}\right) = X_t\beta + \varepsilon_t$$

and regression techniques can be used. Further, note that $Y_t \neq 0$ or 1 in this representation.

e. Ordered or categorical data for Y.

For situations in which more than two ordered outcomes are possible, the user may want to consider the ordered logit or the ordered probit model using the STATA commands

ologit y x's

oprobit y x's

For multiple categorical dependent variables, not necessarily ordered, the multinomial logit model is a possibility

mlogit y x’s

f. Poisson regression

For models in which the dependent variable represents count data, the Poisson regression model may be appropriate. The basic form for the Poisson regression model is as follows:

$$\Pr\left[Y = y_i \middle| X_i\right] = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \quad \lambda_i = e^{X_i\beta},\ y_i = 0,1,2,\ldots\ ,$$

$$E\left(Y_i \middle| X_i\right) = \lambda_i\beta \qquad .$$

The MLE of the unknown parameters are obtained using the STATA command: **poisson y x's** The Poisson model is characterized by the mean being equal to the variance. The negative binomial model relaxes the assumption of the mean and variance being equal.

g. Censored and Grouped or interval Regression Models

(1) Censored Regression

Consider the model

$$Y_t^* = X_t\beta + \varepsilon_t$$

where

$$Y_t = 0 \text{ if } Y_t^* \leq 0$$

$$Y_t = X_t\beta + \varepsilon_t \quad Y_t^* > 0.$$

This model has been used to describe the purchase of consumer durables, hours worked by women in the labor force, arrests after prison, etc.

Note $Y_t = 0$ if $X_t\beta + \varepsilon_t \leq 0$ or $\varepsilon_t \leq -X_t\beta$ and $Y_t > 0$ if $\varepsilon_t > -X_t\beta$. In the <u>censored regression model</u> observations are available for (X_t, Y_t) for $Y_t = 0$ **and** $Y_t > 0$. In <u>truncated regression</u> models observations are only available for $Y_t > 0$.

It should be noted that

$$E(Y_t \mid X_t) = X_t\beta + E(\varepsilon_t \mid \varepsilon_t < -X_t\beta) \neq X_t\beta.$$

This result suggests one approach to estimation based on estimating the last term involving the conditional expected value of the error term. This approach has been discussed by Heckman and involves what is called the Mill's ratio to represent $E(\varepsilon_t|\varepsilon_t<-X_t\beta)$.

Rather than using the Heckman correction, we consider regular maximum likelihood estimation of the censored regression model. If $F(\varepsilon;\theta)$ denotes the cdf for the random disturbance, then

$$\begin{aligned} \Pr(Y_t = 0 \mid X_t) &= \Pr(\varepsilon_t \leq -X_t\beta) \\ &= F(-X_t\beta;\theta). \end{aligned}$$

Let

$$d_i = \begin{cases} 1 & Y_t > 0 \\ 0 & Y_t = 0 \end{cases}$$

The log likelihood function for a random sample is given by

$$\boxed{\ell = \sum_t \{d_t \ln f(Y_t - X_t\beta;\theta) + (1 - d_t)\ln F(-X_t\beta;\theta)\}}$$

ℓ is maximized over β and θ. In cases where the density of the random disturbance is assumed to be symmetric about zero, the likelihood function can be rewritten as

$$\boxed{\ell = \sum_t \{d_t \ln f(Y_t - X_t\beta;\theta) + (1 - d_t)\ln(1 - F(X_t\beta;\theta))\}}$$

If f( ) is assumed to be the normal, then the corresponding censored regression model is known as the Tobit model.

The STATA command for the Tobit model is

tobit y x’s

If the Tobit models is estimated, but the true residuals are not normal, then the estimators may be **inconsistent**. Ignoring the observations for Y=0 and just running least squares on the observations (Y>0) yields inconsistent estimators. Flexible parametric and kernel based formulations have been considered for the censored regression model. Nonlinearities in the index as well as the presence of heteroskedasticity can lead to inconsistent estimates.

Amemiya, T. (1979) considers estimation of simultaneous equation models involving Tobit-like variables and Smith and Blundell (1986) outline an exogeniety test for simultaneous equation Tobit models. For Tobit models with endogenous regressors, the reader may want to consider using the **ivtobit** in STATA.

Powell (1984, 1986) introduces the censored least absolute deviations (CLAD) and symmetrically trimmed least squares (SCLS) estimators of the censored regression model which are robust to heteroskedasticity and distributional assumptions, except for assuming the error distribution is symmetric. Newey (1984) outlines how the SCLS and Tobit estimators can be combined, using a Hausman test, to check the homoskedasticity and normality assumptions. Cosslett (2004) outlines some semi-parametric approaches to estimating censored regression models, including the use of the Kaplan-Meier estimator of the cdf , to circumvent possible problems with the Tobit model. Lewbell and Linton (2002) outline other semi-parametric estimation methods for censored regression models.

Partially adaptive estimation can be modified to model different forms of heteroskedasticity and make adjustments for non-normal heteroskedastic errors. The following table summarizes the results of a simulation study with various estimators with normal, thick tailed, and skewed error terms along with two types of heteroskedasticity. Note that heteroskedasticity as well as non-normality results in Tobit estimators performing poorly.

Table 2 – Simulation Results for 25% Censoring with Homoskedastic Errors

25% Censoring y = a + b*x Slope: b=1	Normal			Mixed Normal			ST		
	bias	std. dev.	RMSE	bias	std. dev.	RMSE	bias	std. dev	RMSE
OLS	-0.238	0.102	0.258	-0.194	0.096	0.216	-0.209	0.070	0.220
SCLS	0.036	0.199	0.202	0.014	0.103	0.104	0.101	0.136	0.170
CLAD	0.025	0.189	0.191	0.008	0.082	0.082	0.016	0.141	0.141
Tobit	**-0.002**	**0.128**	**0.128**	0.175	0.139	0.223	0.050	0.086	0.100
GED - hom	**-0.001**	**0.131**	**0.131**	**-0.003**	**0.077**	**0.077**	0.014	0.091	0.093
SGED - hom	**-0.001**	**0.134**	**0.134**	**-0.002**	**0.078**	**0.078**	**0.006**	**0.078**	**0.078**
IHS - hom	**-0.006**	**0.129**	**0.129**	**-0.002**	**0.059**	**0.059**	**0.002**	**0.073**	**0.073**
Tobit-het	**0.001**	**0.132**	**0.132**	0.184	0.120	0.220	0.090	0.080	0.120
GED -het	**0.002**	**0.135**	**0.135**	**-0.003**	**0.078**	**0.078**	0.047	0.093	0.104
SGED - het	**0.005**	**0.140**	**0.140**	**-0.001**	**0.079**	**0.079**	**0.007**	**0.080**	**0.080**
IHS - het	**-0.007**	**0.133**	**0.133**	**0.000**	**0.060**	**0.060**	**-0.003**	**0.078**	**0.078**

N = 200. T = 1000 Simulations

Table 3 – Simulation Results for 25% censoring with Heteroskedastic Errors (Type I)

25% Censoring y = a + b*x Slope: b=1	Normal			Mixed Normal			ST		
	bias	std. dev	RMSE	bias	std. dev	RMSE	bias	std. dev.	RMSE
OLS	0.173	0.243	0.298	0.043	0.213	0.218	-0.050	0.164	0.171
SCLS	0.001	0.353	0.353	0.039	0.213	0.217	-0.061	0.289	0.295
CLAD	-0.011	0.324	0.324	0.022	0.168	0.169	-0.119	0.221	0.251
Tobit	0.277	0.303	0.411	0.463	0.269	0.535	0.221	0.220	0.312
GED - hom	0.060	0.318	0.323	-0.012	0.164	0.164	-0.105	0.225	0.249
SGED - hom	0.281	0.404	0.492	-0.008	0.167	0.167	-0.163	0.256	0.303
IHS - hom	0.109	0.349	0.365	0.025	0.135	0.137	-0.152	0.242	0.286
Tobit-het	**-0.005**	**0.260**	**0.260**	0.260	0.216	0.338	0.014	0.173	0.173
GED -het	**0.000**	**0.264**	**0.264**	-0.016	0.161	0.161	-0.080	0.181	0.198
SGED - het	**0.001**	**0.275**	**0.275**	-0.023	0.159	0.161	**-0.004**	**0.164**	**0.164**
IHS - het	**-0.003**	**0.258**	**0.258**	**0.004**	**0.115**	**0.115**	**-0.008**	**0.159**	**0.160**

N = 200, T = 1000 Simulations

Table 4 – Simulation Results for 25% censoring with Heteroskedastic Errors(Type 2)

25% Censoring y = a + b*x Slope: b=1	Normal			Mixed Normal			ST		
	bias	std. dev.	RMSE	bias	std. dev.	RMSE	bias	std. dev.	RMSE
OLS	0.219	0.264	0.343	0.074	0.232	0.243	-0.029	0.177	0.180
SCLS	-0.007	0.358	0.358	0.038	0.215	0.218	-0.081	0.291	0.303
CLAD	-0.002	0.343	0.343	0.020	0.165	0.166	-0.148	0.225	0.270
Tobit	0.289	0.329	0.438	0.464	0.271	0.538	0.205	0.236	0.312
GED - hom	0.071	0.352	0.359	-0.020	0.175	0.176	-0.141	0.225	0.265
SGED - hom	0.253	0.432	0.501	-0.014	0.181	0.182	-0.194	0.277	0.338
IHS - hom	0.110	0.381	0.396	0.024	0.137	0.139	-0.200	0.262	0.329
Tobit-het	**0.000**	**0.289**	**0.289**	0.265	0.224	0.347	-0.008	0.186	0.186
GED -het	**0.005**	**0.293**	**0.293**	-0.025	0.171	0.173	-0.106	0.190	0.218
SGED - het	**0.007**	**0.310**	**0.310**	-0.032	0.171	0.174	-0.025	0.183	0.185
IHS - het	**0.000**	**0.286**	**0.286**	**0.011**	**0.119**	**0.120**	**-0.027**	**0.172**	**0.174**

N = 200, T = 1000 Simulations

(2) Grouped Regression

A regression model where Y is reported as being in different groups, e.g., in income intervals, (a_i, b_i) . The corresponding log-likelihood function is given by

$$\sum_i \ln\left(F\left(b_i - X_i\beta;\theta\right) - F\left(a_i - X_i\beta;\theta\right)\right)$$

The stata command

intreg depvar1 depvar2 x's

fits a model of y=[depvar1, depvar2] on independent variables, where y for each observation is point data, interval data, left-censored data, or right-censored data.

depvar1 and depvar2 should have the following form:

type of data		depvar1	depvar2
point data	a = [a,a]	a	a
interval data	[a,b]	a	b
left-censored data	(-inf,b]	.	b
right-censored data	[a,inf)	a	.

Additional options available with the intreg command include het( ) and vce, among others.

The "intreg" command assumes homoskedasticity and normally distributed error terms. These estimators will be inconsistent if the errors are not normally distributed or are characterized by heteroskedasticity. Semi-parametric or partially adaptive estimators may be useful if either of these assumptions are violated.

References:

Amemiya, T. (1979). "The Estimation of a Simultaneous-Equation Tobit Model." International Economic Review, 20, 1969-181

Cosslett, S.R. (2004). "Efficient Semiparametric Estimaion of Censored and Truncated Regressions via a Smoothed Self-consistency Equation." Econometrica, 72, 1277-1293.

Lewbell, A. And O. Linton (2002). "Nonparametric censored and truncated regression." Econometrica 70, 765-779.

Newey, W. K. (1984). "Specification Tests for Distributional Assumptions in the Tobit Model." Journal of Econometics, 34: 125-145.

Powell, J.L. (1984). "Least Absolute Deviations Estimation for the Censored Regression Model." Journal of Econometrics. 25: 303-325.

Powell, J.L. (1986). "Symmetrically Trimmed Least Squares Estimation for Tobit Models." Econometrica, 54: 1435-1460.

Smith, R.J. and R.W. Blundell (1986). "An Exogeniety Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply." Econometrica, 54, 679-85.

h. Truncated Regression Models

In truncated regression models (X_t, Y_t) is only observed (utilized) if $Y_t > a$,

$$Y_t = X_t\beta + \varepsilon_t \geq a, \text{ or}$$

$$\varepsilon_t \geq a - X_t\beta.$$

The conditional pdf of ε_t is given by

$$\frac{f(Y_t - X_t\beta;\theta)}{1 - F(a - X_t\beta;\theta)}$$

It is important to note that

$$\boxed{E(Y_t|X_t) = X_t\beta + E(\varepsilon_t|\varepsilon_t \geq a - X_t\beta)}$$

The log likelihood function for a random sample is given by

$$\ell(\beta,\theta) = \sum_{i=1}^{n}\left\{\ell n\left(f_\varepsilon\left(y_i - X_i\beta;\theta\right)\right) - \ell n\left[1 - F_\varepsilon\left(a - X_i\beta;\theta\right)\right]\right\}$$

which is maximized over β and θ. In the case of upper (b) and lower threshold values (a), the log-likelihood function can be written as

.

$$\ell(\beta,\theta) = \sum_{i=1}^{n}\left\{\ell n\left(f_\varepsilon\left(y_i - X_i\beta;\theta\right)\right) - \ell n\left[F_\varepsilon\left(b - X_i\beta;\theta\right) - F_\varepsilon\left(a - X_i\beta;\theta\right)\right]\right\}$$

The STATA command for normal truncation is

truncreg y x's, ll(#l) ul(#u)

where the a= #l and b= #u denote the lower and upper censoring values.

Incorrectly specifying the error distribution can lead to inconsistent estimators. Flexible parametric pdf/cdf's or non-parametric estimators of the pdf/cdf can be used

Cosslett's (2004) approach using the Kaplan-Meier non-parametric kernel estimator of the cdf can be applied to the truncated regression model.

Cosslett, S.R. (2004). "Efficient Semiparametric Estimaion of Censored and Truncated Regressions via a Smoothed Self-consistency Equation." Econometrica, 72, 1277-1293.

Lewbell, A. And O. Linton (2002). "Nonparametric censored and truncated regression." Econometrica 70, 765-779.

Summary of qualitative response, censored, truncated, and interval regression model specifications

	Data	Model	ℓ (Log-likelihood)	Stata Commands
Qualitative response	$Y = 0, 1$	$\Pr\left(Y_i = 1 \mid X_i\right) = F(X_i\beta;\theta)$	$\sum_i \begin{pmatrix} y_i \ln\left(F(X_i\beta;\theta)\right) + \\ \left(1-y_i\right)\ln\left(1-F(X_i\beta;\theta)\right) \end{pmatrix}$	**probit y x's** **probit y x's,vce(robust)** **hetprob y x's** **logit y x's** **margins, dydx(*) atmeans**
Censored regression	$0 \le Y$	$Y = X\beta + \varepsilon \ge 0$	$\sum_i \begin{pmatrix} d_i \ln\left(f(\varepsilon_i = y_i - X_i\beta;\theta)\right) + \\ (1-d_i)\ln\left(F(-X_i\beta;\theta)\right) \end{pmatrix}$ $d_i = 1 \text{ if } y_i > 0$	**tobit y x's** **tobit y x's, vce(robust)** **tobit y x's, vce(boot)**
Truncated regression	$L < Y < U$	$L < Y = X\beta + \varepsilon < U$ $f_\varepsilon\left(\varepsilon;\theta\right) = \dfrac{f\left(\varepsilon;\theta\right)}{F\left(U - X\beta;\theta\right) - F\left(L - X\beta;\theta\right)}$	$\sum_i \ln\left(f_\varepsilon\left(y_i - X_i\beta;\theta\right)\right)$	**truncreg y x's, ll(L) UL(U)** **margins dydx(*) atmeans**
Interval regression	$Y_i \in \left(a_i, b_i\right)$ **only know endpoints**	$a_i < Y_i = X_i\beta + \varepsilon_i < b_i$	$\sum_i \ln\begin{pmatrix} F\left(b_i - X_i\beta;\theta\right) \\ - F\left(a_i - X_i\beta;\theta\right) \end{pmatrix}$	**intreg depvar1 depvar2 x's, options** **depvar1=a's, depvar2=b's**

Estimators of β are inconsistent if the error pdf is misspecified, heteroskedasticity exist, or if the ind$\left(X\beta\right)$ should be nonlinear

Alternative estimators can be obtained using flexible or kernel specifications for the error distribution.

i. Quantile regression

A popular estimation procedure when outliers may be present is that of least absolute deviations (LAD) estimation defined by

$$\hat{\beta}_{LAD} = \arg\min_{\beta}\left\{\sum_{t=1}^{n}\left|y_t - X_t\beta\right|\right\}$$

the solution of which involves what is referred to as a linear programming problem.

The corresponding estimated relationship is given by $\hat{y}_t = X_t\hat{\beta}_{LAD}$. This actually estimate median corresponding to X_t ; the prothus, for e: X

If the relationship between the dependent variable and the explanatory variables is desired corresponding to different quantiles (e.g. top 75%), the quantile regression methods can be used. The quantile regression model is a generaliztion of the LAD and is defined by

$$\begin{aligned}\Pr(y_t \le \tau \left|X_t\right.) &= F_{\varepsilon_\theta}\left(X_t\beta_\theta + \varepsilon_t \le \tau \left|X_t\right.\right) = F\left(\varepsilon_t \le \tau - X_t\beta_\theta \left|X_t\right.\right)\\ &= F_\varepsilon\left(\tau - X_t\beta_\theta\right) = \theta\end{aligned}$$

The θ^{th} quantile is defined $Quantile_\theta\left(y_t \left|X_t\right.\right) = X_t\beta_\theta$

The corresponding β_θ can be obtained by

$$\begin{aligned}\hat{\beta}_\theta &= \arg\min_{\beta}\left\{\sum_{t:y_t \ge X_t\beta}^{n}\theta\left|y_t - X_t\beta\right| + \sum_{t:y_t < X_t\beta}^{n}(1-\theta)\left|y_t - X_t\beta\right|\right\}\\ &= \arg\min_{\beta}\sum_{t=1}^{n}\left\{\left(\theta - \frac{1}{2} + \frac{1}{2}sign(y_t - X_t\beta)\right)(y_t - X_t\beta)\right\}\end{aligned}$$

Stata's *qreg* command facilitates the evaluation of quantiles with the command

qreg y x's, quantile(θ ***)***

- ***qreg y x's, quantile(.5)* is the equivalent to *qreg y x's — LAD***
- ***qreg y x's, quantile(.9)* reports the 90% quantile**

MLE using the Laplace pdf yields the LAD results

The skewed Laplace pdf with the scale parameter expressed interms of the regressors could accommodate skewed error distributions with heteroskedasticity.

j. Hazard Functions

Hazard functions are important characterizations of random variables in labor, quality control, and in demography. For example, let f(s; θ) denote the probability density function for a completed spell of unemployment. Let the cumulative distribution function be denoted by

$$F(s;\theta) = \Pr(S \le s)$$

$$= \int_0^s f(t;\theta)dt$$

and the probability that an unemployment spell will have length less than s.

The mean length of unemployment spells is given by

$$\mu_s = E(S) = \int_0^\infty sf(s;\theta)ds.$$

The probability of a spell of length greater than s is called the survivor function

$$S(s;\theta) = 1 - F(s;\ \theta).$$

The hazard function gives the conditional probability that a person unemployed for length s will find employment at time s,

$$h(s;\theta) = \Pr(S = s | S > s)$$

$$= \frac{f(s;\theta)}{1 - F(s;\theta)} \quad .$$

The parameters θ can be estimated from the pdf, f(s;θ), or from the hazard functions, h(s;θ). To see this, consider a random sample which includes completed spells and interrupted spells. The pdf for completed spells is $f(s_i;\theta)$ and $1\text{-}F(s_i;\theta)$ for interrupted spells. Let $d_i = 1$ for a complete spell and 0 for an uninterrupted spell.

The likelihood function for the sample can be written as

$$L = \prod_i (f(s_i;\theta))^{d_i}(1 - F(s_i;\theta))^{1-d_i};$$

hence, the log likelihood function is given by

$$\boxed{\ell = \ln L = \sum_i \{d_i \ln f(s_i;\theta) + (1 - d_i)\ln(1 - F(s_i;\theta))\}}$$

Note the relationship between this log likelihood function and that for the qualitative responsive model. Regrouping terms in the log likelihood function yields

$$\boxed{\ell = \sum_i d_i \ln\left(\frac{f(s_i;\theta)}{1 - F(s_i;\theta)}\right) + \sum_i \ln(1 - F(s_i;\theta))}$$

Note the first sum is structured in terms of the hazard functions.

k. Generalized Method of Moments (GMM)

The Regression Section of the notes outlines the formulation of the generalized method of moments (GMM) as minimizing a quadratic form

$$q\left(\hat{\theta}\right)' VAR^{-1}\left(q\left(\hat{\theta}\right)\right)q\left(\hat{\theta}\right)$$

where $q\left(\hat{\theta}\right)$ is a vector whose elements are the difference between sample moments and theoretical moments. In many applications of GMM this optimization involves solving nonlinear optimization problems. The $VAR\left(q\left(\hat{\theta}\right)\right)$ matrix may depend on unknown parameters. Frequently these are estimated and then the corresponding quadratic form is minimized to obtain GMM. If the minimization is over the parameters in q( ) and in the variance matrix, the corresponding estimators are referred to as continuous updating estimators (CUE) and tend to have smaller bias than without updating the variance matrix.

l. The Bootstrap

The bootstrap attempts to approximate the distribution of the actual estimator. The bootstrap is performed by generating repeated samples (with replacement) from the original sample and estimating the unknown parameters. This gives us a "bootstrapped" sample of estimates of the desired parameters or distributional characteristics. The multiple parameter estimates, or distributional characteristics, are used to approximate the exact distribution and construct confidence intervals.

As a simple example, consider using the bootstrap to estimate the standard errors of the OLS regression coefficient estimator in the presence of heteroskedastic errors. For pedagogical purposes the estimated standard errors of the OLS could be compared to the robust standard errors and bootstrap standard errors. These could be obtained using the STATA commands

reg y x's

reg y x's, robust

reg y x's, vce(boot)

The estimated standard errors obtained from the resampling with the bootstrap methodology could be obtained using the commands:

bootstrap _b _se, reps(#r) size(#s): reg y x's

where reps(#r) indicates the number or random samples (with replacement) and #s denotes the sample size (#s ≤ n); hence, #r is the number of sample estimates of the desired entity that are generated.

6. SOME EXERCISES USING NONLINEAR MODELS

Production functions:

Three commonly used production functions are the:

COBB DOUGLAS:

(1) $Y(t) = A(t)\, L_t^{\beta_1} K_t^{\beta_2} \bullet \varepsilon_t$

CONSTANT ELASTICITY OF SUBSTITUTION (CES):

(2) $Y(t) = A(t)\, [\delta L^{\rho} + (1\text{-}\delta)K^{\rho}]^{M/\rho} \bullet \varepsilon_t$

VARIABLE ELASTICITY OF SUBSTITUTION (VES):

(3) $Y(t) = A(t)\, [L_t + (\rho\text{-}1)K_t]^{\alpha\delta\rho} K_t^{\alpha(1-\delta\rho)} \bullet \varepsilon_t$

The A(t) in (1), (2), and (3) can be used as a proxy for technological change allowing output (Y_t) to change without changes in labor (L_t) and capital (K) inputs. Some characteristics and interrelationships between these functions are given by

	CD	CES	VES
Returns to scale	$\beta_1 + \beta_2$	M	α
Elasticity of substitution $\%\Delta(K/L)/\%\Delta(W_L/W_K)$	1	$\dfrac{1}{1-\rho}$	$1+\left(\dfrac{\rho-1}{1-\delta}\right)\left(\dfrac{K}{L}\right)$
Labor's share of output (if constant returns to scale)	β_1	$\dfrac{\delta}{\delta+(1-\delta)\left(\dfrac{K}{L}\right)^{\rho}}$	$\dfrac{\delta P}{1+(\rho-1)\dfrac{K}{L}}$

The Cobb Douglas is a special case of the CES and the VES production functions, in particular

The VES production function with $\rho = 1$ is of a Cobb Douglas form.

The CES production function with $\rho \rightarrow 0$ is also of a Cobb Douglas form.

Consider the following data set:

B.	t	Output (Y_t)	Labor (L_t)	Capital (K_t)
	1	40.26	64.63	133.14
	2	40.84	66.30	139.24
	3	42.83	65.27	141.64
	4	43.89	67.32	148.77
	5	46.10	67.20	151.02
	6	44.45	65.18	143.38
	7	43.87	65.57	148.19
	8	49.99	71.42	167.12
	9	52.64	77.52	171.33
	10	57.93	79.46	176.41

This set consists of the first ten observations used by Solow in his study dealing with estimation of technological change.

Exercises

1. Use Solow's data and assume the possibility of technological change as modeled by

$$A(t)=e^{\beta_1+\beta_2 t}$$

$$\ln\varepsilon_t \sim N[0, \sigma^2],$$

in the production functions, obtain nonlinear least squares estimates of the parameters in the CD, CES, and VES production functions. Also estimate the Cobb Douglas production function under the assumption of constant returns to scale.

Test the following hypotheses:

(a) The CES production function has constant returns to scale

(b) The Cobb Douglas production function has constant returns to scale

(c) CES production function is a Cobb Douglas production function

(d) The VES production function is a Cobb Douglas production function with constant returns to scale. Perform the test using the WALD Test and the likelihood ratio test.

2. Many economists feel that the assumption of normally distributed random disturbances is too restrictive and that in particular, the tails of the normal density are "too thin" to describe many economic series. One alternative is provided by the following:

$$GED(\varepsilon;\sigma,p) = \frac{p\ e^{-|\varepsilon/\sigma|^p}}{2\sigma\Gamma(1/p)}$$

(a) Assume that no autocorrelation exists. Obtain the log likelihood functions corresponding to $y_t = x_t\beta + \varepsilon_t$.

(b) Discuss how one might obtain MLE estimators of the parameters in each case.

(c) How could you test the assumption of Normality of the error distribution?

3. Consider the model

$$Y_t = X_t\beta + \varepsilon_t$$

where $X_t = (1, X_{t2}, ..., X_{tk})$, and $\beta = (\beta_1, \beta_2, ..., \beta_k)$. Assume that $\varepsilon = (\varepsilon_1,...,\varepsilon_N) \sim N[0, \sigma^2 I]$. Show that the Wald, LR and LM tests for H_0: $\beta_2 = ... = \beta_k = 0$ can be expressed in terms of the coefficient of determination as

$$W = N\, R^2/(1 - R^2)$$

$$LR = -N \ln(1 - R^2)$$

$$LM = N\, R^2$$

4. Fourteen applicants to a graduate program had the following quantitative and verbal scores on the GRE examination. Six students were admitted to the program.

	GRE Scores		
	Quantitative	Verbal	Admitted
Student number	Q	V	Yes = 1
1	760	550	1
2	600	350	0
3	720	320	0
4	710	630	1
5	530	430	0
6	650	570	0
7	800	500	1
8	650	680	1
9	520	660	0
10	800	250	0
11	670	480	0
12	670	520	1
13	780	710	1

Source: Donald F. Morrison, Applied Linear Statistical Methods, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1983, p. 279.

(a) Use a linear probability model to "predict" admissions into the graduate program based upon the quantitative and verbal scores.

(b) Repeat (4a) using a probit model.

(c) Repeat (4a) using a logit model.

5. Consider the following data [Hampel, Robust Statistics]:

Y	X
15.7	17.6
44.9	20.0
18.0	20.9
19.9	21.6
23.4	26.0
19.7	27.1
23.1	27.6
23.8	27.8
24.9	32.6
26.1	33.4
27.6	35.1
26.1	37.0
31.3	38.7

(a) Plot the data.

(b) Estimate $Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t$ using least squares with

(1) the entire data set

(2) the second observation deleted.

(c) Use the “qreg y x’s’ in STATA or the “robust y x’s/LAE” command in SHAZAM and compare these results with your answer in (b).

(d) Estimate this model ($Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t$) under the assumption that the ε_t are distributed as a GED (generalized error distribution). .

STATA Commands for the normal and GED pdf’s:

```
cap prog drop normal
program define normal
version 1.0
args lnf mu sigma (used in estimation)
quietly replace `lnf'=ln(normalden($ML_y1,`mu’,`sigma')) STATA 10-calls normal density
end
use “e:\Econ Data\hbj.dta, clear or use infile command
ml search
ml maximize, difficult

cap prog drop GED
program define GED
version 1.0
args lnf xb s p
quietly replace `lnf'=ln(`p')-((abs($ML_y1-`xb')/`s')^(abs(`p')))-ln(2*`s')-lngamma((abs(1/`p')))
        This is the log-likelihood of theGED
end
clear
infile y x using g:\american.dat

ml model lf normal (y=x) (sigma:), technique(dfp)
ml search
ml maximize
ml model lf GED (y=x) (s:) (p:), technique(dfp)
ml search
ml maximize
```

(e) Perform LR and W tests of the hypothesis H_0: p=2, i.e., the errors are normal.

7. Using the data in problem 6, calculate the estimated standard error for the OLS slope coefficient using (a) the OLS results, (b) the robust standard errors (using reg y x,robust), and (c) the estimated standard error based upon a bootstrap with 100 replications.