Constrained Maximum Likelihood Estimation

for GAUSS™

Version 2.0

Aptech Systems, Inc.

Documentation Revision: 1504 March 8, 2010

Part Number: 001860

Contents

Chapter 1

Installation

1.1 UNIX/Linux/Mac

If you are unfamiliar with UNIX/Linux/Mac, see your system administrator or system documentation for information on the system commands referred to below.

1.1.1 Download

1. Copy the `.tar.gz` or `.zip` file to `/tmp`.

2. If the file has a `.tar.gz` extension, unzip it using `gunzip`. Otherwise skip to step 3.

   `gunzip app_`*appname*`_`*vernum*`.`*revnum*`_UNIX.tar.gz`

3. `cd` to your **GAUSS** or **GAUSS Engine** installation directory. We are assuming `/usr/local/gauss` in this case.

   ```
   cd /usr/local/gauss
   ```

4. Use `tar` or `unzip`, depending on the file name extension, to extract the file.

   `tar xvf /tmp/app_`*appname*`_`*vernum*`.`*revnum*`_UNIX.tar`

   – or –

   `unzip /tmp/app_`*appname*`_`*vernum*`.`*revnum*`_UNIX.zip`

1.1.2 CD

1. Insert the Apps CD into your machine's CD-ROM drive.

2. Open a terminal window.

3. `cd` to your current **GAUSS** or **GAUSS Engine** installation directory. We are assuming `/usr/local/gauss` in this case.

   ```
   cd /usr/local/gauss
   ```

4. Use `tar` or `unzip`, depending on the file name extensions, to extract the files found on the CD. For example:

   ```
   tar xvf /cdrom/apps/app_appname_vernum.revnum_UNIX.tar
   ```
   – or –
   ```
   unzip /cdrom/apps/app_appname_vernum.revnum_UNIX.zip
   ```
   However, note that the paths may be different on your machine.

1.2 Windows

1.2.1 Download

Unzip the `.zip` file into your **GAUSS** or **GAUSS Engine** installation directory.

1.2.2 CD

1. Insert the Apps CD into your machine's CD-ROM drive.

2. Unzip the `.zip` files found on the CD to your **GAUSS** or **GAUSS Engine** installation directory.

1.3 Difference Between the UNIX and Windows Versions

- If the functions can be controlled during execution by entering keystrokes from the keyboard, it may be necessary to press *Enter* after the keystroke in the UNIX version.

Chapter 2

Constrained Maximum Likelihood Estimation

written by

Ronald Schoenberg

This module contains a set of procedures for the solution of the constrained maximum likelihood problem

2.1 Getting Started

GAUSS 3.6.23+ is required to use these routines.

2.1.1 README Files

The file **README.cml** contains any last minute information on this module. Please read it before using the procedures in this module.

2.1.2 Setup

In order to use the procedures in the *CONSTRAINED MAXIMUM LIKELIHOOD* Module, the **CML** library must be active. This is done by including `cml` in the **LIBRARY** statement at the top of your program or command file:

```
library cml,pgraph;
```

This enables **GAUSS** to find the *CONSTRAINED MAXIMUM LIKELIHOOD* procedures. If you plan to make any right hand references to the global variables (described in the *REFERENCE* section), you also need the statement:

```
#include cml.ext;
```

Finally, to reset global variables in succeeding executions of the command file the following instruction can be used:

```
cmlset;
```

This could be included with the above statements without harm and would insure the proper definition of the global variables for all executions of the command file.

The version number of each module is stored in a global variable:

__cml_version 3×1 matrix, the first element contains the major version number of the *CONSTRAINED MAXIMUM LIKELIHOOD* Module, the second element the minor version number, and the third element the revision number.

If you call for technical support, you may be asked for the version number of your copy of this module.

2.1.3 Converting MAXLIK Command Files

The **CML** module includes a utility for processing command files to change **MAXLIK** global names to **CML** global names and vice versa. This utility is a standalone executable program that is called outside of **GAUSS**. The format is:

chgvar *control_file target_directory file...*

The *control_file* is an ASCII file containing a list of the symbols to change in the first column and the new symbol names in the second column. The **CML** module comes with three control_files:

cmltoml4	**CML** to **MAXLIK 4.x**
ml4tocml	**MAXLIK 4.x** to **CML**
ml3tocml	**MAXLIK 3.x** to **CML**

CHGVAR processes each file and writes a new file with the same name in the target directory.

A common use for **CHGVAR** is translating a command file that had been used before with **MAXLIK 3.x** to one that can be run with **CML**. For example:

```
mkdir new
chgvar ml3tocml new max*.cmd
```

This would convert every file matching `max*.cmd` in the current directory and create a new file with the same name in the new directory.

The reverse translation is also possible. However, there are many global names in **CML** that don't have a corresponding global in **MAXLIK**, and in these cases no translation occurs.

Further editing of the file may be necessary after processing by **CHGVAR**.

You may edit the control files or create your own. They are ASCII files with each line containing a pair of names, the first column being the old name, and the second column the new name.

2.2 New in CML

2.2.1 Fast Procedures

CML 2.0 contains fast versions of several of its procedures including the primary procedure, **CML**. These procedures are **fastCML**, **fastCMLBoot**, **fastCMLBayes**, **fastCMLProfile**, and **fastCMLPflClimits**. The fast procedures have identical input and output arguments and thus their use requires only the change in name.

These new fast procedures gain their speed at the loss of some features. The data must be completely storable in RAM. During the iterations no keyboard input is allowed, nor is there any iteration information printed to the screen.

You may want to test your program using the slower version of the procedure since diagnostic information will not be available in the faster version.

2.2.2 New Random Numbers

You may now choose either a linear congruential type random number, which is the kind always available in **GAUSS**, or the new "KISS-Monster" random number generator The latter generator has a period of 2^{3859} which is long enough for any serious Monte Carlo work, while the former is faster but with a shorter, 2^{32} period.

The current state of the generators is stored in a **CML** global variable, **__cml_state**. The seed for the random number sequence is the default value of **__cml_state** which is set to 345678. For a different sequence set **__cml_state** to a different integer value.

If you wish to restart a bootstrap from where a former bootstrap ended, save the contents of **__cml_state** in a **GAUSS** matrix file on the disk. Then in your command file load the contents of the matrix file into **__cml_state** before calling **CMLBoot** or **fastCMLBoot**.

2.2.3 Trust Region Method

With a poor starting point the minimization of the log-likelihood can often end up in far regions of the parameter space. This is a problem in particular for the Newton descent method. The trust region is an effective method for preventing this. This is essentially a linear constraint placed on the direction taken at each iteration. It is turned off by default. To turn it on, set the global variable **_cml_TrustRegion** to the desired maximum size of the direction, for example,

```
_cml_TrustRegion = .1;
```

This global can also be turned on or off or modified from the keyboard during the iterations by pressing 'T'.

2.2.4 Switching Algorithms

A new algorithm switching method has been introduced into **CML**. This method allows switching between two algorithms depending on three measures of progress, change in function value, number of iterations, or change in line search step length. For example, to have **CML** switch between the BFGS and Newton methods whenever (a) the function changes less than .001, or (b) the number of iterations since the last change of algorithms is 10, or (c), the line search step length is less than .0001, enter the following statement into your command file before the call to **CML**,

```
_cml_Switch = {   1    3,
                .001 .001,
                 10   10,
               .0001 .0001 };
```

Entering a single column with one algorithm number rather than two columns causes **CML** to change algorithms to that specified in the first row only once when one of the conditions is met.

2.2.5 Grid Search

After a direction is computed **CML** attempts a line search. When the default method STEPBT fails it runs through all of the available methods looking for one that will succeed. If all of them fail **CML** has two alternatives, first, to give up the search for optimal parameters and return with their current value, or second, attempt a grid search for a new direction from which to continue the optimization. **CML** adopts the latter method by default with the global **_cml_GridSearch** set to a nonzero value. The former method will happen when it is set to zero.

The radius of the grid search is set by the global **_cml_GridSearchRadius**. By default it is set to .001.

2.2.6 Multiple Point Numerical Gradients

By default numerical first derivatives are computed using two evaluations of the function and second derivatives three evaluations. For most problems this produces derivatives with enough accuracy to converge in a reasonable number of iterations. The accuracy might not be enough for some difficult problems however. In these cases greater accuracy can be achieved by using a greater number of points in computing the numerical derivatives. This is done by setting **_cml_GradOrder** to some nonzero integer value,

```
_cml_GradOrder = 3;
```

2.2.7 Calling CML without a Dataset

Usually the log-likelihood procedure provided by the user takes the data as the second input argument and returns a vector of log-likelihoods computed by observation. This may not be convenient for some problems. If you wish to handle the data in a global variable outside the call to **CML**, or if it is inconvient to return the log-likelihood by observation, **CML** still produces maximum likelihood estimates.

Handling data outside of CML

The data are normally passed to **CML** in the first argument in the call to **CML**. Alternatively, a scalar zero or a scalar missing value may be entered into the first argument. In this case **CML** will assume that you are handling the data in the log-likelihood function yourself.

Returning a scalar log-likelihood

When you either choose not to return a vector of log-likelihoods computed by observation or if the problem doesn't permit it, you must provide the proper number of observations used for statistical inference in the global **_cml_NumObs**.

The following features are not available when the log-likelihood procedure returns a scalar value:

QML standard errors

BHHH descent method

Bootstrap estimates and standard errors

Bayesian estimates and standard errors

weighted estimates

2.3 The Log-likelihood Function

CONSTRAINED MAXIMUM LIKELIHOOD is a set of procedures for the estimation of the parameters of models via the maximum likelihood method with general constraints on the parameters, along with an additional set of procedures for statistical inference.

CONSTRAINED MAXIMUM LIKELIHOOD solves the general weighted maximum likelihood problem

$$L = \sum_{i=1}^{N} \log P(Y_i; \theta)^{w_i}$$

where N is the number of observations, w_i is a weight. $P(Y_i, \theta)$ is the probability of Y_i given θ, a vector of parameters, subject to the linear constraints,

$$A\theta = B$$
$$C\theta \geq D$$

the nonlinear constraints

$$G(\theta) = 0$$
$$H(\theta) \geq 0$$

and bounds

$$\theta_l \leq \theta \leq \theta_u$$

$G(\theta)$ and $H(\theta)$ are functions provided by the user and must be differentiable at least once with respect to θ.

The *CONSTRAINED MAXIMUM LIKELIHOOD* procedure **CML** finds values for the parameters in θ such that L is maximized. In fact **CML** minimizes $-L$. It is important to note, however, that the user must specify the log-probability to be *maximized*. **CML** transforms the function into the form to be minimized.

CML has been designed to make the specification of the function and the handling of the data convenient. The user supplies a procedure that computes $\log P(Y_i; \theta)$, i.e., the log-likelihood, given the parameters in θ, for either an individual observation or set of observations (i.e., it must return either the log-likelihood for an individual observation or a vector of log-likelihoods for a matrix of observations; see discussion of the global variable **__row** below). **CML** uses this procedure to construct the function to be minimized.

2.4 Algorithm

CONSTRAINED MAXIMUM LIKELIHOOD uses the Sequential Quadratic Programming method. In this method the parameters are updated in a series of iterations beginning with a starting values that you provide. Let θ_t be the current parameter values. Then the succeeding values are

$$\theta_{t+1} = \theta_t + \rho\delta$$

where δ is a $K \times 1$ *direction* vector, and ρ a scalar *step length*.

DIRECTION

Define

$$\begin{aligned} \Sigma(\theta) &= \frac{\partial^2 L}{\partial\theta\partial\theta'} \\ \Psi(\theta) &= \frac{\partial L}{\partial\theta} \end{aligned}$$

and the Jacobians

$$\begin{aligned} \dot{G}(\theta) &= \frac{\partial G(\theta)}{\partial\theta} \\ \dot{H}(\theta) &= \frac{\partial H(\theta)}{\partial\theta} \end{aligned}$$

For the purposes of this exposition, and without loss of generality, we may assume that the linear constraints and bounds have been incorporated into G and H.

The direction, δ is the solution to the quadratic program

$$minimize\ \frac{1}{2}\delta'\Sigma(\theta_t)\delta + \Psi(\theta_t)\delta$$

$$\begin{aligned} subject\ to\quad & \dot{G}(\theta_t)\delta + G(\theta_t) = 0 \\ & \dot{H}(\theta_t)\delta + H(\theta_t) \geq 0 \end{aligned}$$

This solution requires that Σ be positive semi-definite.

In practice, linear constraints are specified separately from the G and H because their Jacobians are known and easy to compute. And the bounds are more easily handled separately from the linear inequality constraints.

LINE SEARCH

Define the merit function

$$m(\theta) = L + \max \mid \kappa \mid \sum_j \mid g_j(\theta) \mid - \max \mid \lambda \mid \sum_\ell \min(0, h_\ell(\theta))$$

where g_j is the j-th row of G, h_ℓ is the ℓ-th row of H, κ is the vector of Lagrangean coefficients of the equality constraints, and λ the Lagrangean coefficients of the inequality constraints.

The line search finds a value of ρ that minimizes or decreases $m(\theta_t + \rho\delta)$.

2.4.1 Derivatives

The SQP method requires the calculation of a Hessian, Σ, and various gradients and Jacobians, Ψ, $\dot{G}(\theta)$, and $\dot{H}(\theta)$. **CML** computes these numerically if procedures to compute them are not supplied.

If you provide a proc for computing Ψ, the first derivative of L, **CML** uses it in computing Σ, the second derivative of L, i.e., Σ is computed as the Jacobian of the gradient. This improves the computational precision of the Hessian by about four places. The accuracy of the gradient is improved and thus the iterations converge in fewer iterations. Moreover, the convergence takes less time because of a decrease in function calls - the numerical gradient requires k function calls while an analytical gradient reduces that to one.

2.4.2 The Secant Algorithms

The Hessian may be very expensive to compute at every iteration, and poor start values may produce an ill-conditioned Hessian. For these reasons alternative algorithms are provided in **CML** for updating the Hessian rather than computing it directly at each iteration. These algorithms, as well as step length methods, may be modified during the execution of **CML**.

Beginning with an initial estimate of the Hessian, or a conformable identity matrix, an update is calculated. The update at each iteration adds more "information" to the estimate of the Hessian, improving its ability to project the direction of the descent. Thus after several iterations the secant algorithm should do nearly as well as Newton iteration with much less computation.

There are two basic types of secant methods, the BFGS (Broyden, Fletcher, Goldfarb, and Shanno), and the DFP (Davidon, Fletcher, and Powell). They are both rank two updates, that is, they are analogous to adding two rows of new data to a previously computed moment matrix. The Cholesky factorization of the estimate of the Hessian is updated using the functions **CHOLUP** and **CHOLDN**.

Secant Methods (BFGS and DFP)

BFGS is the method of Broyden, Fletcher, Goldfarb, and Shanno, and DFP is the method of Davidon, Fletcher, and Powell. These methods are complementary (Luenberger 1984, page 268). BFGS and DFP are like the NEWTON method in that they use both first and second derivative information. However, in DFP and BFGS the Hessian is approximated, reducing considerably the computational requirements. Because they do not explicitly calculate the second derivatives they are sometimes called *quasi-Newton* methods. While it takes more iterations than the NEWTON method, the use of an approximation produces a gain because it can be expected to converge in less overall time (unless analytical second derivatives are available in which case it might be a toss-up).

The secant methods are commonly implemented as updates of the *inverse* of the Hessian. This is not the best method numerically for the BFGS algorithm (Gill and Murray, 1972). This version of **CML**, following Gill and Murray (1972), updates the Cholesky factorization of the Hessian instead, using the functions **CHOLUP** and **CHOLDN** for BFGS. The new direction is then computed using **CHOLSOL**, a Cholesky solve, as applied to the updated Cholesky factorization of the Hessian and the gradient.

2.4.3 Line Search Methods

Given a direction vector d, the updated estimate of the parameters is computed

$$\theta_{t+1} = \theta_t + \rho\delta$$

where ρ is a constant, usually called the *step length*, that increases the descent of the function given the direction. **CML** includes a variety of methods for computing ρ. The value of the function to be minimized as a function of ρ is

$$m(\theta_t + \rho\delta)$$

Given θ and d, this is a function of a single variable ρ. Line search methods attempt to find a value for ρ that decreases m. STEPBT is a polynomial fitting method, BRENT and HALF are iterative search methods. A fourth method called ONE forces a step length of 1. The default line search method is STEPBT. If this, or any selected method, fails, then BRENT is tried. If BRENT fails, then HALF is tried. If all of the line search methods fail, then a random search is tried (provided **_cml_RandRadius** is greater than zero).

STEPBT

STEPBT is an implementation of a similarly named algorithm described in Dennis and Schnabel (1983). It first attempts to fit a quadratic function to $m(\theta_t + \rho\delta)$ and

computes an ρ that minimizes the quadratic. If that fails it attempts to fit a cubic function. The cubic function more accurately portrays the F which is not likely to be very quadratic, but is, however, more costly to compute. STEPBT is the default line search method because it generally produces the best results for the least cost in computational resources.

BRENT

This method is a variation on the *golden section* method due to Brent (1972). In this method, the function is evaluated at a sequence of test values for ρ. These test values are determined by extrapolation and interpolation using the constant, $(\sqrt{5}-1)/2 = .6180....$ This constant is the inverse of the so-called "golden ratio" $((\sqrt{5}+1)/2 = 1.6180...$ and is why the method is called a golden section method. This method is generally more efficient than STEPBT but requires significantly more function evaluations.

HALF

This method first computes $m(x+d)$, i.e., sets $\rho = 1$. If $m(x+d) < m(x)$ then the step length is set to 1. If not, then it tries $m(x+.5d)$. The attempted step length is divided by one half each time the function fails to decrease, and exits with the current value when it does decrease. This method usually requires the fewest function evaluations (it often only requires one), but it is the least efficient in that it is not very likely to find the step length that decreases m the most.

BHHHSTEP

This is a variation on the golden search method. A sequence of step lengths are computed, interpolating or extrapolating using a golden ratio, and the method exits when the function decreases by an amount determined by **__cml_Interp**.

2.4.4 Weighted Maximum Likelihood

Weights are specified by setting the **GAUSS** global, **___weight** to a weighting vector, or by assigning it the name of a column in the **GAUSS** data set being used in the estimation. Thus if a data matrix is being analyzed, **___weight** must be assigned to a vector.

CML assumes that the weights sum to the number of observations, i.e, that the weights are frequencies. This will be an issue only with statistical inference. Otherwise, any multiple of the weights will produce the same results.

2.4.5 Active and Inactive Parameters

The **CML** global **_cml_Active** may be used to fix parameters to their start values. This allows estimation of different models without having to modify the function procedure. **_cml_Active** must be set to a vector of the same length as the vector of start values. Elements of **_cml_Active** set to zero will be fixed to their starting values, while nonzero elements will be estimated.

This feature may also be used for model testing. **_cml_NumObs** times the difference between the function values (the second return argument in the call to **CML**) is chi-squared distributed with degrees of freedom equal to the number of fixed parameters in **_cml_Active**.

2.5 Managing Optimization

The critical elements in optimization are scaling, starting point, and the condition of the model. When the data are scaled, the starting point is reasonably close to the solution, and the data and model go together well, the iterations converge quickly and without difficulty.

For best results therefore, you want to prepare the problem so that model is well-specified, the data scaled, and that a good starting point is available.

The tradeoff among algorithms and step length methods is between speed and demands on the starting point and condition of the model. The less demanding methods are generally time consuming and computationally intensive, whereas the quicker methods (either in terms of time or number of iterations to convergence) are more sensitive to conditioning and quality of starting point.

2.5.1 Scaling

For best performance, the diagonal elements of the Hessian matrix should be roughly equal. If some diagonal elements contain numbers that are very large and/or very small with respect to the others, **CML** has difficulty converging. How to scale the diagonal elements of the Hessian may not be obvious, but it may suffice to ensure that the constants (or "data") used in the model are about the same magnitude.

2.5.2 Condition

The specification of the model can be measured by the condition of the Hessian. The solution of the problem is found by searching for parameter values for which the

gradient is zero. If, however, the Jacobian of the gradient (i.e., the Hessian) is very small for a particular parameter, then **CML** has difficulty determining the optimal values since a large region of the function appears virtually flat to **CML**. When the Hessian has very small elements, the inverse of the Hessian has very large elements and the search direction gets buried in the large numbers.

Poor condition can be caused by bad scaling. It can also be caused by a poor specification of the model or by bad data. Bad models and bad data are two sides of the same coin. If the problem is highly nonlinear, it is important that data be available to describe the features of the curve described by each of the parameters. For example, one of the parameters of the Weibull function describes the shape of the curve as it approaches the upper asymptote. If data are not available on that portion of the curve, then that parameter is poorly estimated. The gradient of the function with respect to that parameter is very flat, elements of the Hessian associated with that parameter is very small, and the inverse of the Hessian contains very large numbers. In this case it is necessary to respecify the model in a way that excludes that parameter.

2.5.3 Starting Point

When the model is not particularly well-defined, the starting point can be critical. When the optimization doesn't seem to be working, try different starting points. A closed form solution may exist for a simpler problem with the same parameters. For example, ordinary least squares estimates may be used for nonlinear least squares problems or nonlinear regressions like probit or logit. There are no general methods for computing start values and it may be necessary to attempt the estimation from a variety of starting points.

2.5.4 Diagnosis

When the optimization is not proceeding well, it is sometimes useful to examine the function, the gradient Ψ , the direction δ, the Hessian Σ, the parameters θ_t, or the step length ρ, during the iterations. The current values of these matrices can be printed out or stored in the global **_cml_Diagnostic** by setting **_cml_Diagnostic** to a nonzero value. Setting it to 1 causes **CML** to print them to the screen or output file, 2 causes **CML** to store then in **_cml_Diagnostic**, and 3 does both.

When you have selected **_cml_Diagnostic** = 2 or 3, **CML** inserts the matrices into **_cml_Diagnostic** using the **VPUT** command. The matrices are extracted using the **VREAD** command. For example,

```
_cml_Diagnostic = 2;
call CMLPrt(CML("tobit",0,&lpr,x0));
h = vread(_cml_Diagnostic,"hessian");
d = vread(_cml_Diagnostic,"direct");
```

The following table contains the strings to be used to retrieve the various matrices in the **VREAD** command:

θ	"params"
δ	"direct"
Σ	"hessian"
Ψ	"gradient"
ρ	"step"

2.6 Constraints

There are two general types of constraints, nonlinear equality constraints and nonlinear inequality constraints. However, for computational convenience they are divided into five types: linear equality, linear inequality, nonlinear equality, nonlinear inequality, and bounds.

2.6.1 Linear Equality Constraints

Linear constraints are of the form:

$$A\theta = B$$

where A is an $m_1 \times k$ matrix of known constants, and B an $m_1 \times 1$ vector of known constants, and θ the vector of parameters.

The specification of linear equality constraints is done by assigning the A and B matrices to the **CML** globals, **_cml_A** and **_cml_B**, respectively. For example, to constrain the first of four parameters to be equal to the third,

```
_cml_A = { 1 0 -1 0 };
_cml_B = { 0 };
```

2.6.2 Linear Inequality Constraints

Linear constraints are of the form:

$$C\theta \geq D$$

where C is an $m_2 \times k$ matrix of known constants, and D an $m_2 \times 1$ vector of known constants, and θ the vector of parameters.

The specification of linear equality constraints is done by assigning the C and D matrices to the **CML** globals, **_cml_C** and **_cml_D**, respectively. For example, to constrain the first of four parameters to be greater than the third, and as well the second plus the fourth greater than 10:

```
_cml_C = { 1 0 -1 0,
           0 1  0 1 };
_cml_D = { 0,
          10 };
```

2.6.3 Nonlinear Equality

Nonlinear equality constraints are of the form:

$$G(\theta) = 0$$

where θ is the vector of parameters, and $G(\theta)$ is an arbitrary, user-supplied function. Nonlinear equality constraints are specified by assigning the pointer to the user-supplied function to the **GAUSS** global, **_cml_EqProc**.

For example, suppose you wish to constrain the norm of the parameters to be equal to 1:

```
proc eqp(b);
    retp(b'b - 1);
endp;
_cml_EqProc = &eqp;
```

2.6.4 Nonlinear Inequality

Nonlinear inequality constraints are of the form:

$$H(\theta) \geq 0$$

where θ is the vector of parameters, and $H(\theta)$ is an arbitrary, user-supplied function. Nonlinear equality constraints are specified by assigning the pointer to the user-supplied function to the **GAUSS** global, **_cml_IneqProc**.

For example, suppose you wish to constrain a covariance matrix to be positive definite, the lower left nonredundant portion of which is stored in elements r:r+s of the parameter vector:

```
proc ineqp(b);
    local v;
    v = xpnd(b[r:r+s]); /* r and s defined elsewhere */
    retp(minc(eigh(v)) - 1e-5);
endp;
_cml_IneqProc = &ineqp;
```

This constrains the minimum eigenvalue of the covariance matrix to be greater than a small number (1e-5). This guarantees the covariance matrix to be positive definite.

2.6.5 Bounds

Bounds are a type of linear inequality constraint. For computational convenience they may be specified separately from the other inequality constraints. To specify bounds, the lower and upper bounds respectively are entered in the first and second columns of a matrix that has the same number of rows as the parameter vector. This matrix is assigned to the **CML** global, **_cml_Bounds**.

If the bounds are the same for all of the parameters, only the first row is necessary.

To bound four parameters:

```
_cml_Bounds = { -10 10,
                -10  0,
                  1 10,
                  0  1 };
```

Suppose all of the parameters are to be bounded between -50 and +50, then,

```
_cml_Bounds = { -50 50 };
```

is all that is necessary.

2.6.6 Example

The following example illustrates the estimation of a tobit model with linear equality constraints, nonlinearly inequality constraints, and bounds on the parameters. The nonlinear inequality constraint constraints the norm of the coefficients to be greater than three. The bounds are provided essentially to constrain the variance parameter to be greater than zero. The linear equality constraints constrain the first and second parameters to be equal.

```
library cml;
#include cml.ext;
cmlset;

proc lpr(x,z);
    local t,s,m,u;
    s = x[4];
    m = z[.,2:4]*x[1:3,.];
    u = z[.,1] ./= 0;
    t = z[.,1]-m;
    retp(u.*(-(t.*t)./(2*s)-.5*ln(2*s*pi)) + (1-u).*(ln(cdfnc(m/sqrt(s)))));
endp;
```

```
x0 = { 1, 1, 1, 1 };

_cml_A = { 1 -1 0 0 };
_cml_B = { 0 };

proc ineq(x);
    local b;
    b = x[1:3];
    retp(b'b - 3);
endp;
_cml_IneqProc = &ineq;

_cml_Bounds = { -10 10,
                -10 10,
                -10 10,
                .01 10 };

{ x,f,g,cov,ret } = CMLPrt(CML("tobit",0,&lpr,x0));

print "linear equality Lagrangeans";
print vread(_cml_Lagrange,"lineq");
print;
print "nonlinear inequality Lagrangeans";
print vread(_cml_Lagrange,"nlinineq");
print;
print "bounds Lagangreans";
print vread(_cml_Lagrange,"bounds");
```

and the output looks like this:

```
================================================================
 CML Version 2.0.0                              02/08/2001   9:51 am
================================================================
                         Data Set:  tobit
----------------------------------------------------------------

return code =    0
normal convergence

Mean log-likelihood        -1.34034
Number of cases     100

Covariance of the parameters computed by the following method:
Inverse of computed Hessian
```

```
Parameters    Estimates     Std. err.    Gradient
-------------------------------------------------------------------
P01             -0.1832        0.0710     -0.2073
P02             -0.1832        0.0710      0.1682
P03              1.7126        0.0152      0.1825
P04              1.0718        0.1589     -0.0000

Number of iterations    8
Minutes to convergence     0.06683

linear equality Lagrangeans
-0.1877

nonlinear inequality Lagrangeans
 0.0533

bounds Lagangreans
.
```

The scalar missing value for the bounds Lagrangeans indicate that they are inactive. The linear equality and nonlinear inequality constraints are active.

At times the Lagrangeans will not be scalar missing values but yet will be equal to zero and thus inactive. This indicates that the constraints became active at some point during the iterations.

2.7 Gradients

2.7.1 Analytical Gradient

To increase accuracy and reduce time, you may supply a procedure for computing the gradient, $\Psi(\theta) = \partial L/\partial\theta$, analytically.

This procedure has two input arguments, a $K \times 1$ vector of parameters and an $N_i \times L$ submatrix of the input data set. The number of rows of the data set passed in the argument to the call of this procedure may be less than the total number of observations when the data are stored in a **GAUSS** data set and there was not enough space to store the data set in RAM in its entirety. In that case subsets of the data set are passed to the procedure in sequence. The gradient procedure must be written to return a gradient (or more accurately, a "Jacobian") with as many rows as the input submatrix of the data set. Thus the gradient procedure returns an $N_i \times K$ matrix of gradients of the N_i observations with respect to the K parameters. The **CML** global, **_cml_GradProc** is then set to the pointer to that procedure. For example,

```
library cml;
#include cml.ext;
cmlset;

proc lpsn(b,z);   /*  Function - Poisson Regression */
   local m;
   m = z[.,2:4]*b;
   retp(z[.,1].*m-exp(m));
endp;

proc lgd(b,z);    /*  Gradient */
   retp((z[.,1]-exp(z[.,2:4]*b)).*z[.,2:4]);
endp;

x0 = { .5, .5, .5 };
_cml_GradProc = &lgd;
_cml_GradCheckTol = 1e-3;

{ x,f0,g,h,retcode } = CML("psn",0,&lpsn,x0);
call CMLPrt(x,f0,g,h,retcode);
```

In practice, unfortunately, much of the time spent on writing the gradient procedure is devoted to debugging. To help in this debugging process, **CML** can be instructed to compute the numerical gradient along with your prospective analytical gradient for comparison purposes. In the example above this is accomplished by setting **_cml_GradCheckTol** to 1e-3.

2.7.2 User-Supplied Numerical Gradient

You may substitute your own numerical gradient procedure for the one used by **CML** by default. This is done by setting the **CML** global, **_cml_UserGrad** to a pointer to the procedure.

CML includes some numerical gradient functions in `gradient.src` which can be invoked using this global. One of these procedures, **GRADRE**, computes numerical gradients using the Richardson Extrapolation method. To use this method set

```
_cml_UserNumGrad = &gradre;
```

2.7.3 Algorithmic Derivatives

ALGORITHMIC DERIVATIVES is a program that can be used to generate a *GAUSS* procedure to compute derivatives of the log-likelihood function. If you have *ALGORITHMIC DERIVATIVES*, be sure to read its manual for details on doing this.

First, copy the procedure computing the log-likelihood to a separate file. Second, from the command line enter

```
ad file_name d_file_name
```

where `file_name` is the name of the file containing the input function procedure, and `d_file_name` is the name of the file containing the output derivative procedure.

If the input function procedure is named `lpr`, the output derivative procedure has the name `d_1_lpr` where the addition to the "`_1_`" indicates that the derivative is with respect to the first of the two arguments.

For example, put the following function into a file called lpr.fct

```
proc lpr(x,z);
    local s,m,u;
    s = x[4];
    m = z[.,2:4]*x[1:3,.];
    u = z[.,1] ./= 0;
    retp(u.*lnpdfmvn(z[.,1]-m,s) + (1-u).*(lncdfnc(m/sqrt(s))));
endp;
```

Then enter the following at the *GAUSS* command line

```
library ad;
ad lpr.fct d_lpr.fct;
```

If successful, the following is printed to the screen

```
java -jar d:\gauss6.0\src\GaussAD.jar lpr.fct d_lpr.fct
```

and the derivative procedure is written to file named `d_lpr.fct`:

```
/* Version:1.0 - May 15, 2004 */
/* Generated from:lpr.fct */

/* Taking derivative with respect to argument 1 */
Proc(1)=d_1_lpr(x, z);
    Clearg _AD_fnValue;
        Local s, m, u;
        s =  x[(4)] ;
        Local _AD_t1;
        _AD_t1 =  x[(1):(3),.] ;
        m =  z[.,(2):(4)]  * _AD_t1;
        u =  z[.,(1)]  ./= 0;
```

```
        _AD_fnValue = (u .* lnpdfmvn( z[.,(1)]  - m, s)) + ((1 - u) .*
lncdfnc(m / sqrt(s)));
        /* retp(_AD_fnValue); */
        /* endp; */
        struct _ADS_optimum _AD_d__AD_t1 ,_AD_d_x ,_AD_d_s ,_AD_d_m
,_AD_d__AD_fnValue;
        /* _AD_d__AD_t1 = 0; _AD_d_s = 0; _AD_d_m = 0;  */
    _AD_d__AD_fnValue = _ADP_d_x_dx(_AD_fnValue);
    _AD_d_s = _ADP_DtimesD(_AD_d__AD_fnValue,
_ADP_DplusD(_ADP_DtimesD(_ADP_d_xplusy_dx(u .* lnpdfmvn( z[.,(1)]  - m, s),
(1 - u) .* lncdfnc(m / sqrt(s))), _ADP_DtimesD(_ADP_d_ydotx_dx(u, lnpdfmvn(
 z[.,(1)]  - m, s)), _ADP_DtimesD(_ADP_internal(d_2_lnpdfmvn( z[.,(1)]  - m,
s)), _ADP_d_x_dx(s)))), _ADP_DtimesD(_ADP_d_yplusx_dx(u .* lnpdfmvn(
 z[.,(1)]  - m, s), (1 - u) .* lncdfnc(m / sqrt(s))),
_ADP_DtimesD(_ADP_d_ydotx_dx(1 - u, lncdfnc(m / sqrt(s))),
_ADP_DtimesD(_ADP_d_lncdfnc(m / sqrt(s)), _ADP_DtimesD(_ADP_d_ydivx_dx(m,
sqrt(s)), _ADP_DtimesD(_ADP_d_sqrt(s), _ADP_d_x_dx(s)))))));
    _AD_d_m = _ADP_DtimesD(_AD_d__AD_fnValue,
_ADP_DplusD(_ADP_DtimesD(_ADP_d_xplusy_dx(u .* lnpdfmvn( z[.,(1)]  - m, s),
(1 - u) .* lncdfnc(m / sqrt(s))), _ADP_DtimesD(_ADP_d_ydotx_dx(u, lnpdfmvn(
 z[.,(1)]  - m, s)), _ADP_DtimesD(_ADP_internal(d_1_lnpdfmvn( z[.,(1)]  - m,
s)), _ADP_DtimesD(_ADP_d_yminusx_dx( z[.,(1)] , m), _ADP_d_x_dx(m)))),
_ADP_DtimesD(_ADP_d_yplusx_dx(u .* lnpdfmvn( z[.,(1)]  - m, s), (1 - u) .*
lncdfnc(m / sqrt(s))), _ADP_DtimesD(_ADP_d_ydotx_dx(1 - u, lncdfnc(m / sqrt(s)
)), _ADP_DtimesD(_ADP_d_lncdfnc(m / sqrt(s)), _ADP_DtimesD(_ADP_d_xdivy_dx(m,
sqrt(s)), _ADP_d_x_dx(m)))));
    /* u =  z[.,(1)]  ./= 0; */
    _AD_d__AD_t1 = _ADP_DtimesD(_AD_d_m, _ADP_DtimesD(_ADP_d_yx_dx(
 z[.,(2):(4)] , _AD_t1), _ADP_d_x_dx(_AD_t1)));
    Local _AD_sr_x, _AD_sc_x;
    _AD_sr_x = _ADP_seqaMatrixRows(x);
    _AD_sc_x = _ADP_seqaMatrixCols(x);
    _AD_d_x = _ADP_DtimesD(_AD_d__AD_t1, _ADP_d_x2Idx_dx(x,
 _AD_sr_x[(1):(3)] ,  _AD_sc_x[0] ));
    Local _AD_s_x;
    _AD_s_x = _ADP_seqaMatrix(x);
    _AD_d_x = _ADP_DplusD(_ADP_DtimesD(_AD_d_s, _ADP_d_xIdx_dx(x,
 _AD_s_x[(4)] )), _AD_d_x);
    retp(_ADP_external(_AD_d_x));
endp;
```

If there's a syntax error in the input function procedure, the following is written to the screen

```
java -jar d:\gauss6.0\src\GaussAD.jar lpr.fct d_lpr.fct
Command 'java -jar d:\gauss6.0\src\GaussAD.jar lpr.fct d_lpr.fct' exit status 1
```

the **exit status 1** indicating that an error has occurred. The output file then contains the reason for the error:

```
/* Version:1.0 - May 15, 2004 */
/* Generated from:lpr.fct */

/* Taking derivative with respect to argument 1 */

proc lpr(x,z);
    local s,m,u;
    s = x[4];
    m = z[.,2:4]*x[1:3,.];
    u = z[.,1] ./= 0;
    retp(u.*lnpdfmvn(z[.,1]-m,s) + (1-u).*(lncdfnc(m/sqrt(s)));
Error: lpr.fct:12:63: expecting ')', found ';'
```

Finally, set the global, **__cml_GradProc** equal to a pointer to this above procedure, for example,

```
library cml,ad;

#include ad.sdf

x0 = { 1, 1, 1, 1 };
__title = "tobit example";

_cml_Bounds = { -10 10,
                -10 10,
                -10 10,
                .1 10 };

_cml_GradProc = &d_1_lpr;

CML("tobit",0,&lpr,x0);
```

Speeding Up the Algorithmic Derivative

A slightly faster derivative procedure can be generated by modifying the log-likelihood proc to return a scalar sum of the log-likelihoods in the input file in the call to **AD**. It is important to note that this derivative function based on a scalar return cannot be used for computing the QML covariance matrix of the parameters. Thus if you want both a derivative procedure based on a scalar return and QML standard errors you will need to provide both types of gradient procedures. To accomplish this first copy both

versions of the log-likelihood procedure into separate files and run **AD** on both of them with different output files. Then copy both of these derivatives procedures to the command file. Note: the log-likelihood procedure that returns a vector of log-likelihoods should remain in the command file, i.e., don't use the version of the log-likelihood that returns a scalar in the command file.

For example, enlarging on the example in the previous section, put the following into a separate file,

```
proc lpr2(x,z);
    local s,m,u,logl;
    s = x[4];
    m = z[.,2:4]*x[1:3,.];
    u = z[.,1] ./= 0;
    logl = u.*lnpdfmvn(z[.,1]-m,s) + (1-u).*(lncdfnc(m/sqrt(s)));
    retp(sumc(logl));
endp;
```

Then enter on the command line

```
ad lpr2.src d_lpr2.src
```

and copy the contents of d˙lpr2.src into the command file.

Our comand file now contains two derivative procedures, one based on a scalar result and another on a vector result. The one in the previous section `d_1_lpr` is our vector result derivative, and the from run above, `d_1_lpr2` is our scalar result derivative. We want to use `d_1_lpr2` for the iterations because it will be faster (it is computing a $1 \times K$ vector gradient), and for the QML covariance matrix of the parameters we will use `d_1_lpr` which returns a $N \times K$ matrix of derivatives as required for the QML covariance matrix.

Our command file will be

```
library cml,ad;

#include ad.sdf

x0 = { 1, 1, 1, 1 };
__title = "tobit example";

_cml_Bounds = { -10 10,
                -10 10,
                -10 10,
                .1 10 };
```

```
    _cml_QMLProc = &d_1_lpr;
    _cml_GradProc = &d_1_lpr2;

    CML("tobit",0,&lpr,x0);
```

in addition to the two derivative procedures.

2.7.4 Analytical Hessian

You may provide a procedure for computing the Hessian, $\Sigma(\theta) = \partial^2 L/\partial\theta\partial\theta'$. This procedure has two arguments, the $K \times 1$ vector of parameters, an $N_i \times L$ submatrix of the input data set (where N_i may be less than N), and returns a $K \times K$ symmetric matrix of second derivatives of the objection function with respect to the parameters.

The pointer to this procedure is stored in the global variable **_cml_HessProc**.

In practice, unfortunately, much of the time spent on writing the Hessian procedure is devoted to debugging. To help in this debugging process, **CML** can be instructed to compute the numerical Hessian along with your prospective analytical Hessian for comparison purposes. To accomplish this **_cml_GradCheckTol** is set to a small nonzero value.

```
    library cml;
    #include cml.ext;

    proc lnlk(b,z);
        local dev,s2;
        dev = z[.,1] - b[1] * exp(-b[2]*z[.,2]);
        s2 = dev'dev/rows(dev);
        retp(-0.5*(dev.*dev/s2 + ln(2*pi*s2)));
    endp;

    proc grdlk(b,z);
        local d,s2,dev,r;
        d = exp(-b[2]*z[.,2]);
        dev = z[.,1] - b[1]*d;
        s2 = dev'dev/rows(dev);
        r = dev.*d/s2;
/*      retp(r~(-b[1]*z[.,2].*r));        correct gradient */
        retp(r~(z[.,2].*r));         /* incorrect gradient */
    endp;
```

```
proc hslk(b,z);
    local d,s2,dev,r, hss;
    d = exp(-b[2]*z[.,2]);
    dev = z[.,1] - b[1]*d;
    s2 = dev'dev/rows(dev);
    r = z[.,2].*d.*(b[1].*d - dev)/s2;
    hss = -d.*d/s2~r~-b[1].*z[.,2].*r;
    retp(xpnd(sumc(hss)));
endp;

cmlset;
_cml_HessProc = &hslk;
_cml_GradProc = &grdlk;
_cml_Bounds = { 0 10, 0 10 };   /* constrain parameters to */
                                /* be positive             */
_cml_GradCheckTol = 1e-3;

startv = { 2, 1 };

{ x,f0,g,cov,retcode } = CML("nlls",0,&lnlk,startv);
call CMLPrt(x,f0,g,cov,retcode);
```

The gradient is incorrectly computed, and **CML** responds with an error message. It is clear that the error is in the calculation of the gradient for the second parameter.

```
analytical and numerical gradients differ
       numerical       analytical
    -0.015387035     -0.015387035
     0.031765317     -0.015882659

================================================================================
              analytical Hessian and analytical gradient
================================================================================
 CML Version 2.0.0                                  02/08/2001  10:10 am
================================================================================
                            Data Set:  nlls
--------------------------------------------------------------------------------

return code =    7
function cannot be evaluated at initial parameter values

Mean log-likelihood         1.12119
Number of cases     150

The covariance of the parameters failed to invert
```

```
Parameters     Estimates      Gradient
-----------------------------------------------
P01            2.000000     -0.015387
P02            1.000000     -0.015883

Number of iterations             .
Minutes to convergence           .
```

2.7.5 User-Supplied Numerical Hessian

You may substitute your own numerical Hessian procedure for the one used by **CML** by default. This done by setting the **CML** global, **_cml_UserHess** to a pointer to the procedure. This procedure has three input arguments, a pointer to the log-likelihood function, a $K \times 1$ vector of parameters, and an $N_i \times K$ matrix containing the data. It must return a K $\times$ K matrix which is the estimated Hessian evaluated at the parameter vector.

2.7.6 Analytical Nonlinear Constraint Jacobians

When nonlinear equality or inequality constraints have been placed on the parameters, the convergence can be improved by providing a procedure for computing their Jacobians, i.e., $\dot{G}(\theta) = \partial G(\theta)/\partial\theta$ and $\dot{H}(\theta) = \partial H(\theta)/\partial\theta$.

These procedures have one argument, the $K \times 1$ vector of parameters, and return an $M_j \times K$ matrix, where M_j is the number of constraints computed in the corresponding constraint function. Then the **CML** globals, **_cml_EqJacobian** and **_cml_IneqJacobian** are set to pointers to the nonlinear equality and inequality Jacobians, respectively. For example,

```
library cml;
#include cml.ext;
cmlset;

proc lpr(c,x);       /* ordinary least squares model */
    local s,t;
    t = x[.,1] - x[.,2:4]*c[1:3];
    s = c[4];
    retp( -(t.*t)./(2*s)-.5*ln(2*s*pi)) );
endp;

proc ineq(c);         /* constrain parameter norm to be > 1 */
    local b;
    b = c[1:3];
```

```
    retp(b'b - 1);
endp;

proc ineqj(c);        /* constraint Jacobian */
    local b;
    b = c[1:3];
    retp((2*b')~0);
endp;

_cml_Bounds = { -1e256 1e256,   /* bound residual to be larger */
                -1e256 1e256,   /* than a small number         */
                -1e256 1e256,
                  1e-3 1e256 };

_cml_IneqProc = &ineq;           /* set pointers */
_cml_IneqJacobian = &ineqj;

x0 = { 1, 1, 1, 1 };         /* Y and X defined elsewhere */
{ x,f,g,cov,ret } = CMLPrt(CML(Y~X,0,&lpr,x0));
```

2.8 Inference

CML includes four broad classes of methods for analyzing the distributions of the estimated parameters:

- Taylor Series covariance matrix of the parameters. This includes two types: the inverted Hessian and the heteroskedastic- consistent covariance matrix computed from both the Hessian and the cross-product of the first derivatives.
- Confidence limits computed by inversion of the Wald and likelihood ratio statistics that take into account constraints
- Bootstrap with additional procedures for kernel density plots, histograms, surface plots, and confidence limits
- Likelihood profile and profile t traces

CML computes a Taylor-series covariance matrix of the parameters that includes the sampling distributions of the Lagrangean coefficients. However, when the model includes inequality constraints, confidence limits computed from the usual t-statistics, i.e., by simply dividing the parameter estimates by their standard errors, are incorrect because they do not account for boundaries placed on the distributions of the parameters by the inequality constraints.

The likelihood ratio statistic becomes a mixture of chi-squared distributions in the region of constraint boundaries (Gourieroux et al., 1982). For a parameter of interest only in the region of a boundary this is a simple mixture which can be corrected by **CML**. If all other parameters near constraint boundaries are correlated less than about .6 with the parameter of interest, then statistical inference is mostly unaffected (Schoenberg, 1997). There is unfortunately no method for the correction of confidence limits if the converse is true.

For correct statistical inference, provided the condition just described hold, use either **CMLClimits** which computes confidence limits by inversion of the Wald statistic, or **CMLPflClimits** which does the same by inversion of the likelihood ratio statistic. Both of these compute limits correctly that are in the region of constraint boundaries provided there are no other parameters in the model near constraint boundaries with which the parameter of interest is correlated more than .6.

If there are no parameters with limits near constraint boundaries, bootstrapping will suffice. Taylor-series methods assume that it is reasonable to truncate the Taylor-series approximation to the distribution of the parameters at the second order. If this is not reasonable, bootstrapping is an alternative not requiring this assumption. It is important to note that if the limit of the parameter of interest or any other parameters with which it is correlated more than .6 are near constraint boundaries, then bootstrapping will not produce correct inference (Andrews, 1999).

The procedure **CMLBoot** generates the mean vector and covariance matrix of the bootstrapped parameters, and **CMLHist** which produces histograms and surface plots. The likelihood profile and profile t traces explicated by Bates and Watts (1988) provide diagnostic material for evaluating parameter distributions. **CMLProfile** generates trace plots which are used for this evaluation.

2.8.1 Covariance Matrix of the Parameters

An argument based on a Taylor-series approximation to the likelihood function (e.g., Amemiya, 1985, page 111) shows that

$$\hat{\theta} \rightarrow N(\theta, A^{-1}BA^{-1})$$

where

$$\begin{aligned} A &= E\left[\frac{\partial^2 L}{\partial\theta\partial\theta'}\right] \\ B &= E\left[\left(\frac{\partial L}{\partial\theta}\right)'\left(\frac{\partial L}{\partial\theta}\right)\right] \end{aligned}$$

Estimates of A and B are

$$\begin{aligned}\hat{A} &= \frac{1}{N}\sum_{i}^{N}\frac{\partial^2 L_i}{\partial\theta\partial\theta'} \\ \hat{B} &= \frac{1}{N}\sum_{i}^{N}\left(\frac{\partial L_i}{\partial\theta}\right)'\left(\frac{\partial L_i}{\partial\theta}\right)\end{aligned}$$

Assuming the correct specification of the model plim(A) = plim(B) and thus

$$\hat{\theta} \rightarrow N(\theta, \hat{A}^{-1})$$

Without loss of generality we may consider two types of constraints, the nonlinear equality and the nonlinear inequality constraints (the linear constraints are included in nonlinear, and the bounds are regarded as a type of linear inequality). Furthermore, the inequality constraints may be treated as equality constraints with the introduction of "slack" parameters into the model:

$$H(\theta) \geq 0$$

is changed to

$$H(\theta) = \zeta^2$$

where ζ is a conformable vector of slack parameters.

Further distinguish *active* from *inactive* inequality constraints. Active inequality constraints have nonzero Lagrangeans, γ_j, and zero slack parameters, ζ_j, while the reverse is true for inactive inequality constraints. Keeping this in mind, define the diagonal matrix, Z, containing the slack parameters, ζ_j, for the inactive constraints, and another diagonal matrix, Γ, containing the Lagrangean coefficients. Also, define $H_{\oplus}(\theta)$ representing the active constraints, and $H_{\ominus}(\theta)$ the inactive.

The likelihood function augmented by constraints is then

$$\begin{aligned}L_A = L + \lambda_1 g(\theta)_1 + \cdots + \lambda_I g(\theta)^I + \gamma_1 h_{\oplus 1}(\theta) + \cdots + \gamma_J h_{\oplus J}(\theta) + \\ h_{\ominus 1}(\theta)_i - \zeta_1^2 + \cdots + h_{\ominus K}(\theta) - \zeta_K^2\end{aligned}$$

and the Hessian of the augmented likelihood is

$$E(\frac{\partial^2 L_A}{\partial\theta\partial\theta'}) = \begin{bmatrix} \Sigma & 0 & 0 & \dot{G}' & \dot{H}'_{\oplus} & \dot{H}'_{\ominus} \\ 0 & 2\Gamma & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2Z \\ \dot{G} & 0 & 0 & 0 & 0 & 0 \\ \dot{H}_{\oplus} & 0 & 0 & 0 & 0 & 0 \\ \dot{H}_{\ominus} & 0 & 2Z & 0 & 0 & 0 \end{bmatrix}$$

where the dot represents the Jacobian with respect to θ, $L = \sum_{i=1}^{N} \log P(Y_i; \theta)$, and $\Sigma = \partial^2 L / \partial\theta\partial\theta'$. The covariance matrix of the parameters, Lagrangeans, and slack parameters is the Moore-Penrose inverse of this matrix. Usually, however, we are interested only in the covariance matrix of the parameters, as well as the covariance matrices of the Lagrange coefficients associated with the active inequality constraints and the equality constraints.

These matrices may be computed without requiring the storage and manipulation of the entire Hessian. Construct the partitioned array

$$\tilde{B} == \begin{bmatrix} \dot{G} \\ \dot{H}_{\oplus} \\ \dot{H}_{\ominus} \end{bmatrix}$$

and denote the i-th row of $\tilde{B}$ as $\tilde{b}_i$. Then the $k \times k$ upper left portion of the inverse, that is, that part associated with the estimated parameters, is calculated recursively. First, compute

$$\Omega_1 = \Sigma^{-1} - \frac{1}{\tilde{b}_1 \Sigma^{-1} \tilde{b}_1'} \Sigma^{-1} \tilde{b}_1' \tilde{b}_1 \Sigma^{-1}$$

then continue to compute for all rows of $\tilde{B}$:

$$\Omega_i = \Omega_{i-1} - \frac{1}{\tilde{b}_i \Omega_{i-1} \tilde{b}_i'} \Omega_{i-1} \tilde{b}_i' \tilde{b}_i \Omega_{i-1}$$

Rows associated with the inactive inequality constraints in $\tilde{B}$, i.e., with $\dot{H}_{\ominus}$, drop out and therefore they need not be considered.

Standard errors for some parameters associated with active inequality constraints may not be available, i.e., the rows and columns of Ω associated with those parameters may be all zeros.

2.8.2 Testing Constraints

Equality Constraints

When equality constraints are present in the model, their associated Lagrange coefficients may be tested to determine their reasonableness. An estimate of the covariance matrix of the joint distribution of the Lagrange coefficients associated with the equality constraints is $\dot{G}\Sigma^{-1}\dot{G}'$ and therefore

$$\hat{\lambda}' \dot{G} \Sigma^{-1} \dot{G}' \hat{\lambda}$$

is asymptotically $\chi^2(p)$ where p is the length of $\hat{\lambda}$. Individual constraints may be tested using their associated t-statistics.

When appropriate, **CML** inserts $\dot{G}\Sigma^{-1}\dot{G}'$ as "eqcov" in the global, **__cml_Lagrange**, using the **GAUSS VPUT** command.

Active Inequality Constraints

When inequality constraints are active, their associated Lagrange coefficients are nonzero. The expected value of their Lagrange coefficients is zero (assuming correct specification of the model), and they are active only in occasional samples. How many samples this occurs in depends on their covariance matrix, which is estimated by $\dot{H}_{\oplus}\Sigma^{-1}\dot{H}'_{\oplus}$.

When appropriate **CML** inserts $\dot{H}_{\oplus}\Sigma^{-1}\dot{H}_{\oplus}$ as "ineqcov" in the global, **_cml_Lagrange**, using the **GAUSS VPUT** command.

2.8.3 Heteroskedastic-consistent Covariance Matrix

When **_cml_CovPar** is set to 3, **CML** returns heteroskedastic-consistent covariance matrices of the parameters, and as well as the corresonding heteroskedastic-consistent covariance matrices of the Lagrange coefficients in the global, **_cml_Lagrange**.

Define

$$B = \left(\frac{\partial F}{\partial \theta}\right)' \left(\frac{\partial L}{\partial \theta}\right)$$

evaluated at the estimates. Then the covariance matrix of the parameters is

$$\Omega B \Omega$$

2.8.4 Confidence Limits by Inversion

When the model includes inequality constraints, confidence limits computed as the ratio of the parameter estimate to its standard error are not correct because they do not take into account that the distribution of the parameter is restricted by its boundaries.

Inversion of the Likelihood Ratio Statistic. Partition a k-vector of parameters, $\theta = (\theta_1\ \theta_2)$, and let $\tilde{\theta}$ be a maximum likelihood estimate of θ, where θ_1 is fixed to some value. A $100(1-\alpha)\%$ confidence region for the parameters in θ_1 is defined by

$$-2 * log(L(\tilde{\theta})/L(\hat{\theta})) \leq \chi^2_{(1-\alpha,k)}.$$

Let

$$F_{lr}(\phi) = \min(-2 * log(L(\tilde{\theta})/L(\hat{\theta})) \mid \eta_i'\theta = \phi, H(\theta) \geq 0)$$

where η is a vector with a one in the i-th position and zeros elsewhere, and $H(\theta)$ is a function describing the constraints. The lower limit of the $(1-\alpha)$ interval for θ_i is the value of ϕ such that

$$F_{lr}(\phi) = \chi^2_{(1-\alpha,k)}. \tag{2.1}$$

A modified secant method is used to find the value of ϕ that satisfies (2.1). The upper limit is found by defining F_{lr} as a maximum.

The **CML** procedure **CMLPflClimits** solves this problem. Corrections are made by **CMLPflClimits** when the limits are near constraint boundaries.

Inversion of the Wald Statistic. A $1-\alpha$ joint confidence region for θ is the hyper-ellipsoid

$$JF(J, N-K; \alpha) = (\theta - \hat{\theta})' V^{-1} (\theta - \hat{\theta}) \tag{2.2}$$

where V is the covariance matrix of the parameters, J is the number of parameters involved in the hypothesis, and $F(J, N-K; \alpha)$ is the upper α area of the F-distribution with J, N-K degrees of freedom.

If there are no constraints in the model, the $1-\alpha$ confidence interval for any selected parameter is

$$\hat{\theta} \pm \sqrt{\eta_k' V^{-1} \eta_k}\; t(N-K; \alpha/2)$$

where η_k is a vector of zeros with the k-th element corresponding to the parameter being tested set to one.

When there are constraints no such simple description of the interval is possible. Instead it is necessary to state the confidence limit problem as a parametric nonlinear programming problem.

The lower limit of the confidence limit is the solution to

$$\min \left\{ \eta_k' \theta \mid (\theta - \hat{\theta})' V^{-1} (\theta - \hat{\theta}) \geq JF(J, N-K; \alpha), G(\theta) = 0, H(\theta) \geq 0) \right\}$$

where now η can be an arbitrary vector of constants and $J = \sum \eta_k \neq 0$, and where again we have assumed that the linear constraints and bounds have been folded in among nonlinear constraints. The upper limit is the maximum of the this same function.

In this form, the minimization is not convex and can't be solved by the usual methods. However, the problem can be re-stated as a parametric nonlinear programming problem (Rust and Burrus, 1972). Define the function

$$F(\phi) = \min((\theta - \hat{\theta})' V^{-1} (\theta - \hat{\theta}) \mid \eta_k' \theta = \phi, G(\theta) = 0, H(\theta) \geq 0)$$

The upper and lower limits of the $1 - \alpha$ interval are the values of ϕ such that

$$F(\phi) = JF(J, N - K; \alpha)$$

To find this value it is necessary to iteratively refine ϕ by interpolation until 2.8.4 is satisfied. The **CML** procedure **CMLClimits** solves this problem.

Corrections are made by **CMLClimits** when the parameter is near a constraint boundary.

2.8.5 Bootstrap

The bootstrap method is used to generate empirical distributions of the parameters, thus avoiding the difficulties with the usual methods of statistical inference described above.

CMLBoot

Rather than randomly sample with replacement from the data set, **CMLBoot** performs **_cml_NumSample** weighted maximum likelihood estimations where the weights are Poisson pseudo-random numbers with expected value equal to the the number of observations. This is asymptotically equivalent to simple random sampling with replacement. **_cml_NumSample** is set by the **CMLBoot** global variable. The default is 50 re-samplings. Efron and Tibshirani (1993:52) suggest that 100 is satisfactory, 50 is often enough to give a good estimate, and rarely are more than 200 needed.

The mean and covariance matrix of the bootstrapped parameters is returned by **CMLBoot**. In addition **CMLBoot** writes the bootstrapped parameter estimates to a **GAUSS** data set for use with **CMLHist**, which produces histograms and surface plots, **CMLDensity**, which produces kernel density plots, and **CMLBlimits**, which produces confidence limits based on the bootstrapped coefficients. The data set name can be specified by the user in the global **_cml_BootFname**. However, if not specified, **CMLBoot** selects the name BOOTxxxx, where xxxx starts at 0000 and increments by 1 until a name is found that is not already in use.

CMLDensity

CMLDensity is a procedure for computing kernel type density plots. The global, **_cml_Kernel** permits you to select from a variety of kernels, normal, Epanechnikov, biweight, triangular, rectangular, and truncated normal. For each selected parameter, a plot is generated of a smoothed density. The smoothing coefficients may be specified using the global, **_cml_Smoothing**, or **CMLDensity** will compute them.

CMLHist

CMLHist is a procedure for visually displaying the results of the bootstrapping in univariate histograms and bivariate surface plots for selected parameters. The univariate discrete distributions of the parameters used for the histograms are returned by **CMLHist** in a matrix.

Example

To bootstrap the example in Section 2.6.6, the only necessary alteration is the change the call to **CML** to a call to **CMLBoot**:

```
_cml_BootFname = "bootdata";

call CMLPrt(cmlboot("tobit",0,&lpr,x0));

call CMLDensity("bootdata",0);
call CMLHist("bootdata",0);
```

2.8.6 Profiling

The **CML** proc, **CMLProfile** generates profile t plots as well as plots of the likelihood profile traces for all of the parameters in the model in pairs. The profile t plots are used to assess the nonlinearity of the distributions of the individual parameters, and the likelihood profile traces are used to assess the bivariate distributions. The input and output arguments to **CMLProfile** are identical to those of **CML**. But in addition to providing the maximum likelihood estimates and covariance matrix of the parameters, a series of plots are printed to the screen using **GAUSS**' Publication Quality Graphics. A screen is printed for each possible pair of parameters. There are three plots, a profile t plot for each parameter, and a third plot containing the likelihood profile traces for the two parameters.

The discussion in this section is based on Bates and Watts (1988), pages 205-216, which is recommended reading for the interpretation and use of profile t plots and likelihood profile traces.

The Profile t Plot

Define

$$\tilde{\theta_k} = (\tilde{\theta}_1, \tilde{\theta}_2, ..., \tilde{\theta}_{k-1}, \theta_k, \tilde{\theta}_{k+1}, ..., \tilde{\theta}_K)$$

This is the vector of maximum likelihood estimates *conditional* on θ_k, i.e., where θ_k is fixed to some value. Further define the profile t function

$$\tau(\theta_k) = sign(\theta_k - \hat{\theta}_k)(N-K)\sqrt{2N\left[L(\tilde{\theta}_k) - L(\hat{\theta}_k)\right]}$$

For each parameter in the model, τ is computed over a range of values for θ_k. These plots provide exact likelihood intervals for the parameters, and reveal how nonlinear the estimation is. For a linear model, τ is a straight line through the origin with unit slope. For nonlinear models, the amount of curvature is diagnostic of the nonlinearity of the estimation. High curvature suggests that the usual statistical inference using the t-statistic is hazardous.

The Likelihood Profile Trace

The likelihood profile traces provide information about the bivariate likelihood surfaces. For nonlinear models the profile traces are curved, showing how the parameter estimates affect each other and how the projection of the likelihood contours onto the (θ_k, θ_ℓ) plane might look. For the (θ_k, θ_ℓ) plot, two lines are plotted, $L(\tilde{\theta}_k)$ against θ_k and $L(\tilde{\theta}_\ell)$ against θ_ℓ.

If the likelihood surface contours are long and thin, indicating the parameters to be collinear, the profile traces are close together. If the contours are fat, indicating the parameters to be more uncorrelated, the profile traces tend to be perpendicular. And if the contours are nearly elliptical, the profile traces are straight. The surface contours for a linear model would be elliptical and thus the profile traces would be straight and perpendicular to each other. Significant departures of the profile traces from straight, perpendicular lines, therefore, indicate difficulties with the usual statistical inference.

To generate profile t plots and likelihood profile traces from the example in Section 2.6.6, it is necessary only to change the call to **CML** to a call to **CMLProfile**:

```
call CMLPrt(cmlprofile("tobit",0,&lpr,x0));
```

CMLProfile produces the same output as **CML** which can be printed out using a call to **CMLPRT**.

For each pair of parameters a plot is generated containing an xy plot of the likelihood profile traces of the two parameters, and two profile t plots, one for each parameter.

The likelihood profile traces indicate that the distributions of parameters 1 and 2 are highly correlated. Ideally, the traces would be perpendicular and the trace in this example is far from ideal.

The profile t plots indicate that the parameter distributions are somewhat nonlinear. Ideally the profile t plots would be straight lines and this example exhibits significant nonlinearity. It is clear that any interpretations of the parameters of this model must be made quite carefully.

2.9 Run-Time Switches

If the user presses **Alt-H** during the iterations, a help table is printed to the screen which describes the run-time switches. By this method, important global variables may be modified during the iterations.

On most systems the Alt key can be ignored, especially if the Alt key is under control by the operating system for some other purpose. The case may also be ignored, that is, either upper or lower case letters suffice.

Alt-A	Change Algorithm
Alt-C	Force Exit
Alt-E	Edit Parameter Vector
Alt-G	Toggle **_cml_GradMethod**
Alt-H	Help Table
Alt-I	Compute Hessian
Alt-K	toggle Grid Search
Alt-L	set **_cml_GridRadius**
Alt-M	Maximum Tries
Alt-N	Set **_cml_GradOrder**
Alt-O	Toggle **__output**
Alt-R	Set **_cml_CovPar**, type of covariance matrix
Alt-S	Set line search Method
Alt-T	toggle Trust Region Method
Alt-U	set **_cml_TrustRadius**
Alt-V	Set **_cml_DirTol**

The algorithm may be switched during the iterations either by pressing **Alt-A**, or by pressing one of the following:

Alt-1	Broyden-Fletcher-Goldfarb-Shanno (BFGS)
Alt-2	Davidon-Fletcher-Powell (DFP)
Alt-3	Newton-Raphson (NEWTON) or (NR)
Alt-4	Berndt, Hall, Hall & Hausman (BHHH)
Alt-5	scaled BFGS
Alt-5	scaled DFP

The line search method may be switched during the iterations either by pressing **Alt-S**, or by pressing one of the following:

Shift-1	no search (1.0 or 1 or ONE)
Shift-2	cubic or quadratic method (STEPBT)
Shift-3	step halving method (HALF)
Shift-4	Brent's method (BRENT)
Shift-5	BHHH step method (BHHHSTEP)

Keyboard polling can be turned off completely by setting the global **_cml_key** to zero.

2.10 Error Handling

2.10.1 Return Codes

The fourth argument in the return from **CML** contains a scalar number that contains information about the status of the iterations upon exiting **CML**. The following table describes their meanings:

0	normal convergence
1	forced exit
2	maximum iterations exceeded
3	function calculation failed
4	gradient calculation failed
5	Hessian calculation failed
6	line search failed
7	function cannot be evaluated at initial parameter values
8	error with gradient
9	error with constraints
10	secant update failed
11	maximum time exceeded
12	error with weights
13	quadratic program failed
14	equality Jacobian failed
15	inequality Jacobian failed
20	Hessian failed to invert
34	data set could not be opened
35	number of observations not set
99	termination condition unknown

2.10.2 Error Trapping

Setting the global **__output** = 0 turns off all printing to the screen. Error codes, however, still are printed to the screen unless error trapping is also turned on. Setting the trap flag to 4 causes **CML** to *not* send the messages to the screen:

```
trap 4;
```

Whatever the setting of the trap flag, **CML** discontinues computations and returns with an error code. The trap flag in this case only affects whether messages are printed to the screen or not. This is an issue when the **CML** function is embedded in a larger program, and you want the larger program to handle the errors.

2.11 References

Andrews, D.W.K, 1999. "Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space", *Econometrica*,99.

Amemiya, Takeshi, 1985. *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Bates, Douglas M. and Watts, Donald G., 1988. *Nonlinear Regression Analysis and Its Applications*. New York: John Wiley & Sons.

Berndt, E., Hall, B., Hall, R., and Hausman, J. 1974. "Estimation and inference in nonlinear structural models". *Annals of Economic and Social Measurement* 3:653-665.

Brent, R.P., 1972. *Algorithms for Minimization Without Derivatives*. Englewood Cliffs, NJ: Prentice-Hall.

Dennis, Jr., J.E., and Schnabel, R.B., 1983. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Englewood Cliffs, NJ: Prentice-Hall.

Efron, Gradley, Robert J. Tibshirani, 1993. *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Fletcher, R., 1987. *Practical Methods of Optimization*. New York: Wiley.

Gill, P. E. and Murray, W. 1972. "Quasi-Newton methods for unconstrained optimization." *J. Inst. Math. Appl.*, 9, 91-108.

Gourieroux, Christian, Holly, Alberto, and Monfort, Alain (1982). "Likelihood Ratio test, Wald Test, and Kuhn-Tucker test in linear models with inequality constraints on the regression parameters", *Econometrica* 50: 63-80.

Han, S.P., 1977. "A globally convergent method for nonlinear programming." *Journal of Optimization Theory and Applications*, 22:297-309.

Hock, Willi and Schittkowski, Klaus, 1981. *Lecture Notes in Economics and Mathematical Systems*. New York: Springer-Verlag.

Jamshidian, Mortaza and Bentler, P.M., 1993. "A modified Newton method for constrained estimation in covariance structure analysis." *Computational Statistics & Data Analysis*, 15:133-146.

Judge, G.G., R.C. Hill, W.E. Griffiths, H. Lütkepohl and T.C. Lee. 1988. *Introduction to the Theory and Practice of Econometrics.* 2nd Edition. New York:Wiley.

Judge, G.G., W.E. Griffiths, R.C. Hill, H. Lütkepohl and T.C. Lee. 1985. *The Theory and Practice of Econometrics.* 2nd Edition. New York:Wiley.

Schoenberg, Ronald, 1997. "Constrained Maximum Likelihood". *Computational Economics*, 1997:251-266.

White, H. 1981. "Consequences and detection of misspecified nonlinear regression models." *Journal of the American Statistical Association* 76:419-433.

White, H. 1982. "Maximum likelihood estimation of misspecified models." *Econometrica* 50:1-25.

Chapter 3

Constrained Maximum Likelihood Reference

Purpose

Computes estimates of parameters of a constrained maximum likelihood function.

Library

`cml`

Format

{ *x*,*f*,*g*,*cov*,*retcode* } = **CML**(*dataset*,*vars*,&*fct*,*start*)

Input

dataset string containing name of **GAUSS** data set
– or –
$N \times NV$ matrix, data
If there is no data or if the data is to be handled outside of **CML**, a scalar zero or a scalar missing value may be entered.

vars $NV \times 1$ character vector, labels of variables selected for analysis
– or –
$NV \times 1$ numeric vector, indices of variables selected for analysis.
If *dataset* is a matrix, *vars* may be a character vector containing either the standard labels created by **CML** (i.e., either V1, V2,..., or V01, V02,..... See discussion of the global variable **__vpad** below, or the user-provided labels in **__altnam**).

&*fct* a pointer to a procedure that returns either the log-likelihood for one observation or a vector of log-likelihoods for a matrix of observations (see discussion of the global variable **__row** in global variable section below).
If this function returns a scalar value and ¨row /= 1, the BHHH descent method and QML standard errors will not be available. It is also necessary to set **_cml_NumObs** to the number of observations.

start $K \times 1$ vector, start values.

Output

x $K \times 1$ vector, estimated parameters

f scalar, function at minimum (the mean log-likelihood)

g $K \times 1$ vector, gradient evaluated at *x*

h $K \times K$ matrix, covariance matrix of the parameters (see discussion of the global variable **_cml_CovPar** below).

retcode scalar, return code. If normal convergence is achieved, then *retcode* = 0, otherwise a positive integer is returned indicating the reason for the abnormal termination:

0 normal convergence
1 forced exit.
2 maximum iterations exceeded.
3 function calculation failed.
4 gradient calculation failed.
5 Hessian calculation failed.
6 line search failed.
7 function cannot be evaluated at initial parameter values.
8 error with gradient
9 error with constraints
10 secant update failed
11 maximum time exceeded
12 error with weights
13 quadratic program failed
20 Hessian failed to invert
34 data set could not be opened.
35 number of observations not set.
99 termination condition unknown.

Globals

The globals variables used by **CML** can be organized in the following categories according to which aspect of the optimization they affect:

Options **_cml_Options**

Constraints **_cml_A**, **_cml_B**, **_cml_C**, **_cml_D**, **_cml_EqProc**, **_cml_IneqProc**, **_cml_EqJacobian**, **_cml_IneqJacobian**, **_cml_Bounds**, **_cml_Lagrange**

Descent and Line Search **_cml_Algorithm**, **_cml_Delta**, **_cml_LineSearch**, **_cml_Maxtry**, **_cml_Extrap**, **_cml_Interp**, **_cml_UserSearch** **_cml_Switch**, **_cml_Trust**, **_cml_TrustRadius**, **_cml_GridsSearch**, **_cml_GridSearchRadius**, **_cml_FeasibleTest**

Covariance Matrix of Parameters **_cml_CovPar**, **_cml_XprodCov**, **_cml_HessCov**, **_cml_FinalHess**

Gradient **_cml_GradMethod**, **_cml_GradProc**, **_cml_UserNumGrad**, **_cml_HessProc**, **_cml_UserNumHess**, **_cml_GradStep**, **_cml_GradCheckTol** **_cml_GradOrder**

Termination Conditions **_cml_DirTol**, **_cml_DFTol**, **_cml_MaxIters**, **_cml_MaxTime**

Data **_cml_Lag**, **_cml_NumObs**, **__weight**, **__row**, **__rowfac**

Parameters **_cml_Active**, **_cml_ParNames**

Miscellaneous **__title**, **_cml_IterData**, **_cml_Diagnostic** **_cml_State** **_cml_Alpha** **_cml_Key**

The list below contains an alphabetical listing of each global with a complete description.

_cml_A $M_1 \times K$ matrix, linear equality constraint coefficient matrix **_cml_A** is used with **_cml_B** to specify linear equality constraints:

```
_cml_A * X = _cml_B
```

where X is the $K \times 1$ unknown parameter vector.

_cml_Alpha scalar, sets alpha level for statistical inference. Default = .05.

_cml_Active vector, defines fixed/active coefficients. This global allows you to fix a parameter to its starting value. This is useful, for example, when you wish to try different models with different sets of parameters without having to re-edit the function. When it is to be used, it must be a vector of the same length as the starting vector. Set elements of **_cml_Active** to 1 for an active parameter, and to zero for a fixed one.

_cml_Algorithm scalar, selects optimization method:

1	BFGS - Broyden, Fletcher, Goldfarb, Shanno method
2	DFP - Davidon, Fletcher, Powell method
3	NEWTON - Newton-Raphson method
4	BHHH - Berndt, Hall, Hall, Hausman method

Default = 3

_cml_Delta scalar, floor for eigenvalues of Hessian in the NEWTON algorithm. When nonzero, the eigenvalues of the Hessian are augmented to this value.

_cml_B $M_1 \times 1$ vector, linear equality constraint constant vector **_cml_B** is used with **_cml_A** to specify linear equality constraints:

```
_cml_A * X = _cml_B
```

where X is the $K \times 1$ unknown parameter vector.

_cml_Bounds $K \times 2$ matrix, bounds on parameters. The first column contains the lower bounds, and the second column the upper bounds. If the bounds for all the coefficients are the same, a 1x2 matrix may be used. Default = { -1e256 1e256 }.

_cml_C $M_3 \times K$ matrix, linear inequality constraint coefficient matrix **_cml_C** is used with **_cml_D** to specify linear inequality constraints:

```
_cml_C * X = _cml_D
```

where X is the $K \times 1$ unknown parameter vector.

_cml_CovPar scalar, type of covariance matrix of parameters

0	not computed
1	computed from inverse of Hessian calculated after the iterations
2	computed from cross-product of Jacobian after the iterations
3	heteroskedastic-consistent covariance matrix of the parameters

Default = 1;

_cml_D $M_3 \times 1$ vector, linear inequality constraint constant vector **_cml_D** is used with **_cml_C** to specify linear inequality constraints:

```
_cml_C * X = _cml_D
```

where X is the $K \times 1$ unknown parameter vector.

_cml_DFTol scalar. Iterations are halted when the absolute value of the change in the function is less than this amount. Default = 0;

_cml_Diagnostic scalar.

0	nothing is stored or printed
1	current estimates, gradient, direction, function value, Hessian, and step length are printed to the screen
2	the current quantities are stored in **_cml_Diagnostic** using the **VPUT** command. Use the following strings to extract from **_cml_Diagnostic** using **VREAD**:

function	"function"
estimates	"params"
direction	"direct"
Hessian	"hessian"
gradient	"gradient"
step	"step"

When **_cml_Diagnostic** is nonzero, **__output** is forced to 1.

_cml_DirTol scalar, convergence tolerance for gradient of estimated coefficients. When this criterion has been satisifed CML exits the iterations. Default = 1e-5.

_cml_EqJacobian scalar, pointer to a procedure that computes the Jacobian of the nonlinear equality constraints with respect to the parameters. The procedure has one input argument, the $K \times 1$ vector of parameters, and one output argument, the $M_2 \times 1$ vector of derivatives of the constraints with respect to the parameters. For example, if the nonlinear equality constraint procedure was,

```
proc eqproc(p);
    retp(p[1]*p[2]-p[3]);
endp;
```

the Jacobian procedure and assignment to the global would be,

```
proc eqj(p);
    retp(p[2]~p[1]~-1);
endp;

_cml_EqJacobian = &eqj;
```

_cml_EqProc scalar, pointer to a procedure that computes the nonlinear equality constraints. For example, the statement:

```
_cml_EqProc = &eqproc;
```

tells **CML** that nonlinear equality constraints are to be placed on the parameters and where the procedure computing them is to be found. The procedure must have one input argument, the $K \times 1$ vector of parameters, and one output argument, the $M_2 \times K$ matrix of computed constraints that are to be equal to zero. For example, suppose that you wish to place the following constraint:

```
P[1] * P[2] = P[3]
```

The proc for this is:

```
proc eqproc(p);
    retp(p[1]*[2]-p[3]);
endp;
```

_cml_Extrap scalar, extrapolation constant in BRENT. Default = 2.

_cml_FeasibleTest scalar, if nonzero a test for the feasibility of the parameter vector is tested before calling the log-likelihood function. This is important if feasibility is necessary for the calculation of the log-likelihood, for example if the procedure is computing the log of a parameter which is being constrained to be positive. By default this function is turned on, but if feasibility is not required for calculating the function then set **_cml_FeasibleTest** to zero if the iterations are being halted as a result of the test. Default = 1.

_cml_FinalHess $K \times K$ matrix, the Hessian used to compute the covariance matrix of the parameters is stored in **_cml_FinalHess**. This is most useful if the inversion of the hessian fails, which is indicated when **CML** returns a missing value for the covariance matrix of the parameters. An analysis of the Hessian stored in **_cml_FinalHess** can then reveal the source of the linear dependency responsible for the singularity.

_cml_GradCheckTol scalar. Tolerance for the deviation of numerical and analytical gradients when proc's exist for the computation of analytical gradients, Hessians, and/or Jacobians. If set to zero, the analytical gradients will not be compared to their numerical versions. When adding procedures for computing analytical gradients it is highly recommended that you perform the check. Set **_cml_GradCheckTol** to some small value, 1e-3, say when checking. It may have to be set larger if the numerical gradients are poorly computed to make sure that **CML** doesn't fail when the analytical gradients are being properly computed.

_cml_GradMethod scalar, method for computing numerical gradient.

0 central difference

1 forward difference (default)

_cml_GradOrder scalar, If set to zero, standard two point method is used. If greater than two, sets number of points for computing multiple point numerical gradient.

_cml_GradProc scalar, pointer to a procedure that computes the gradient of the function with respect to the parameters. For example, the statement:

```
_cml_GradProc=&gradproc;
```

tells **CML** that a gradient procedure exists as well where to find it. The user-provided procedure has two input arguments, an $K \times 1$ vector of parameter values and an N×K matrix of data. The procedure returns a single output argument, an $N \times K$ matrix of gradients of the log-likelihood function with respect to the parameters evaluated at the vector of parameter values.

For example, suppose the log-likelihood function is for a Poisson regression, then the following would be added to the command file:

```
proc lgd(b,z);
    retp((z[.,1]-exp(z[.,2:4]*b)).*z[.,2:4]);
endp;

_cml_GradProc = &lgd;
```

Default = 0, i.e., no gradient procedure has been provided.

_cml_GradStep scalar, increment size for computing gradient. When the numerical gradient is performing well, set to a larger value (1e-3, say). Default is the cube root of machine precision.

_cml_GridSearch scalar, if nonzero a grid search for a new direction will be attempted whenever the line search fails. It is nonzero by default.

_cml_GridSearchRadius scalar, set to the radius of the grid search. Default = .01

_cml_HessCov $K \times K$ matrix. When **_cml_CovPar** is set to 3 the information matrix covariance matrix of the parameters, i.e., the inverse of the matrix of second order partial derivatives of the log-likelihood by observations, is returned in **_cml_HessCov**.

_cml_HessProc scalar, pointer to a procedure that computes the hessian, i.e., the matrix of second order partial derivatives of the function with respect to the parameters. For example, the instruction:

```
_cml_HessProc = &hessproc;
```

tells **CML** that a procedure has been provided for the computation of the hessian and where to find it. The procedure that is provided by the user must have two input arguments, a $K \times 1$ vector of parameter values and an N×P data matrix. The procedure returns a single output argument, the $K \times K$ symmetric matrix of second order derivatives of the function evaluated at the parameter values.

_cml_IneqJacobian scalar, pointer to a procedure that computes the Jacobian of the nonlinear equality constraints with respect to the parameters. The procedure has one input argument, the $K \times 1$ vector of parameters, and one output argument, the $M_4 \times K$ matrix of derivatives of the constraints with respect to the parameters. For example, if the nonlinear equality constraint procedure was,

```
proc ineqproc(p);
    retp(p[1]*p[2]-p[3]);
endp;
```

the Jacobian procedure and assignment to the global would be,

```
proc ineqj(p);
    retp(p[2]~p[1]~-1);
endp;

_cml_IneqJacobian = &ineqj;
```

_cml_IneqProc scalar, pointer to a procedure that computes the nonlinear inequality constraints. For example the statement:

```
_cml_IneqProc = &ineqproc;
```

tells **CML** that nonlinear equality constraints are to be placed on the parameters and where the procedure computing them is to be found. The procedure must have one input argument, the $K \times 1$ vector of parameters, and one output argument, the $M_4 \times K$ matrix of computed constraints that are to be equal to zero. For example, suppose that you wish to place the following constraint:

```
P[1] * P[2] >= P[3]
```

The proc for this is:

```
proc ineqproc(p);
    retp(p[1]*[2]-p[3]);
endp;
```

_cml_Interp scalar, interpolation constant in BRENT. Default = .25.

_cml_IterData 3x1 vector, contains information about the iterations. The first element contains the # of iterations, the second element contains the elapsed time in minutes of the iterations, and the third element contains a character variable indicating the type of covariance matrix of the parameters.

_cml_Key scalar, if nonzero, the keyboard is polled for run-time modification of globals governing the interation process (See Section 2.9). Default is nonzero.

_cml_Lag scalar, if the function includes lagged values of the variables **_cml_Lag** may be set to the number of lags. When **_cml_Lag** is set to a nonzero value then **__row** is set to 1 (that is, the function must evaluated one observation at a time), and **CML** passes a matrix to the user-provided function and gradient procedures. The first row in this matrix is the (i - **_cml_Lag**)-th observation and the last row is the i-th observation. The read loop begins with the (**_cml_Lag**+1)-th observation. Default = 0.

_cml_Lagrange vector, created using **VPUT**. Contains the Lagrangean coefficients for the constraints. They may be extracted with the **VREAD** command using the following strings:

"lineq"	linear equality constraints
"nlineq"	nonlinear equality constraints
"linineq"	linear inequality constraints
"nlinineq"	nonlinear inequality constraints
"bounds"	bounds
"eqcov"	covariance matrix of equality Lagrangeans
"ineqcov"	covariance matrix of inequality Lagrangeans
"boundcov"	covariance matrix of bounds' Lagrangeans

When an inequality or bounds constraint is active, its associated Lagrangean is nonzero. The linear Lagrangeans preceed the nonlinear Lagrangeans in the covariance matrices.

_cml_LineSearch scalar, selects method for conducting line search. The result of the line search is a *step length*, i.e., a number which reduces the function value when multiplied times the direction..

1 step length = 1.

2 cubic or quadratic step length method (STEPBT)

3 step halving (HALF)

4 Brent's step length method (BRENT)

5 BHHH step length method (BHHHSTEP)

Default = 2.

Usually **_cml_LineSearch** = 2 is best. If the optimization bogs down, try setting **_cml_LineSearch** = 1, 4 or 5. **_cml_LineSearch** = 3 generates slower iterations but faster convergence and **_cml_LineSearch** = 1 generates faster iterations but slower convergence.

When any of these line search methods fails, **CML** attempts a random search of radius **_cml_RandRadius** times the truncated log to the base 10 of the gradient when **_cml_RandRadius** is set to a nonzero value. If **_cml_UserSearch** is set to 1, **CML** enters an interactive line search mode.

_cml_MaxIters scalar, maximum number of iterations.

_cml_MaxTime scalar, maximum time in iterations in minutes. This global is most useful in bootstrapping. You might want 100 re-samples, but would be happy with anything more than 50 depending on the time it took. Set **_cml_NumSample** = 100, and **_cml_MaxTime** to maximum time you would be willing to wait for results. Default = 1e+5, about 10 weeks.

_cml_MaxTry scalar, maximum number of tries to find step length that produces a descent.

_cml_NumObs scalar, number of observations. By default the number of observations is determined from the dataset and this number is returned in the global after the iterations. If a different number is required for correct statistical inference, set **_cml_NumObs** to this value. It may also be necessary to set this global if a dataset is not passed to **CML** and if the log-likelihood returns a scalar.

_cml_Options character vector, specification of options. This global permits setting various **CML** options in a single global using identifiers. The following

```
_cml_Options = { newton stepbt forward screen };
```

the descent method to NEWTON, the line search method to STEPBT, the numerical gradient method to forward differences, and **__OUTPUT** = 2.

The following is a list of the identifiers:

Algorithms BFGS, DFP, NEWTON, BHHH

Line Search ONE, STEPBT, HALF, BRENT, BHHHSTEP

Covariance Matrix NOCOV, INFO, XPROD, HETCON

Gradient method CENTRAL, FORWARD, BACKWARD

Output method NONE, FILE, SCREEN

__output scalar, determines printing of intermediate results. Generally when **__output** is nonzero, i.e., where there some kind of printing during the iterations, the time of the iterations is degraded.

0	nothing is written
1	serial ASCII output format suitable for disk files or printers
2	output is suitable for screen only. ANSI.SYS must be active.
$\geq$**5**	same as **__output = 1** except that information is printed only every **__output**-th iteration.

When **_cml_Diagnostic** is nonzero, **__output** is forced to 1.

_cml_ParNames $K \times 1$ character vector, parameter labels.

_cml_State scalar,

_cml_Trust scalar, if nonzero, a trust region is imposed on the direction vector in the iterations. The trust region method is helpful when the starting values are poor. By default it is turned off.

__cml__TrustRadius scalar, sets the radius of the trust region method. Default = .001.

__cml__UserNumGrad scalar, pointer to user provided numerical gradient procedure. The instruction

```
_cml_UserNumGrad = &userproc;
```

tells **CML** that a procedure for computing the numerical gradients exists. The user-provided procedure has three input arguments, a pointer to a function that computes the log-likelihood function, a $K \times 1$ vector of parameter values, and an $K \times P$ matrix of data. The procedure returns a single output argument, an $N \times K$ matrix of gradients of each row of the input data matrix with respect to each parameter.

___row scalar, specifies how many rows of the data set are read per iteration of the read loop. See the *REMARKS* Section for a more detailed discussion of how to set up your log-likelihood to handle more than one row of your data set. By default, the number of rows to be read is calculated by **CML**.

___rowfac scalar, "row factor". If **CML** fails due to insufficient memory while attempting to read a **GAUSS** data set, then **___rowfac** may be set to some value between 0 and 1 to read a *proportion* of the original number of rows of the **GAUSS** data set. For example, setting

```
__rowfac = 0.8;
```

causes **GAUSS** to read in 80% of the rows of the **GAUSS** data set that were read when **CML** failed due to insufficient memory.

This global has an affect only when **___row** = 0. Default = 1.

__cml__Switch 4×1 or 4×2 vector,

controls algorithm switching. If 4×1 set

_cml_Switch[1] algorithm number to switch to

_cml_Switch[2] CML switches if functions changes less than this amount

_cml_Switch[3] CML switches if this number of iterations is exceeded.

_cml_Switch[4] CML switches if line search step changes less than this amount

or else if 4×2 set each column as in the 4×1 case with different algorithms in the first row. CML switches between the algorithm in column 1 and column 2.

___title string title of run

__cml__UserNumHess scalar, pointer to user provided numerical Hessian procedure. The instruction

```
_cml_UserHess = &hessproc;
```

tells **CML** that a procedure for computing the numerical Hessian exists. The user-provided procedure three input arguments, a pointer to a function that computes the log-likelihood function, a $K \times 1$ vector of parameter values, and an N×P matrix of data. The procedure returns a single output argument, a $K \times K$ Hessian matrix of the function with respect to the parameters.

_cml_UserSearch scalar, if nonzero and if all other line search methods fail **CML** enters an interactive mode in which the user can select a line search parameter

___weight vector, frequency of observations. By default all observations have a frequency of 1. zero frequencies are allowed. It is assumed that the elements of **___weight** sum to the number of observations.

_cml_XprodCov $K \times K$ matrix. When **_cml_CovPar** is set to 3 the cross-product matrix covariance matrix of the parameters, i.e., the inverse of the cross-product of the first derivatives of the log-likelihood computed by observations, is is returned in **_cml_XprodCov**.

Remarks

Specifying Constraints.

There are five types of constraints: linear equality, linear inequality, nonlinear equality, nonlinear inequality, bounds Linear constraints are specified by initializing the appropriate **CML** globals to known matrices of constants. The linear equality constraint matrices are **_cml_A** and **_cml_B**, and they assume the following relationship with the parameter vector:

```
_cml_A * x = _cml_B
```

where `x` is the parameter vector.

Similarly, the linear *in*equality constraint matrices are **_cml_C** and **_cml_D**, and assume the following relationship with the parameter vector:

```
_cml_C * x >= _cml_D
```

The nonlinear constraints are specified by providing procedures and assigning their pointers to **CML** globals. These procedures take a single argument, the vector of parameters, and return a column vector of evaluations of the constraints at the parameters. Each element of the column vector is a separate constraint.

For example, suppose you wish to constrain the product of the first and third coefficients to be equal to 10, and the squared second and fourth coefficients to be equal to the squared fifth coefficient:

```
proc eqp(x);
    local c;
    c = zeros(2,1);
    c[1] = x[1] * x[3] - 10;
    c[2] = x[2] * x[2] + x[4] * x[4] - x[5] * x[5];
    retp(c);
endp;

_cml_EqProc = &eqp;
```

The nonlinear equality constraint procedure causes **CML** to find estimates for which its evaluation is equal to a conformable vector of zeros.

The nonlinear *in*equality constraint procedure is similar to the equality procedure. **CML** finds estimates for which the evaluation of the procedure is greater than or equal to zero. The nonlinear inequality constraint procedure is assigned to the global **_cml_IneqProc**. For example, suppose you wish to constrain the norm of the coefficients to be greater than one:

```
proc ineqp(x);
    retp(x'x-3);
endp;

_cml_IneqProc = &eqp;
```

Bounds are a type of linear inequality constraint. They are specified separately for computational and notational convenience. To declare bounds on the parameters assign a two column vector with rows equal to the number of parameters to the **CML** global, **_cml_Bounds**. The first column is the lower bounds and the second column the upper bounds. For example,

```
_cml_Bounds = {  0 10,
               -10  0
               -10 20 };
```

If the bounds are the same for all of the parameters, only the first row is required.

Writing the Log-likelihood Function

CML requires a procedure for computing the log-likelihood.

The procedure has two input arguments: first, a vector of parameter values, and second, one or more rows of the data matrix. The output argument is the log-likelihood for the observation or observations in the second argument evaluated at the parameter values in the first argument.

Suppose that the function procedure has been named *pfct*, the following considerations apply:

The format of the procedure is:

logprob **=** *pfct***(***x***,***y***);**

where

x	column vector of parameters of model
y	one or more rows of the data set (if the data set has been transformed, or if *vars* $\neq 0$, i.e., there is selection, then y is a transformed, selected observation). if a scalar zero or scalar missing value was passed to **CML** in its first argument not data will be passed to *pfct* and the user will required to provide the necessary data for computing the log-likelihood. if **__row** $= n$, then n rows of the data set are read at a time if **__row** $= 0$, the maximum number of rows that fit in memory is computed by **CML**.

For most uses, the output from this procedure is a vector of log-likelihoods where the vector has the same number of rows as the input dataset in the second argument. A scalar log-likelihood may also be returned but it is important to keep in mind that in this case certain features of **CML** will not be available such as QML standard errors or bootstrapping. Also if a scalar log-likelihood is returned by this procedure you must also set **_cml_NumObs** to its appropriate value.

CML can be forced, if necessary, to send the data n observation at a time by setting $\_row = n$. By default $\_row = 0$ causing **CML** to send the entire matrix to *pfct* if it is stored entirely in memory, or to compute the maximum number of rows if it is a **GAUSS** data set stored on disk (Note that even if the data starts out in a **GAUSS** data set, **CML** determines whether the data set fits in memory, and if it does, then it reads the data set into an array in memory). If you are getting *insufficient memory* messages, then set **__rowfac** to a positive value less than 1.

Supplying an Analytical GRADIENT Procedure

To decrease the time of computation, the user may provide a procedure for the calculation of the gradient of the log-likelihood. The global variable **_cml_GradProc** must contain the pointer to this procedure. Suppose the name of this procedure is *gradproc*. Then,

g **=** *gradproc***(***x***,***y***);**

where the input arguments are

x	vector of coefficients
y	one or more rows of data set. if a scalar zero or scalar missing value was passed to **CML** in its first argument not data will be passed to *gradproc* and the user will required to provide the necessary data for computing the gradient.

and the output argument is

g	row vector of gradients of log-likelihood with respect to coefficients, or a matrix of gradients (i.e., a Jacobian) if the data passed in y is a matrix (unless **_cml_Lag** ≥ 1 in which case the data passed in y is a matrix of lagged values but a row vector of gradients is passed back in g).

It is important to note that the gradient is row oriented. Thus if the function that computes the log-likelihood returns a scalar value (**__row** = 1), then a row vector of the first derivatives of the log-likelihood with respect to the coefficients must be returned, but if the procedure that computes the log-likelihood returns a column vector, then **_cml_GradProc** must return a matrix of first derivatives in which rows are associated with observations and columns with coefficients.

Providing a procedure for the calculation of the first derivatives also has a significant effect on the calculation time of the Hessian. The calculation time for the numerical computation of the Hessian is a quadratic function of the size of the matrix. For large matrices, the calculation time can be very significant. This time can be reduced to a linear function of size if a procedure for the calculation of analytical first derivatives is available. When such a procedure is available, **CML** automatically uses it to compute the numerical Hessian.

The major problem one encounters when writing procedures to compute gradients and Hessians is in making sure that the gradient is being properly computed. **CML** checks the gradients and Hessian when **_cml_GradCheckTol** is nonzero. **CML** generates both numerical and analytical gradients, and viewing the discrepancies between them can help in debugging the analytical gradient procedure.

Supplying an Analytical HESSIAN Procedure.

Selection of the NEWTON algorithm becomes feasible if the user supplies a procedure to compute the Hessian. If such a procedure is provided, the global variable **_cml_HessProc** must contain a pointer to this procedure. Suppose this procedure is called *hessproc*, the format is

h **=** *hessproc***(**x,y**);**

The input arguments are

x	$K \times 1$ vector of coefficients

y	one or more rows of data set if a scalar zero or scalar missing value was passed to **CML** in its first argument not data will be passed to *hessproc* and the user will required to provide the necessary data for computing the Hessian.

and the output argument is

h	$K \times K$ matrix of second order partial derivatives evaluated at the coefficients in x.

To compare numerical and analytical Hessians set **_cml_GradCheckTol** to a nonzero value.

Supplying Analytical Jacobians of the Nonlinear Constraints.

At each iteration the Jacobians of the nonlinear constraints, if they exist, are computed numerically. This is time-consuming and generates a loss of precision. For models with a large number of inequality constraints a significant speed-up can be achieved by providing analytical Jacobian procedures. The improved accuracy can also have a significant effect on convergence.

The Jacobian procedures take a single argument, the vector of parameters, and return a matrix of derivatives of each constraint with respect to each parameter. The rows are associated with the constraints and the columns with the parameters. The pointer to the nonlinear equality Jacobian procedure is assigned to **_cml_EqJacobian**. The pointer to the nonlinear *in*equality Jacobian procedure is assigned to **_cml_IneqJacobian**.

For example, suppose the following procedure computes the equality constraints:

```
proc eqp(x);
    local c;
    c = zeros(2,1);
    c[1] = x[1] * x[3] - 10;
    c[2] = x[2] * x[2] + x[4] * x[4] - x[5] * x[5];
    retp(c);
endp;
cml_EqProc = &eqp;
```

Then the Jacobian procedure would look like this:

```
proc eqJacob(x);
    local c;
    c = zeros(2,5);
    c[1,1] = x[3];
    c[1,3] = x[1];
```

```
    c[2,2] = 2*x[2];
    c[2,4] = 2*x[4];
    c[3,5] = -2*x[5];
    retp(c);
endp;
_cml_EqJacobian = &eqJacob;
```

The Jacobian procedure for the nonlinear inequality constraints is specified similarly, except that the associated global containing the pointer to the procedure is **_cml_IneqJacobian**.

Source

cml.src

Purpose

Produces kernel Density plots of bootstrapped parameters in **GAUSS** data set

Library

`cml`

Format

cl = **CMLBlimits(***dataset***)**

Input

dataset	string containing name of **GAUSS** data set – or – N×K matrix, data

Output

cl	$K \times 2$ matrix, lower (first column) and upper (second column) confidence limits of the selected parameters

Globals

_cml_Alpha (1 - **_cml_Alpha**)% confidence limits are computed. Default = .05

_cml_Select selection vector for selecting coefficients to be included in profiling, for example

```
_cml_Select = { 1, 3, 4 };
```

selects the 1st, 3rd, and 4th parameters for profiling.

Remarks

CMLBlimits sorts each column of the parameter data set and computes (1-**_cml_Alpha**)% confidence limits by measuring back **_cml_Alpha**/2 times the number of rows from each end of the columns. The confidence limits are the values in those elements. If amount to be measured back from each end of the columns doesn't fall exactly on an element of the column, the confidence limit is interpolated from the bordering elements.

Source

`cmlblim.src`

Purpose

Computes confidence limits by inversion of Wald statistic

Library

`cml`

Format

climits = **CMLClimits**(*b*, *V*)

Input

b	$K \times 1$ vector, parameter estimates
V	$K \times K$ matrix, covariance matrix of parameters in *b*

Output

climits	$K \times 2$ matrix, lower confidence limits in the first column and upper limits in second column

Globals

The **CML** procedure global variables are also applicable.

_cml_Alpha scalar, **CMLClimits** computes (1-**_cml_Alpha**)% confidence intervals, where **_cml_Alpha** varies between zero and one. Default = .05.

Remarks

Confidence limits are computed by inversion of the Wald statistic. For inequality constrained models they are the solutions to a parametric nonlinear programming problem. **CMLClimits** solves this problem given a covariance matrix, the vector of parameter estimates, and given the model constraints.

The calculation of confidence limits for large models can be time consuming. In that case it might be necessary to select parameters for analysis. This can be done using the **CML** global, **_cml_Select**.

The global **_cml_NumObs** must be set. If **CMLClimits** is called immediately after a call to **CML**, **_cml_NumObs** will be set by **CML**.

Source

`cmlclim.src`

Purpose

Computes confidence limits by inversion of the likelihood ratio statistic.

Library

`cml`

Format

pflclimits = **CMLPflClimits(***b*,*f*,*dataset*,*vars*,**&***fct***)**

Input

b	$K \times 1$ vector, parameter estimates
f	scalar, function at minimum (mean log-likelihood)
dataset	string containing name of **GAUSS** data set – or – $N \times NV$ matrix, data If there is no data or if the data is to be handled outside of **CMLPflClimits**, a scalar zero or a scalar missing value may be entered.
vars	$NV \times 1$ character vector, labels of variables selected for analysis – or – $NV \times 1$ numeric vector, indices of variables selected for analysis. If *dataset* is a matrix, *vars* may be a character vector containing either the standard labels created by **CML** (i.e., either V1, V2,..., or V01, V02,..... See discussion of the global variable **__vpad** below, or the user-provided labels in **__altnam**).
&*fct*	a pointer to a procedure that returns either the log-likelihood for one observation or a vector of log-likelihoods for a matrix of observations (see discussion of the global variable **__row** in global variable section below). If this function returns a scalar value and ¨row /= 1, the BHHH descent method and QML standard errors will not be available. It is also necessary to set **_cml_NumObs** to the number of observations.

Output

pflclimits	$K \times 2$ matrix, lower confidence limits in the first column and upper limits in second column

Globals

The **CML** procedure global variables are also applicable.

_cml_Alpha scalar, **CMLPflClimits** computes (1-**_cml_Alpha**)% confidence intervals, where **_cml_Alpha** varies between zero and one. Default = .05.

Remarks

Confidence limits are computed by inversion of the likelihood ratio statistic. b and f should be returns from a call to **CML**. This will also properly set up the global **_cml_NumObs**.

For inequality constrained models the limits are the solutions to a parametric nonlinear programming problem. **CMLPflClimits** solves this problem given the vector of parameter estimates and the model constraints.

The calculation of confidence limits for large models can be time consuming. In that case it might be necessary to select parameters for analysis. This can be done using the **CML** global, **_cml_Select**.

The global **_cml_NumObs** must be set. If **CMLPflClimits** is called immediately after a call to **CML**, **_cml_NumObs** will be set by **CML**.

Source

`cmlpflcl.src`

Purpose

Computes confidence limits based on t-statistics

Library

cml

Format

cl = **CMLTlimits**(*b*,*cov*)

Input

b	$K \times 1$ vector, parameter estimates
cov	$K \times K$ matrix, covariance matrix of parameter estimates

Output

cl	$K \times 2$ matrix, lower (first column) and upper (second column) confidence limits of the selected parameters

Globals

_cml_Alpha (1-**_cml_Alpha**)% confidence limits are computed.
Default = .05

_cml_NumObs scalar, number of observations. Must be set.

_cml_Select selection vector for selecting coefficients to be included in profiling, for example

```
_cml_Select = { 1, 3, 4 };
```

selects the 1st, 3rd, and 4th parameters for profiling.

Remarks

CMLTlimits returns $b[i] \pm t(\_cml\_NumObs - K; \_cml\_Alpha/2) \times \sqrt{cov[i,i]}$

_cml_NumObs must be set.

The global **_cml_NumObs** must be set. If **CMLTlimits** is called immediately after a call to **CML**, **_max_NumObs** will be set by **CML**.

Source

cml.src

Purpose

Baysian Inference using weighted maximum likelihood bootstrap

Library

`cml`

Format

`{` x,f,g,cov,$retcode$ `}` = **CMLBayes(***dataset*,*vars*,**&***fct*,*start***)**

Input

dataset	string containing name of **GAUSS** data set – or – $N \times NV$ matrix, data If there is no data or if the data is to be handled outside of **CMLBayes**, a scalar zero or a scalar missing value may be entered.
vars	$NV \times 1$ character vector, labels of variables selected for analysis – or – $NV \times 1$ numeric vector, indices of variables selected for analysis. If *dataset* is a matrix, *vars* may be a character vector containing either the standard labels created by **CMLBayes** (i.e., either V1, V2,..., or V01, V02,..... See discussion of the global variable **__vpad** below, or the user-provided labels in **__altnam**).
&*fct*	a pointer to a procedure that returns either the log-likelihood for one observation or a vector of log-likelihoods for a matrix of observations (see discussion of the global variable **__row** in global variable section below).
start	$K \times 1$ vector, start values.

Output

x	$K \times 1$ vector, mean of simulated posterior
f	scalar, mean weighted bootstrap log-likelihood
g	$K \times 1$ vector, means gradient of weighted bootstrap
h	$K \times K$ matrix, covariance matrix of the simulated posterior
retcode	scalar, return code. If normal convergence is achieved, then *retcode* = 0, otherwise a positive integer is returned indicating the reason for the abnormal termination:

0	normal convergence
1	forced exit
2	maximum iterations exceeded
3	function calculation failed
4	gradient calculation failed
5	Hessian calculation failed
6	line search failed
7	function cannot be evaluated at initial parameter values
8	error with gradient
9	error with constraints
10	secant update failed
11	maximum time exceeded
12	error with weights
13	quadratic program failed
34	data set could not be opened
35	number of observations not set
99	termination condition unknown

Globals

The **CML** procedure global variables are also applicable.

_cml_BayesAlpha scalar, exponent of the Dirichlet random variates used in the weights for the weighted bootstrap. See Newton and Raftery, "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap", J.R.Statist. Soc. B (1994), 56:3-48. Default = 1.4.

_cml_PriorProc scalar, pointer to proc for computing prior. This proc takes the parameter vector as its only argument, are returns a scalar probability. If a proc is not provided, a uniform prior is assumed.

_cml_BootFname string, file name of **GAUSS** data set (do not include .DAT extension) containing bootstrapped parameter estimates. If not specified, **CMLBayes** selects a temporary filename, BAYESxxxx where xxxx is 0000 incremented by 1 until a name is found that doesn't exist on the current directory.

_cml_MaxTime scalar, maximum amount of time spent in re-sampling. Default = 1e5 (about 10 weeks).

_cml_NumSample scalar, number of samples to be drawn. Default = 100.

Remarks

CMLBayes implements random sampling with replacement by computing **_cml_NumObs** pseudo-random Poisson variates and using them as weights in a call to **CML**. **CMLBayes** returns the mean vector of the estimates in the first argument and the covariance matrix of the estimates in the third argument.

A **GAUSS** data set is also generated containing the bootstrapped parameter estimates. The file name of the data set is either the name found in the global **_cml_BootFname**, or a temporary name. If **CMLBayes** selects a file name, it returns that file name in **_cml_BootFname**. The coefficients in this data set may be used as input to the **CML** procedures **CMLHist** and **CMLDensity** for further analysis.

Source

`cmlbayes.src`

Purpose

Computes bootstrapped estimates of parameters of a constrained maximum likelihood function.

Library

`cml`

Format

`{` *x*,*f*,*g*,*cov*,*retcode* `}` = **CMLBoot(***dataset*,*vars*,**&***fct*,*start***)**

Input

dataset string containing name of **GAUSS** data set
– or –
$N \times NV$ matrix, data
If there is no data or if the data is to be handled outside of **CMLBoot**, a scalar zero or a scalar missing value may be entered.

vars $NV \times 1$ character vector, labels of variables selected for analysis
– or –
$NV \times 1$ numeric vector, indices of variables selected for analysis.
If *dataset* is a matrix, *vars* may be a character vector containing either the standard labels created by **CMLBoot** (i.e., either V1, V2,..., or V01, V02,..... See discussion of the global variable **__vpad** below, or the user-provided labels in **__altnam**).

&*fct* a pointer to a procedure that returns either the log-likelihood for one observation or a vector of log-likelihoods for a matrix of observations (see discussion of the global variable **__row** in global variable section below).

start $K \times 1$ vector, start values.

Output

x $K \times 1$ vector, means of re-sampled parameters

f scalar, mean re-sampled function at minimum (the mean log-likelihood)

g $K \times 1$ vector, means of re-sampled gradients evaluated at the estimates

h $K \times K$ matrix, covariance matrix of the re-sampled parameters

retcode scalar, return code. If normal convergence is achieved, then *retcode* = 0, otherwise a positive integer is returned indicating the reason for the abnormal termination:

0	normal convergence
1	forced exit
2	maximum iterations exceeded
3	function calculation failed
4	gradient calculation failed
5	Hessian calculation failed
6	line search failed
7	function cannot be evaluated at initial parameter values
8	error with gradient
9	error with constraints
10	secant update failed
11	maximum time exceeded
12	error with weights
13	quadratic program failed
34	data set could not be opened
35	number of observations not set
99	termination condition unknown

Globals

The **CML** procedure global variables are also applicable.

_cml_BootFname string, file name of **GAUSS** data set (do not include .DAT extension) containing bootstrapped parameter estimates. If not specified, **CMLBoot** selects a temporary filename, BOOTxxxx where xxxx is 0000 incremented by 1 until a name is found that doesn't exist on the current directory.

_cml_MaxTime scalar, maximum amount of time spent in re-sampling. Default = 1e5 (about 10 weeks).

_cml_NumSample scalar, number of samples to be drawn. Default = 100.

Remarks

CMLBoot implements random sampling with replacement by computing **_cml_NumObs** pseudo-random Poisson variates and using them as weights in a call to

CML. **CMLBoot** returns the mean vector of the estimates in the first argument and the covariance matrix of the estimates in the third argument.

A **GAUSS** data set is also generated containing the bootstrapped parameter estimates. The file name of the data set is either the name found in the global **_cml_BootFname**, or a temporary name. If **CMLBoot** selects a file name, it returns that file name in **_cml_BootFname**. The coefficients in this data set may be used as input to the **CML** procedures **CMLHist** and **CMLDensity** for further analysis.

Source

`cmlboot.src`

Purpose

Generates kernel density plots from **GAUSS** data sets

Library

`cml, pgraph`

Format

`{` *px*, *py*, *smth* `}` = **CMLDensity(***dataset*,*vars***)**

Input

dataset	string containing name of **GAUSS** data set – or – N×K matrix, data
vars	$K \times 1$ character vector, labels of variables selected for analysis – or – $K \times 1$ numeric vector, indices of variables selected for analysis. If *dataset* is a matrix, *vars* may be a character vector containing either the standard labels created by **CMLDensity** (i.e., either V1, V2,..., or V01, V02,..... See discussion of the global variable **__vpad** below, or the user-provided labels in **__altnam**).

Output

px	**_cml_NumPoints** × K matrix, abscissae of plotted points
py	**_cml_NumPoints** × K matrix, ordinates of plotted points
smth	K × 1 vector, smoothing coefficients

Globals

The **CML** procedure global variables are also applicable.

_cml_Kernel K × 1 character vector, type of kernel:

NORMAL normal kernel
EPAN Epanechnikov kernel
BIWGT biweight kernel
TRIANG triangular kernel
RECTANG rectangular kernel

TNORMAL truncated normal kernel

If **_cml_Kernel** is scalar, the kernel is the same for all parameter densities. Default = NORMAL.

_cml_NumPoints scalar, number of points to be computed for plots

_cml_EndPoints K × 2 matrix, lower (in first column) and upper (in second column) endpoints of density. Default is minimum and maximum, respectively, of the parameter values. If 1 × 2 matrix, endpoints are the same for all parameters.

_cml_Smoothing K × 1 vector, smoothing coefficients for each plot. If scalar, smoothing coefficient is the same for each plot. If zero, smoothing coefficient is computed by **CMLDensity**. Default = 0.

_cml_Truncate K × 2 matrix, lower (in first column) and upper (in second column) truncation limits for truncated normal kernel. If 1x2 matrix, truncations limits are the same for all plots. Default is minimum and maximum, respectively.

__output If nonzero, K density plots are printed to the screen, otherwise no plots are generated.

Source

`cmldens.src`

Purpose

Generates histograms and surface plots from **GAUSS** data sets

Library

`cml, pgraph`

Format

`{` *tab*, *cut* `}` = **CMLHist(***dataset*,*vars***)**

Input

dataset	string containing name of **GAUSS** data set – or – N×K matrix, data
vars	$K \times 1$ character vector, labels of variables selected for analysis – or – $K \times 1$ numeric vector, indices of variables selected for analysis. If *dataset* is a matrix, *vars* may be a character vector containing either the standard labels created by **CMLHist** (i.e., either V1, V2,..., or V01, V02,..... See discussion of the global variable **__vpad** below, or the user-provided labels in **__altnam**).

Output

tab	**_cml_NumCat** × K matrix, univariate distributions of bootstrapped parameters
cut	**_cml_NumCat** × K matrix, cutting points

Globals

The **CML** procedure global variables are also applicable.

_cml_Center $K \times 1$ value of center category in histograms. Default is initial coefficient estimates.

_cml_CutPoint **_cml_NumCat** × 1 vector, output, cutting points for histograms

_cml_Increment $K \times 1$ vector, increments for cutting points of the histograms. Default is 2 * **_cml_Width** * std dev / **_cml_NumCat**.

_cml_NumCat scalar, number of categories in the histograms

_cml_Width scalar, width of histograms, default = 2

__output If nonzero, K density plots are printed to the screen, otherwise no plots are generated.

Remarks

If **__output** is nonzero, $K(K-1)/2$ plots are printed to the screen displaying univariate histograms and bivariate surface plots of the bootstrapped parameter distributions in pairs.

The globals, **_cml_Center**, **_cml_Width**, and **_cml_Increment** may be used to establish cutting points (which is stored in **_cml_Increment**) for the tables of re-sampled coefficients in *tab* The numbers in **_cml_Center** fix the center categories, **_cml_Width** is a factor which when multiplied times the standard deviation of the estimate determines the increments between categories. Alternatively, the increments between categories can be fixed directly by supplying them in **_cml_Increment**.

Source

`cmlhist.src`

Library

`cml, pgraph`

Purpose

Computes profile t plots and likelihood profile traces for constrained maximum likelihood models

Format

{ *x*,*f*,*g*,*cov*,*retcode* } = **CMLProfile**(*dataset*,*vars*,&*fct*,*start*)

Input

dataset string containing name of **GAUSS** data set
– or –
$N \times NV$ matrix, data
If there is no data or if the data is to be handled outside of **CMLProfile**, a scalar zero or a scalar missing value may be entered.

vars $NV \times 1$ character vector, labels of variables selected for analysis
– or –
$NV \times 1$ numeric vector, indices of variables selected for analysis.
If *dataset* is a matrix, *vars* may be a character vector containing either the standard labels created by **CMLProfile** (i.e., either V1, V2,..., or V01, V02,..... See discussion of the global variable **__vpad** below, or the user-provided labels in **__altnam**).

&*fct* a pointer to a procedure that returns either the log-likelihood for one observation or a vector of log-likelihoods for a matrix of observations (see discussion of the global variable **__row** in global variable section below).

start $K \times 1$ vector, start values.

Output

x $K \times 1$ vector, parameter estimates

f scalar, log-likelihood at maximum

g $K \times 1$ vector, gradients evaluated at the estimates

h $K \times K$ matrix, covariance matrix of the parameters

retcode scalar, return code. If normal convergence is achieved, then *retcode* = 0, otherwise a positive integer is returned indicating the reason for the abnormal termination:

0	normal convergence
1	forced exit
2	maximum iterations exceeded
3	function calculation failed
4	gradient calculation failed
5	Hessian calculation failed
6	line search failed
7	function cannot be evaluated at initial parameter values
8	error with gradient
9	error with constraints
10	secant update failed
11	maximum time exceeded
12	error with weights
13	quadratic program failed
34	data set could not be opened.
35	number of observations not set
99	termination condition unknown

Globals

The **CML** procedure global variables are also relevant.

_cml_NumCat scalar, number of categories in profile table. Default = 16.

_cml_Increment K × 1 vector, increments for cutting points, default is 2 * **_cml_Width** * std dev / **_cml_NumCat**. If scalar zero, increments are computed by **CMLProfile**.

_cml_Center K × 1 vector, value of center category in profile table. Default values are coefficient estimates.

_cml_Select selection vector for selecting coefficients to be included in profiling, for example

```
_cml_Select = { 1, 3, 4 };
```

selects the 1st, 3rd, and 4th parameters for profiling.

_cml_Width scalar, width of profile table in units of the standard deviations of the parameters. Default = 2.

Remarks

For each pair of the selected parameters, three plots are printed to the screen. Two of the are the profile t trace plots that describe the univariate profiles of the parameters, and one of them is the profile likelihood trace describing the joint distribution of the two parameters. Ideally distributed parameters would have univariate profile t traces that are straight lines, and bivariate likelihood profile traces that are two straight lines intersecting at right angles. This ideal is generally not met by nonlinear models, however, large deviations from the ideal indicate serious problems with the usual statistical inference.

Source

`cmlprof.src`

Purpose

Resets *CONSTRAINED MAXIMUM LIKELIHOOD* global variables to default values.

Library

`cml`

Format

CMLSet;

Input

None

Output

None

Remarks

Putting this instruction at the top of all command files that invoke **CML** is generally good practice. This prevents globals from being inappropriately defined when a command file is run several times or when a command file is run after another command file has executed that calls **CML**.

Source

`cml.src`

Purpose

Formats and prints the output from a call to **CML**.

Library

`cml`

Format

{ *x*,*f*,*g*,*h*,*retcode* } = **CMLPrt**(*x*,*f*,*g*,*h*,*retcode*);

Input

x	$K \times 1$ vector, parameter estimates
f	scalar, value of function at minimum
g	$K \times 1$ vector, gradient evaluated at x
h	$K \times K$ matrix, covariance matrix of parameters
retcode	scalar, return code.

Output

The input arguments are returned unchanged.

Globals

__header string. This is used by the printing procedure to display information about the date, time, version of module, etc. The string can contain one or more of the following characters:

"t"	print title (see **__title**)	
"l"	bracket title with lines	
"d"	print date and time	Example:
"v"	print version number of program	
"f"	print file name being analyzed	

```
__header = "tld";
```

Default = "tldvf".

__title string, message printed at the top of the screen and printed out by **CMLPrt**. Default = "".

Remarks

The call to **CML** can be nested in the call to **CMLPrt**:

```
{ x,f,g,h,retcode } = CMLPrt(CML(dataset,vars,&fct,start));
```

Source

`cml.src`

Purpose

Formats and prints the output from a call to **CML** along with confidence limits

Library

`cml`

Format

`{` x,f,g,cl,$retcode$ `}` = **CMLCLPrt(**x,f,g,cl,$retcode$**);**

Input

x	$K \times 1$ vector, parameter estimates
f	scalar, value of function at minimum
g	$K \times 1$ vector, gradient evaluated at x
cl	$K \times 2$ matrix, lower and upper confidence limits The lower limits are in the first column and the upper limits are in the second column.
$retcode$	scalar, return code.

Output

The input arguments are returned unchanged.

Globals

__header string. This is used by the printing procedure to display information about the date, time, version of module, etc. The string can contain one or more of the following characters:

"t"	print title (see **__title**)	
"l"	bracket title with lines	
"d"	print date and time	Example:
"v"	print version number of program	
"f"	print file name being analyzed	

```
__header = "tld";
```

Default = "tldvf".

__title string, message printed at the top of the screen and printed out by **CMLPrt**. Default = "".

Remarks

Confidence limits computed by **CMLBlimits**, **CMLClimits**, or **CMLTlimits** may be passed in the fourth argument in the call to **CMLCLPrt**:

```
{ b,f,g,cov,ret } = CMLBoot("tobit",0,&lpr,x0);
 cl = CMLBlimits(_cml_BootFname);
call CMLCLPrt(b,f,g,cl,ret);
```

Purpose

Computes estimates of parameters of a constrained maximum likelihood function

Library

`cml`

Format

`{` *x*,*f*,*g*,*cov*,*retcode* `}` = **fastCML(***dataset*,*vars*,**&***fct*,*start***)**

Input

dataset string containing name of **GAUSS** data set
– or –
$N \times NV$ matrix, data
If there is no data or if the data is to be handled outside of **fastCML**, a scalar zero or a scalar missing value may be entered.

vars $NV \times 1$ character vector, labels of variables selected for analysis
– or –
$NV \times 1$ numeric vector, indices of variables selected for analysis.
If *dataset* is a matrix, *vars* may be a character vector containing either the standard labels created by **CML** (i.e., either V1, V2,..., or V01, V02,..... See discussion of the global variable **__vpad** below, or the user-provided labels in **__altnam**).

&*fct* a pointer to a procedure that returns either the log-likelihood for one observation or a vector of log-likelihoods for a matrix of observations (see discussion of the global variable **__row** in global variable section below).
If this function returns a scalar value and ¨row /= 1, the BHHH descent method and QML standard errors will not be available. It is also necessary to set **_cml_NumObs** to the number of observations.

start $K \times 1$ vector, start values.

Output

x $K \times 1$ vector, estimated parameters

f scalar, function at minimum (the mean log-likelihood)

g $K \times 1$ vector, gradient evaluated at x

h $K \times K$ matrix, covariance matrix of the parameters (see discussion of the global variable **_cml_CovPar** below).

retcode scalar, return code. If normal convergence is achieved, then *retcode* = 0, otherwise a positive integer is returned indicating the reason for the abnormal termination:

0	normal convergence
1	forced exit.
2	maximum iterations exceeded.
3	function calculation failed.
4	gradient calculation failed.
5	Hessian calculation failed.
6	line search failed.
7	function cannot be evaluated at initial parameter values.
8	error with gradient
9	error with constraints
10	secant update failed
11	maximum time exceeded
12	error with weights
13	quadratic program failed
20	Hessian failed to invert
34	data set could not be opened.
35	number of observations not set.
99	termination condition unknown.

Globals

The globals variables used by **CML** can be organized in the following categories according to which aspect of the optimization they affect:

Options **_cml_Options**

Constraints **_cml_A**, **_cml_B**, **_cml_C**, **_cml_D**, **_cml_EqProc**, **_cml_IneqProc**, **_cml_EqJacobian**, **_cml_IneqJacobian**, **_cml_Bounds**, **_cml_Lagrange**

Descent and Line Search **_cml_Algorithm**, **_cml_Delta**, **_cml_LineSearch**, **_cml_Maxtry**, **_cml_Extrap**, **_cml_Interp**, **_cml_Switch**, **_cml_Trust**, **_cml_TrustRadius**, **_cml_GridsSearch**, **_cml_GridSearchRadius**, **_cml_FeasibleTest**

Covariance Matrix of Parameters **_cml_CovPar**, **_cml_XprodCov**, **_cml_HessCov**, **_cml_FinalHess**

Gradient **_cml_GradMethod**, **_cml_GradProc**, **_cml_HessProc**, **_cml_GradStep**, **_cml_GradCheckTol** **_cml_GradOrder**,

Termination Conditions **_cml_DirTol**, **_cml_DFTol**, **_cml_MaxIters**, **_cml_MaxTime**

Data **_cml_NumObs**, **__weight**,

Parameters **_cml_Active**, **_cml_ParNames**

Miscellaneous **_cml_IterData**, **_cml_State**, **_cml_Alpha**,

The list below contains an alphabetical listing of each global with a complete description.

_cml_A $M_1 \times K$ matrix, linear equality constraint coefficient matrix **_cml_A** is used with **_cml_B** to specify linear equality constraints:

```
_cml_A * X = _cml_B
```

where X is the $K \times 1$ unknown parameter vector.

_cml_Alpha scalar, sets alpha level for statistical inference. Default = .05.

_cml_Active vector, defines fixed/active coefficients. This global allows you to fix a parameter to its starting value. This is useful, for example, when you wish to try different models with different sets of parameters without having to re-edit the function. When it is to be used, it must be a vector of the same length as the starting vector. Set elements of **_cml_Active** to 1 for an active parameter, and to zero for a fixed one.

_cml_Algorithm scalar, selects optimization method:

1	BFGS - Broyden, Fletcher, Goldfarb, Shanno method
2	DFP - Davidon, Fletcher, Powell method
3	NEWTON - Newton-Raphson method
4	BHHH - Berndt, Hall, Hall, Hausman method

Default = 3

_cml_Delta scalar, floor for eigenvalues of Hessian in the NEWTON algorithm. When nonzero, the eigenvalues of the Hessian are augmented to this value.

_cml_B $M_1 \times 1$ vector, linear equality constraint constant vector **_cml_B** is used with **_cml_A** to specify linear equality constraints:

```
_cml_A * X = _cml_B
```

where X is the $K \times 1$ unknown parameter vector.

_cml_Bounds $K \times 2$ matrix, bounds on parameters. The first column contains the lower bounds, and the second column the upper bounds. If the bounds for all the coefficients are the same, a 1x2 matrix may be used. Default = { -1e256 1e256 }.

_cml_C $M_3 \times K$ matrix, linear inequality constraint coefficient matrix **_cml_C** is used with **_cml_D** to specify linear inequality constraints:

```
_cml_C * X = _cml_D
```

where X is the $K \times 1$ unknown parameter vector.

_cml_CovPar scalar, type of covariance matrix of parameters

0	not computed
1	computed from inverse of Hessian calculated after the iterations
2	computed from cross-product of Jacobian after the iterations
3	heteroskedastic-consistent covariance matrix of the parameters

Default = 1;

_cml_D $M_3 \times 1$ vector, linear inequality constraint constant vector **_cml_D** is used with **_cml_C** to specify linear inequality constraints:

```
_cml_C * X = _cml_D
```

where X is the $K \times 1$ unknown parameter vector.

_cml_DFTol scalar. Iterations are halted when the absolute value of the change in the function is less than this amount. Default = 0;

_cml_DirTol scalar, convergence tolerance for gradient of estimated coefficients. When this criterion has been satisifed CML exits the iterations. Default = 1e-5.

_cml_EqJacobian scalar, pointer to a procedure that computes the Jacobian of the nonlinear equality constraints with respect to the parameters. The procedure has one input argument, the $K \times 1$ vector of parameters, and one output argument, the $M_2 \times 1$ vector of derivatives of the constraints with respect to the parameters. For example, if the nonlinear equality constraint procedure was,

```
proc eqproc(p);
    retp(p[1]*p[2]-p[3]);
endp;
```

the Jacobian procedure and assignment to the global would be,

```
proc eqj(p);
    retp(p[2]~p[1]~-1);
endp;

_cml_EqJacobian = &eqj;
```

_cml_EqProc scalar, pointer to a procedure that computes the nonlinear equality constraints. For example, the statement:

```
_cml_EqProc = &eqproc;
```

tells **CML** that nonlinear equality constraints are to be placed on the parameters and where the procedure computing them is to be found. The procedure must have one input argument, the $K \times 1$ vector of parameters, and one output argument, the $M_2 \times K$ matrix of computed constraints that are to be equal to zero. For example, suppose that you wish to place the following constraint:

```
P[1] * P[2] = P[3]
```

The proc for this is:

```
proc eqproc(p);
    retp(p[1]*[2]-p[3]);
endp;
```

_cml_Extrap scalar, extrapolation constant in BRENT. Default = 2.

_cml_FeasibleTest scalar, if nonzero a test for the feasibility of the parameter vector is tested before calling the log-likelihood function. This is important if feasibility is necessary for the calculation of the log-likelihood, for example if the procedure is computing the log of a parameter which is being constrained to be positive. By default this function is turned on, but if feasibility is not required for calculating the function then set **_cml_FeasibleTest** to zero if the iterations are being halted as a result of the test. Default = 1.

_cml_FinalHess $K \times K$ matrix, the Hessian used to compute the covariance matrix of the parameters is stored in **_cml_FinalHess**. This is most useful if the inversion of the hessian fails, which is indicated when **CML** returns a missing value for the covariance matrix of the parameters. An analysis of the Hessian stored in **_cml_FinalHess** can then reveal the source of the linear dependency responsible for the singularity.

_cml_GradMethod scalar, method for computing numerical gradient.

0	central difference
1	forward difference (default)

_cml_GradOrder scalar, If set to zero, standard two point method is used. If greater than two, sets number of points for computing multiple point numerical gradient.

_cml_GradProc scalar, pointer to a procedure that computes the gradient of the function with respect to the parameters. For example, the statement:

```
_cml_GradProc=&gradproc;
```

tells **CML** that a gradient procedure exists as well where to find it. The user-provided procedure has two input arguments, an $K \times 1$ vector of parameter values and an N×K matrix of data. The procedure returns a single output argument, an $N \times K$ matrix of gradients of the log-likelihood function with respect to the parameters evaluated at the vector of parameter values.

For example, suppose the log-likelihood function is for a Poisson regression, then the following would be added to the command file:

```
proc lgd(b,z);
    retp((z[.,1]-exp(z[.,2:4]*b)).*z[.,2:4]);
endp;

_cml_GradProc = &lgd;
```

Default = 0, i.e., no gradient procedure has been provided.

_cml_GradStep scalar, increment size for computing gradient. When the numerical gradient is performing well, set to a larger value (1e-3, say). Default is the cube root of machine precision.

_cml_GridSearch scalar, if nonzero a grid search for a new direction will be attempted whenever the line search fails. It is nonzero by default.

_cml_GridSearchRadius scalar, set to the radius of the grid search. Default = .01

_cml_HessCov $K \times K$ matrix. When **_cml_CovPar** is set to 3 the information matrix covariance matrix of the parameters, i.e., the inverse of the matrix of second order partial derivatives of the log-likelihood by observations, is returned in **_cml_HessCov**.

_cml_HessProc scalar, pointer to a procedure that computes the hessian, i.e., the matrix of second order partial derivatives of the function with respect to the parameters. For example, the instruction:

```
_cml_HessProc = &hessproc;
```

tells **CML** that a procedure has been provided for the computation of the hessian and where to find it. The procedure that is provided by the user must have two input arguments, a $K \times 1$ vector of parameter values and an N×P data matrix. The procedure returns a single output argument, the $K \times K$ symmetric matrix of second order derivatives of the function evaluated at the parameter values.

_cml_IneqJacobian scalar, pointer to a procedure that computes the Jacobian of the nonlinear equality constraints with respect to the parameters. The procedure has one input argument, the $K \times 1$ vector of parameters, and one output argument, the $M_4 \times K$ matrix of derivatives of the constraints with respect to the parameters. For example, if the nonlinear equality constraint procedure was,

```
proc ineqproc(p);
    retp(p[1]*p[2]-p[3]);
endp;
```

the Jacobian procedure and assignment to the global would be,

```
proc ineqj(p);
    retp(p[2]~p[1]~-1);
endp;

_cml_IneqJacobian = &ineqj;
```

_cml_IneqProc scalar, pointer to a procedure that computes the nonlinear inequality constraints. For example the statement:

```
_cml_IneqProc = &ineqproc;
```

tells **CML** that nonlinear equality constraints are to be placed on the parameters and where the procedure computing them is to be found. The procedure must have one input argument, the $K \times 1$ vector of parameters, and one output argument, the $M_4 \times K$ matrix of computed constraints that are to be equal to zero. For example, suppose that you wish to place the following constraint:

```
P[1] * P[2] >= P[3]
```

The proc for this is:

```
proc ineqproc(p);
    retp(p[1]*[2]-p[3]);
endp;
```

_cml_Interp scalar, interpolation constant in BRENT. Default = .25.

_cml_IterData 3x1 vector, contains information about the iterations. The first element contains the # of iterations, the second element contains the elapsed time in minutes of the iterations, and the third element contains a character variable indicating the type of covariance matrix of the parameters.

_cml_Lagrange vector, created using **VPUT**. Contains the Lagrangean coefficients for the constraints. They may be extracted with the **VREAD** command using the following strings:

"lineq"	linear equality constraints
"nlineq"	nonlinear equality constraints
"linineq"	linear inequality constraints
"nlinineq"	nonlinear inequality constraints
"bounds"	bounds
"eqcov"	covariance matrix of equality Lagrangeans
"ineqcov"	covariance matrix of inequality Lagrangeans
"boundcov"	covariance matrix of bounds' Lagrangeans

When an inequality or bounds constraint is active, its associated Lagrangean is nonzero. The linear Lagrangeans preceed the nonlinear Lagrangeans in the covariance matrices.

__cml__LineSearch scalar, selects method for conducting line search. The result of the line search is a *step length*, i.e., a number which reduces the function value when multiplied times the direction..

1 step length = 1.

2 cubic or quadratic step length method (STEPBT)

3 step halving (HALF)

4 Brent's step length method (BRENT)

5 BHHH step length method (BHHHSTEP)

Default = 2.

Usually **__cml__LineSearch** = 2 is best. If the optimization bogs down, try setting **__cml__LineSearch** = 1, 4 or 5. **__cml__LineSearch** = 3 generates slower iterations but faster convergence and **__cml__LineSearch** = 1 generates faster iterations but slower convergence.

When any of these line search methods fails, **CML** attempts a random search of radius **__cml__RandRadius** times the truncated log to the base 10 of the gradient when **__cml__RandRadius** is set to a nonzero value. If **__cml__UserSearch** is set to 1, **CML** enters an interactive line search mode.

__cml__MaxIters scalar, maximum number of iterations.

__cml__MaxTime scalar, maximum time in iterations in minutes. This global is most useful in bootstrapping. You might want 100 re-samples, but would be happy with anything more than 50 depending on the time it took. Set **__cml__NumSample** = 100, and **__cml__MaxTime** to maximum time you would be willing to wait for results. Default = 1e+5, about 10 weeks.

_cml_MaxTry scalar, maximum number of tries to find step length that produces a descent.

_cml_NumObs scalar, number of observations. By default the number of observations is determined from the dataset and this number is returned in the global after the iterations. If a different number is required for correct statistical inference, set **_cml_NumObs** to this value. It may also be necessary to set this global if a dataset is not passed to **CML** and if the log-likelihood returns a scalar.

_cml_Options character vector, specification of options. This global permits setting various **CML** options in a single global using identifiers. The following

```
_cml_Options = { newton stepbt forward screen };
```

the descent method to NEWTON, the line search method to STEPBT, the numerical gradient method to forward differences.

The following is a list of the identifiers:

Algorithms BFGS, DFP, NEWTON, BHHH

Line Search ONE, STEPBT, HALF, BRENT, BHHHSTEP

Covariance Matrix NOCOV, INFO, XPROD, HETCON

Gradient method CENTRAL, FORWARD, BACKWARD

_cml_ParNames $K \times 1$ character vector, parameter labels.

_cml_State scalar,

_cml_Trust scalar, if nonzero, a trust region is imposed on the direction vector in the iterations. The trust region method is helpful when the starting values are poor. By default it is turned off.

_cml_TrustRadius scalar, sets the radius of the trust region method. Default = .001.

_cml_Switch 4×1 or 4×2 vector,

controls algorithm switching. If 4×1 set

_cml_Switch[1] algorithm number to switch to

_cml_Switch[2] CML switches if functions changes less than this amount

_cml_Switch[3] CML switches if this number of iterations is exceeded.

_cml_Switch[4] CML switches if line search step changes less than this amount

or else if 4×2 set each column as in the 4×1 case with different algorithms in the first row. CML switches between the algorithm in column 1 and column 2.

__weight vector, frequency of observations. By default all observations have a frequency of 1. zero frequencies are allowed. It is assumed that the elements of **__weight** sum to the number of observations.

_cml_XprodCov $K \times K$ matrix. When **_cml_CovPar** is set to 3 the cross-product matrix covariance matrix of the parameters, i.e., the inverse of the cross-product of the first derivatives of the log-likelihood computed by observations, is is returned in **_cml_XprodCov**.

Remarks

Specifying Constraints.

There are five types of constraints: linear equality, linear inequality, nonlinear equality, nonlinear inequality, bounds Linear constraints are specified by initializing the appropriate **CML** globals to known matrices of constants. The linear equality constraint matrices are **_cml_A** and **_cml_B**, and they assume the following relationship with the parameter vector:

```
_cml_A * x = _cml_B
```

where `x` is the parameter vector.

Similarly, the linear *in*equality constraint matrices are **_cml_C** and **_cml_D**, and assume the following relationship with the parameter vector:

```
_cml_C * x >= _cml_D
```

The nonlinear constraints are specified by providing procedures and assigning their pointers to **CML** globals. These procedures take a single argument, the vector of parameters, and return a column vector of evaluations of the constraints at the parameters. Each element of the column vector is a separate constraint.

For example, suppose you wish to constrain the product of the first and third coefficients to be equal to 10, and the squared second and fourth coefficients to be equal to the squared fifth coefficient:

```
proc eqp(x);
    local c;
    c = zeros(2,1);
    c[1] = x[1] * x[3] - 10;
    c[2] = x[2] * x[2] + x[4] * x[4] - x[5] * x[5];
    retp(c);
endp;

_cml_EqProc = &eqp;
```

The nonlinear equality constraint procedure causes **CML** to find estimates for which its evaluation is equal to a conformable vector of zeros.

The nonlinear *in*equality constraint procedure is similar to the equality procedure. **CML** finds estimates for which the evaluation of the procedure is greater than or equal to zero. The nonlinear inequality constraint procedure is assigned to the global **_cml_IneqProc**. For example, suppose you wish to constrain the norm of the coefficients to be greater than one:

```
proc ineqp(x);
    retp(x'x-3);
endp;

_cml_IneqProc = &eqp;
```

Bounds are a type of linear inequality constraint. They are specified separately for computational and notational convenience. To declare bounds on the parameters assign a two column vector with rows equal to the number of parameters to the **CML** global, **_cml_Bounds**. The first column is the lower bounds and the second column the upper bounds. For example,

```
_cml_Bounds = {  0 10,
               -10  0
               -10 20 };
```

If the bounds are the same for all of the parameters, only the first row is required.

Writing the Log-likelihood Function

CML requires a procedure for computing the log-likelihood.

The procedure has two input arguments: first, a vector of parameter values, and second, one or more rows of the data matrix. The output argument is the log-likelihood for the observation or observations in the second argument evaluated at the parameter values in the first argument.

Suppose that the function procedure has been named *pfct*, the following considerations apply:

The format of the procedure is:

logprob = *pfct*(*x*,*y*);

where

x column vector of parameters of model

y	one or more rows of the data set (if the data set has been transformed, or if *vars* $\neq 0$, i.e., there is selection, then *y* is a transformed, selected observation). if a scalar zero or scalar missing value was passed to **CML** in its first argument not data will be passed to *pfct* and the user will required to provide the necessary data for computing the log-likelihood. if **__row** $= n$, then n rows of the data set are read at a time if **__row** $= 0$, the maximum number of rows that fit in memory is computed by **CML**.

For most uses, the output from this procedure is a vector of log-likelihoods where the vector has the same number of rows as the input dataset in the second argument. A scalar log-likelihood may also be returned but it is important to keep in mind that in this case certain features of **CML** will not be available such as QML standard errors or bootstrapping. Also if a scalar log-likelihood is returned by this procedure you must also set **_cml_NumObs** to its appropriate value. ;

CML can be forced, if necessary, to send the data n observation at a time by setting *__row* $= n$. By default *__row* $= 0$ causing **CML** to send the entire matrix to *pfct* if it is stored entirely in memory, or to compute the maximum number of rows if it is a **GAUSS** data set stored on disk (Note that even if the data starts out in a **GAUSS** data set, **CML** determines whether the data set fits in memory, and if it does, then it reads the data set into an array in memory). If you are getting *insufficient memory* messages, then set **__rowfac** to a positive value less than 1.

Supplying an Analytical GRADIENT Procedure

To decrease the time of computation, the user may provide a procedure for the calculation of the gradient of the log-likelihood. The global variable **_cml_GradProc** must contain the pointer to this procedure. Suppose the name of this procedure is *gradproc*. Then,

g **=** *gradproc***(***x*,*y***);**

where the input arguments are

x	vector of coefficients
y	one or more rows of data set. if a scalar zero or scalar missing value was passed to **CML** in its first argument not data will be passed to *gradproc* and the user will required to provide the necessary data for computing the gradient.

and the output argument is

g	row vector of gradients of log-likelihood with respect to coefficients, or a matrix of gradients (i.e., a Jacobian) if the data passed in y is a matrix (unless **__cml_Lag** ≥ 1 in which case the data passed in y is a matrix of lagged values but a row vector of gradients is passed back in g).

It is important to note that the gradient is row oriented. Thus if the function that computes the log-likelihood returns a scalar value (**___row** $= 1$), then a row vector of the first derivatives of the log-likelihood with respect to the coefficients must be returned, but if the procedure that computes the log-likelihood returns a column vector, then **__cml_GradProc** must return a matrix of first derivatives in which rows are associated with observations and columns with coefficients.

Providing a procedure for the calculation of the first derivatives also has a significant effect on the calculation time of the Hessian. The calculation time for the numerical computation of the Hessian is a quadratic function of the size of the matrix. For large matrices, the calculation time can be very significant. This time can be reduced to a linear function of size if a procedure for the calculation of analytical first derivatives is available. When such a procedure is available, **CML** automatically uses it to compute the numerical Hessian.

The major problem one encounters when writing procedures to compute gradients and Hessians is in making sure that the gradient is being properly computed. **CML** checks the gradients and Hessian when **__cml_GradCheckTol** is nonzero. **CML** generates both numerical and analytical gradients, and viewing the discrepancies between them can help in debugging the analytical gradient procedure.

Supplying an Analytical HESSIAN Procedure.

Selection of the NEWTON algorithm becomes feasible if the user supplies a procedure to compute the Hessian. If such a procedure is provided, the global variable **__cml_HessProc** must contain a pointer to this procedure. Suppose this procedure is called *hessproc*, the format is

h **=** *hessproc***(**x**,**y**);**

The input arguments are

x	$K \times 1$ vector of coefficients
y	one or more rows of data set if a scalar zero or scalar missing value was passed to **CML** in its first argument not data will be passed to *hessproc* and the user will required to provide the necessary data for computing the Hessian.

and the output argument is

h	$K \times K$ matrix of second order partial derivatives evaluated at the coefficients in x.

To compare numerical and analytical Hessians set **_cml_GradCheckTol** to a nonzero value.

Supplying Analytical Jacobians of the Nonlinear Constraints.

At each iteration the Jacobians of the nonlinear constraints, if they exist, are computed numerically. This is time-consuming and generates a loss of precision. For models with a large number of inequality constraints a significant speed-up can be achieved by providing analytical Jacobian procedures. The improved accuracy can also have a significant effect on convergence.

The Jacobian procedures take a single argument, the vector of parameters, and return a matrix of derivatives of each constraint with respect to each parameter. The rows are associated with the constraints and the columns with the parameters. The pointer to the nonlinear equality Jacobian procedure is assigned to **_cml_EqJacobian**. The pointer to the nonlinear *in*equality Jacobian procedure is assigned to **_cml_IneqJacobian**.

For example, suppose the following procedure computes the equality constraints:

```
proc eqp(x);
    local c;
    c = zeros(2,1);
    c[1] = x[1] * x[3] - 10;
    c[2] = x[2] * x[2] + x[4] * x[4] - x[5] * x[5];
    retp(c);
endp;
cml_EqProc = &eqp;
```

Then the Jacobian procedure would look like this:

```
proc eqJacob(x);
    local c;
    c = zeros(2,5);
    c[1,1] = x[3];
    c[1,3] = x[1];
    c[2,2] = 2*x[2];
    c[2,4] = 2*x[4];
    c[3,5] = -2*x[5];
    retp(c);
 endp;
 _cml_EqJacobian = &eqJacob;
```

The Jacobian procedure for the nonlinear inequality constraints is specified similarly, except that the associated global containing the pointer to the procedure is **_cml_IneqJacobian**.

Source

`cml.src`

Purpose

Baysian Inference using weighted maximum likelihood bootstrap

Library

`cml`

Format

`{` x,f,g,cov,$retcode$ `}` = **fastCMLBayes(**$dataset$,$vars$,**&**fct,$start$**)**

Input

dataset	string containing name of **GAUSS** data set – or – $N \times NV$ matrix, data If there is no data or if the data is to be handled outside of **fastCMLBayes**, a scalar zero or a scalar missing value may be entered.
vars	$NV \times 1$ character vector, labels of variables selected for analysis – or – $NV \times 1$ numeric vector, indices of variables selected for analysis. If *dataset* is a matrix, *vars* may be a character vector containing either the standard labels created by **fastCMLBayes** (i.e., either V1, V2,..., or V01, V02,..... See discussion of the global variable **__vpad** below, or the user-provided labels in **__altnam**).
&*fct*	a pointer to a procedure that returns either the log-likelihood for one observation or a vector of log-likelihoods for a matrix of observations (see discussion of the global variable **__row** in global variable section below).
start	$K \times 1$ vector, start values.

Output

x	$K \times 1$ vector, mean of simulated posterior
f	scalar, mean weighted bootstrap log-likelihood
g	$K \times 1$ vector, means gradient of weighted bootstrap
h	$K \times K$ matrix, covariance matrix of the simulated posterior
retcode	scalar, return code. If normal convergence is achieved, then $retcode = 0$, otherwise a positive integer is returned indicating the reason for the abnormal termination:

0	normal convergence
1	forced exit
2	maximum iterations exceeded
3	function calculation failed
4	gradient calculation failed
5	Hessian calculation failed
6	line search failed
7	function cannot be evaluated at initial parameter values
8	error with gradient
9	error with constraints
10	secant update failed
11	maximum time exceeded
12	error with weights
13	quadratic program failed
34	data set could not be opened
35	number of observations not set
99	termination condition unknown

Globals

The **CML** procedure global variables are also applicable.

_cml_BayesAlpha scalar, exponent of the Dirichlet random variates used in the weights for the weighted bootstrap. See Newton and Raftery, "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap", J.R.Statist. Soc. B (1994), 56:3-48. Default = 1.4.

_cml_PriorProc scalar, pointer to proc for computing prior. This proc takes the parameter vector as its only argument, are returns a scalar probability. If a proc is not provided, a uniform prior is assumed.

_cml_BootFname string, file name of **GAUSS** data set (do not include .DAT extension) containing bootstrapped parameter estimates. If not specified, **fastCMLBayes** selects a temporary filename, BAYESxxxx where xxxx is 0000 incremented by 1 until a name is found that doesn't exist on the current directory.

_cml_MaxTime scalar, maximum amount of time spent in re-sampling. Default = 1e5 (about 10 weeks).

_cml_NumSample scalar, number of samples to be drawn. Default = 100.

Remarks

fastCMLBayes implements random sampling with replacement by computing **_cml_NumObs** pseudo-random Poisson variates and using them as weights in a call to **CML**. **fastCMLBayes** returns the mean vector of the estimates in the first argument and the covariance matrix of the estimates in the third argument.

A **GAUSS** data set is also generated containing the bootstrapped parameter estimates. The file name of the data set is either the name found in the global **_cml_BootFname**, or a temporary name. If **fastCMLBayes** selects a file name, it returns that file name in **_cml_BootFname**. The coefficients in this data set may be used as input to the **CML** procedures **CMLHist** and **CMLDensity** for further analysis.

Source

`fastcbayes.src`

Purpose

Computes bootstrapped estimates of parameters of a constrained maximum likelihood function.

Library

`cml`

Format

`{` *x*,*f*,*g*,*cov*,*retcode* `}` = **fastCMLBoot(***dataset*,*vars*,**&***fct*,*start***)**

Input

dataset string containing name of **GAUSS** data set
– or –
$N \times NV$ matrix, data
If there is no data or if the data is to be handled outside of **fastCMLBoot**, a scalar zero or a scalar missing value may be entered.

vars $NV \times 1$ character vector, labels of variables selected for analysis
– or –
$NV \times 1$ numeric vector, indices of variables selected for analysis.
If *dataset* is a matrix, *vars* may be a character vector containing either the standard labels created by **CMLBoot** (i.e., either V1, V2,..., or V01, V02,..... See discussion of the global variable **__vpad** below, or the user-provided labels in **__altnam**).

&*fct* a pointer to a procedure that returns either the log-likelihood for one observation or a vector of log-likelihoods for a matrix of observations (see discussion of the global variable **__row** in global variable section below).

start $K \times 1$ vector, start values.

Output

x $K \times 1$ vector, means of re-sampled parameters

f scalar, mean re-sampled function at minimum (the mean log-likelihood)

g $K \times 1$ vector, means of re-sampled gradients evaluated at the estimates

h $K \times K$ matrix, covariance matrix of the re-sampled parameters

retcode scalar, return code. If normal convergence is achieved, then *retcode* = 0, otherwise a positive integer is returned indicating the reason for the abnormal termination:

0	normal convergence
1	forced exit
2	maximum iterations exceeded
3	function calculation failed
4	gradient calculation failed
5	Hessian calculation failed
6	line search failed.
7	function cannot be evaluated at initial parameter values
8	error with gradient
9	error with constraints
10	secant update failed
11	maximum time exceeded
12	error with weights
13	quadratic program failed
34	data set could not be opened
35	number of observations not set
99	termination condition unknown

Globals

The **CML** procedure global variables are also applicable.

_cml_BootFname string, file name of **GAUSS** data set (do not include .DAT extension) containing bootstrapped parameter estimates. If not specified, **CMLBoot** selects a temporary filename, BOOTxxxx where xxxx is 0000 incremented by 1 until a name is found that doesn't exist on the current directory.

_cml_MaxTime scalar, maximum amount of time spent in re-sampling. Default = 1e5 (about 10 weeks).

_cml_NumSample scalar, number of samples to be drawn. Default = 100.

Remarks

CMLBoot implements random sampling with replacement by computing **_cml_NumObs** pseudo-random Poisson variates and using them as weights in a call to

CML. **CMLBoot** returns the mean vector of the estimates in the first argument and the covariance matrix of the estimates in the third argument.

A **GAUSS** data set is also generated containing the bootstrapped parameter estimates. The file name of the data set is either the name found in the global **_cml_BootFname**, or a temporary name. If **CMLBoot** selects a file name, it returns that file name in **_cml_BootFname**. The coefficients in this data set may be used as input to the **CML** procedures **CMLHist** and **CMLDensity** for further analysis.

Source

`fastcboot.src`

Purpose

Computes confidence limits by inversion of the likelihood ratio statistic.

Library

`cml`

Format

pflclimits = **fastCMLPflClimts(***b*,*f*,*dataset*,*vars*,**&***fct***)**

Input

b	$K \times 1$ vector, parameter estimates
f	scalar, function at minimum (mean log-likelihood)
dataset	string containing name of **GAUSS** data set – or – $N \times NV$ matrix, data If there is no data or if the data is to be handled outside of **fastCMLPflClimts**, a scalar zero or a scalar missing value may be entered.
vars	$NV \times 1$ character vector, labels of variables selected for analysis – or – $NV \times 1$ numeric vector, indices of variables selected for analysis. If *dataset* is a matrix, *vars* may be a character vector containing either the standard labels created by **CML** (i.e., either V1, V2,..., or V01, V02,..... See discussion of the global variable **__vpad** below, or the user-provided labels in **__altnam**).
&*fct*	a pointer to a procedure that returns either the log-likelihood for one observation or a vector of log-likelihoods for a matrix of observations (see discussion of the global variable **__row** in global variable section below). If this function returns a scalar value and ¨row /= 1, the BHHH descent method and QML standard errors will not be available. It is also necessary to set **_cml_NumObs** to the number of observations.

Output

pflclimits	$K \times 2$ matrix, lower confidence limits in the first column and upper limits in second column

Globals

The **CML** procedure global variables are also applicable.

_cml_Alpha scalar, **fastCMLPflClimts** computes (1-**_cml_Alpha**)% confidence intervals, where **_cml_Alpha** varies between zero and one. Default = .05.

Remarks

Confidence limits are computed by inversion of the likelihood ratio statistic. b and f should be returns from a call to **CML**. This will also properly set up the global **_cml_NumObs**.

For inequality constrained models the limits are the solutions to a parametric nonlinear programming problem. **fastCMLPflClimts** solves this problem given the vector of parameter estimates and the model constraints.

The calculation of confidence limits for large models can be time consuming. In that case it might be necessary to select parameters for analysis. This can be done using the **CML** global, **_cml_Select**.

The global **_cml_NumObs** must be set. If **fastCMLPflClimts** is called immediately after a call to **CML**, **_cml_NumObs** will be set by **CML**.

Source

`fastcpflcl.src`

Library

`cml, pgraph`

Purpose

Computes profile t plots and likelihood profile traces for constrained maximum likelihood models

Format

{ *x*,*f*,*g*,*cov*,*retcode* } = **fastCMLProfile**(*dataset*,*vars*,&*fct*,*start*)

Input

dataset string containing name of **GAUSS** data set
– or –
$N \times NV$ matrix, data
If there is no data or if the data is to be handled outside of **fastCMLProfile**, a scalar zero or a scalar missing value may be entered.

vars $NV \times 1$ character vector, labels of variables selected for analysis
– or –
$NV \times 1$ numeric vector, indices of variables selected for analysis.
If *dataset* is a matrix, *vars* may be a character vector containing either the standard labels created by **fastCMLProfile** (i.e., either V1, V2,..., or V01, V02,..... See discussion of the global variable **__vpad** below, or the user-provided labels in **__altnam**).

&*fct* a pointer to a procedure that returns either the log-likelihood for one observation or a vector of log-likelihoods for a matrix of observations (see discussion of the global variable **__row** in global variable section below).

start $K \times 1$ vector, start values.

Output

x $K \times 1$ vector, parameter estimates

f scalar, log-likelihood at maximum

g $K \times 1$ vector, gradients evaluated at the estimates

h $K \times K$ matrix, covariance matrix of the parameters

retcode scalar, return code. If normal convergence is achieved, then *retcode* = 0, otherwise a positive integer is returned indicating the reason for the abnormal termination:

0	normal convergence
1	forced exit
2	maximum iterations exceeded
3	function calculation failed
4	gradient calculation failed
5	Hessian calculation failed
6	line search failed
7	function cannot be evaluated at initial parameter values
8	error with gradient
9	error with constraints
10	secant update failed
11	maximum time exceeded
12	error with weights
13	quadratic program failed
34	data set could not be opened
35	number of observations not set
99	termination condition unknown

Globals

The **CML** procedure global variables are also relevant.

_cml_NumCat scalar, number of categories in profile table. Default = 16.

_cml_Increment K × 1 vector, increments for cutting points, default is 2 * **_cml_Width** * std dev / **_cml_NumCat**. If scalar zero, increments are computed by **fastCMLProfile**.

_cml_Center K × 1 vector, value of center category in profile table. Default values are coefficient estimates.

_cml_Select selection vector for selecting coefficients to be included in profiling, for example

```
_cml_Select = { 1, 3, 4 };
```

selects the 1st, 3rd, and 4th parameters for profiling.

_cml_Width scalar, width of profile table in units of the standard deviations of the parameters. Default = 2.

Remarks

For each pair of the selected parameters, three plots are printed to the screen. Two of the are the profile t trace plots that describe the univariate profiles of the parameters, and one of them is the profile likelihood trace describing the joint distribution of the two parameters. Ideally distributed parameters would have univariate profile t traces that are straight lines, and bivariate likelihood profile traces that are two straight lines intersecting at right angles. This ideal is generally not met by nonlinear models, however, large deviations from the ideal indicate serious problems with the usual statistical inference.

Source

`fastcprof.src`

Chapter 4

Constrained Event Count and Duration Regression

by

Gary King

Department of Government

Harvard University

This module contains procedures for estimating statistical models of event count or duration data with general nonlinear equality and inequality constraints on the parameters

The programs included in this module implement maximum likelihood estimators for parametric statistical models of events data. Data based on events come in two forms: *event counts* and *durations* between events. Event counts are dependent variables that take on only nonnegative integer values, such as the number of wars in a year, the number of medical consultations in a month, the number of patents per firm, or even the frequency in the cell of a contingency table. Dependent variables that are measured as durations between events measure time and may take on any nonnegative real number; examples include the duration of parliamentary coalitions or time between coups d'etat. Note that the *same* underlying phenomena may be represented as either event counts (e.g., number of wars) or durations (time between wars), and some of the programs included in the *CONSTRAINED COUNT* module enable you to estimate exactly the same parameters with either form of data.

A variety of statistical models have been proposed to analyze events data, and the programs here provide some that I have developed, along with others I have found particularly useful in my research. I list here the specific programs included in this module, the models each program can estimate, and citations to the work for which I wrote each program. More complete references to the literature on event count and duration models appear at the end of this document.

CMLPoisson	Poisson regression (King, 1988, 1987), truncated Poisson regression (1989d: Section 5), and log-linear and log-proportion models for contingency tables (1989a: Chapter 6).
CMLNegbin	Negative binomial regression (1989b), truncated negative binomial regression (1989d: Section 5), truncated or untruncated variance function models (1989d: Section 5), overdispersed log-linear and log-proportion models for contingency tables (1989a: Chapter 6).
CMLHurdlep	Hurdle Poisson regression model (1989d: Section 4).
CMLSupreme	Seemingly unrelated Poisson regression model (1989c).
CMLSupreme2	Poisson regression model with unobserved dependent variables (1989d: Section 6).
CMLExpon	Exponential duration model with or without censoring (King, Alt, Burns, and Laver, 1989).
CMLExpgam	Exponential-Gamma duration model with or without censoring (King, Alt, Burns, and Laver, 1989).
CMLPareto	Pareto duration model with or without censoring (King, Alt, Burns, and Laver, 1989).

4.1 Getting Started

GAUSS 3.1.0+ is required to use these routines.

4.1.1 README Files

The file **README.ccn** contains any last minute information on this module. Please read it before using the procedures in this module.

4.1.2 Setup

In order to use the procedures in the *CONSTRAINED COUNT* Module, the **CMLCount** library must be active. This is done by including `count` in the **LIBRARY** statement at the top of your program or command file:

```
library cmlcount,quantal,pgraph;
```

This enables **GAUSS** to find the *CONSTRAINED COUNT* and required *CONSTRAINED MAXIMUM LIKELIHOOD* procedures. If you plan to make any right hand references to the global variables (which are described in a later section), you also need the statement:

```
#include cmlcount.ext;
```

To reset global variables in succeeding executions of the command file, the following instruction can be used:

```
cmlcountset;
```

This could be included with the above statements without harm and would insure the proper definition of the global variables for all executions of the command file.

The version number of each module is stored in a global variable. For the *CONSTRAINED COUNT* Module, this global is:

_cmlc_version 3×1 matrix, the first element contains the major version number of the *CONSTRAINED COUNT* Module, the second element the minor version number, and the third element the revision number.

If you call for technical support, you may be asked for the version number of your copy of this module.

4.2 About the *CONSTRAINED COUNT* Procedures

The format of the programs included in this module are all very similar:

```
{ b,vc,llik } =   CMLExpon(dataset,dep,ind);
{ b,vc,llik } =  CMLExpgam(dataset,dep,ind);
{ b,vc,llik } =  CMLPareto(dataset,dep,ind);
{ b,vc,llik } = CMLPoisson(dataset,dep,ind);
{ b,vc,llik } =  CMLNegbin(dataset,dep,ind1,ind2);
{ b,vc,llik } = CMLHurdlep(dataset,dep,ind1,ind2);
{ b,vc,llik } = CMLSupreme(dataset,dep1,dep2,ind1,ind2);
{ b,vc,llik } = CMLSupreme2(dataset,dep1,dep2,ind1,ind2,ind3);
```

An example program file looks like this:

```
library cmlcount;

CMLCountSet;

dep = { wars };
ind = { age, party, unem };
dataset = "sample";

_cml_A = { 0 1 -1 0 };
_cml_B = { 0 };
_cml_Bounds = { 0 10 };

call CMLPoisson(dataset,dep,ind);
```

This run constrains the coefficients of `age` and `party` to be equal, and bounds the coefficients to be positive and less than 10.

You may run these lines, or ones like them, from the **GAUSS** editor or interactively in command mode.

4.2.1 Inputs

The variable *dataset* is always the first argument. This may either be a matrix or a string containing the name of a **GAUSS** data set.

The dependent variable (or variables) is specified in each program by naming a symbol or a column number. For example,

```
dep = { durat };
```

or

```
dep = 7;
```

The independent variable vector (or vectors) is also specified in each program with variable names or column numbers. For example,

```
ind = { age, sex, race, height, size, iq };
```

or

```
ind = { 2, 4, 5, 6, 7 };
```

For each procedure, the data set and dependent variables must be specified. However, since constant terms are automatically included as part of independent variable vectors, you may occasionally wish to include no additional independent variables. You may do this easily by setting the relevant vector to zero. For example, $ind = 0$. For another example, you may wish to run the negative binomial regression model with a scalar dispersion parameter rather than a variance function: $ind2 = 0$.

4.2.2 Outputs

Printed output is controlled by the global **__output**, described in the section below. This section describes the outputs *b*, *vc*, and *llik* on the left hand side of the expressions above.

b vector, the maximum likelihood estimates for all the parameters. The mean vector comes first; the variance function, other mean vectors, and scalar dispersion parameters, if any, come next.

vc matrix, the variance-covariance matrix evaluated at the maximum. The standard errors are **SQRT(DIAG(***vc***))**. If you choose the **CML** global **_cml_CovPar** = 3, *vc* contains heteroskedastic-consistent parameter estimates.. See Section 2.8 for more discussion of options for statistical inference in constrained maximum likelihood models.

llik scalar, the value of the log-likelihood function at the maximum.

4.2.3 Global Control Variables

_cmlc_Inference string, determines the type of statistical inference.

BOOT generates bootstrapped estimates and covariance matrix of estimates

CML generates maximum likelihood estimates

Setting **_cmlc_Inference** to BOOT generates a **GAUSS** data set containing the bootstrapped parameters. The file name of this data set is either the default BOOTx, where x is a four digit number starting with 1 and increasing until a unique name is found, or the name in the **CML** global variable, **_cml_BootFname**. This data set can be used with **CMLBlimits** for generating confidence limits, with **CMLDensity** for generating density estimates and plots of the bootstrapped parameters, or with **CMLHist** for generating histogram and surface plots.

_cmlc_Censor scalar, allows you to include a variable indicating which observations are censored. This is used by the exponential, exponential-gamma, and Pareto models of duration data. Alternatively, you may set it to a symbol **_cmlc_Censor** = "varname" if you are using a **GAUSS** data set, or a number (**_cmlc_Censor** = 11) if the data set is a matrix in memory. The censoring variable should be 0 for censored observations and 1 for others.

By default, no observations are censored.

_cmlc_Fix scalar, name of index number of logged variable among the regressors with coefficient fixed to 1.0. By default, no logged variables are included.

In some of the programs, you have the option of including the log of a variable and fixing its coefficient to 1.0. To include the variable (the program takes the log), set **_cmlc_Fix** to a variable name or number (**_cmlc_Fix** = "totals" or **_cmlc_Fix** = 12). The default (**_cmlc_Fix** = 0) includes no additional variable. In most event count data, the observation period is the same length for all i (a year, month, etc.). However, in others, the observation period varies. For example, suppose one observed the number of times a citizen was contacted by a candidate in the interval between two public opinion polls; since polls typically take some time to administer, the observation period would vary over the individuals. In still other situations, the observation period may be the same length but the population of potential events varies. For example, if one observed the number of suicides per state, one would need some way to include information on differing state sizes in the analysis. It turns out that both of these situations can be dealt with in the same way by including an additional variable in the stochastic portion of the model. But (as explained in King, 1989, Section 5.8), this procedure turns out to be mathematically equivalent to including the log of this additional variable in the regression component, and constraining its coefficient to 1.0. There is often little harm in just including the log of this variable and estimating its coefficient with all the others, but several of the programs allow one to make this constraint.

_cmlc_Dispersion scalar, set this to a value to change the starting value for only the dispersion parameter in the negative binomial (**CMLNegbin**), generalized event count (**CMLHurdlep**), exponential-gamma (**CMLExpgam**), Pareto (**CMLPareto**), and seemingly-unrelated Poisson models (**CMLSupreme**, **CMLSupreme2**). By default, a special starting value is not used for the dispersion parameter.

_cmlc_Precision scalar, the number of digits printed to the right of the decimal point on output. Default = 4.

_cmlc_Start scalar, selects method of calculating starting values. Possible values are:

- **0** calculates them by regressing $\ln(y + 0.5)$ on the explanatory variables.
- **1** uses a vector of user supplied start values stored in the global variable **_cmlc_StartValue**.
- **2** uses a vector of zeros.
- **3** uses random uniform numbers on the interval $[-\frac{1}{2}, \frac{1}{2}]$.

Default = 0.

_cmlc_StartValue L×1 vector, start values if **_cmlc_Start** = 1.

_cmlc_ZeroTruncate scalar, specifies whether or not the model is a truncated model. For the Poisson and negative binomial models, **_cmlc_ZeroTruncate** = 0 estimates a truncated-at-zero version of the model. By default, the regular untruncated model is estimated.

__altnam K×1 vector, alternate names for variables when a matrix is passed to a **CMLCount** procedure. When a data matrix is passed to a **CMLCount** procedure and the user is selecting from that matrix, the global variable **__altnam**, if it is used, must contain names for the columns of the original matrix.

__output scalar, determines printing of intermediate results.

0	nothing is written.
1	serial ASCII output format suitable for disk files or printers.
2	(DOS only) output is suitable for screen only. ANSI.SYS must be active.

Default = 2.

__row scalar, specifies how many rows of the data set are read per iteration of the read loop. By default, the number of rows to be read is calculated automatically.

__rowfac scalar, row factor. If a *CONSTRAINED COUNT* procedure fails due to insufficient memory while attempting to read a **GAUSS** data set, then **__rowfac** may be set to some value between 0 and 1 to read a *proportion* of the original number of rows of the **GAUSS** data set. For example, setting

```
__rowfac = 0.8;
```

causes **GAUSS** to read in 80% of the rows originally calculated.

This global has an affect only when **__row** = 0.

Default = 1.

__title string, message printed at the top of the screen and printed out by **CMLCountPrt**. Default = "".

__vpad scalar, if *dataset* is a matrix in memory, the variable names are automatically created by **CML**. Two types of names can be created:

0	Variable names automatically created by **CML** are not padded to give them equal length. For example, V1, V2,...V10, V11,....

1 Variable names created by the procedure are padded with zeros to give them an equal number of characters. For example, V01, V02, ..., V10, V11,.... This is useful if you want the variable names to sort properly.

Default = 1.

4.2.4 Adding Constraints

There are two general types of constraints, nonlinear equality constraints and nonlinear inequality constraints. However, for computational convenience they are divided into five types: linear equality, nonlinear equality, linear inequality, nonlinear inequality, and bounds. For a discussion of specifying constraints, see Section 2.6.

The specification of constraints requires knowledge of the order of the parameters. For all models, the first parameter is a constant term, then one parameter for each explanatory variable, and then a dispersion parameter. For **CMLHurdlep** and **CMLSupreme2** another constant term and set of explanatory parameters follows the dispersion parameter. For example, suppose there are four explanatory variables, and you wish to constrain the coefficients and the dispersion parameter to be positive:

```
_cml_Bounds = { 0 1e200 };
```

To constrain the coefficients of the first two explanatory variables to be equal:

```
_cml_A = { 0 1 -1 0 0 0 };
_cml_B = { 0 };
```

To constrain the norm of the coefficients of the explanatory variables to be greater than 2:

```
proc eqp(b);
    local c;
    c = b[2:4];
    retp(c'c - 2);
endp;
_cml_EqProc = &eqp;
```

4.2.5 Statistical Inference

CML statistical inference features may be accessed through the **COUNT** global, **_cmlc_Inference**. **_cmlc_Inference** has the following settings:

CML	constrained maximum likelihood estimates (default)
BOOT	bootstrapped estimates

That is to generate bootstrapped estimates, set

```
_cmlc_Inference = "boot";
```

Confidence limits for inequality constrained parameters are generated by first leaving **_cmlc_Inference** at its default setting and then calling **CMLClimits** with the covariance matrix of the parameters and the parameter estimates as arguments.

Confidence Limits of Constrained Parameters

When inequality constraints are present, the considerations discussed in Section 2.8 are relevant. You may need to use **CMLClimits** for correct confidence limits. In this case, use the covariance matrix and estimates returned from the **CMLCount** procedure as input to **CMLClimits** as well as any globals and procedures used for the constraints. For example,

```
library cmlcount;
CMLCountSet;
dep = { wars };
ind = { age, party, unem };
dataset = "sample";

_cml_C = { 0 1 -1 0 };   /* constrains coefficient of age */
_cml_D = { 0 };          /* to be greater than that of party */

_cml_Bounds = { 0 10 };  /* bounds coefficients to be positive */
                         /* and less than 10 */

{ b,vc,llik } = CMLPoisson(dataset,dep,ind);

cl = CMLClimits(b,vc);

print "confidence limits of PARTY coefficient"
print cl[2,1] " <--> " cl[2,2];
```

Bootstrapping

In addition to the usual standard errors, you may generate bootstrap standard errors. Setting **_cmlc_Inference** to BOOT causes **CMLCOUNT** to call **CMLBoot**. This generates bootstrapped estimates and covariance matrices of the estimates.

The bootstrapped parameters are also stored in a **GAUSS** data set. The name of the data set can be determined by setting **_cml_BootFname** to a file name, or by default it will be set to BOOTx where x is a four digit number incremented from 0001 until a name not in use is found. For further details about the bootstrap, see Section 2.8.5.

The data set thus generated can be used for computing confidence intervals of the coefficients using **CMLBlimits**. Also, density estimates and plots can be generated using **CMLDensity**, and histograms and surface plots of the coefficients can be produced using **CMLHist**. For further details about **CMLDensity**, see Section 2.8.5, and for further details about **CMLHist** see Section 2.8.5.

4.2.6 Problems with Convergence

All the programs use maximum likelihood estimation by numerically maximizing a different likelihood function. As with virtually all nonlinear iterative procedures, convergence works most of the time, but not every time. Problems to be aware of include the following:

1. The explanatory variables in each regression function should not be highly collinear among themselves.
2. The model should have more observations than parameters; indeed, the more observations, the better.
3. Starting values should not be too far from the optimal values.
4. The model specified should fit the data.
5. The Poisson hurdle model must have at least some observations with $y_i = 0$ and should take on at least two other values greater than zero.
6. The truncated models should have no observations with zeros (if inadvertently included, a message appears and the program stops).
7. The models with scalar dispersion parameters and variance functions should have maximum likelihood estimates that are bounded so that, for example, in the negative binomial model $\hat{\sigma}^2 > 1$

If you avoid the potential problems listed in the last paragraph, you should have little problem with convergence. Of course, avoiding these problems with difficult data sets is not always easy nor obvious. In these cases, problems may be indicated by the following situations:

1. iterations sending the parameters off in unreasonable directions or creating very large numbers.

2. the program actually bombing out.

3. a single iteration taking an extraordinarily long time.

4. the program taking more than 40 or 50 iterations with no convergence.

If one of these problems occur, you have several options. First, look over the list in the last paragraph. To verify that the problem does indeed exist, you might try running your data on the Poisson regression model if you have event count data, or the exponential regression model if you have duration data. Both are known to be globally concave and tend to converge very easily. If this model works, but another does not, you probably do have a problem.

In the case of problems, you must consider iteration a participatory process. When **CML** is iterating, you can press **Alt-H** to receive a list of options that may be changed during iteration. See *CML REFERENCE* for a full explanation of each. I find that the following practices tend to work well:

1. If the program has produced many iterations without much progress, try pressing **Alt-I** every few iterations to force the program to calculate the information matrix or switch Newton-Raphson iterations. Either of these may not work if the iterations are not far enough along.

2. The number of zeros to the right of the decimal point on the relative gradients (printed on the screen while the program is iterating) is the approximate precision of your final estimates. If the program is having trouble converging, but the gradients are small enough (i.e., you have sufficient precision for your substantive problem), press **Alt-C** to force the program to converge.

3. If the program bombs out very quickly, changing the starting values are your best bet (with the global **_cmlc_Start**). The default starting values created with least squares, **_cmlc_Start** = 0, usually works best. If that does not work, you can also try creating them yourself, by thinking about what the answer is likely to be or by running a simpler model. For example, the exponential-gamma model is sometimes problematic; however, the exponential model often provides good starting values for the effect parameters. Thus if the other methods do not work, you might try the following:

```
library cmlcount;
CMLCountSet;
dep = { durat };
ind = { unem, infl, age };
dataset = "datafile";
{ b,vc,llik } = CMLExpon(dataset,dep,ind);
_cmlc_StartValue = b;
_cmlc_Start = 1;
call CMLExpgam(dataset,dep,ind);
```

You can also choose one of the other methods of creating starting values by changing the **_cmlc_Start** global (described above).

4.3 Annotated Bibliography

Allison, Paul. 1984. *Event History Analysis*. Beverly Hills: Sage. [A simple overview of event history methods for duration data.]

Bishop, Yvonne M.M.; Stephen E. Fienberg; and Paul W. Holland. 1975. *Discrete Multivariate Analysis* Cambridge, Mass.: M.I.T. Press. [Models for contingency tables.]

Cameron, A. Colin and Pravin K. Trivedi. 1986. "Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests," *Journal of Applied Econometrics* 1, 29–53. [Review of the econometric literature on event counts.]

Grogger, Jeffrey T. and Richard T. Carson. 1988. "Models for Counts from Choice Based Samples," Discussion Paper 88-9, Department of Economics, University of California, San Diego. [Truncated event count models.]

Gourieroux, C.; A. Monfort; and A. Trognon. 1984. "Pseudo Maximum Likelihood Methods: Applications to Poisson Models," *Econometrica* 52: 701–720. [A three-stage robust estimation method for count data.]

Hall, Bronwyn H.; Zvi Griliches; and Jerry A. Hausman. 1986. "Patents and R and D: Is there a Lag?" *International Economic Review*. 27, 2 (June): 265–83. [Nice example of a applying a variety of different estimators to single equation count models.]

Hausman, Jerry; Bronwyn H. Hall; and Zvi Griliches. 1984. "Econometrics Models for Count Data with An Application to the Patents-R&D Relationship," *Econometrica*. 52, 4 (July): 909-938. [Count models for time-series cross sectional panels.]

Holden, Robert T. 1987. "Time Series Analysis of a Contagious Process," *Journal of the American Statistical Association*. 82, 400 (December): 1019–1026. [A time series model of count data applied to airline hijack attempts.]

Jorgenson, Dale W. 1961. "Multiple Regression Analysis of a Poisson Process," *Journal of the American Statistical Association* 56,294 (June): 235–45. [The Poisson regression model.]

Kalbfleisch, J.D. and R.L. Prentice. 1980. *The Statistical Analysis of Failure Time Data*. New York: Wiley. [Summary of research on many models of duration data.]

King, Gary. 1989a. *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. New York: Cambridge University Press. [Introduction to likelihood, maximum likelihood, and a large variety of statistical models as special cases; Chapter 5 is discrete regression models.]

_____. 1989b. "Variance Specification in Event Count Models: From Restrictive Assumptions to a Generalized Estimator," *American Journal of Political Science*, vol. 33, no. 3 (August):762-784. [Poisson-based models with over- and under-dispersion.]

_____. 1989c. "A Seemingly Unrelated Poisson Regression Model," *Sociological Methods and Research*. 17, 3 (February, 1989): 235–255. [A model for simultaneously analyzing a pair of event count variables in a SURM framework.]

_____. 1989d. "Event Count Models for International Relations: Generalizations and Applications," *International Studies Quarterly*, vol. 33, no. 3 (June):123-147. [Hurdle models, truncated models, and models with unobserved dependent variables, all for event count data.]

_____. 1988. "Statistical Models for Political Science Event Counts: Bias in Conventional Procedures and Evidence for The Exponential Poisson Regression Model," *American Journal of Political Science*, 32, 3 (August): 838–863. [Introduction to count models; analytical and Monte Carlo comparisons of LS, logged-LS, and Poisson regression models.]

_____. 1987. "Presidential Appointments to the Supreme Court: Adding Systematic Explanation to Probabilistic Description," *American Politics Quarterly*, 15, 3 (July): 373–386. [An application of the Poisson regression model.]

King, Gary; James Alt; Nancy Burns; Michael Laver. 1990. "A Unified Model of Cabinet Duration in Parliamentary Democracies," American Journal of Political Science, vol. 34, no. 3 (August):846-871. [Exponential model of duration data with censoring.]

McCullagh, P. And J.A. Nelder 1983. *Generalized Linear Models*. London: Chapman and Hall. [A unified approach to specifying and estimating this class of models. Some count and duration models are covered.]

Mullahy, John. 1986. "Specification and Testing of Some Modified Count Data Models," *Journal of Econometrics*. 33: 341–65. [Several hurdle-type models of event count data.]

Tuma, Nancy Brandon and Michael T. Hannan. 1984. *Social Dynamics*. New York: Academic Press.

Chapter 5

CMLCount Reference

Purpose

Formats and prints the output from calls to *CONSTRAINED COUNT* procedures.

Library

`cmlcount`

Format

`{` *b*,*vc*,*llik* `}` = **CMLCountPrt(***b*,*vc*,*llik*);

Input

b	(K+1)×1 vector, maximum likelihood estimates of the effect parameters stacked on top of the dispersion parameter.
vc	(K+1)×(K+1) matrix, variance-covariance matrix
llik	scalar, value of the log-likelihood function at the maximum.

Output

The input arguments are returned unchanged.

Remarks

The call to *CONSTRAINED COUNT* procedures can be nested in the call to the **CMLCountPrt**:

```
{ b,vc,llik } = cmlcountprt(CMLExpgam(dataset,dep,ind));
```

Purpose

Formats and prints the output from calls to *CONSTRAINED COUNT* procedures with confidence limits

Library

`cmlcount`

Format

`{` *b*,*cl*,*llik* `}` = **CMLCountCLPrt(***b*,*cl*,*llik***);**

Input

b	(K+1)×1 vector, maximum likelihood estimates of the effect parameters stacked on top of the dispersion parameter.
vc	$(K+1) \times 2$ matrix, confidence limits
llik	scalar, value of the log-likelihood function at the maximum.

Output

The input arguments are returned unchanged.

Remarks

Confidence limits computed by **CMLBlimits**, **CMLClimits**, or **CMLTlimits** may be passed in the fourth argument in the call to **CMLCountCLPrt**:

```
{ b,vc,llik } = CMLExpgam(dataset,dep,ind);
 cl = CMLBlimits(_cml_BootFname);
 call CMLCountCLPrt(b,cl,llik);
```

Source

`ccount.src`

Purpose

Resets *CONSTRAINED COUNT* global variables to default values.

Library

`cmlcount`

Format

CMLCountSet;

Input

None

Output

None

Remarks

Putting this instruction at the top of all command files that invoke *CONSTRAINED COUNT* procedures is generally good practice. This prevents globals from being inappropriately defined when a command file is run several times or when a command file is run after another command file has executed that calls a *CONSTRAINED COUNT* procedure.

CMLCountSet calls **CMLSet** which calls **GAUSSET**.

Source

`ccount.src`

Purpose

Estimates an exponential-gamma regression model, for the analysis of duration data, with maximum likelihood.

Library

`cmlcount`

Format

`{` *b*,*vc*,*llik* `}` = **CMLExpgam(***dataset*,*dep*,*ind***);**

Input

dataset	string, name of **GAUSS** data set. – or – N×K matrix, data
dep	string, the name of the dependent variable. – or – scalar, the index of the dependent variable.
ind	K×1 character vector, names of the independent variables. – or – K×1 numeric vector, indices of independent variables. Set to 0 to include only a constant term.

If *dataset* is a matrix, *dep* or *ind* may be a string or character variable containing either the standard labels created by **CML** (V1, V2,..., or V01, V02,...., depending on the value of **__vpad**), or the user-provided labels in **__altnam**.

Output

b	(K+1)×1 vector, maximum likelihood estimates of the effect parameters stacked on top of the dispersion parameter.
vc	(K+1)×(K+1) matrix, variance-covariance matrix of the estimated parameters evaluated at the maximum. If you choose the **CML** global **__cml_CovPar** = 3, *vc* contains heteroskedastic-consistent parameter estimates.
llik	scalar, value of the log-likelihood function at the maximum.

Globals

CML globals are also relevant, including constraint matrices and procedures.

_cmlc_Inference string, determines the type of statistical inference.

BOOT generates bootstrapped estimates and covariance matrix of estimates

CML generates maximum likelihood estimates

Setting **_cmlc_Inference** to BOOT generates a **GAUSS** data set containing the bootstrapped parameters. The file name of this data set is either the default BOOTx, where x is a four digit number starting with 1 and increasing until a unique name is found, or the name in the **CML** global variable, **_cml_BootFname**. This data set can be used with **CMLBlimits** for generating confidence limits, with **CMLDensity** for generating density estimates and plots of the boostrapped parameters, or with **CMLHist** for generating histogram and surface plots.

_cmlc_Censor string, the name of the censor variable from *dataset*.
– or –
scalar, the index of the censor variable from *dataset*.
By default, no censoring is used.

_cmlc_Start scalar, selects method of calculating starting values. Possible values are:

0 calculates them by regressing $\ln(y + 0.5)$ on the explanatory variables.

1 uses a vector of user supplied start values stored in the global variable **_cmlc_StartValue**.

2 uses a vector of zeros.

3 uses random uniform numbers on the interval $[-\frac{1}{2}, \frac{1}{2}]$.

Default = 0.

_cmlc_StartValue (K+1)×1 vector, start values if **_cmlc_Start** = 1.

_cmlc_Precision scalar, number of decimal points to print on output. Default = 4.

__altnam K×1 vector, alternate names for variables when a matrix is passed to **CMLExpgam**. When a data matrix is passed to **CMLExpgam** and when the user is selecting from that matrix, the global variable **__altnam**, if it is used, must contain names for the columns of the original matrix.

__miss scalar, determines how missing data will be handled.

0 Missing values will not be checked for, and so the data set must not have any missings. This is the fastest option.

1 Listwise deletion. Removes from computation any observation with a missing value on any variable included in the analysis.

Default = 0.

__output scalar, determines printing of intermediate results.

0	nothing is written.
1	serial ASCII output format suitable for disk files or printers.
2	(DOS only) output is suitable for screen only. ANSI.SYS must be active.

Default = 2.

__row scalar, specifies how many rows of the data set will be read per iteration of the read loop. By default, the number of rows to be read will be calculated automatically.

__rowfac scalar, row factor. If **CMLExpgam** fails due to insufficient memory while attempting to read a **GAUSS** data set, then **__rowfac** may be set to some value between 0 and 1 to read a *proportion* of the original number of rows of the **GAUSS** data set. For example, setting

```
__rowfac = 0.8;
```

will cause **GAUSS** to read in 80% of the rows originally calculated.

This global has an affect only when **__row** = 0.

Default = 1.

__title string, message printed at the top of the screen and printed out by **CMLCountPrt**. Default = "".

__vpad scalar, if *dataset* is a matrix in memory, the variable names are automatically created by **CML**. Two types of names can be created:

0	Variable names automatically created by **CML** are not padded to give them equal length. For example, V1, V2,...V10, V11,....
1	Variable names created by the procedure are padded with zeros to give them an equal number of characters. For example, V01, V02, ..., V10, V11,.... This is useful if you want the variable names to sort properly.

Default = 1.

Remarks

Let the n duration observations (nonnegative real numbers) for the dependent variable be denoted as $y_1, \ldots, y_n$. Assume that y_i follows a gamma distribution with expected

value μ_i and variance $\mu_i^2\sigma^2$. Let the mean μ_i be an exponential-linear function of a vector of explanatory variables, x_i:

$$E(y_i) \equiv \mu_i = \exp(x_i\beta) \tag{5.1}$$

The program includes a constant term as the first column of x_i and allows one to include any number of explanatory variables. Note that μ_i from a duration model equals $1/\lambda_i$ from an event count model; thus, one need only change the sign of the effect parameters to get estimates of the same parameters from these different kinds of data.

The dispersion σ^2 is parametrized as follows:

$$\sigma_i^2 = \exp(\gamma) \tag{5.2}$$

EXPGAM reports estimates of β and γ.

For an introduction to the exponential gamma regression model see King, Alt, Burns, and Laver (1989) or Kalbfleisch and Prentice (1980).

Example

Constrained Exponential-Gamma Regression Model of Duration Data

```
library cmlcount;
#include cmlcount.ext;
CMLCountset;
dataset = "wars";
dep = { wars };
ind = { unem, poverty, allianc };
_cml_Bounds = { 0 10 };  /* constrains coefficients */
                         /* to be positive */
{ b,vc,llik } = CMLExpgam(dataset,dep,ind);
output file = cmlcount.out reset;
call CMLCountPrt(b,vc,llik);
output off;
```

A vector of effect parameters and a scalar dispersion parameter are estimated. The vector includes one element corresponding to each explanatory variable named in *ind* and a constant term. Five parameters are estimated in this example.

Constrained Censored Exponential-Gamma Regression Model of Duration Data

```
library cmlcount;
#include cmlcount.ext;
CMLCountset;
```

```
dataset = "wars";
dep = { wars };
ind = { unem, poverty, allianc };
_Censor = { v12 };
{ b,vc,llik } = CMLExpgam(dataset,dep,ind);
output file = cmlcount.out reset;
call CMLCountPrt(b,vc,llik);
output off;
```

A vector of effect parameters and a scalar dispersion parameter are estimated. The vector includes one element corresponding to each explanatory variable named in *ind* and a constant term. Five parameters are estimated in this example.

Source

`cmlexpgm.src`

Purpose

Estimates a constrained exponential regression model or censored exponential regression model with maximum likelihood.

Library

`cmlcount`

Format

`{` *b*,*vc*,*llik* `}` = **CMLExpon(***dataset*,*dep*,*ind***);**

Input

dataset	string, name of **GAUSS** data set. – or – N×K matrix, data
dep	string, the name of the dependent variable – or – scalar, the index of the dependent variable
ind	K×1 character vector, names of the independent variables – or – K×1 numeric vector, indices of independent variables Set to 0 to include only a constant term.

If *dataset* is a matrix, *dep* or *ind* may be a string or character variable containing either the standard labels created by **CML** (V1, V2,..., or V01, V02,...., depending on the value of **__vpad**), or the user-provided labels in **__altnam**.

Output

b	K×1 vector, maximum likelihood estimates of the effect parameters.
vc	K×K matrix, variance-covariance matrix of the estimated parameters evaluated at the maximum. If the **CML** global **__cml_CovPar** is set to 3, *vc* will contain heteroskedastic-consistent parameter estimates.
llik	scalar, value of the log-likelihood function at the maximum.

Globals

CML globals are also relevant, including constraint matrices and procedures.

_cmlc_Inference string, determines the type of statistical inference.

BOOT generates bootstrapped estimates and covariance matrix of estimates

CML generates maximum likelihood estimates

Setting **_cmlc_Inference** to BOOT generates a **GAUSS** data set containing the bootstrapped parameters. The file name of this data set is either the default BOOTx, where x is a four digit number starting with 1 and increasing until a unique name is found, or the name in the **CML** global variable, **_cml_BootFname**. This data set can be used with **CMLBlimits** for generating confidence limits, with **CMLDensity** for generating density estimates and plots of the boostrapped parameters, or with **CMLHist** for generating histogram and surface plots.

_cmlc_Censor string, the name of the censor variable from *dataset*
– or –
scalar, the index of the censor variable from *dataset*
By default, no censoring is used.

_cmlc_Start scalar, selects method of calculating starting values. Possible values are:

0 calculates them by regressing $\ln(y + 0.5)$ on the explanatory variables.

1 will use a vector of user supplied start values stored in the global variable **_cmlc_StartValue**.

2 uses a vector of zeros.

3 uses random uniform numbers on the interval $[-\frac{1}{2}, \frac{1}{2}]$.

Default = 0.

_cmlc_StartValue K×1 vector, start values if **_cmlc_Start** = 1.

_cmlc_Precision scalar, number of decimal points to print on output. Default = 4.

__altnam K×1 vector, alternate names for variables when a matrix is passed to **CMLExpon**. When a data matrix is passed to **CMLExpon** and the user is selecting from that matrix, the global variable **__altnam**, if it is used, must contain names for the columns of the original matrix.

__miss scalar, determines how missing data will be handled.

0 Missing values will not be checked for, and so the data set must not have any missings. This is the fastest option.

1 Listwise deletion. Removes from computation any observation with a missing value on any variable included in the analysis.

Default = 0.

__output scalar, determines printing of intermediate results.

0	nothing is written.
1	serial ASCII output format suitable for disk files or printers.
2	(DOS only) output is suitable for screen only. ANSI.SYS must be active.

Default = 2.

__row scalar, specifies how many rows of the data set will be read per iteration of the read loop. By default, the number of rows to be read will be calculated automatically.

__rowfac scalar, row factor. If **EXPON** fails due to insufficient memory while attempting to read a **GAUSS** data set, then **__rowfac** may be set to some value between 0 and 1 to read a *proportion* of the original number of rows of the **GAUSS** data set. For example, setting

```
__rowfac = 0.8;
```

will cause **GAUSS** to read in 80% of the rows originally calculated.

This global only has an affect when **__row** = 0.

Default = 1.

__title string, message printed at the top of the screen and printed out by **CMLCountPrt**. Default = "".

__vpad scalar, if *dataset* is a matrix in memory, the variable names are automatically created by **CML**. Two types of names can be created:

0	Variable names automatically created by **CML** are not padded to give them equal length. For example, V1, V2,...V10, V11,....
1	Variable names created by the procedure are padded with zeros to give them an equal number of characters. For example, V01, V02, ..., V10, V11,.... This is useful if you want the variable names to sort properly.

Default = 1.

Remarks

Let y_i $(i = 1, \ldots, n)$ take on any non-negative real number representing a duration. Often y_i is only measured as an integer, such as the number of days or months. Even so, if your dependent variable is a measure of time, duration models, and not event count models, are called for. Let y_i be distributed exponentially with mean μ_i. Also let

$E(y_i) \equiv \mu_i = \exp(x_i\beta)$. Note that μ_i from a duration model equals $1/\lambda_i$ from an event count model; thus, one need only change the sign of the effect parameters to get estimates of the same parameters from these different kinds of data.

For an introduction to the exponential regression model and the censored exponential regression model see Kalbfleisch and Prentice (1980) and King, Alt, Burns, and Laver (1989).

Example

Constrained Exponential Regression Model

```
library cmlcount;
#include cmlcount.ext;
CMLCountset;
dataset = "wars";
dep = { wars };
ind = { unem, poverty, allianc };
_cml_Bounds = { 0 10 };  /* constrains coefficients */
                         /* to be positive */
{ b,vc,llik } = CMLExpon(dataset,dep,ind);
output file = cmlcount.out on;
call CMLCountPrt(b,vc,llik);
output off;
```

A single vector of effect parameters are estimated. This vector includes one element corresponding to each explanatory variable named in *ind* and a constant term.

Constrained Censored Exponential Regression Model

```
library cmlcount;
#include cmlcount.ext;
CMLCountset;
dataset = "wars";
dep = { wars };
ind = { unem, poverty, allianc };
_cmlc_Censor = { notseen };
{ b,vc,llik } = CMLExpon(dataset,dep,ind);
output file = cmlcount.out reset;
call CMLCountPrt(b,vc,llik);
output off;
```

A single vector of effect parameters are estimated. This vector includes one element corresponding to each explanatory variable named in *ind* and a constant term.

Source

cmlexpon.src

Purpose

Estimates a constrained hurdle Poisson regression model, for the analysis of event counts, with maximum likelihood.

Library

`cmlcount`

Format

`{` *b*,*vc*,*llik* `}` = **CMLHurdlep(***dataset*,*dep*,*ind*);

Input

dataset	string, name of **GAUSS** data set. – or – N×K matrix, data
dep	string, the name of the dependent variable – or – scalar, the index of the dependent variable
ind1	K×1 character vector, names of first event independent variables – or – K×1 numeric vector, indices of first event independent variables
ind2	K×1 character vector, names of second event independent variables – or – K×1 numeric vector, indices of second event independent variables

If *dataset* is a matrix, *dep*, *ind1*, or *ind2* may be a string or character variable containing either the standard labels created by **CML** (V1, V2,..., or V01, V02,...., depending on the value of **__vpad**), or the user-provided labels in **__altnam**.

Output

b	(K+L)×1 vector, maximum likelihood estimates of the effect parameters stacked on top of the dispersion parameter.
vc	(K+L)×(K+L) matrix, variance-covariance matrix of the estimated parameters evaluated at the maximum. If you choose the **CML** global **__cml_CovPar** = 3, *vc* will contain heteroskedastic-consistent parameter estimates.
llik	scalar, value of the log-likelihood function at the maximum.

Globals

CML globals are also relevant, including constraint matrices and procedures.

__cmlc_Inference string, determines the type of statistical inference.

BOOT generates bootstrapped estimates and covariance matrix of estimates

CML generates maximum likelihood estimates

Setting **__cmlc_Inference** to BOOT generates a **GAUSS** data set containing the bootstrapped parameters. The file name of this data set is either the default BOOTx, where x is a four digit number starting with 1 and increasing until a unique name is found, or the name in the **CML** global variable, **__cml_BootFname**. This data set can be used with **CMLBlimits** for generating confidence limits, with **CMLDensity** for generating density estimates and plots of the boostrapped parameters, or with **CMLHist** for generating histogram and surface plots.

__cmlc_Start scalar, selects method of calculating starting values. Possible values are:

0 calculates them by regressing $\ln(y + 0.5)$ on the explanatory variables.

1 will use a vector of user supplied start values stored in the global variable **__cmlc_StartValue**.

2 uses a vector of zeros.

3 uses random uniform numbers on the interval $[-\frac{1}{2}, \frac{1}{2}]$.

Default = 0.

__cmlc_StartValue (K+L)×1 vector, start values if **__cmlc_Start** = 1.

__cmlc_Precision scalar, number of decimal points to print on output. Default = 4.

___altnam K×1 vector, alternate names for variables when a matrix is passed to **CMLHurdlep**. When a data matrix is passed to **CMLHurdlep** and the user is selecting from that matrix, the global variable **___altnam**, if it is used, must contain names for the columns of the original matrix.

___miss scalar, determines how missing data will be handled.

0 Missing values will not be checked for, and so the data set must not have any missings. This is the fastest option.

1 Listwise deletion. Removes from computation any observation with a missing value on any variable included in the analysis.

Default = 0.

__output scalar, determines printing of intermediate results.

0	nothing is written.
1	serial ASCII output format suitable for disk files or printers.
2	(DOS only) output is suitable for screen only. ANSI.SYS must be active.

Default = 2.

__row scalar, specifies how many rows of the data set will be read per iteration of the read loop. By default, the number of rows to be read will be calculated automatically.

__rowfac scalar, row factor. If **CMLHurdlep** fails due to insufficient memory while attempting to read a **GAUSS** data set, then **__rowfac** may be set to some value between 0 and 1 to read a *proportion* of the original number of rows of the **GAUSS** data set. For example, setting

```
__rowfac = 0.8;
```

will cause **GAUSS** to read in 80% of the rows originally calculated.

This global only has an affect when **__row** = 0.

Default = 1.

__title string, message printed at the top of the screen and printed out by **CMLCountPrt**. Default = "".

__vpad scalar, if *dataset* is a matrix in memory, the variable names are automatically created by **CML**. Two types of names can be created:

0	Variable names automatically created by **CML** are not padded to give them equal length. For example, V1, V2,...V10, V11,....
1	Variable names created by the procedure are padded with zeros to give them an equal number of characters. For example, V01, V02, ..., V10, V11,.... This is useful if you want the variable names to sort properly.

Default = 1.

Remarks

Let the n event count observations (nonnegative integers) for the dependent variable be denoted as $y_1, \ldots, y_n$. y_i is then a random dependent variable representing the number of events that have occurred during observation period i. Let λ_{0i} be the rate of the first event occurrence and λ_{+i} be the rate for all additional events after the first. If these are

the expected values of two separate Poisson processes, we have the hurdle Poisson regression model. These means are parametrized as usual:

$$\lambda_{0i} = \exp(x_i\beta) \tag{5.3}$$

and

$$\lambda_{+i} = \exp(z_i\gamma) \tag{5.4}$$

where x_i and z_i are (possibly) different vectors of explanatory variables. The program produces estimates of β and γ. If $\beta = \gamma$ and $x = z$, this model reduces to the Poisson.

For an introduction to the Hurdle Poisson regression model see Mullahy (1986) and King (1989d).

Example

Constrained Hurdle Poisson Regression Model:

```
library cmlcount;
#include cmlcount.ext;
CMLCountset;
dataset = "wars";
dep = { wars };
ind1 = { unem, poverty, allianc };
ind2 = { race, sex, age, partyid, x4, v5 };
_cml_Bounds = { 0 10 };  /* constrains coefficients */
                         /* to be positive */
{ b,vc,llik } =  CMLHurdlep(dataset,dep,ind1,ind2);
output file = cmlcount.out reset;
call CMLCountPrt(b,vc,llik);
output off;
```

Two vectors of effect parameters are estimated. Each includes one element corresponding to each explanatory variable plus a constant term (in the example, four parameters appear in the first regression function and seven in the second).

Source

`cmlhurdl.src`

Purpose

Estimates a constrained negative binomial regression model or truncated-at-zero negative binomial regression model with maximum likelihood.

Library

`cmlcount`

Format

`{` *b*,*vc*,*llik* `}` = **CMLNegbin(***dataset*,*dep*,*ind1*,*ind2*);

Input

dataset string, name of **GAUSS** data set.
– or –
N×K matrix, data

dep string, the name of the dependent variable
– or –
scalar, the index of the dependent variable

ind1 K×1 character vector, names of first event independent variables
– or –
K×1 numeric vector, indices of first event independent variables
Set to 0 to include only a constant term.

ind2 K×1 character vector, names of second event independent variables
– or –
K×1 numeric vector, indices of second event independent variables
Set to 0 for a scalar dispersion parameter.

If *dataset* is a matrix, *dep*, *ind1*, or *ind2* may be a string or character variable containing either the standard labels created by **CML** (V1, V2,..., or V01, V02,....., depending on the value of **__vpad**), or the user-provided labels in **__altnam**.

Output

b (K+1)×1 or (K+L)×1 vector, maximum likelihood estimates of the effect parameters stacked on top of either the dispersion parameter or the coefficients of the variance function.

vc (K+1)×(K+1) or (K+L)×(K+L) matrix, variance-covariance matrix of the estimated parameters evaluated at the maximum. If you choose the **CML** global **__cml_CovPar** = 3, *vc* will contain heteroskedastic-consistent parameter estimates.

llik scalar, value of the log-likelihood function at the maximum.

Globals

CML globals are also relevant, including constraint matrices and procedures.

_cmlc_Inference string, determines the type of statistical inference.

BOOT generates bootstrapped estimates and covariance matrix of estimates

CML generates maximum likelihood estimates

Setting **_cmlc_Inference** to BOOT generates a **GAUSS** data set containing the bootstrapped parameters. The file name of this data set is either the default BOOTx, where x is a four digit number starting with 1 and increasing until a unique name is found, or the name in the **CML** global variable, **_cml_BootFname**. This data set can be used with **CMLBlimits** for generating confidence limits, with **CMLDensity** for generating density estimates and plots of the boostrapped parameters, or with **CMLHist** for generating histogram and surface plots.

_cmlc_Fix scalar, name of index number of logged variable among the regressors with coefficient constrained to 1.0 By default, no logged variables are included.

_cmlc_Start scalar, selects method of calculating starting values. Possible values are:

0 calculates them by regressing $\ln(y + 0.5)$ on the explanatory variables.

1 will use a vector of user supplied start values stored in the global variable **_cmlc_StartValue**.

2 uses a vector of zeros.

3 uses random uniform numbers on the interval $[-\frac{1}{2}, \frac{1}{2}]$.

Default = 0.

_cmlc_StartValue (K+1)×1 or (K+L)×1 vector, start values if **_cmlc_Start** = 1.

_cmlc_Dispersion scalar, start value for scalar dispersion parameter. Default = 3.

_cmlc_Precision scalar, number of decimal points to print on output. Default = 4.

_cmlc_ZeroTruncate scalar, specifies which model is used:

0 truncated-at-zero negative binomial model

1 negative binomial model is used.

__altnam K×1 vector, alternate names for variables when a matrix is passed to **CMLNegbin**. When a data matrix is passed to **CMLNegbin** and the user is selecting from that matrix, the global variable **__altnam**, if it is used, must contain names for the columns of the original matrix.

__miss scalar, determines how missing data will be handled.

0 Missing values will not be checked for, and so the data set must not have any missings. This is the fastest option.

1 Listwise deletion. Removes from computation any observation with a missing value on any variable included in the analysis.

Default = 0.

__output scalar, determines printing of intermediate results.

0 nothing is written.

1 serial ASCII output format suitable for disk files or printers.

2 (DOS only) output is suitable for screen only. ANSI.SYS must be active.

Default = 2.

__row scalar, specifies how many rows of the data set will be read per iteration of the read loop. By default, the number of rows to be read will be calculated automatically.

__rowfac scalar, row factor. If **CMLNegbin** fails due to insufficient memory while attempting to read a **GAUSS** data set, then **__rowfac** may be set to some value between 0 and 1 to read a *proportion* of the original number of rows of the **GAUSS** data set. For example, setting

```
__rowfac = 0.8;
```

will cause **GAUSS** to read in 80% of the rows originally calculated.

This global only has an affect when **__row** = 0.

Default = 1.

__title string, message printed at the top of the screen and printed out by **CMLCountPrt**. Default = "".

__vpad scalar, if *dataset* is a matrix in memory, the variable names are automatically created by **CML**. Two types of names can be created:

0 Variable names automatically created by **CML** are not padded to give them equal length. For example, V1, V2,...V10, V11,....

1 Variable names created by the procedure are padded with zeros to give them an equal number of characters. For example, V01, V02, ..., V10, V11,.... This is useful if you want the variable names to sort properly.

Default = 1.

Remarks

Let y_i be a random dependent variable representing the number of events that have occurred during observation period i ($i = 1, \ldots, n$). Assume that y_i follows a negative binomial distribution with expected value λ_i and variance $\lambda_i\sigma^2$. Let the mean λ_i (the rate of event occurrence, which must be greater than zero) be an exponential-linear function of a vector of explanatory variables, x_i:

$$E(y_i) \equiv \lambda_i = \exp(x_i\beta) \tag{5.5}$$

The program includes a constant term as the first column of x_i and allows one to include any number of explanatory variables.

σ^2 is parametrized as follows:

$$\sigma_i^2 = 1 + \exp(z_i\gamma) \tag{5.6}$$

where $z_i = 1$, if estimating a scalar dispersion parameter, or a vector of explanatory variables, if estimating a variance function. The program calculates estimates of β and γ.

For an introduction to the negative binomial regression model, see Hausman, Hall, and Griliches (1984) and King (1989b); for information on the truncated negative binomial model, see Grogger and Carson (1988), and on the variance function model with or without truncation see King (1989d: Section 5)

Example

Constrained Negative Binomial Regression Model

```
library cmlcount;
#include cmlcount.ext;
CMLCountset;
dataset = "wars";
dep = { wars };
ind1 = { unem, poverty, allianc };
_cml_Bounds = { 0 10 };  /* constrains coefficients */
                         /* to be positive */
{ b,vc,llik } = CMLNegbin(dataset,dep,ind1,0);
output file = cmlcount.out reset;
call CMLCountPrt(b,vc,llik);
output off;
```

A single vector of effect parameters and one scalar dispersion parameter are estimated. The vector of effect parameters includes one element corresponding to each explanatory variable and a constant term. In the example, five parameters are estimated.

Constrained Negative Binomial Variance Function Regression Model

```
library cmlcount;
#include cmlcount.ext;
CMLCountset;
dataset = "wars";
dep1 = { wars };
ind1 = { unem, poverty, allianc };
ind2 = { partyid, x4 };
{ b,vc,llik } = CMLNegbin(dataset,dep,ind1,ind2);
output file = cmlcount.out reset;
call CMLCountPrt(b,vc,llik);
output off;
```

Two vectors of effect parameters are estimated, one for the mean *ind1* and one for the variance function *ind2*. Each vector includes a constant term and one element corresponding to each explanatory variable. The example estimates seven parameters.

Constrained Truncated-at-zero Negative Binomial Regression Model

```
library cmlcount;
#include cmlcount.ext;
CMLCountset;
dataset = "wars";
dep1 = { wars };
ind1 = { unem, poverty, allianc };
_cmlc_ZeroTruncate = 0;
{ b,vc,llik } = CMLNegbin(dataset,dep,ind1,0);
output file = cmlcount.out reset;
call CMLCountPrt(b,vc,llik);
output off;
```

A single vector of effect parameters and one scalar dispersion parameter are estimated. The vector of effect parameters includes one element corresponding to each explanatory variable and a constant term. In the example, five parameters are estimated.

Constrained Truncated-at-zero Negative Binomial Variance Function Regression Model

```
library cmlcount;
#include cmlcount.ext;
CMLCountset;
dataset = "wars";
dep1 = { wars };
ind1 = { unem, poverty, allianc };
ind2 = { partyid, x4 };
_cmlc_ZeroTruncate = 0;
{ b,vc,llik } = CMLNegbin(dataset,dep,ind1,0);
output file = cmlcount.out reset;
call CMLCountPrt(b,vc,llik);
output off;
```

Two vectors of effect parameters are estimated, one for the mean and one for the variance function. Each vector includes a constant term and one element corresponding to each explanatory variable. In the example, the variables specified in *ind1* pertain to the expected value and *ind2* to the variance. Seven parameters are estimated.

Source

`cmlnegbn.src`

Purpose

Estimates a constrained Pareto regression model, for the analysis of duration data, with maximum likelihood.

Library

`cmlcount`

Format

`{` *b*,*vc*,*llik* `}` = **CMLPareto(***dataset*,*dep*,*ind***);**

Input

dataset	string, name of **GAUSS** data set. – or – N×K matrix, data
dep	string, the name of the dependent variable – or – scalar, the index of the dependent variable
ind	K×1 character vector, names of the independent variables – or – K×1 numeric vector, indices of independent variables Set to 0 to include only a constant term.

If *dataset* is a matrix, *dep* and *ind* may be a string or character variable containing either the standard labels created by **CML** (V1, V2,..., or V01, V02,...., depending on the value of **__vpad**), or the user-provided labels in **__altnam**.

Output

b	(K+1)×1 vector, maximum likelihood estimates of the effect parameters stacked on top of the dispersion parameter.
vc	(K+1)×(K+1) matrix, variance-covariance matrix of the estimated parameters evaluated at the maximum. If the **CML** global **__cml_CovPar** is set to 3, *vc* will contain heteroskedastic-consistent parameter estimates.
llik	scalar, value of the log-likelihood function at the maximum.

Globals

CML globals are also relevant, including constraint matrices and procedures.

_cmlc_Inference string, determines the type of statistical inference.

BOOT generates bootstrapped estimates and covariance matrix of estimates

CML generates maximum likelihood estimates

Setting **_cmlc_Inference** to BOOT generates a **GAUSS** data set containing the bootstrapped parameters. The file name of this data set is either the default BOOTx, where x is a four digit number starting with 1 and increasing until a unique name is found, or the name in the **CML** global variable, **_cml_BootFname**. This data set can be used with **CMLBlimits** for generating confidence limits, with **CMLDensity** for generating density estimates and plots of the boostrapped parameters, or with **CMLHist** for generating histogram and surface plots.

_cmlc_Censor string, the name of the censor variable from *dataset*
– or –
scalar, the index of the censor variable from *dataset*

Each element of censor variable is 0 if censored, or 1 if not.

By default, no censoring is used.

_cmlc_Start scalar, selects method of calculating starting values. Possible values are:

0 calculates them by regressing $\ln(y + 0.5)$ on the explanatory variables.

1 will use a vector of user supplied start values stored in the global variable **_cmlc_StartValue**.

2 uses a vector of zeros.

3 uses random uniform numbers on the interval $[-\frac{1}{2}, \frac{1}{2}]$.

Default = 0.

_cmlc_StartValue (K+1)×1 vector, start values if **_cmlc_Start** = 1.

_cmlc_Dispersion scalar, start value for scalar dispersion parameter. Default = 3.

_cmlc_Precision scalar, number of decimal points to print on output. Default = 4.

___altnam K×1 vector, alternate names for variables when a matrix is passed to **CMLPareto**. When a data matrix is passed to **CMLPareto** and the user is selecting from that matrix, the global variable **___altnam**, if it is used, must contain names for the columns of the original matrix.

___miss scalar, determines how missing data will be handled.

0 Missing values will not be checked for, and so the data set must not have any missings. This is the fastest option.

1 Listwise deletion. Removes from computation any observation with a missing value on any variable included in the analysis.

Default = 0.

__output scalar, determines printing of intermediate results.

0 nothing is written.

1 serial ASCII output format suitable for disk files or printers.

2 (DOS only) output is suitable for screen only. ANSI.SYS must be active.

Default = 2.

__row scalar, specifies how many rows of the data set will be read per iteration of the read loop. By default, the number of rows to be read will be calculated automatically.

__rowfac scalar, row factor. If **CMLPareto** fails due to insufficient memory while attempting to read a **GAUSS** data set, then **__rowfac** may be set to some value between 0 and 1 to read a *proportion* of the original number of rows of the **GAUSS** data set. For example, setting

```
__rowfac = 0.8;
```

will cause **GAUSS** to read in 80% of the rows originally calculated.

This global only has an affect when **__row** = 0.

Default = 1.

__title string, message printed at the top of the screen and printed out by **CMLCountPrt**. Default = "".

__vpad scalar, if *dataset* is a matrix in memory, the variable names are automatically created by **CML**. Two types of names can be created:

0 Variable names automatically created by **CML** are not padded to give them equal length. For example, V1, V2,...V10, V11,....

1 Variable names created by the procedure are padded with zeros to give them an equal number of characters. For example, V01, V02, ..., V10, V11,.... This is useful if you want the variable names to sort properly.

Default = 1.

Remarks

Let the n duration observations (non-negative real numbers) for the dependent variable be denoted as $y_1, \ldots, y_n$. Assume that y_i follows a Pareto distribution with expected value μ_i and variance $\mu_i\sigma^2 + \mu_i^2$. Let the mean μ_i be an exponential-linear function of a vector of explanatory variables, x_i:

$$E(y_i) \equiv \mu_i = \exp(x_i\beta) \tag{5.7}$$

The program includes a constant term as the first column of x_i and allows one to include any number of explanatory variables. Note that μ_i from a duration model equals $1/\lambda_i$ from an event count model; thus, one need only change the sign of the effect parameters to get estimates of the same parameters from these different kinds of data.

The dispersion σ^2 is parametrized as follows:

$$\sigma_i^2 = \exp(\gamma) \tag{5.8}$$

The program gives estimates of β and γ.

For an introduction to the Pareto regression model see Hannan and Tuma (1984) and King, Alt, Burns, and Laver (1989).

Example

Pareto Regression Model of Duration Data

```
library cmlcount;
#include cmlcount.ext;
CMLCountset;
dataset = "wars";
dep = { wars };
ind = { unem, poverty, allianc };
_cml_Bounds = { 0 10 };  /* constrains coefficients */
                         /* to be positive */
{ b,vc,llik } = CMLPareto(dataset,dep,ind);
output file = cmlcount.out reset;
call CMLCountPrt(b,vc,llik);
output off;
```

A vector of effect parameters and a scalar dispersion parameter are estimated. The vector includes one element corresponding to each explanatory variable named in *ind* and a constant term. Five parameters are estimated in this example.

Constrained Censored Pareto Regression Model of Duration Data

```
library cmlcount;
#include cmlcount.ext;
CMLCountset;
dataset = "wars";
dep = { wars };
ind = { unem, poverty };
_cmlc_Censor = { cvar };
_cml_Bounds = { 0 10 };  /* constrains coefficients */
                         /* to be positive */
{ b,vc,llik } = CMLPareto(dataset,dep,ind);
output file = cmlcount.out reset;
call CMLCountPrt(b,vc,llik);
output off;
```

A vector of effect parameters and a scalar dispersion parameter are estimated. The vector includes one element corresponding to each explanatory variable named in *ind* and a constant term. Five parameters are estimated in this example.

Source

cmlparet.src

Purpose

Estimates a constrained Poisson regression model or truncated-at-zero Poisson regression model with maximum likelihood.

Library

`cmlcount`

Format

`{` *b*,*vc*,*llik* `}` = **CMLPoisson(***dataset*,*dep*,*ind*);

Input

dataset	string, name of **GAUSS** data set. – or – N×K matrix, data
dep	string, the name of the dependent variable – or – scalar, the index of the dependent variable
ind	K×1 character vector, names of the independent variables – or – K×1 numeric vector, indices of independent variables Set to 0 to include only a constant term.

If *dataset* is a matrix, *dep* and *ind* may be a string or character variable containing either the standard labels created by **CML** (V1, V2,..., or V01, V02,...., depending on the value of **__vpad**), or the user-provided labels in **__altnam**.

Output

b	K×1 vector, maximum likelihood estimates of the effect parameters.
vc	K×K matrix, variance-covariance matrix of the estimated parameters evaluated at the maximum. If you choose the **CML** global **__cml_CovPar** = 3, *vc* will contain heteroskedastic-consistent parameter estimates.
llik	scalar, value of the log-likelihood function at the maximum.

Globals

CML globals are also relevant, including constraint matrices and procedures.

_cmlc_Inference string, determines the type of statistical inference.

BOOT generates bootstrapped estimates and covariance matrix of estimates

CML generates maximum likelihood estimates

Setting **_cmlc_Inference** to BOOT generates a **GAUSS** data set containing the bootstrapped parameters. The file name of this data set is either the default BOOTx, where x is a four digit number starting with 1 and increasing until a unique name is found, or the name in the **CML** global variable, **_cml_BootFname**. This data set can be used with **CMLBlimits** for generating confidence limits, with **CMLDensity** for generating density estimates and plots of the boostrapped parameters, or with **CMLHist** for generating histogram and surface plots.

_cmlc_Fix scalar, name of index number of logged variable among the regressors with coefficient constrained to 1.0 By default, no logged variables are included.

_cmlc_Start scalar, selects method of calculating starting values. Possible values are:

0 calculates them by regressing $\ln(y + 0.5)$ on the explanatory variables.

1 will use a vector of user supplied start values stored in the global variable **_cmlc_StartValue**.

2 uses a vector of zeros.

3 uses random uniform numbers on the interval $[-\frac{1}{2}, \frac{1}{2}]$.

Default = 0.

_cmlc_StartValue K×1 vector, start values if **_cmlc_Start** = 1.

_cmlc_Precision scalar, number of decimal points to print on output. Default = 4.

_cmlc_ZeroTruncate scalar, specifies which model is used:

0 truncated-at-zero negative binomial model

1 negative binomial model is used.

Default = 1.

__altnam K×1 vector, alternate names for variables when a matrix is passed to **CMLPoisson**. When a data matrix is passed to **CMLPoisson** and the user is selecting from that matrix, the global variable **__altnam**, if it is used, must contain names for the columns of the original matrix.

__miss scalar, determines how missing data will be handled.

0 Missing values will not be checked for, and so the data set must not have any missings. This is the fastest option.

1 Listwise deletion. Removes from computation any observation with a missing value on any variable included in the analysis.

Default = 0.

__output scalar, determines printing of intermediate results.

0 nothing is written.

1 serial ASCII output format suitable for disk files or printers.

2 (DOS only) output is suitable for screen only. ANSI.SYS must be active.

Default = 2.

__row scalar, specifies how many rows of the data set will be read per iteration of the read loop. By default, the number of rows to be read will be calculated automatically.

__rowfac scalar, row factor. If **POISSON** fails due to insufficient memory while attempting to read a **GAUSS** data set, then **__rowfac** may be set to some value between 0 and 1 to read a *proportion* of the original number of rows of the **GAUSS** data set. For example, setting

```
__rowfac = 0.8;
```

will cause **GAUSS** to read in 80% of the rows originally calculated.

__title string, message printed at the top of the screen and printed out by **CMLCountPrt**. Default = "".

__vpad scalar, if *dataset* is a matrix in memory, the variable names are automatically created by **CML**. Two types of names can be created:

0 Variable names automatically created by **CML** are not padded to give them equal length. For example, V1, V2,...V10, V11,....

1 Variable names created by the procedure are padded with zeros to give them an equal number of characters. For example, V01, V02, ..., V10, V11,.... This is useful if you want the variable names to sort properly.

Default = 1.

Remarks

Let the n event count observations (non-negative integers) for the dependent variable be denoted as $y_1, \ldots, y_n$. y_i is then a random dependent variable representing the number of events that have occurred during observation period i. By assuming that the events occurring within each period are independent and have constant rates of occurrence, y_i can be shown to follow a Poisson distribution:

$$f_p(y_i|\lambda_i) = \begin{cases} \frac{e^{-\lambda_i}(\lambda_i)^{y_i}}{y_i!} & \text{for } \lambda_i > 0 \text{ and } y_i = 0, 1, \ldots \\ 0 & \text{otherwise} \end{cases} \tag{5.9}$$

with expected value and variance λ_i. Under the Poisson regression model, λ_i (the rate of event occurrence, which must be greater than zero) is assumed to be an exponential-linear function of a vector of explanatory variables, x_i:

$$E(y_i) \equiv \lambda_i = \exp(x_i\beta) \tag{5.10}$$

The program includes a constant term as the first element of x_i and allows one to include any number of explanatory variables.

For an introduction to the Poisson regression model see King (1988); on the truncated model, see Grogger and Carson (1988) and King (1989d).

Example

Constrained Poisson Regression Model

```
library cmlcount;
#include cmlcount.ext;
CMLCountset;
dataset = "wars";
dep = { wars };
ind = { unem, poverty, allianc };
_cml_Bounds = { 0 10 };  /* constrains coefficients */
                         /* to be positive */
{ b,vc,llik } = CMLPoisson(dataset,dep,ind);
output file = cmlcount.out reset;
call CMLCountPrt(b,vc,llik);
output off;
```

Constrained Truncated-at-zero Poisson Regression Model

```
library cmlcount;
#include cmlcount.ext;
```

```
CMLCountset;
dataset = "wars";
dep = { wars };
ind = { unem, poverty, allianc };
_cmlc_ZeroTruncate = 0;
{ b,vc,llik } = CMLPoisson(dataset,dep,ind);
output file = cmlcount.out reset;
call CMLCountPrt(b,vc,llik);
output off;
```

Source

cmlpoiss.src

Purpose

Estimates a constrained seemingly unrelated Poisson regression model, for the analysis of two event *CONSTRAINED COUNT* variables, with maximum likelihood.

Library

`cmlcount`

Format

`{` *b*,*vc*,*llik* `}` = **CMLSupreme(***dataset*,*dep1*,*dep2*,*ind1*,*ind2***);**

Input

dataset	string, name of **GAUSS** data set. – or – N×K matrix, data
dep1	string, name of the first dependent variable – or – scalar, index of the first dependent variable
dep2	string, name of the second dependent variable – or – scalar, index of the second dependent variable
ind1	K×1 character vector, names of first event independent variables – or – K×1 numeric vector, indices of first event independent variables Set to 0 to include only a constant term.
ind2	K×1 character vector, names of second event independent variables – or – K×1 numeric vector, indices of second event independent variables Set to 0 to include only a constant term.

If *dataset* is a matrix, *dep1*, *dep2*, *ind1* and *ind2* may be a string or character variable containing either the standard labels created by **CML** (V1, V2,..., or V01, V02,....., depending on the value of **__vpad**), or the user-provided labels in **__altnam**.

Output

b	(K+L+2)×1 vector, maximum likelihood estimates of the effect parameters of β and γ stacked on top of the covariance parameter ξ.

vc (K+L+2)×(K+L+2) matrix, variance-covariance matrix of the estimated parameters evaluated at the maximum. If you choose the **CML** global **_cml_CovPar** = 3, *vc* will contain heteroskedastic-consistent parameter estimates.

llik scalar, value of the log-likelihood function at the maximum.

Globals

CML globals are also relevant, including constraint matrices and procedures.

_cmlc_Inference string, determines the type of statistical inference.

BOOT generates bootstrapped estimates and covariance matrix of estimates

CML generates maximum likelihood estimates

Setting **_cmlc_Inference** to BOOT generates a **GAUSS** data set containing the bootstrapped parameters. The file name of this data set is either the default BOOTx, where x is a four digit number starting with 1 and increasing until a unique name is found, or the name in the **CML** global variable, **_cml_BootFname**. This data set can be used with **CMLBlimits** for generating confidence limits, with **CMLDensity** for generating density estimates and plots of the boostrapped parameters, or with **CMLHist** for generating histogram and surface plots.

_cmlc_Start scalar, selects method of calculating starting values. Possible values are:

0 calculates them by regressing $\ln(y + 0.5)$ on the explanatory variables.

1 will use a vector of user supplied start values stored in the global variable **_cmlc_StartValue**.

2 uses a vector of zeros.

3 uses random uniform numbers on the interval $[-\frac{1}{2}, \frac{1}{2}]$.

Default = 0.

_cmlc_StartValue (K+L+2)×1 vector, start values if **_cmlc_Start** = 1.

_cmlc_Precision scalar, number of decimal points to print on output. Default = 4.

__altnam K×1 vector, alternate names for variables when a matrix is passed to **CMLSupreme**. When a data matrix is passed to **CMLSupreme** and the user is selecting from that matrix, the global variable **__altnam**, if it is used, must contain names for the columns of the original matrix.

__miss scalar, determines how missing data will be handled.

0 Missing values will not be checked for, and so the data set must not have any missings. This is the fastest option.

1 Listwise deletion. Removes from computation any observation with a missing value on any variable included in the analysis.

Default = 0.

__output scalar, determines printing of intermediate results.

0 nothing is written.

1 serial ASCII output format suitable for disk files or printers.

2 (DOS only) output is suitable for screen only. ANSI.SYS must be active.

Default = 2.

__row scalar, specifies how many rows of the data set will be read per iteration of the read loop. By default, the number of rows to be read will be calculated automatically.

__rowfac scalar, row factor. If **CMLSupreme** fails due to insufficient memory while attempting to read a **GAUSS** data set, then **__rowfac** may be set to some value between 0 and 1 to read a *proportion* of the original number of rows of the **GAUSS** data set. For example, setting

```
__rowfac = 0.8;
```

will cause **GAUSS** to read in 80% of the rows originally calculated.

This global only has an affect when **__row** = 0.

Default = 1.

__title string, message printed at the top of the screen and printed out by **CMLCountPrt**. Default = "".

__vpad scalar, if *dataset* is a matrix in memory, the variable names are automatically created by **CML**. Two types of names can be created:

0 Variable names automatically created by **CML** are not padded to give them equal length. For example, V1, V2,...V10, V11,....

1 Variable names created by the procedure are padded with zeros to give them an equal number of characters. For example, V01, V02, ..., V10, V11,.... This is useful if you want the variable names to sort properly.

Default = 1.

Remarks

Suppose we observe two event count dependent variables y_{1i} and y_{2i} for n observations. Let these variables be distributed as a bivariate Poisson with $E(y_{1i}) = \lambda_{1i}$ and $E(y_{2i}) = \lambda_{2i}$. These means are parametrized as follows:

$$\lambda_{0i} = \exp(x_i\beta) \tag{5.11}$$

and

$$\lambda_{+i} = \exp(z_i\gamma) \tag{5.12}$$

where x_i and z_i are (possibly) different vectors of explanatory variables. The covariance parameter is ξ.

If you have convergence problems, you might try **CMLSupreme2** with argument *ind3* = 0 instead.

For details about this model, see King (1989c).

Example

Constrained Seemingly Unrelated Poisson Regression Model (CMLSupreme)

```
library cmlcount;
#include cmlcount.ext;
CMLCountset;
dataset = "wars";
dep1 = { wars };
ind1 = { unem, poverty, allianc };
dep2 = { coups };
ind2 = { unem, age, sex, race };
_cml_Bounds = { 0 10 };  /* constrains coefficients */
                         /* to be positive */
{ b,vc,llik } = CMLSupreme(dataset,dep1,dep2,ind1,ind2);
output file = cmlcount.out reset;
call CMLCountPrt(b,vc,llik);
output off;
```

Two vectors of effect parameters and one scalar covariance parameter are estimated. The vectors of effect parameters each include one element corresponding to each explanatory variable and a constant term. In the example, ten parameters are estimated.

Source

cmlsupr.src

Purpose

Estimates a constrained Poisson regression model with unobserved dependent variables, for the analysis of two observed (and three unobserved) event count variables, with maximum likelihood.

Library

`cmlcount`

Format

`{` *b*,*vc*,*llik* `}` = **CMLSupreme2(***dataset*,*dep1*,*dep2*,*ind1*,*ind2*,*ind3***);**

Input

dataset string, name of **GAUSS** data set.
– or –
N×K matrix, data

dep1 string, name of the first dependent variable
– or –
scalar, index of the first dependent variable

dep2 string, name of the second dependent variable
– or –
scalar, index of the second dependent variable

ind1 K×1 character vector, names of first event independent variables
– or –
K×1 numeric vector, indices of first event independent variables
Set to 0 to include only a constant term.

ind2 L×1 character vector, names of second event independent variables
– or –
L×1 numeric vector, indices of second event independent variables
Set to 0 to include only a constant term.

ind3 M×1 character vector, names of second event independent variables
– or –
M×1 numeric vector, indices of second event independent variables
Set to 0 to include only a constant term.

If *dataset* is a matrix, *dep1*, *dep2*, *ind1*, *ind2*, or *ind3* may be a string or character variable containing either the standard labels created by **CML** (V1, V2,..., or V01, V02,...., depending on the value of **__vpad**), or the user-provided labels in **__altnam**.

Output

b (K+L+M)×1 vector, maximum likelihood estimates of the effect parameters of β and γ stacked on top of the covariance parameter ξ.

vc (K+L+M)×(K+L+M) matrix, variance-covariance matrix of the estimated parameters evaluated at the maximum. If you choose the **CML** global **_cml_CovPar** = 3, *vc* will contain heteroskedastic-consistent parameter estimates.

llik scalar, value of the log-likelihood function at the maximum.

Globals

CML globals are also relevant, including constraint matrices and procedures.

_cmlc_Inference string, determines the type of statistical inference.

BOOT generates bootstrapped estimates and covariance matrix of estimates

CML generates maximum likelihood estimates

Setting **_cmlc_Inference** to BOOT generates a **GAUSS** data set containing the bootstrapped parameters. The file name of this data set is either the default BOOTx, where x is a four digit number starting with 1 and increasing until a unique name is found, or the name in the **CML** global variable, **_cml_BootFname**. This data set can be used with **CMLBlimits** for generating confidence limits, with **CMLDensity** for generating density estimates and plots of the boostrapped parameters, or with **CMLHist** for generating histogram and surface plots.

_cmlc_Start scalar, selects method of calculating starting values. Possible values are:

0 calculates them by regressing $\ln(y + 0.5)$ on the explanatory variables.

1 will use a vector of user supplied start values stored in the global variable **_cmlc_StartValue**.

2 uses a vector of zeros.

3 uses random uniform numbers on the interval $[-\frac{1}{2}, \frac{1}{2}]$.

Default = 0.

_cmlc_StartValue (K+L+M)×1 vector, start values if **_cmlc_Start** = 1.

_cmlc_Precision scalar, number of decimal points to print on output. Default = 4.

__altnam K×1 vector, alternate names for variables when a matrix is passed to **CMLSupreme2**. When a data matrix is passed to **CMLSupreme2** and the user is selecting from that matrix, the global variable **__altnam**, if it is used, must contain names for the columns of the original matrix.

__miss scalar, determines how missing data will be handled.

0	Missing values will not be checked for, and so the data set must not have any missings. This is the fastest option.
1	Listwise deletion. Removes from computation any observation with a missing value on any variable included in the analysis.

Default = 0.

__output scalar, determines printing of intermediate results.

0	nothing is written.
1	serial ASCII output format suitable for disk files or printers.
2	(DOS only) output is suitable for screen only. ANSI.SYS must be active.

Default = 2.

__row scalar, specifies how many rows of the data set will be read per iteration of the read loop. By default, the number of rows to be read will be calculated automatically.

__rowfac scalar, row factor. If **CMLSupreme2** fails due to insufficient memory while attempting to read a **GAUSS** data set, then **__rowfac** may be set to some value between 0 and 1 to read a *proportion* of the original number of rows of the **GAUSS** data set. For example, setting

```
__rowfac = 0.8;
```

will cause **GAUSS** to read in 80% of the rows originally calculated.

This global only has an affect when **__row** = 0.

Default = 1.

__title string, message printed at the top of the screen and printed out by **CMLCountPrt**. Default = "".

__vpad scalar, if *dataset* is a matrix in memory, the variable names are automatically created by **CML**. Two types of names can be created:

0	Variable names automatically created by **CML** are not padded to give them equal length. For example, V1, V2,...V10, V11,....
1	Variable names created by the procedure are padded with zeros to give them an equal number of characters. For example, V01, V02, ..., V10, V11,.... This is useful if you want the variable names to sort properly.

Default = 1.

Remarks

This model assumes the existence of three independent unobserved variables, y^*_{1i}, y^*_{2i}, and y^*_{3i}, with means $E(y^*_{ji}) = \lambda_{ji}$, for $j = 1, 2, 3$. Although these are not observed, we do observe y_{1i} and y_{2i}, which are functions of these three variables:

$$\begin{aligned} y_{1i} &= y^*_{1i} + y^*_{3i} \\ y_{2i} &= y^*_{2i} + y^*_{3i} \end{aligned}$$

The procedure estimates three separate regression functions, one for the expected value of each of the unobserved variables:

$$\lambda_{1i} = \exp(x_{1i}\beta_1) \tag{5.13}$$

$$\lambda_{2i} = \exp(x_{2i}\beta_2)$$

$$\lambda_{3i} = \exp(x_{3i}\beta_3)$$

where x_{1i}, x_{2i} and x_{3i} are (possibly) different sets of explanatory variables and β_1, β_2, and β_3 are separate parameter vectors. This option produces maximum likelihood estimates for these three parameter vectors.

Example

Poisson Regression Model with Unobserved Dependent Variables

```
library cmlcount;
#include cmlcount.ext;
CMLCountset;
dataset = "wars";
dep1 = { wars };
ind1 = { unem, poverty, allianc };
dep2 = { coups };
ind2 = { unem, age, sex, race };
ind3 = { us, sov };
_cml_Bounds = { 0 10 };  /* constrains coefficients */
                         /* to be positive */
{ b,vc,llik } = CMLSupreme2(dataset,dep1,dep2,ind1,ind2,ind3);
output file = cmlcount.out reset;
call CMLCountPrt(b,vc,llik);
output off;
```

Three vectors of effect parameters are estimated. Each includes one element corresponding to each explanatory variable plus a constant term. In the example, twelve parameters are estimated.

Source

cmlsupr2.src

Index

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W