Overview: HPQLIM Procedure

The HPQLIM (high-performance qualitative and limited dependent variable model) procedure is a high-performance version of the QLIM procedure in SAS/ETS software, which analyzes univariate limited dependent variable models in which dependent variables are observed only in a limited range of values. Unlike the QLIM procedure, which can be run only on an individual workstation, the HPQLIM procedure takes advantage of a computing environment that enables it to distribute the optimization task to one or more nodes. In addition, each node can use one or more threads to perform the optimization on its subset of the data. When several nodes are used and each node uses several threads to carry out its part of the work, the result is a highly parallel computation that provides a dramatic gain in performance.

With the HPQLIM procedure you can read and write data in distributed form and perform analyses in distributed mode and single-machine mode. For more information about how to affect the execution mode of SAS high-performance analytical procedures, see the section “Processing Modes” (Chapter 2, SAS/ETS User’s Guide: High-Performance Procedures).

The HPQLIM procedure is specifically designed to operate in the high-performance distributed environment. It can use maximum likelihood or Bayesian methods. In both cases, the likelihood evaluation is performed in a distributed environment. By default, PROC HPQLIM uses multiple threads to perform computations.

The HPQLIM procedure is similar in use to the other SAS procedures that support regression or simultaneous equations models. For example, the standard model with censoring or truncation is estimated by specifying the endogenous variable to be truncated or censored. When the data are limited by specific values or variables, the limits of the dependent variable can be specified with the CENSORED or TRUNCATED option in the ENDOGENOUS or MODEL statement. For example, the two-limit censored model requires two variables: one that contains the lower (bottom) bound and one that contains the upper (top) bound. The following statements execute the model in the distributed computing environment with two threads and four nodes:

```sas
proc hpqlim data=a;
  model y = x1 x2 x3;
  endogenous y ~ censored(lb=bottom ub=top);
  performance nthreads=2 nodes=4 details;
run;
```

The bounds can be numbers if they are fixed for all observations in the data set. For example, the standard Tobit model can be specified as follows:

```sas
proc hpqlim data=a;
  model y = x1 x2 x3;
  endogenous y ~ censored(lb=0);
  performance nthreads=2 nodes=4 details;
run;
```
PROC HPQLIM Features

The HPQLIM procedure supports the following models:

- linear regression models with heteroscedasticity
- Tobit models (censored and truncated) with heteroscedasticity
- stochastic frontier production and cost models

In linear regression models with heteroscedasticity, the assumption that error variance is constant across observations is relaxed. The HPQLIM procedure allows for a number of different linear and nonlinear variance specifications.

The HPQLIM procedure also offers a class of models in which the dependent variable is censored or truncated from below or above or both. When a continuous dependent variable is observed only within a certain range, and values outside this range are not available, the HPQLIM procedure offers a class of models that adjust for truncation. In some cases, the dependent variable is continuous only in a certain range, and all values outside this range are reported as being on its boundary. For example, if it is not possible to observe negative values, the value of the dependent variable is reported as equal to 0. Because the data are censored, ordinary least squares (OLS) results are inconsistent, and it cannot be guaranteed that the predicted values from the model will fall in the appropriate region.

Stochastic frontier production and cost models allow for random shocks of the production or cost. They include a systematic positive component in the error term that adjusts for technical or cost inefficiency.

The HPQLIM procedure can use maximum likelihood or Bayesian methods. Initial starting values for the nonlinear optimizations are typically calculated by OLS. Initial values for the Bayesian sampling are typically calculated by maximum likelihood.

Getting Started: HPQLIM Procedure

This example illustrates the use of the HPQLIM procedure. The data were originally published by Mroz (1987), and the following statements show a subset of that data set:

```sas
title1 'Estimating a Tobit model';

data subset;
  input Hours Yrs_Ed Yrs_Exp @@;
  if Hours eq 0 then Lower=.;
  else Lower=Hours;
datalines;
  0 8 9 0 8 12 0 9 10 0 10 15 0 11 4 0 11 6
  1000 12 1 1960 12 29 0 13 3 2100 13 36
  3686 14 11 1920 14 38 0 15 14 1728 16 3
  1568 16 19 1316 17 7 0 17 15
;```

In these data, Hours is the number of hours that the wife worked outside the household in a given year, Yrs_Ed is the years of education, and Yrs_Exp is the years of work experience.
By the nature of the data it is clear that there are a number of women who committed some positive number of hours to outside work ($y_i > 0$ is observed). There are also a number of women who did not work outside the home at all ($y_i = 0$ is observed). This yields the following model:

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

$$y_i = \begin{cases} 
  y_i^* & \text{if } y_i^* > 0 \\
  0 & \text{if } y_i^* \leq 0 
\end{cases}$$

where $\epsilon_i \sim \text{iid} N(0, \sigma^2)$ and the set of explanatory variables is denoted by $x_i$. The following statements fit a Tobit model to the hours worked with years of education and years of work experience as covariates:

```plaintext
/*--- Tobit Model ---*/
proc hpqlim data=subset;
  model hours = yrs_ed yrs_exp;
  endogenous hours ~ censored(lb=0);
  performance nthreads=2 nodes=4 details;
run;
```

The output of the HPQLIM procedure is shown in Output 21.1.

**Figure 21.1** Tobit Analysis Results

**Estimating a Tobit model**

The HPQLIM Procedure

<table>
<thead>
<tr>
<th>Model Fit Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Endogenous Variables</td>
</tr>
<tr>
<td>Endogenous Variable</td>
</tr>
<tr>
<td>Number of Observations</td>
</tr>
<tr>
<td>Log Likelihood</td>
</tr>
<tr>
<td>Maximum Absolute Gradient</td>
</tr>
<tr>
<td>Number of Iterations</td>
</tr>
<tr>
<td>Optimization Method</td>
</tr>
<tr>
<td>AIC</td>
</tr>
<tr>
<td>Schwarz Criterion</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>Yrs_Ed</td>
</tr>
<tr>
<td>Yrs_Exp</td>
</tr>
<tr>
<td>_Sigma</td>
</tr>
</tbody>
</table>

The “Parameter Estimates” table contains four rows. The first three rows correspond to the vector estimate of the regression coefficients $\beta$. The last row is called _Sigma, which corresponds to the estimate of the error variance $\sigma$. 
Syntax: HPQLIM Procedure

The following statements are available in the HPQLIM procedure:

```
PROC HPQLIM options ;
  BAYES < options > ;
  BOUNDS bound1 < , bound2 . . . > ;
  BY variables ;
  FREQ variable ;
  ENDOGENOUS variables ~ options ;
  HETERO dependent-variables ~ exogenous-variables / options ;
  INIT initvalue1 < , initvalue2 . . . > ;
  MODEL dependent-variables = regressors / options ;
  OUTPUT OUT=SAS-data-set < output-options > ;
  PRIOR _REGRESSORS | parameter-list ~ distribution ;
  RESTRICT restriction1 < , restriction2 . . . > ;
  TEST options ;
  WEIGHT variable < / option > ;
  PERFORMANCE < performance-options > ;
```

One MODEL statement is required. If a FREQ or WEIGHT statement is specified more than once, the variable that is specified in the first instance is used.

Functional Summary

Table 21.1 summarizes the statements and options used with the HPQLIM procedure.

<table>
<thead>
<tr>
<th>Description</th>
<th>Statement</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Set Options</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specifies the input data set</td>
<td>PROC HPQLIM</td>
<td>DATA=</td>
</tr>
<tr>
<td>Writes parameter estimates to an output data set</td>
<td>PROC HPQLIM</td>
<td>OUTEST=</td>
</tr>
<tr>
<td>Writes predictions to an output data set</td>
<td>OUTPUT</td>
<td>OUT=</td>
</tr>
<tr>
<td><strong>Declaring the Role of Variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specifies BY-group processing</td>
<td>BY</td>
<td></td>
</tr>
<tr>
<td>Specifies a frequency variable</td>
<td>FREQ</td>
<td></td>
</tr>
<tr>
<td>Specifies a weight variable</td>
<td>WEIGHT</td>
<td>NONORMALIZE</td>
</tr>
<tr>
<td><strong>Printing Control Options</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Requests all printing options</td>
<td>PROC HPQLIM</td>
<td>PRINTALL</td>
</tr>
<tr>
<td>Prints the correlation matrix of the estimates</td>
<td>PROC HPQLIM</td>
<td>CORRB</td>
</tr>
<tr>
<td>Prints the covariance matrix of the estimates</td>
<td>PROC HPQLIM</td>
<td>COVB</td>
</tr>
<tr>
<td>Suppresses the normal printed output</td>
<td>PROC HPQLIM</td>
<td>NOPRINT</td>
</tr>
</tbody>
</table>
### Table 21.1 continued

<table>
<thead>
<tr>
<th>Description</th>
<th>Statement</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Plotting Options</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Displays plots</td>
<td>PROC HPQLIM</td>
<td>PLOTS=</td>
</tr>
<tr>
<td><strong>Optimization Process Control Options</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selects the iterative minimization method to use</td>
<td>PROC HPQLIM</td>
<td>METHOD=</td>
</tr>
<tr>
<td>Specifies the maximum number of iterations allowed</td>
<td>PROC HPQLIM</td>
<td>MAXITER=</td>
</tr>
<tr>
<td>Specifies the maximum number of function calls</td>
<td>PROC HPQLIM</td>
<td>MAXFUNC=</td>
</tr>
<tr>
<td>Specifies the upper limit of CPU time in seconds</td>
<td>PROC HPQLIM</td>
<td>MAXTIME=</td>
</tr>
<tr>
<td>Specifies an absolute convergence criterion</td>
<td>PROC HPQLIM</td>
<td>ABSCONV=</td>
</tr>
<tr>
<td>Specifies an absolute function convergence criterion</td>
<td>PROC HPQLIM</td>
<td>ABSFCONV=</td>
</tr>
<tr>
<td>Specifies an absolute gradient convergence criterion</td>
<td>PROC HPQLIM</td>
<td>ABSGCONV=</td>
</tr>
<tr>
<td>Specifies a relative function convergence criterion</td>
<td>PROC HPQLIM</td>
<td>FCONV=</td>
</tr>
<tr>
<td>Specifies a relative gradient convergence criterion</td>
<td>PROC HPQLIM</td>
<td>GCONV=</td>
</tr>
<tr>
<td>Specifies an absolute parameter convergence criterion</td>
<td>PROC HPQLIM</td>
<td>ABSXCONV=</td>
</tr>
<tr>
<td>Specifies a matrix singularity criterion</td>
<td>PROC HPQLIM</td>
<td>SINGULAR=</td>
</tr>
<tr>
<td>Sets boundary restrictions on parameters</td>
<td>BOUNDS</td>
<td></td>
</tr>
<tr>
<td>Sets initial values for parameters</td>
<td>INIT</td>
<td></td>
</tr>
<tr>
<td>Sets linear restrictions on parameters</td>
<td>RESTRICT</td>
<td></td>
</tr>
<tr>
<td><strong>Model Estimation Options</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Suppresses the intercept parameter</td>
<td>MODEL</td>
<td>NOINT</td>
</tr>
<tr>
<td>Specifies the method to calculate parameter covariance</td>
<td>PROC HPQLIM</td>
<td>COVEST=</td>
</tr>
<tr>
<td><strong>Bayesian MCMC Options</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specifies the initial values of the MCMC</td>
<td>INIT</td>
<td></td>
</tr>
<tr>
<td>Specifies the maximum number of tuning phases</td>
<td>BAYES</td>
<td>MAXTUNE=</td>
</tr>
<tr>
<td>Specifies the minimum number of tuning phases</td>
<td>BAYES</td>
<td>MINTUNE=</td>
</tr>
<tr>
<td>Specifies the number of burn-in iterations</td>
<td>BAYES</td>
<td>NBI=</td>
</tr>
<tr>
<td>Specifies the number of iterations during the sampling phase</td>
<td>BAYES</td>
<td>NMC=</td>
</tr>
<tr>
<td>Specifies the number of iterations during the tuning phase</td>
<td>BAYES</td>
<td>NTU=</td>
</tr>
<tr>
<td>Controls options for constructing the initial proposal covariance matrix</td>
<td>BAYES</td>
<td>PROPCOV</td>
</tr>
<tr>
<td>Specifies the sampling scheme</td>
<td>BAYES</td>
<td>SAMPLING=</td>
</tr>
<tr>
<td>Specifies the random number generator seed</td>
<td>BAYES</td>
<td>SEED=</td>
</tr>
<tr>
<td>Controls the thinning of the Markov chain</td>
<td>BAYES</td>
<td>THIN=</td>
</tr>
<tr>
<td><strong>Bayesian Summary Statistics and Convergence Diagnostic Options</strong></td>
<td>BAYES</td>
<td>DIAGNOSTICS=</td>
</tr>
</tbody>
</table>
### Table 21.1 continued

<table>
<thead>
<tr>
<th>Description</th>
<th>Statement</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>Displays summary statistics of the posterior samples</td>
<td>BAYES</td>
<td>STATISTICS=</td>
</tr>
<tr>
<td><strong>Bayesian Prior and Posterior Sample Options</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specifies a SAS data set for the posterior samples</td>
<td>BAYES</td>
<td>OUTPOST=</td>
</tr>
<tr>
<td><strong>Bayesian Analysis Options</strong></td>
<td>PRIOR</td>
<td>NORMAL(MEAN=, VAR=)</td>
</tr>
<tr>
<td>Specifies the normal prior distribution</td>
<td>PRIOR</td>
<td>GAMMA(SHAPE=, SCALE=)</td>
</tr>
<tr>
<td>Specifies the gamma prior distribution</td>
<td>PRIOR</td>
<td>IGAMMA(SHAPE=, SCALE=)</td>
</tr>
<tr>
<td>Specifies the inverse gamma prior distribution</td>
<td>PRIOR</td>
<td>UNIFORM(MIN=, MAX=)</td>
</tr>
<tr>
<td>Specifies the uniform prior distribution</td>
<td>PRIOR</td>
<td>BETA(SHAPE1=, SHAPE2=, MIN=, MAX=)</td>
</tr>
<tr>
<td>Specifies the $t$ prior distribution</td>
<td>PRIOR</td>
<td>T(LOCATION=, DF=)</td>
</tr>
<tr>
<td><strong>Endogenous Variable Options</strong></td>
<td>ENDOGENOUS</td>
<td>DISCRETE()</td>
</tr>
<tr>
<td>Specifies a discrete variable</td>
<td>ENDOGENOUS</td>
<td>CENSORED()</td>
</tr>
<tr>
<td>Specifies a censored variable</td>
<td>ENDOGENOUS</td>
<td>TRUNCATED()</td>
</tr>
<tr>
<td>Specifies a stochastic frontier variable</td>
<td>ENDOGENOUS</td>
<td>FRONTIER()</td>
</tr>
<tr>
<td><strong>Heteroscedasticity Model Options</strong></td>
<td>HETERO</td>
<td>LINK=</td>
</tr>
<tr>
<td>Specifies the function for heteroscedasticity models</td>
<td>HETERO</td>
<td>SQUARE</td>
</tr>
<tr>
<td>Squares the function for heteroscedasticity models</td>
<td>HETERO</td>
<td>NOCONST</td>
</tr>
<tr>
<td><strong>Output Control Options</strong></td>
<td>OUTPUT</td>
<td>PREDICTED</td>
</tr>
<tr>
<td>Outputs predicted values</td>
<td>OUTPUT</td>
<td>XBETA</td>
</tr>
<tr>
<td>Outputs the structured part</td>
<td>OUTPUT</td>
<td>RESIDUAL</td>
</tr>
<tr>
<td>Outputs residuals</td>
<td>OUTPUT</td>
<td>ERRSTD</td>
</tr>
<tr>
<td>Outputs the error standard deviation</td>
<td>OUTPUT</td>
<td>MARGINAL</td>
</tr>
<tr>
<td>Outputs marginal effects</td>
<td>OUTPUT</td>
<td>PROB</td>
</tr>
<tr>
<td>Outputs probability for the current response</td>
<td>OUTPUT</td>
<td>PROBALL</td>
</tr>
<tr>
<td>Outputs probability for all responses</td>
<td>OUTPUT</td>
<td>EXPECTED</td>
</tr>
<tr>
<td>Outputs the expected value</td>
<td>OUTPUT</td>
<td>CONDITIONAL</td>
</tr>
<tr>
<td>Outputs the conditional expected value</td>
<td>OUTPUT</td>
<td>MILLS</td>
</tr>
<tr>
<td>Outputs inverse Mills ratio</td>
<td>OUTPUT</td>
<td>TE1</td>
</tr>
<tr>
<td>Outputs technical efficiency measures</td>
<td>OUTPUT</td>
<td>TE2</td>
</tr>
</tbody>
</table>
Table 21.1  continued

<table>
<thead>
<tr>
<th>Description</th>
<th>Statement</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>Includes covariances in the OUTEST= data set</td>
<td>PROC HPQLIM</td>
<td>COVOUT</td>
</tr>
<tr>
<td>Includes correlations in the OUTEST= data set</td>
<td>PROC HPQLIM</td>
<td>CORROUT</td>
</tr>
</tbody>
</table>

**Test Request Options**

- Requests Wald, Lagrange multiplier, and likelihood ratio tests
  - TEST  ALL
- Requests the Wald test
  - TEST  WALD
- Requests the Lagrange multiplier test
  - TEST  LM
- Requests the likelihood ratio test
  - TEST  LR

**PROC HPQLIM Statement**

```
PROC HPQLIM options;
```

The PROC HPQLIM statement invokes the HPQLIM procedure. You can specify the following **options**.

**Data Set Options**

```
DATA=SAS-data-set
```

specifies the input SAS data set. If this option is not specified, PROC HPQLIM uses the most recently created SAS data set.

**Output Data Set Options**

```
OUTEST=SAS-data-set
```

writes the parameter estimates to an output data set.

- **COVOUT**

  writes the covariance matrix for the parameter estimates to the OUTEST= data set. This option is valid only if the OUTEST= option is specified.

- **CORROUT**

  writes the correlation matrix for the parameter estimates to the OUTEST= data set. This option is valid only if the OUTEST= option is specified.

**Printing Options**

```
NOPRINT
```

suppresses the normal printed output but does not suppress error listings. If this option is specified, then any other print option is turned off.
PRINTALL

turns on all the printing options. The options that are set by PRINTALL are COVB and CORRB.

CORRB

prints the correlation matrix of the parameter estimates.

COVB

prints the covariance matrix of the parameter estimates.

Model Estimation Options

COVEST=covariance-option

specifies the method for calculating the covariance matrix of parameter estimates. You can specify the following covariance-options:

OP

specifies the covariance from the outer product matrix.

HESSIAN

specifies the covariance from the inverse Hessian matrix.

QML

specifies the covariance from the outer product and Hessian matrices (the quasi-maximum likelihood estimates).

The default is COVEST=HESSIAN.

Optimization Control Options

PROC HPQLIM uses the nonlinear optimization (NLO) subsystem to perform nonlinear optimization tasks. You can specify the following options:

ABSCONV=r

ABSTOL=r

specifies an absolute function value convergence criterion by which minimization stops when

\[ f(\theta^{(k)}) \leq r. \]

The default value of \( r \) is the negative square root of the largest double-precision value, which serves only as a protection against overflows.

ABSFCONV=r

ABSFTOL=r

specifies an absolute function difference convergence criterion by which minimization stops when the function value has a small change in successive iterations:

\[ |f(\theta^{(k-1)}) - f(\theta^{(k)})| \leq r \]

The default value is \( r = 0. \)

ABSGCONV=r

ABSGTOL=r

specifies an absolute gradient convergence criterion. Optimization stops when the maximum absolute gradient element is small:

\[ \max_j |g_j(\theta^{(k)})| \leq r \]

The default value is \( r=1E^{-5}. \)
**Chapter 21: The HPQLIM Procedure**

**ABSXCONV=\(r\)**

**ABSXTOL=\(r\)**

specifies an absolute parameter convergence criterion. Optimization stops when the Euclidean distance between successive parameter vectors is small:

\[ \| \theta^{(k)} - \theta^{(k-1)} \|_2 \leq r \]

The default is 0.

**FCONV=\(r\)**

**FTOL=\(r\)**

specifies a relative function convergence criterion. Optimization stops when a relative change of the function value in successive iterations is small:

\[ \frac{|f(\theta^{(k)}) - f(\theta^{(k-1)})|}{|f(\theta^{(k-1)})|} \leq r \]

The default value is \(r = 2\epsilon\), where \(\epsilon\) denotes the machine precision constant, which is the smallest double-precision floating-point number such that \(1 + \epsilon > 1\).

**GCONV=\(r\)**

**GTOL=\(r\)**

specifies a relative gradient convergence criterion. For all techniques except CONGRA, optimization stops when the normalized predicted function reduction is small:

\[ \frac{g(\theta^{(k)})^T [H^{(k)}]^{-1} g(\theta^{(k)})}{|f(\theta^{(k)})|} \leq r \]

For the CONGRA technique (where a reliable Hessian estimate \(H\) is not available), the following criterion is used:

\[ \frac{\| g(\theta^{(k)}) \|_2 \| s(\theta^{(k)}) \|_2}{\| g(\theta^{(k)}) - g(\theta^{(k-1)}) \|_2 |f(\theta^{(k)})|} \leq r \]

The default value is \(r = 1\times10^{-8}\).

**MAXFUNC=\(i\)**

**MAXFU=\(i\)**

specifies the maximum number of function calls in the optimization process. The default is 1,000.

The optimization can terminate only after completing a full iteration. Therefore, the number of function calls that are actually performed can exceed the number of calls that are specified by this option.

**MAXITER=\(i\)**

**MAXIT=\(i\)**

specifies the maximum number of iterations in the optimization process. The default is 200.

**MAXTIME=\(r\)**

specifies an upper limit of \(r\) seconds of CPU time for the optimization process. The default value is the largest floating-point double representation of your computer. The time that is specified by this option is checked only once at the end of each iteration. Therefore, the actual running time can be much longer than \(r\). The actual running time includes the remaining time needed to finish the iteration and the time needed to generate the output of the results.
METHOD=value
specifies the iterative minimization method to use. The default is METHOD=NEWRAP. You can specify the following values:

- **CONGRA** specifies the conjugate-gradient method.
- **DBLDOG** specifies the double dogleg method.
- **NONE** specifies that no optimization be performed beyond using the ordinary least squares method to compute the parameter estimates.
- **NEWRAP** specifies the Newton-Raphson method (the default).
- **NRRIDG** specifies the Newton-Raphson ridge method.
- **QUANEW** specifies the quasi-Newton method.
- **TRUREG** specifies the trust region method.

**SINGULAR=r**
specifies the general singularity criterion that is applied by the HPQLIM procedure in sweeps and inversions. The default for the optimization is 1E–8.

### Plotting Options

**PLOTS< (global-plot-options) >= plot-request | (plot-requests)**
controls the display of plots. By default, the plots are displayed in panels unless the UNPACK **global-plot-option** is specified. When you specify only one **plot-request**, you can omit the parentheses around it.

#### Global Plot Options
You can specify the following **global-plot-options**:

- **ONLY**
  displays only the requested plot.

- **UNPACKPANEL**
- **UNPACK**
  specifies that all paneled plots be unpacked, meaning that each plot in a panel is displayed separately.

#### Plot Requests
You can specify the following **plot-requests**:

- **ALL**
  specifies all types of available plots.

- **AUTOCORR< (LAGS=n)>**
  displays the autocorrelation function plots for the parameters. The optional **LAGS=n** suboption specifies the number (up to lag n) of autocorrelations to be plotted in the autocorrelation function plot. If this suboption is not specified, autocorrelations are plotted up to lag 50. This **plot-request** is available only for Bayesian analysis.
BAYESDIAG
is equivalent to specifying the TRACE, AUTOCORR, and DENSITY plot-requests.

DENSITY< (FRINGE) >
displays the kernel density plots for the parameters. If you specify the FRINGE suboption, a fringe plot is created on the X axis of the kernel density plot. This plot-request is available only for Bayesian analysis.

NONE
suppresses all diagnostic plots.

TRACE< (SMOOTH) >
displays the trace plots for the parameters. The SMOOTH suboption displays a fitted penalized B-spline curve for each plot. This plot-request is available only for Bayesian analysis.

---

BAYES Statement

BAYES < options > ;

The BAYES statement controls the Metropolis sampling scheme that is used to obtain samples from the posterior distribution of the underlying model and data.

DIAGNOSTICS=ALL | NONE | (keyword-list)

controls which diagnostics are produced. All the following diagnostics are produced when you specify DIAGNOSTICS=ALL. If you do not want any of these diagnostics, specify DIAGNOSTICS=NONE. If you want some but not all of the diagnostics, or if you want to change certain settings of these diagnostics, specify one or more of the following keywords. The default is DIAGNOSTICS=NONE.

AUTOCORR < (LAGS=numeric-list) >
computes the autocorrelations at lags that are specified in the numeric-list. Elements in the numeric-list are truncated to integers, and repeated values are removed. If the LAGS= option is not specified, autocorrelations of lags 1, 5, and 10 are computed.

ESS
computes Carlin’s estimate of the effective sample size, the correlation time, and the efficiency of the chain for each parameter.

GEWEKE < (geweke-options) >
computes the Geweke spectral density diagnostics, which are essentially a two-sample t test between the first \( f_1 \) portion and the last \( f_2 \) portion of the chain. The defaults are \( f_1 = 0.1 \) and \( f_2 = 0.5 \), but you can choose other fractions by using the following geweke-options:

FRAC1=value
specifies the fraction \( f_1 \) for the first window.

FRAC2=value
specifies the fraction \( f_2 \) for the second window.
HEIDELBERGER <(heidel-options)> computes for each variable the Heidelberger and Welch diagnostic, which consists of a stationarity test of the null hypothesis that the sample values form a stationary process. If the stationarity test is not rejected, a halfwidth test is then carried out. Optionally, you can specify one or more of the following heidel-options:

**EPS=value**
specifies a positive number $\epsilon$ such that if the halfwidth is less than $\epsilon$ times the sample mean of the retained iterates, the halfwidth test is passed.

**HALPHA=value**
specifies the $\alpha$ level ($0 < \alpha < 1$) for the halfwidth test.

**SALPHA=value**
specifies the $\alpha$ level ($0 < \alpha < 1$) for the stationarity test.

MCSE
MCERROR
computes the Monte Carlo standard error for each parameter. The Monte Carlo standard error, which measures the simulation accuracy, is the standard error of the posterior mean estimate and is calculated as the posterior standard deviation divided by the square root of the effective sample size.

RAFTERY< (raftery-options)> computes the Raftery and Lewis diagnostics, which evaluate the accuracy of the estimated quantile ($\hat{O}_Q$ for a given $Q \in (0,1)$) of a chain. $\hat{O}_Q$ can achieve any degree of accuracy when the chain is allowed to run for a long time. The computation stops when the estimated probability $\hat{P}_Q = \Pr(\theta \leq \hat{O}_Q)$ reaches within $\pm R$ of the value $Q$ with probability $S$; that is, $\Pr(Q - R \leq \hat{P}_Q \leq Q + R) = S$. The following raftery-options enable you to specify $Q$, $R$, $S$, and a precision level $\epsilon$ for the test:

**QUANTILE | Q=value**
specifies the order (a value between 0 and 1) of the quantile of interest. The default is 0.025.

**ACCURACY | R=value**
specifies a small positive number as the margin of error for measuring the accuracy of the estimation of the quantile. The default is 0.005.

**PROBABILITY | S=value**
specifies the probability of attaining the accuracy of the estimation of the quantile. The default is 0.95.

**EPSILON | EPS=value**
specifies the toleration level (a small positive number) for the stationary test. The default is 0.001.

**MINTUNE=number**
specifies the minimum number of tuning phases. The default is 2.
Chapter 21: The HPQLIM Procedure

MAXTUNE=number
specifies the maximum number of tuning phases. The default is 24.

NBI=number
specifies the number of burn-in iterations before the chains are saved. The default is 1,000.

NMC=number
specifies the number of iterations after the burn-in. The default is 1,000.

NTU=number
specifies the number of samples for each tuning phase. The default is 500.

OUTPOST=SAS-data-set
names the SAS data set to contain the posterior samples. Alternatively, you can create the output data set by specifying an ODS OUTPUT statement as follows:

ODS OUTPUT POSTERIORSAMPLE = <SAS-data-set> ;

PROPCOV=value
specifies the method that is used in constructing the initial covariance matrix for the Metropolis-Hastings algorithm. The QUANEW and NMSIMP methods find numerically approximated covariance matrices at the optimum of the posterior density function with respect to all continuous parameters. The tuning phase starts at the optimized values; in some problems, this can greatly increase convergence performance. If the approximated covariance matrix is not positive definite, then an identity matrix is used instead. You can specify the following values:

CONGRA performs a conjugate-gradient optimization.
DBLDOG performs a version of double-dogleg optimization.
NEWRAP performs a Newton-Raphson optimization that combines a line-search algorithm with ridging.
NMSIMP performs a Nelder-Mead simplex optimization.
NRRIDG performs a Newton-Raphson optimization with ridging.
QUANEW performs a quasi-Newton optimization.
TRUREG performs a trust-region optimization.

SAMPLING=MULTIMETROPOLIS | UNIMETROPOLIS
specifies how to sample from the posterior distribution. SAMPLING=MULTIMETROPOLIS implements a Metropolis sampling scheme on a single block that contains all the parameters of the model. SAMPLING=UNIMETROPOLIS implements a Metropolis sampling scheme on multiple blocks, one for each parameter of the model. The default is SAMPLING=MULTIMETROPOLIS.

SEED=number
specifies an integer seed in the range 1 to $2^{31} - 1$ for the random number generator in the simulation. Specifying a seed enables you to reproduce identical Markov chains for the same specification. If you do not specify the SEED= option, or if you specify a nonpositive seed, a random seed is derived from the time of day.
STATISTICS < (global-options) > = ALL | NONE | keyword | (keyword-list)

STATS < (global-options) > = ALL | NONE | keyword | (keyword-list)

controls the number of posterior statistics that are produced. Specifying STATISTICS=ALL is equivalent to specifying STATISTICS=(CORR COV INTERVAL PRIOR SUMMARY). If you do not want any posterior statistics, specify STATISTICS=NONE. The default is STATISTICS=(SUMMARY INTERVAL). You can specify the following global-options:

ALPHA=value < ,value>...< ,value>
controls the probabilities of the credible intervals. The value, which must be between 0 and 1, produces a pair of 100(1–value)% equal-tail and highest posterior density (HPD) intervals for each parameter. The default is ALPHA=0.05, which yields the 95% credible intervals for each parameter.

PERCENT=value < ,value>...< ,value>
requests the percentile points of the posterior samples. The value must be between 0 and 100. The default is PERCENT=25, 50, 75, which yields the 25th, 50th, and 75th percentile points, respectively, for each parameter.

You can specify the following keywords:

CORR produces the posterior correlation matrix.

COV produces the posterior covariance matrix.

INTERVAL produces equal-tail credible intervals and HPD intervals. The default is to produce the 95% equal-tail credible intervals and 95% HPD intervals, but you can use the ALPHA= global-option to request intervals of any probabilities.

NONE suppresses printing of all summary statistics.

PRIOR produces a summary table of the prior distributions that are used in the Bayesian analysis.

SUMMARY produces the means, standard deviations, and percentile points (25th, 50th, and 75th) for the posterior samples. You can use the PERCENT= global-option to request specific percentile points.

THIN=number
THINNING=number
controls the thinning of the Markov chain. Only one in every k samples is used when THIN=k. If NBI=n₀ and NMC=n, the number of samples that are retained is

\[
\left\lfloor \frac{n_0 + n}{k} \right\rfloor - \left\lfloor \frac{n_0}{k} \right\rfloor
\]

where \([a]\) represents the integer part of the number \(a\). The default is THIN=1.
BOUNDSS Statement

\textit{BOUNDS} \texttt{bound1 < , bound2 . . . > ;}

The BOUNDS statement imposes simple boundary constraints on the parameter estimates. BOUNDS statement constraints refer to the parameters that are estimated by the HPQLIM procedure. You can specify any number of BOUNDS statements.

Each \textit{bound} is composed of parameters, constants, and inequality operators. Parameters that are associated with regressor variables are referred to by the names of the corresponding regressor variables. Specify each bound as follows:

\begin{verbatim}
    item operator item  < operator item < operator item . . . > >
\end{verbatim}

Each \textit{item} is a constant, the name of a parameter, or a list of parameter names. For more information about how parameters are named in the HPQLIM procedure, see the section “Naming of Parameters” on page 1170. Each \textit{operator} is \texttt{<, >, <=, or >=}.

You can use both the BOUNDS statement and the RESTRICT statement to impose boundary constraints; however, the BOUNDS statement provides a simpler syntax for specifying these types of constraints. For more information, see the section “RESTRICT Statement” on page 1155.

The following BOUNDS statement constrains the estimates of the parameters that are associated with the variable \texttt{ttime} and the variables \texttt{x1} through \texttt{x10} to be between 0 and 1. The following example illustrates the use of parameter lists to specify boundary constraints:

\begin{verbatim}
    bounds 0 < ttime x1-x10 < 1;
\end{verbatim}

The following BOUNDS statement constrains the estimates of the correlation (_RHO) and sigma (_SIGMA) in the bivariate model:

\begin{verbatim}
    bounds _rho >= 0, _sigma.y1 > 1, _sigma.y2 < 5;
\end{verbatim}

\textbf{BY Statement}

\textbf{BY} \texttt{variables ;}

A BY statement can be used with PROC HPQLIM to obtain separate analyses on observations in groups defined by the BY variables.

BY statement processing is not supported when the HPQLIM procedure runs alongside the database or alongside the Hadoop Distributed File System (HDFS). These modes are used if the input data are stored in a database or HDFS and the grid host is the appliance that houses the data.
The ENDOGENOUS statement specifies the type of dependent variables that appear on the left-hand side of the equation. The listed endogenous variables refer to the dependent variables that appear on the left-hand side of the equation. Currently, no right-hand-side endogeneity is handled in PROC HPQLIM. All variables that appear on the right-hand side of the equation are treated as exogenous.

Discrete Variable Options

\[
\text{DISCRETE} \ (\text{discrete-options})
\]

specifies that the endogenous variables in this statement be discrete. You can specify the following `discrete-options`:

\[
\text{DISTRIBUTION=} \text{distribution-type}
\]

\[
\text{DIST=} \text{distribution-type}
\]

\[
\text{D=} \text{distribution-type}
\]

specifies the cumulative distribution function that is used to model the response probabilities. You can specify the following `distribution-types`:

\[
\text{LOGISTIC}
\]

specifies the logistic distribution for the logit model.

\[
\text{NORMAL}
\]

specifies the normal distribution for the probit model.

By default, `DISTRIBUTION=NORMAL`.

\[
\text{ORDER=} \text{DATA} \mid \text{FORMATTED} \mid \text{FREQ} \mid \text{INTERNAL}
\]

specifies the sort order for the levels of the discrete variables that are specified in the ENDOGENOUS statement. This ordering determines which parameters in the model correspond to each level in the data. You can specify the following sort orders:

\[
\text{DATA}
\]

sorts levels by order of appearance in the input data set.

\[
\text{FORMATTED}
\]

sorts levels by formatted value. The sort order is machine-dependent.

\[
\text{FREQ}
\]

sorts levels by descending frequency count; levels that have the most observations come first in the order.

\[
\text{INTERNAL}
\]

sorts levels by unformatted value. The sort order is machine-dependent.

By default, `ORDER=FORMATTED`. For more information about sort order, see the chapter on the SORT procedure in the Base SAS Procedures Guide.

Censored Variable Options

\[
\text{CENSORED} \ (\text{censored-options})
\]

specifies that the endogenous variables in this statement be censored. You can specify the following `censored-options`:
Chapter 21: The HPQLIM Procedure

\[ \text{LB=} \text{value} \mid \text{variable} \]
\[ \text{LOWERBOUND=} \text{value} \mid \text{variable} \]
specifies the lower bound of the censored variables. If \text{value} is missing or the value in \text{variable} is missing, no lower bound is set. By default, no lower bound is set.

\[ \text{UB=} \text{value} \mid \text{variable} \]
\[ \text{UPPERBOUND=} \text{value} \mid \text{variable} \]
specifies the upper bound of the censored variables. If \text{value} is missing or the value in \text{variable} is missing, no upper bound is set. By default, no upper bound is set.

**Truncated Variable Options**

\[ \text{TRUNCATED (truncated-options) } \]
You can specify the following truncated-options:

\[ \text{LB=} \text{value} \mid \text{variable} \]
\[ \text{LOWERBOUND=} \text{value} \mid \text{variable} \]
specifies the lower bound of the truncated variables. If \text{value} is missing or the value in \text{variable} is missing, no lower bound is set. By default, no lower bound is set.

\[ \text{UB=} \text{value} \mid \text{variable} \]
\[ \text{UPPERBOUND=} \text{value} \mid \text{variable} \]
specifies the upper bound of the truncated variables. If \text{value} is missing or the value in \text{variable} is missing, no upper bound is set. By default, no upper bound is set.

**Stochastic Frontier Variable Options**

\[ \text{FRONTIER < (frontier-options) > } \]
You can specify the following frontier-options:

\[ \text{TYPE=} \text{HALF} \mid \text{EXPONENTIAL} \mid \text{TRUNCATED} \]
specifies the model type.

\[ \text{HALF} \]
specifies half-normal model.

\[ \text{EXPONENTIAL} \]
specifies exponential model.

\[ \text{TRUNCATED} \]
specifies truncated normal model.

\[ \text{PRODUCTION} \]
specifies that the estimated model be a production function.

\[ \text{COST} \]
specifies that the estimated model be a cost function.

If neither PRODUCTION nor COST is specified, a production function is estimated by default.
**FREQ Statement**

FREQ variable ;

The FREQ statement identifies a variable that contains the frequency of occurrence of each observation. PROC HPQLIM treats each observation as if it appeared \( n \) times, where \( n \) is the value of the FREQ variable for the observation. If the frequency value is not an integer, it is truncated to an integer. If the frequency value is less than 1 or missing, the observation is not used in the model fitting. When the FREQ statement is not specified, each observation is assigned a frequency of 1. If you specify more than one FREQ statement, then the first FREQ statement is used.

**HETERO Statement**

HETERO dependent-variables ~ exogenous-variables / options ;

The HETERO statement specifies variables that are related to the heteroscedasticity of the residuals and the way that these variables are used to model the error variance. PROC HPQLIM supports the following heteroscedastic regression model:

\[
y_i = x_i' \beta + \epsilon_i
\]

\[
\epsilon_i \sim N(0, \sigma_i^2)
\]

For more information about the specification of functional forms, see the section “Heteroscedasticity” on page 1160. The following options specify the functional forms of heteroscedasticity:

**LINK=EXP | LINEAR**

specifies the functional form.

EXP specifies the exponential link function:

\[
\sigma_i^2 = \sigma^2(1 + \exp(x_i' y))
\]

LINEAR specifies the linear link function:

\[
\sigma_i^2 = \sigma^2(1 + z_i' y)
\]

The default is LINK=EXP.

**NOCONST**

specifies that there be no constant in the linear or exponential heteroscedasticity model:

\[
\sigma_i^2 = \sigma^2(z_i' y)
\]

\[
\sigma_i^2 = \sigma^2 \exp(z_i' y)
\]

This option is ignored if you do not specify the LINK= option.
SQUARE

estimates the model by using the square of the linear heteroscedasticity function. For example, you can specify the following heteroscedasticity function:

\[ \sigma_i^2 = \sigma^2 (1 + (z'_i y)^2) \]

```plaintext
model y = x1 x2 / censored(lb=0);
hetero y ~ z1 / link=linear square;
```

The SQUARE option does not apply to the exponential heteroscedasticity function because the square of an exponential function of \( z'_i y \) is the same as the exponential of \( 2z'_i y \). Hence, the only difference is that all \( y \) estimates are divided by two.

This option is ignored if you do not specify the LINK= option. You cannot use the HETERO statement within a Bayesian framework.

---

**INIT Statement**

```plaintext
INIT initvalue1 < , initvalue2 > ;
```

The INIT statement sets initial values for parameters in the optimization. You can specify any number of INIT statements.

Each `initvalue` is written as a parameter or parameter list, followed by an optional equality operator (=), followed by a number:

`parameter <= number`

---

**MODEL Statement**

```plaintext
MODEL dependent-variables = regressors < / options > ;
```

The MODEL statement specifies the dependent variable and independent regressor variables for the regression model.

You can specify the following `option` after a slash (/):

**NOINT**

suppresses the intercept parameter.

You can also specify the following endogenous variable options, which are the same as the options that are specified in the ENDOGENOUS statement. If an endogenous variable option is specified in both the MODEL statement and the ENDOGENOUS statement, the option in the ENDOGENOUS statement is used.
Discrete Variable Options

**DISCRETE** (<discrete-options>)
specifies that the endogenous variables in this statement be discrete. You can specify the following discrete-options:

**DISTRIBUTION**=distribution-type
**DIST**=distribution-type
**D**=distribution-type

specifies the cumulative distribution function that is used to model the response probabilities. You can specify the following distribution-types:

**LOGISTIC** specifies the logistic distribution for the logit model.
**NORMAL** specifies the normal distribution for the probit model.

By default, DISTRIBUTION=NORMAL.

**ORDER**=DATA | FORMATTED | FREQ | INTERNAL
specifies the sort order for the levels of the discrete variables that are specified in the ENDOGENOUS statement. This ordering determines which parameters in the model correspond to each level in the data. You can specify the following sort orders:

**DATA** sorts levels by order of appearance in the input data set.
**FORMATTED** sorts levels by formatted value. The sort order is machine-dependent.
**FREQ** sorts levels by descending frequency count; levels that have the most observations come first in the order.
**INTERNAL** sorts levels by unformatted value. The sort order is machine-dependent.

By default, ORDER=FORMATTED. For more information about sort order, see the chapter on the SORT procedure in the Base SAS Procedures Guide.

Censored Variable Options

**CENSORED** (<censored-options>)
specifies that the endogenous variables in this statement be censored. You can specify the following censored-options:

**LB**=value | variable
**LOWERBOUND**=value | variable

specifies the lower bound of the censored variables. If value is missing or the value in variable is missing, no lower bound is set. By default, no lower bound is set.

**UB**=value | variable
**UPPERBOUND**=value | variable

specifies the upper bound of the censored variables. If value is missing or the value in variable is missing, no upper bound is set. By default, no upper bound is set.
Truncated Variable Options

**TRUNCATED < (truncated-options ) >**

You can specify the following **truncated-options**:

- **LB=value | variable**
- **LOWERBOUND=value | variable**
  
  Specifies the lower bound of the truncated variables. If value is missing or the value in variable is missing, no lower bound is set. By default, no lower bound is set.

- **UB=value | variable**
- **UPPERBOUND=value | variable**
  
  Specifies the upper bound of the truncated variables. If value is missing or the value in variable is missing, no upper bound is set. By default, no upper bound is set.

Stochastic Frontier Variable Options

**FRONTIER < (frontier-options ) >**

You can specify the following **frontier-options**:

- **TYPE=HALF | EXPONENTIAL | TRUNCATED**
  
  Specifies the model type.

  - **HALF** specifies a half-normal model.
  - **EXPONENTIAL** specifies an exponential model.
  - **TRUNCATED** specifies a truncated normal model.

- **PRODUCTION**
  
  Specifies that the estimated model be a production function.

- **COST**
  
  Specifies that the estimated model be a cost function.

If neither PRODUCTION nor COST is specified, a production function is estimated by default.

OUTPUT Statement

**OUTPUT OUT=SAS-data-set < output-options > ;**

The OUTPUT statement creates a new SAS data set to contain variables that are specified with the COPYVAR option and the following data if they are specified by output-options: estimates of $x^\prime \beta$, predicted value, residual, marginal effects, probability, standard deviation of the error, expected value, conditional expected value, technical efficiency measures, and inverse Mills ratio. When the response values are missing for the observation, all output estimates except the residual are still computed as long as none of the explanatory variables are missing. This enables you to compute these statistics for prediction. You can specify only one OUTPUT statement.

You must specify the OUT= option:
OUT= SAS-data-set
    names the output data set.

COPYVAR= SAS-variable-names
COPYVARS= (SAS-variable-names)
    adds SAS variables to the output data set

You can specify one or more of the following output-options:

CONDITIONAL
    outputs estimates of conditional expected values of continuous endogenous variables.

ERRSTD
    outputs estimates of $\sigma_j$, the standard deviation of the error term.

EXPECTED
    outputs estimates of expected values of continuous endogenous variables.

MARGINAL
    outputs marginal effects.

MILLS
    outputs estimates of inverse Mills ratios of censored or truncated continuous, binary discrete, and
    selection endogenous variables.

PREDICTED
    outputs estimates of predicted endogenous variables.

PROB
    outputs estimates of probability of discrete endogenous variables taking the current observed responses.

PROBALL
    outputs estimates of probability of discrete endogenous variables for all possible responses.

RESIDUAL
    outputs estimates of residuals of continuous endogenous variables.

XBETA
    outputs estimates of $x'\beta$.

TE1
    outputs estimates of technical efficiency for each producer in the stochastic frontier model that is
    suggested by Battese and Coelli (1988).

TE2
    outputs estimates of technical efficiency for each producer in the stochastic frontier model that is
    suggested by Jondrow et al. (1982).
PERFORMANCE Statement

PERFORMANCE <performance-options>

The PERFORMANCE statement specifies performance-options to control the multithreaded and distributed computing environment and requests detailed performance results of the HPQLIM procedure. You can also use the PERFORMANCE statement to control whether the HPQLIM procedure executes in single-machine or distributed mode. You can specify the following performance-options:

DETAILS
requests a table that shows a timing breakdown of the procedure steps.

NODES=n
specifies the number of nodes in the distributed computing environment, provided that the data are not processed alongside the database.

NTHREADS=n
specifies the number of threads for analytic computations and overrides the SAS system option THREADS | NOTHREADS. If you do not specify the NTHREADS= option, PROC HPQLIM creates one thread per CPU for the analytic computations.

The PERFORMANCE statement is documented further in the section “PERFORMANCE Statement” (Chapter 2, SAS/ETS User’s Guide: High-Performance Procedures).

PRIOR Statement

PRIOR _REGRESSORS | parameter-list ~ distribution

The PRIOR statement specifies the prior distribution of the model parameters. You must specify one parameter or a list of parameters, a tilde ~, and then a distribution with its parameters. Multiple PRIOR statements are allowed.

You can specify the following distributions:

NORMAL(MEAN=μ, VAR=σ²)
specifies a normal distribution with the parameters MEAN and VAR.

GAMMA(SHAPE=a, SCALE=b)
specifies a gamma distribution with the parameters SHAPE and SCALE.

IGAMMA(SHAPE=a, SCALE=b)
specifies an inverse gamma distribution with the parameters SHAPE and SCALE.

UNIFORM(MIN=m, MAX=M)
specifies a uniform distribution that is defined between MIN and MAX.

BETA(SHAPE1=a, SHAPE2=b, MIN=m, MAX=M)
specifies a beta distribution with the parameters SHAPE1 and SHAPE2 and defined between MIN and MAX.
T(LOCATION=μ, DF=ν)
specifies a noncentral t distribution with DF degrees of freedom and a location parameter equal to LOCATION.

For more information about how to specify distributions, see the section “Standard Distributions” on page 1163.

You can specify the special keyword REGRESSORS to select all the parameters that are used in the linear regression component of the model.

RESTRICT Statement

RESTRICT restriction1 <, restriction2 . . . > ;

The RESTRICT statement imposes linear restrictions on the parameter estimates. You can specify any number of RESTRICT statements, but the number of restrictions that are imposed is limited by the number of regressors.

Each restriction is written as an expression, followed by an equality operator (=) or an inequality operator (<, >, <=, >=), followed by a second expression:

expression operator expression

The operator can be =, <, >, <=, or >=. The operator and second expression are optional.

Restriction expressions can be composed of parameter names; multiplication (*), addition (+), and substitution (-) operators; and constants. Parameters that are named in restriction expressions must be among the parameters that are estimated by the model. Parameters that are associated with a regressor variable are referred to by the name of the corresponding regressor variable. The restriction expressions must be a linear function of the parameters.

The following statements illustrate the use of the RESTRICT statement:

proc hpqlim data=one;
  model y = x1-x10 / censored(lb=0);
  restrict x1*2 <= x2 + x3;
run;

TEST Statement

<’label’>: TEST <’string’>: equation <, equation . . . > / options ;

The TEST statement performs Wald, Lagrange multiplier, and likelihood ratio tests of linear hypotheses about the regression parameters in the preceding MODEL statement. Each equation specifies a linear hypothesis to be tested. All hypotheses in one TEST statement are tested jointly. Variable names in the equations must correspond to regressors in the preceding MODEL statement, and each name represents the coefficient of the corresponding regressor. Use the keyword INTERCEPT for a test that includes a constant.

You can specify the following options after the slash (/):
ALL
requests Wald, Lagrange multiplier, and likelihood ratio tests.

LM
requests the Lagrange multiplier test.

LR
requests the likelihood ratio test.

WALD
requests the Wald test.

The following statements illustrate the use of the TEST statement (note the use of the INTERCEPT keyword in the second TEST statement):

```
proc hpqlim;
   model y = x1 x2 x3;
   test x1 = 0, x2 * .5 + 2 * x3 = 0;
   test _int: test intercept = 0, x3 = 0;
run;
```

The first TEST statement investigates the joint hypothesis that

$$\beta_1 = 0$$

and

$$0.5\beta_2 + 2\beta_3 = 0$$

Only linear equality restrictions and tests are permitted in PROC HPQLIM. Test expressions can be composed only of algebraic operations that involve the addition symbol (+), subtraction symbol (–), and multiplication symbol (*).

The TEST statement accepts labels that are reproduced in the printed output. You can label a TEST statement in two ways: you can specify a label followed by a colon before the TEST keyword, or you can specify a quoted string after the keyword. If you specify both a label before the TEST keyword and a quoted string after the keyword, PROC HPQLIM uses the label that precedes the colon. If no label or quoted string is specified, PROC HPQLIM labels the test automatically.

---

**WEIGHT Statement**

```
WEIGHT variable < / option > ;
```

The WEIGHT statement specifies a variable that supplies weighting values to use for each observation in estimating parameters. The log likelihood for each observation is multiplied by the corresponding weight variable value.

If the weight of an observation is nonpositive, that observation is not used in the estimation.

You can add the following option after a slash (/):
**Details: HPQLIM Procedure**

**Ordinal Discrete Choice Modeling**

**Binary Probit and Logit Model**

The binary choice model is

\[ y_i^* = x_i' \beta + \epsilon_i \]

where the value of the latent dependent variable, \( y_i^* \), is observed only as follows:

\[ y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \]

The disturbance, \( \epsilon_i \), of the probit model has a standard normal distribution with the distribution function (CDF)

\[ \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2) dt \]

The disturbance of the logit model has a standard logistic distribution with the distribution function (CDF)

\[ \Lambda(x) = \frac{\exp(x)}{1 + \exp(x)} = \frac{1}{1 + \exp(-x)} \]

The binary discrete choice model has the following probability that the event \( \{y_i = 1\} \) occurs:

\[ P(y_i = 1) = F(x_i' \beta) = \begin{cases} \Phi(x_i' \beta) & \text{(probit)} \\ \Lambda(x_i' \beta) & \text{(logit)} \end{cases} \]

For more information, see the section “Ordinal Discrete Choice Modeling” on page 1986.

**Ordinal Probit/Logit**

When the dependent variable is observed in sequence with \( M \) categories, binary discrete choice modeling is not appropriate for data analysis. McKelvey and Zavoina (1975) propose the ordinal (or ordered) probit model.

Consider the regression equation

\[ y_i^* = x_i' \beta + \epsilon_i \]
where error disturbances, $\epsilon_i$, have the distribution function $F$. The unobserved continuous random variable, $y^*_i$, is identified as $M$ categories. Suppose there are $M + 1$ real numbers, $\mu_0, \ldots, \mu_M$, where $\mu_0 = -\infty$, $\mu_1 = 0$, $\mu_M = \infty$, and $\mu_0 \leq \mu_1 \leq \cdots \leq \mu_M$. Define

$$R_{i,j} = \mu_j - x'_i \beta$$

The probability that the unobserved dependent variable is contained in the $j$th category can be written as

$$P[\mu_{j-1} < y^*_i \leq \mu_j] = F(R_{i,j}) - F(R_{i,j-1})$$

For more information, see the section “Ordinal Discrete Choice Modeling” on page 1986.

**Limited Dependent Variable Models**

**Censored Regression Models**

When the dependent variable is censored, values in a certain range are all transformed to a single value. For example, the standard Tobit model can be defined as

$$y^*_i = x'_i \beta + \epsilon_i$$

$$y_i = \begin{cases} y^*_i & \text{if } y^*_i > 0 \\ 0 & \text{if } y^*_i \leq 0 \end{cases}$$

where $\epsilon_i \sim \text{iid} N(0, \sigma^2)$.

The Tobit model can be generalized to handle observation-by-observation censoring. The censored model on both the lower and upper limits can be defined as

$$y_i = \begin{cases} R_i & \text{if } y^*_i \geq R_i \\ y^*_i & \text{if } L_i < y^*_i < R_i \\ L_i & \text{if } y^*_i \leq L_i \end{cases}$$

For more information, see Chapter 28.7, “Censored Regression Models.”

**Truncated Regression Models**

In a truncated model, the observed sample is a subset of the population where the dependent variable falls within a certain range. For example, when neither a dependent variable nor exogenous variables are observed for $y^*_i \leq 0$, the truncated regression model can be specified as

$$\ell = \sum_{i \in \{y_i > 0\}} \left\{ -\ln \Phi(x'_i \beta / \sigma) + \ln \left[ \frac{\phi((y_i - x'_i \beta) / \sigma)}{\sigma} \right] \right\}$$

For more information, see the section “Truncated Regression Models” on page 1991.
Stochastic Frontier Production and Cost Models

Stochastic frontier production models were first developed by Aigner, Lovell, and Schmidt (1977); Meeusen and van den Broeck (1977). Specification of these models allow for random shocks of the production or cost but also include a term for technical or cost inefficiency. Assuming that the production function takes a log-linear Cobb-Douglas form, the stochastic frontier production model can be written as

$$\ln(y_i) = \beta_0 + \sum_n \beta_n \ln(x_{ni}) + \epsilon_i$$

where $\epsilon_i = v_i - u_i$. The $v_i$ term represents the stochastic error component, and the $u_i$ term represents the nonnegative, technical inefficiency error component. The $v_i$ error component is assumed to be distributed iid normal and independent from $u_i$. If $u_i > 0$, the error term $\epsilon_i$ is negatively skewed and represents technical inefficiency. If $u_i < 0$, the error term $\epsilon_i$ is positively skewed and represents cost inefficiency. PROC HPQLIM models the $u_i$ error component as a half-normal, exponential, or truncated normal distribution.

The Normal-Half-Normal Model

When $v_i$ is iid $N(0, \sigma_v^2)$ in a normal-half-normal model, $u_i$ is iid $N^+(0, \sigma_u^2)$, with $v_i$ and $u_i$ independent of each other. Given the independence of error terms, the joint density of $v$ and $u$ can be written as

$$f(u, v) = \frac{2}{2\pi \sigma_u \sigma_v} \exp \left\{ -\frac{u^2}{2\sigma_u^2} - \frac{v^2}{2\sigma_v^2} \right\}$$

Substituting $v = \epsilon + u$ into the preceding equation and integrating $u$ out gives

$$f(\epsilon) = \frac{2}{\sigma} \phi \left( \frac{\epsilon}{\sigma} \right) \Phi \left( -\frac{\lambda}{\sigma} \right)$$

where $\lambda = \sigma_u / \sigma_v$ and $\sigma = \sqrt{\sigma_u^2 + \sigma_v^2}$.

In the case of a stochastic frontier cost model, $v = \epsilon - u$ and

$$f(\epsilon) = \frac{2}{\sigma} \phi \left( \frac{\epsilon}{\sigma} \right) \Phi \left( \frac{\epsilon + \lambda}{\sigma} \right)$$

For more information, see the section “Stochastic Frontier Production and Cost Models” on page 1992.

The Normal-Exponential Model

Under the normal-exponential model, $v_i$ is iid $N(0, \sigma_v^2)$ and $u_i$ is iid exponential. Given the independence of error term components $u_i$ and $v_i$, the joint density of $v$ and $u$ can be written as

$$f(u, v) = \frac{1}{\sqrt{2\pi \sigma_u \sigma_v}} \exp \left\{ -\frac{u}{\sigma_u} - \frac{v}{2\sigma_v^2} \right\}$$

The marginal density function of $\epsilon$ for the production function is

$$f(\epsilon) = \left( \frac{1}{\sigma_u} \right) \Phi \left( -\frac{\epsilon}{\sigma_v} \right) \exp \left\{ \frac{\epsilon}{\sigma_u} + \frac{\sigma_v^2}{2\sigma_u^2} \right\}$$
The marginal density function for the cost function is equal to

\[ f(\epsilon) = \left( \frac{1}{\sigma_u} \right) \Phi \left( \frac{\epsilon - \sigma_v}{\sigma_u} \right) \exp \left\{ - \frac{\epsilon}{\sigma_u} + \frac{\sigma_v^2}{2\sigma_u^2} \right\} \]

For more information, see the section “Stochastic Frontier Production and Cost Models” on page 1992.

**The Normal–Truncated Normal Model**

The normal–truncated normal model is a generalization of the normal-half-normal model that allows the mean of \( u_i \) to differ from zero. Under the normal–truncated normal model, the error term component \( v_i \) is iid \( N(0, \sigma_v^2) \) and \( u_i \) is iid \( N(\mu, \sigma_u^2) \). The joint density of \( v_i \) and \( u_i \) can be written as

\[ f(u, v) = \frac{1}{\sqrt{2\pi} \sigma_u \sigma_v} \exp \left\{ - \frac{(u - \mu)^2}{2\sigma_u^2} - \frac{v^2}{2\sigma_v^2} \right\} \]

The marginal density function of \( \epsilon \) for the production function is

\[ f(\epsilon) = \frac{1}{\sigma} \phi \left( \frac{\epsilon + \mu}{\sigma} \right) \Phi \left( \frac{\mu}{\sigma} \right) \Phi \left( \frac{\mu}{\sigma_u} \right)^{-1} \]

The marginal density function for the cost function is

\[ f(\epsilon) = \frac{1}{\sigma} \phi \left( \frac{\epsilon - \mu}{\sigma} \right) \Phi \left( \frac{\mu}{\sigma} \right) \Phi \left( \frac{\mu}{\sigma_u} \right)^{-1} \]

For more information, see the section “Stochastic Frontier Production and Cost Models” on page 1992.

For more information about normal-half-normal, normal-exponential, and normal–truncated normal models, see Kumbhakar and Lovell (2000); Coelli, Prasada Rao, and Battese (1998).

**Heteroscedasticity**

If the variance of regression disturbance, \( (\epsilon_i) \), is heteroscedastic, the variance can be specified as a function of variables

\[ E(\epsilon_i^2) = \sigma_i^2 = f(z_i, y) \]

Table 21.2 shows various functional forms of heteroscedasticity and the corresponding options to request each model.

<table>
<thead>
<tr>
<th>Number</th>
<th>Model</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( f(z_i, y) = \sigma^2(1 + \exp(z_i y)) )</td>
<td>LINK=EXP (default)</td>
</tr>
<tr>
<td>2</td>
<td>( f(z_i, y) = \sigma^2 \exp(z_i y) )</td>
<td>LINK=EXP NOCONST</td>
</tr>
<tr>
<td>3</td>
<td>( f(z_i, y) = \sigma^2(1 + \sum_{i=1}^L y_i z_{ii}) )</td>
<td>LINK=LINEAR</td>
</tr>
<tr>
<td>4</td>
<td>( f(z_i, y) = \sigma^2(1 + (\sum_{i=1}^L y_i z_{ii})^2) )</td>
<td>LINK=LINEAR SQUARE</td>
</tr>
<tr>
<td>5</td>
<td>( f(z_i, y) = \sigma^2(\sum_{i=1}^L y_i z_{ii}) )</td>
<td>LINK=LINEAR NOCONST</td>
</tr>
<tr>
<td>6</td>
<td>( f(z_i, y) = \sigma^2((\sum_{i=1}^L y_i z_{ii})^2) )</td>
<td>LINK=LINEAR SQUARE NOCONST</td>
</tr>
</tbody>
</table>
In models 3 and 5, variances of some observations might be negative. Although the HPQLIM procedure
assigns a large penalty to move the optimization away from such a region, the optimization might not be able
to improve the objective function value and might become locked in the region. Signs of such an outcome
include extremely small likelihood values or missing standard errors in the estimates. In models 2 and 6,
variances are guaranteed to be greater than or equal to zero, but variances of some observations might be very
close to 0. In these scenarios, standard errors might be missing. Models 1 and 4 do not have such problems.
Variances in these models are always positive and never close to 0.

For more information, see the section “Heteroscedasticity and Box-Cox Transformation” on page 1994.

Tests on Parameters

In general, the tested hypothesis can be written as

\[ H_0 : h(\theta) = 0 \]

where \( h(\theta) \) is an \( r \times 1 \) vector-valued function of the parameters \( \theta \) given by the \( r \) expressions that are specified
in the TEST statement.

Let \( \hat{V} \) be the estimate of the covariance matrix of \( \hat{\theta} \). Let \( \hat{\theta} \) be the unconstrained estimate of \( \theta \) and \( \tilde{\theta} \) be the
constrained estimate of \( \theta \) such that \( h(\tilde{\theta}) = 0 \). Let

\[ A(\theta) = \frac{\partial h(\theta)}{\partial \theta} \bigg|_{\tilde{\theta}} \]

Using this notation, the test statistics for the three types of tests are computed as follows.

- The Wald test statistic is defined as

\[ W = h'(\hat{\theta}) (A(\hat{\theta})\hat{V} A'(\hat{\theta}))^{-1} h(\hat{\theta}) \]

- The Lagrange multiplier test statistic is

\[ LM = \lambda' A(\hat{\theta})\hat{V} A'(\hat{\theta})\lambda \]

where \( \lambda \) is the vector of Lagrange multipliers from the computation of the restricted estimate \( \bar{\theta} \).

- The likelihood ratio test statistic is

\[ LR = 2 \left( L(\hat{\theta}) - L(\tilde{\theta}) \right) \]

where \( \tilde{\theta} \) represents the constrained estimate of \( \theta \) and \( L \) is the concentrated log-likelihood value.

The following statements use the TEST statement to perform a likelihood ratio test:

\[
\begin{verbatim}
proc hpqlim;
  model y = x1 x2 x3;
  test x1 = 0, x2 * .5 + 2 * x3 = 0 /lr;
run;
\end{verbatim}
\]

For more information, see the section “Tests on Parameters” on page 2001.
Bayesian Analysis

To perform Bayesian analysis, you must specify a BAYES statement. Unless otherwise stated, all options that are described in this section are options in the BAYES statement.

By default, PROC HPQLIM uses the random walk Metropolis algorithm to obtain posterior samples. For the implementation details of the Metropolis algorithm in PROC HPQLIM, such as the blocking of the parameters and tuning of the covariance matrices, see the sections “Blocking of Parameters” on page 1162 and “Tuning the Proposal Distribution” on page 1162.

The Bayes theorem states that

\[ p(\theta|y) \propto \pi(\theta)L(y|\theta) \]

where \( \theta \) is a parameter or a vector of parameters and \( \pi(\theta) \) is the product of the prior densities that are specified in the PRIOR statement. The term \( L(y|\theta) \) is the likelihood that is associated with the MODEL statement.

Blocking of Parameters

In a multivariate parameter model, all the parameters are updated in one single block (by default or when you specify the SAMPLING=MULTIMETROPOLIS option). This can be inefficient, especially when parameters have vastly different scales. As an alternative, you can update the parameters one at a time (by specifying SAMPLING=UNIMETROPOLIS).

Tuning the Proposal Distribution

One key factor in achieving high efficiency of a Metropolis-based Markov chain is finding a good proposal distribution for each block of parameters. This process is called tuning. The tuning phase consists of a number of loops that are controlled by the options MINTUNE= and MAXTUNE=. The MINTUNE= option controls the minimum number of tuning loops and has a default value of 2. The MAXTUNE= option controls the maximum number of tuning loops and has a default value of 24. Each loop repeats the number of times specified by the NTU= option, which has a default of 500. At the end of every loop, PROC HPQLIM examines the acceptance probability for each block. The acceptance probability is the percentage of NTU proposed values that have been accepted. If this probability does not fall within the acceptance tolerance range (see the following section), the proposal distribution is modified before the next tuning loop.

A good proposal distribution should resemble the actual posterior distribution of the parameters. Large sample theory states that the posterior distribution of the parameters approaches a multivariate normal distribution (Gelman et al. 2004, Appendix B; Schervish 1995, Section 7.4). That is why a normal proposal distribution often works well in practice. The default proposal distribution in PROC HPQLIM is the normal distribution.

For more information, see Chapter 28.7, “Bayesian Analysis.”

Initial Values of the Markov Chains

You can assign initial values to any parameters. For more information, see the INIT statement. If you use the optimization PROPPOV= option, PROC HPQLIM starts the tuning at the optimized values. This option overwrites the provided initial values.
Prior Distributions

The PRIOR statement specifies the prior distribution of the model parameters. You must specify one parameter or a list of parameters, a tilde ~, and then a distribution with its parameters. You can specify multiple PRIOR statements to define independent priors. Parameters that are associated with a regressor variable are referred to by the name of the corresponding regressor variable.

You can specify the special keyword _REGRESSORS to consider all the regressors of a model. If multiple PRIOR statements affect the same parameter, the last PRIOR statement prevails. For example, in a regression with two regressors (X1, X2), the following statements imply that the prior on X1 is NORMAL(MEAN=0, VAR=1) and the prior on X2 is GAMMA(SHAPE=3, SCALE=4):

```plaintext
... prior _Regressors ~ uniform(min=0, max=1);
prior X1 X2 ~ gamma(shape=3, scale=4);
prior X1 ~ normal(mean=0, var=1);
... 
```

If a parameter is not associated with a PRIOR statement or if some of the prior hyperparameters are missing, then the default choices in Table 21.3 are considered.

<table>
<thead>
<tr>
<th>PRIOR Distribution</th>
<th>Hyperparameter1</th>
<th>Hyperparameter2</th>
<th>Min</th>
<th>Max</th>
<th>Parameters Default Choice</th>
</tr>
</thead>
<tbody>
<tr>
<td>NORMAL</td>
<td>MEAN=0</td>
<td>VAR=1E6</td>
<td>−∞</td>
<td>∞</td>
<td>Regression-Location-Threshold</td>
</tr>
<tr>
<td>IGAMMA</td>
<td>SHAPE=2.000001</td>
<td>SCALE=1</td>
<td>&gt; 0</td>
<td>∞</td>
<td>Scale</td>
</tr>
<tr>
<td>GAMMA</td>
<td>SHAPE=1</td>
<td>SCALE=1</td>
<td>0</td>
<td>∞</td>
<td></td>
</tr>
<tr>
<td>UNIFORM</td>
<td></td>
<td></td>
<td>−∞</td>
<td>∞</td>
<td></td>
</tr>
<tr>
<td>BETA</td>
<td>SHAPE1=1</td>
<td>SHAPE2=1</td>
<td>−∞</td>
<td>∞</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>LOCATION=0</td>
<td>DF=3</td>
<td>−∞</td>
<td>∞</td>
<td></td>
</tr>
</tbody>
</table>

For density specification, see the section “Standard Distributions” on page 1163.

Standard Distributions

Table 21.4 through Table 21.9 show all the distribution density functions that PROC HPQLIM recognizes. You specify these distribution densities in the PRIOR statement.

Table 21.4 Beta Distribution

<table>
<thead>
<tr>
<th>PRIOR statement</th>
<th>BETA(SHAPE1=a, SHAPE2=b, MIN=m, MAX=M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density</td>
<td>( \frac{(\theta-m)^{a-1}(M-\theta)^{b-1}}{B(a,b)(M-m)^{a+b-1}} )</td>
</tr>
<tr>
<td>Parameter restriction</td>
<td>( a &gt; 0, \ b &gt; 0, \ -\infty &lt; m &lt; M &lt; \infty )</td>
</tr>
</tbody>
</table>
| Range | \[
| \{ [m, M] \quad \text{when } a = 1, b = 1 \\
| \{ [m, M] \quad \text{when } a = 1, b \neq 1 \\
| \{ (m, M) \quad \text{when } a \neq 1, b = 1 \\
| \{ (m, M) \quad \text{otherwise} \] |
Chapter 21: The HPQLIM Procedure

Mean
\[ \frac{a}{a+b} \times (M - m) + m \]

Variance
\[ \frac{ab}{(a+b)^2(a+b+1)} \times (M - m)^2 \]
\[ \frac{a-1}{a+b-2} \times M + \frac{b-1}{a+b-2} \times m \quad \text{if } a, b > 1 \]
\[ m \text{ and } M \quad \text{if } a < 1, b < 1 \]

Mode
\[ \begin{cases} 
  m & \text{if } a < 1, b \geq 1 \\
  a = 1, b > 1 & \text{if } a = 1, b = 1 \\
  a > 1, b < 1 & \text{if } a > 1, b = 1 \\
  a = b = 1 & \text{if not unique} 
\end{cases} \]

Defaults SHAPE1=SHAPE2=1, MIN \to -\infty, MAX \to \infty

Table 21.5 Gamma Distribution

<table>
<thead>
<tr>
<th>PRIOR statement</th>
<th>GAMMA(SHAPE=a, SCALE=b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density</td>
<td>( \frac{1}{b^a \Gamma(a)} \theta^{a-1} e^{-\theta/b} )</td>
</tr>
<tr>
<td>Parameter restriction</td>
<td>( a &gt; 0, b &gt; 0 )</td>
</tr>
<tr>
<td>Range</td>
<td>([0, \infty))</td>
</tr>
<tr>
<td>Mean</td>
<td>( ab )</td>
</tr>
<tr>
<td>Variance</td>
<td>( ab^2 )</td>
</tr>
<tr>
<td>Mode</td>
<td>( (a - 1)b )</td>
</tr>
<tr>
<td>Defaults</td>
<td>SHAPE=SCALE=1</td>
</tr>
</tbody>
</table>

Table 21.6 Inverse Gamma Distribution

<table>
<thead>
<tr>
<th>PRIOR statement</th>
<th>IGAMMA(SHAPE=a, SCALE=b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density</td>
<td>( \frac{b^a}{\Gamma(a)} \theta^{-(a+1)} e^{-b/\theta} )</td>
</tr>
<tr>
<td>Parameter restriction</td>
<td>( a &gt; 0, b &gt; 0 )</td>
</tr>
<tr>
<td>Range</td>
<td>( 0 &lt; \theta &lt; \infty )</td>
</tr>
<tr>
<td>Mean</td>
<td>( \frac{b}{a-1}, \quad a &gt; 1 )</td>
</tr>
<tr>
<td>Variance</td>
<td>( \frac{b^2}{(a-1)^2(a-2)}, \quad a &gt; 2 )</td>
</tr>
<tr>
<td>Mode</td>
<td>( \frac{b}{a+1} )</td>
</tr>
<tr>
<td>Defaults</td>
<td>SHAPE=2.000001, SCALE=1</td>
</tr>
</tbody>
</table>
**Table 21.7** Normal Distribution

<table>
<thead>
<tr>
<th>PRIOR statement</th>
<th>NORMAL(MEAN=μ, VAR=σ^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density</td>
<td>( \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(\theta-\mu)^2}{2\sigma^2}\right) )</td>
</tr>
<tr>
<td>Parameter restriction</td>
<td>( \sigma^2 &gt; 0 )</td>
</tr>
<tr>
<td>Range</td>
<td>(-\infty &lt; \theta &lt; \infty)</td>
</tr>
<tr>
<td>Mean</td>
<td>( \mu )</td>
</tr>
<tr>
<td>Variance</td>
<td>( \sigma^2 )</td>
</tr>
<tr>
<td>Mode</td>
<td>( \mu )</td>
</tr>
<tr>
<td>Defaults</td>
<td>MEAN=0, VAR=1000000</td>
</tr>
</tbody>
</table>

**Table 21.8** t Distribution

<table>
<thead>
<tr>
<th>PRIOR statement</th>
<th>T(LOCATION=μ, DF=ν)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density</td>
<td>( \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi \nu} \Gamma\left(\frac{\nu}{2}\right)} \left[ 1 + \frac{(\theta-\mu)^2}{\nu} \right]^{-\frac{\nu+1}{2}} )</td>
</tr>
<tr>
<td>Parameter restriction</td>
<td>( \nu &gt; 0 )</td>
</tr>
<tr>
<td>Range</td>
<td>(-\infty &lt; \theta &lt; \infty)</td>
</tr>
<tr>
<td>Mean</td>
<td>( \mu ), for ( \nu &gt; 1 )</td>
</tr>
<tr>
<td>Variance</td>
<td>( \frac{\nu}{\nu-2} ), for ( \nu &gt; 2 )</td>
</tr>
<tr>
<td>Mode</td>
<td>( \mu )</td>
</tr>
<tr>
<td>Defaults</td>
<td>LOCATION=0, DF=3</td>
</tr>
</tbody>
</table>

**Table 21.9** Uniform Distribution

<table>
<thead>
<tr>
<th>PRIOR statement</th>
<th>UNIFORM(MIN=m, MAX=M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density</td>
<td>( \frac{1}{M-m} )</td>
</tr>
<tr>
<td>Parameter restriction</td>
<td>(-\infty &lt; m &lt; M &lt; \infty)</td>
</tr>
<tr>
<td>Range</td>
<td>( \theta \in [m, M] )</td>
</tr>
<tr>
<td>Mean</td>
<td>( \frac{m+M}{2} )</td>
</tr>
<tr>
<td>Variance</td>
<td>( \frac{(M-m)^2}{12} )</td>
</tr>
<tr>
<td>Mode</td>
<td>Not unique</td>
</tr>
<tr>
<td>Defaults</td>
<td>MIN\rightarrow -\infty, MAX\rightarrow \infty</td>
</tr>
</tbody>
</table>
Output to SAS Data Set

XBeta, Predicted, and Residual

XBeta is the structural part on the right-hand side of the model. The predicted value is the predicted dependent variable value. For censored variables, if the predicted value is outside the boundaries, it is reported as the closest boundary. The residual is defined only for continuous variables and is defined as

\[ \text{Residual} = \text{Observed} - \text{Predicted} \]

Error Standard Deviation

The error standard deviation is \( \sigma_i \) in the model. It varies only when the HETERO statement is used.

Marginal Effects

A marginal effect is defined as a contribution of one control variable to the response variable. For a binary choice model with two response categories, \( \mu_0 = -\infty \) and \( \mu_1 = 0, \mu_2 = \infty \). For an ordinal response model with \( M \) response categories \( \mu_0, \ldots, \mu_M \), define

\[ R_{i,j} = \mu_j - x_i^T \beta \]

The probability that the unobserved dependent variable is contained in the \( j \)th category can be written as

\[ P[\mu_{j-1} < y_i^* \leq \mu_j] = F(R_{i,j}) - F(R_{i,j-1}) \]

The marginal effect of changes in the regressors on the probability of \( y_i = j \) is then

\[ \frac{\partial \text{Prob}[y_i = j]}{\partial x} = [f(\mu_{j-1} - x_i^T \beta) - f(\mu_j - x_i^T \beta)] \beta \]

where \( f(x) = \frac{dF(x)}{dx} \). In particular,

\[ f(x) = \frac{dF(x)}{dx} = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} & \text{(probit)} \\ \frac{e^{-x}}{[1+e^{-x}]^2} & \text{(logit)} \end{cases} \]

The marginal effects in the truncated regression model are

\[ \frac{\partial E[y_i|L_i < y_i^* < R_i]}{\partial x} = \beta \left[ 1 - \frac{(\phi(a_i) - \phi(b_i))^2}{(\Phi(b_i) - \Phi(a_i))^2} + \frac{a_i \phi(a_i) - b_i \phi(b_i)}{\Phi(b_i) - \Phi(a_i)} \right] \]

where \( a_i = \frac{L_i - x_i^T \beta}{\sigma_i} \) and \( b_i = \frac{R_i - x_i^T \beta}{\sigma_i} \).

The marginal effects in the censored regression model are

\[ \frac{\partial E[y|x_i]}{\partial x} = \beta \times \text{Prob}[L_i < y_i^* < R_i] \]
Expected and Conditionally Expected Values

The expected value is the unconditional expectation of the dependent variable. For a censored variable, it is

\[ E[y_i] = \Phi(a_i) L_i + (x'_i \beta + \lambda \sigma_i) (\Phi(b_i) - \Phi(a_i)) + (1 - \Phi(b_i)) R_i \]

For a left-censored variable \((R_i = \infty)\), this formula is

\[ E[y_i] = \Phi(a_i) L_i + (x'_i \beta + \lambda \sigma_i) (1 - \Phi(a_i)) \]

where \(\lambda = \frac{\phi(a_i)}{1 - \Phi(a_i)}\).

For a right-censored variable \((L_i = -\infty)\), this formula is

\[ E[y_i] = (x'_i \beta + \lambda \sigma_i) \Phi(b_i) + (1 - \Phi(b_i)) R_i \]

where \(\lambda = -\frac{\phi(b_i)}{\Phi(b_i)}\).

For a noncensored variable, this formula is

\[ E[y_i] = x'_i \beta \]

The conditional expected value is the expectation when the variable is inside the boundaries:

\[ E[y_i | L_i < y_i < R_i] = x'_i \beta + \lambda \sigma_i \]

Technical Efficiency

Technical efficiency for each producer is computed only for stochastic frontier models.

In general, the stochastic production frontier can be written as

\[ y_i = f(x_i; \beta) \exp\{v_i\} T E_i \]

where \(y_i\) denotes producer \(i\)'s actual output, \(f(\cdot)\) is the deterministic part of the production frontier, \(\exp\{v_i\}\) is a producer-specific error term, and \(T E_i\) is the technical efficiency coefficient, which can be written as

\[ T E_i = \frac{y_i}{f(x_i; \beta) \exp\{v_i\}} \]

For a Cobb-Douglas production function, \(T E_i = \exp\{-u_i\}\). For more information, see the section “Stochastic Frontier Production and Cost Models” on page 1159.

The cost frontier can be written in general as

\[ E_i = c(y_i, w_i; \beta) \exp\{v_i\} / C E_i \]

where \(w_i\) denotes producer \(i\)'s input prices, \(c(\cdot)\) is the deterministic part of the cost frontier, \(\exp\{v_i\}\) is a producer-specific error term, and \(C E_i\) is the cost efficiency coefficient, which can be written as

\[ C E_i = \frac{c(x_i, w_i; \beta) \exp\{v_i\}}{E_i} \]
For a Cobb-Douglas cost function, \( CE_i = \exp(-u_i) \). For more information, see the section “Stochastic Frontier Production and Cost Models” on page 1159. Hence, both technical and cost efficiency coefficients are the same. The estimates of technical efficiency are provided in the following subsections.

Normal-Half-Normal Model

Define \( \mu_* = -\epsilon \sigma_u^2 / \sigma^2 \) and \( \sigma_*^2 = \sigma_u^2 \sigma_v^2 / \sigma^2 \). Then, as shown by Jondrow et al. (1982), conditional density is as follows:

\[
\begin{align*}
    f(u|\epsilon) &= \frac{1}{\sqrt{2\pi}\sigma_*} \exp\left\{ -\frac{(u - \mu_*)^2}{2\sigma_*^2} \right\} / \left[ 1 - \Phi\left( -\frac{\mu_*}{\sigma_*} \right) \right] \\

    \text{Hence, } f(u|\epsilon) \text{ is the density for } N^+(\mu_*, \sigma_*^2).
\end{align*}
\]

From this result, it follows that the estimate of technical efficiency (Battese and Coelli 1988) is

\[
TE1_i = E(\exp\{-u_i|\epsilon_i\}) = \left[ \frac{1 - \Phi(\sigma_* - \mu_* / \sigma_*)}{1 - \Phi(-\mu_* / \sigma_*)} \right] \exp\left\{ -\mu_* + \frac{1}{2} \sigma_*^2 \right\}
\]

The second version of the estimate (Jondrow et al. 1982) is

\[
TE2_i = \exp\{-E(u_i|\epsilon_i)\}
\]

where

\[
E(u_i|\epsilon_i) = \mu_* + \sigma_* \left[ \frac{\phi(-\mu_* / \sigma_*)}{1 - \Phi(-\mu_* / \sigma_*)} \right] = \sigma_* \left[ \frac{\phi(\epsilon_i \lambda / \sigma)}{1 - \Phi(\epsilon_i \lambda / \sigma)} - \left( \frac{\epsilon_i \lambda}{\sigma} \right) \right]
\]

Normal-Exponential Model

Define \( A = -\bar{\mu} / \sigma_v \) and \( \bar{\mu} = -\epsilon - \sigma_v^2 / \sigma_u \). Then, as shown by Kumbhakar and Lovell (2000), conditional density is as follows:

\[
\begin{align*}
    f(u|\epsilon) &= \frac{1}{\sqrt{2\pi}\sigma_v \Phi(-\bar{\mu} / \sigma_v)} \exp\left\{ -\frac{(u - \bar{\mu})^2}{2\sigma_v^2} \right\} \\

    \text{Hence, } f(u|\epsilon) \text{ is the density for } N^+(\bar{\mu}, \sigma_v^2).
\end{align*}
\]

From this result, it follows that the estimate of technical efficiency is

\[
TE1_i = E(\exp\{-u_i|\epsilon_i\}) = \left[ \frac{1 - \Phi(\sigma_v - \bar{\mu}_i / \sigma_v)}{1 - \Phi(-\bar{\mu}_i / \sigma_v)} \right] \exp\left\{ -\bar{\mu}_i + \frac{1}{2} \sigma_v^2 \right\}
\]

The second version of the estimate is

\[
TE2_i = \exp\{-E(u_i|\epsilon_i)\}
\]

where

\[
E(u_i|\epsilon_i) = \bar{\mu}_i + \sigma_v \left[ \frac{\phi(-\bar{\mu}_i / \sigma_v)}{1 - \Phi(-\bar{\mu}_i / \sigma_v)} \right] = \sigma_v \left[ \frac{\phi(A)}{\Phi(-A) - A} \right]
\]
Normal–Truncated Normal Model

Define \( \mu = (-\sigma^2_\epsilon \epsilon_i + \mu \sigma^2_\sigma^2) / \sigma^2 \) and \( \sigma^2_\sigma^2 = \sigma^2_\sigma^2 \sigma^2_\epsilon / \sigma^2 \). Then, as shown by Kumbhakar and Lovell (2000), conditional density is as follows:

\[
 f(u|\epsilon) = \frac{1}{\sqrt{2\pi\sigma_*[1 - \Phi(-\mu/\sigma_*)]}} \exp\left\{ -\frac{(u - \mu)^2}{2\sigma^2_*}\right\}
\]

Hence, \( f(u|\epsilon) \) is the density for \( N^+(\mu, \sigma^2_\sigma^2) \).

From this result, it follows that the estimate of technical efficiency is

\[
 TE_{1i} = E(\exp\{-u_i\}|\epsilon_i) = \frac{1 - \Phi(\sigma_* - \mu_i/\sigma_*)}{1 - \Phi(-\mu_i/\sigma_*)} \exp\left\{ -\mu_i + \frac{1}{2}\sigma^2_\sigma^2\right\}
\]

The second version of the estimate is

\[
 TE_{2i} = \exp\{-E(u_i|\epsilon_i)\}
\]

where

\[
 E(u_i|\epsilon_i) = \mu_i + \sigma_* \left[ \frac{\phi(\mu_i/\sigma_*)}{1 - \Phi(-\mu_i/\sigma_*)} \right]
\]

OUTEST= Data Set

The OUTEST= data set contains all the parameters that are estimated by a MODEL statement. Each parameter contains the estimate for the corresponding parameter in the corresponding model. In addition, the OUTEST= data set contains the following variables:

_\_NAME\_ indicates the name of the independent variable.
_\_TYPE\_ indicates the type of observation. PARM indicates the row of coefficients; STD indicates the row of standard deviations of the corresponding coefficients.
_\_STATUS\_ indicates the convergence status for optimization.

The rest of the columns correspond to the explanatory variables.

The OUTEST= data set contains one observation for the MODEL statement, which shows the parameter estimates for that model. If you specify the COVOUT option in the PROC HPQLIM statement, the OUTEST= data set includes additional observations for the MODEL statement, which show the rows of the covariance matrix of parameter estimates. For covariance observations, the value of the _\_TYPE\_ variable is COV, and the _\_NAME\_ variable identifies the parameter that is associated with that row of the covariance matrix. If you specify the CORROUT option in the PROC HPQLIM statement, the OUTEST= data set includes additional observations for the MODEL statement, which show the rows of the correlation matrix of parameter estimates. For correlation observations, the value of the _\_TYPE\_ variable is CORR, and the _\_NAME\_ variable identifies the parameter that is associated with that row of the correlation matrix.
Naming

Naming of Parameters

The parameters are named in the same way as in other SAS procedures such as the REG and PROBIT procedures. The constant in the regression equation is called Intercept. The coefficients of independent variables are named by the independent variables. The standard deviation of the errors is called _Sigma. If the HETERO statement is included, the coefficients of the independent variables in the HETERO statement are called _H.x, where x is the name of the independent variable.

Naming of Output Variables

Table 21.10 shows the options in the OUTPUT statement, with the corresponding variable names and their explanations.

<table>
<thead>
<tr>
<th>output-option</th>
<th>Variable Name</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONDITIONAL</td>
<td>CEXPCT_y</td>
<td>Conditional expected value of y, conditioned on the truncation</td>
</tr>
<tr>
<td>ERRSTD</td>
<td>ERRSTD_y</td>
<td>Standard deviation of error term</td>
</tr>
<tr>
<td>EXPECTED</td>
<td>EXPCT_y</td>
<td>Unconditional expected value of y</td>
</tr>
<tr>
<td>MARGINAL</td>
<td>MEFF_x</td>
<td>Marginal effect of x on y (dy/dx) with single equation</td>
</tr>
<tr>
<td>PREDICTED</td>
<td>P_y</td>
<td>Predicted value of y</td>
</tr>
<tr>
<td>RESIDUAL</td>
<td>RESID_y</td>
<td>Residual of y, (y – PredictedY)</td>
</tr>
<tr>
<td>PROB</td>
<td>PROB_y</td>
<td>Probability that y is taking the observed value in this observation (discrete y only)</td>
</tr>
<tr>
<td>PROBALL</td>
<td>PROBi_y</td>
<td>Probability that y is taking the ith value (discrete y only)</td>
</tr>
<tr>
<td>MILLS</td>
<td>MILLS_y</td>
<td>Inverse Mills ratio for y</td>
</tr>
<tr>
<td>TE1</td>
<td>TE1</td>
<td>Technical efficiency estimate for each producer proposed by Battese and Coelli (1988)</td>
</tr>
<tr>
<td>TE2</td>
<td>TE2</td>
<td>Technical efficiency estimate for each producer proposed by Jondrow et al. (1982)</td>
</tr>
<tr>
<td>XBETA</td>
<td>XBETA_y</td>
<td>Structure part (x'β) of y equation</td>
</tr>
</tbody>
</table>

If you prefer to name the output variables differently, you can use the RENAME option in the data set. For example, the following statements rename the residual of y as Resid:

```sas
proc hpqlim data=one;
    model y = x1-x10 / censored;
    output out=outds(rename=(resid_y=resid)) residual;
run;
```
**ODS Table Names**

PROC HPQLIM assigns a name to each table that it creates. You can use these names to refer to the table when you use the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in Table 21.11.

<table>
<thead>
<tr>
<th>ODS Table Name</th>
<th>Description</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ODS Tables Created by the MODEL Statement and TEST Statement</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResponseProfile</td>
<td>Response profile</td>
<td>Default</td>
</tr>
<tr>
<td>FitSummary</td>
<td>Summary of nonlinear estimation</td>
<td>Default</td>
</tr>
<tr>
<td>ParameterEstimates</td>
<td>Parameter estimates</td>
<td>Default</td>
</tr>
<tr>
<td>SummaryContResponse</td>
<td>Summary of continuous response</td>
<td>Default</td>
</tr>
<tr>
<td>CovB</td>
<td>Covariance of parameter estimates</td>
<td>COVB</td>
</tr>
<tr>
<td>CorrB</td>
<td>Correlation of parameter estimates</td>
<td>CORRB</td>
</tr>
<tr>
<td><strong>ODS Tables Created by the BAYES Statement</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AutoCorr</td>
<td>Autocorrelation statistics for each parameter</td>
<td>Default</td>
</tr>
<tr>
<td>Corr</td>
<td>Correlation matrix of the posterior samples</td>
<td>STATS=COR</td>
</tr>
<tr>
<td>Cov</td>
<td>Covariance matrix of the posterior samples</td>
<td>STATS=COV</td>
</tr>
<tr>
<td>ESS</td>
<td>Effective sample size for each parameter</td>
<td>Default</td>
</tr>
<tr>
<td>MCSE</td>
<td>Monte Carlo standard error for each parameter</td>
<td>Default</td>
</tr>
<tr>
<td>Geweke</td>
<td>Geweke diagnostics for each parameter</td>
<td>Default</td>
</tr>
<tr>
<td>Heidelberger</td>
<td>Heidelberger-Welch diagnostics for each parameter</td>
<td>DIAGNOSTICS=HEIDEL</td>
</tr>
<tr>
<td>PostIntervals</td>
<td>Equal-tail and HPD intervals for each parameter</td>
<td>Default</td>
</tr>
<tr>
<td>PosteriorSample</td>
<td>Posterior samples</td>
<td>(ODS output data set only)</td>
</tr>
<tr>
<td>PostSummaries</td>
<td>Posterior summaries</td>
<td>Default</td>
</tr>
<tr>
<td>PriorSummaries</td>
<td>Prior summaries</td>
<td>STATS=PRIOR</td>
</tr>
<tr>
<td>Raftery</td>
<td>Raftery-Lewis diagnostics for each parameter</td>
<td>DIAGNOSTICS=RAFTERY</td>
</tr>
<tr>
<td><strong>ODS Tables Created by the TEST Statement</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TestResults</td>
<td>Test results</td>
<td>Default</td>
</tr>
</tbody>
</table>

**ODS Graphics**

You can use a name to reference every graph that is produced through ODS Graphics. The names of the graphs that PROC HPQLIM generates are listed in Table 21.12.
Table 21.12  Graphs Produced by PROC HPQLIM When a
BAYES Statement Is Included

<table>
<thead>
<tr>
<th>ODS Graph Name</th>
<th>Plot Description</th>
<th>Statement and Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian Diagnostic Plots</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADPanel</td>
<td>Autocorrelation function and density panel</td>
<td>PLOTS=(AUTOCORR DENSITY)</td>
</tr>
<tr>
<td>AutocorrPanel</td>
<td>Autocorrelation function panel</td>
<td>PLOTS=AUTOCORR</td>
</tr>
<tr>
<td>AutocorrPlot</td>
<td>Autocorrelation function plot</td>
<td>PLOTS(UNPACK)=AUTOCORR</td>
</tr>
<tr>
<td>DensityPanel</td>
<td>Density panel</td>
<td>PLOTS=DENSITY</td>
</tr>
<tr>
<td>DensityPlot</td>
<td>Density plot</td>
<td>PLOTS(UNPACK)=DENSITY</td>
</tr>
<tr>
<td>TAPanel</td>
<td>Trace and autocorrelation function panel</td>
<td>PLOTS=(TRACE AUTOCORR)</td>
</tr>
<tr>
<td>TADPanel</td>
<td>Trace, density, and autocorrelation function panel</td>
<td>PLOTS=(TRACE AUTOCORR DENSITY)</td>
</tr>
<tr>
<td>TDPanel</td>
<td>Trace and density panel</td>
<td>PLOTS=(TRACE DENSITY)</td>
</tr>
<tr>
<td>TracePanel</td>
<td>Trace panel</td>
<td>PLOTS=TRACE</td>
</tr>
<tr>
<td>TracePlot</td>
<td>Trace plot</td>
<td>PLOTS(UNPACK)=TRACE</td>
</tr>
</tbody>
</table>

Examples: The HPQLIM Procedure

Example 21.1: High-Performance Model with Censoring

This example shows the use of the HPQLIM procedure with an emphasis on processing a large data set and on the performance improvements that are achieved by executing in the high-performance distributed environment.

The following DATA step generates 5 million replicates from a censored model. The model contains seven variables:

```r
data simulate;
  call streaminit(12345);
  array vars x1-x7;
  array parms{7} (3 4 2 4 -3 -5 -3);
  intercept=2;
  do i=1 to 5000000;
    sum_xb=0;
    do j=1 to 7;
      vars[j]=rand('NORMAL',0,1);
      sum_xb=sum_xb+parms[j]*vars[j];
    end;
    y=intercept+sum_xb+400*rand('NORMAL',0,1);
```plaintext
if y>400 then y=400;
if y<0 then y=0;
output;
end;
keep y x1-x7;
run;
```

The following statements estimate a censored model. The model is executed in the distributed computing environment with two threads and only one node. These settings are used to obtain a hypothetical environment that might resemble running the HPQLIM procedure on a desktop workstation with a dual-core CPU. To run these statements successfully, you need to set the macro variables `GRIDHOST` and `GRIDINSTALLLOC` to resolve to appropriate values, or you can replace the references to the macro variables in the example with the appropriate values.

```plaintext
option set=GRIDHOST="&GRIDHOST";
option set=GRIDINSTALLLOC="&GRIDINSTALLLOC";

proc hpqlim data=simulate;
performance nthreads=2 nodes=1 details
    host="&GRIDHOST" install="&GRIDINSTALLLOC";
    model y=x1-x7 /censored(lb=0 ub=400);
run;
```

Output 21.1.1 shows that the censored model was estimated on the grid, defined in a macro variable named `GRIDHOST`, in a distributed environment on only one node with two threads.

Output 21.1.1 Censored Model with One Node and Two Threads: Performance Table

**Estimating a Tobit model**

<table>
<thead>
<tr>
<th>Performance Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Host Node</td>
</tr>
<tr>
<td>Install Location</td>
</tr>
<tr>
<td>Execution Mode</td>
</tr>
<tr>
<td>Number of Compute Nodes</td>
</tr>
<tr>
<td>Number of Threads per Node</td>
</tr>
</tbody>
</table>

Output 21.1.2 shows the estimation results for the censored model. The “Model Fit Summary” table shows detailed information about the model and indicates that all 5 million observations were used to fit the model. All parameter estimates in the “Parameter Estimates” table are highly significant and correspond to their theoretical values that were set during the data generating process. The optimization of the model with 5 million observations took 45.4 seconds.

Output 21.1.2 Censored Model with One Node and Two Threads: Summary

<table>
<thead>
<tr>
<th>Model Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Source</td>
</tr>
<tr>
<td>Response Variable</td>
</tr>
<tr>
<td>Optimization Technique</td>
</tr>
</tbody>
</table>
Output 21.1.2 continued

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard Error</th>
<th>Type</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
<th>N Obs Lower Bound</th>
<th>N Obs Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>127.0</td>
<td>159.491090</td>
<td>Censored</td>
<td>0</td>
<td>400.0</td>
<td>249E4</td>
<td>8E5</td>
</tr>
</tbody>
</table>

Convergence criterion (FCONV=2.220446E-16) satisfied.

Model Fit Summary

<table>
<thead>
<tr>
<th>Number of Endogenous Variables</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endogenous Variable</td>
<td>y</td>
</tr>
<tr>
<td>Number of Observations</td>
<td>5000000</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-15268972</td>
</tr>
<tr>
<td>Maximum Absolute Gradient</td>
<td>0.0003291</td>
</tr>
<tr>
<td>Number of Iterations</td>
<td>11</td>
</tr>
<tr>
<td>Optimization Method</td>
<td>Quasi-Newton</td>
</tr>
<tr>
<td>AIC</td>
<td>30537962</td>
</tr>
<tr>
<td>Schwarz Criterion</td>
<td>30538083</td>
</tr>
</tbody>
</table>

Parameter Estimates

| Parameter | DF | Estimate | Standard Error | t Value | Approx Pr > |t| |
|-----------|----|----------|----------------|---------|-------------|---|
| Intercept | 1  | 2.220379 | 0.222201       | 9.99    | <.0001      |   |
| x1        | 1  | 3.055533 | 0.201620       | 15.15   | <.0001      |   |
| x2        | 1  | 4.000176 | 0.201570       | 19.85   | <.0001      |   |
| x3        | 1  | 1.852740 | 0.201555       | 9.19    | <.0001      |   |
| x4        | 1  | 4.170266 | 0.201533       | 20.69   | <.0001      |   |
| x5        | 1  | -3.010679| 0.201458       | -14.94  | <.0001      |   |
| x6        | 1  | -5.176016| 0.201541       | -25.68  | <.0001      |   |
| x7        | 1  | -2.695948| 0.201671       | -13.37  | <.0001      |   |
| Sigma     | 1  | 399.997845| 0.261930       | 1527.12 | <.0001      |   |

Procedure Task Timing

<table>
<thead>
<tr>
<th>Task</th>
<th>Seconds</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading and Levelizing Data</td>
<td>1.43</td>
<td>3.06%</td>
</tr>
<tr>
<td>Communication to Client</td>
<td>0.04</td>
<td>0.08%</td>
</tr>
<tr>
<td>Optimization</td>
<td>45.18</td>
<td>96.86%</td>
</tr>
<tr>
<td>Post-optimization</td>
<td>0.00</td>
<td>0.00%</td>
</tr>
</tbody>
</table>
In the following statements, the PERFORMANCE statement is modified to use a grid with 10 nodes, with each node capable of spawning eight threads:

```plaintext
proc hpqlim data=simulate ;
   performance nthreads=8 nodes=10 details
       host="&GRIDHOST" install="&GRIDINSTALLLOC";
   model y=x1-x7 /censored(lb=0 ub=400);
run;
```

The second model which was run on a grid with 10 nodes and eight threads each (Output 21.1.3) took only 1.2 seconds instead of 45.4 seconds to optimize.

**Output 21.1.3**  Censored Model on Ten Nodes with Eight Threads Each: Performance Table

<table>
<thead>
<tr>
<th>Estimating a Tobit model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance Information</td>
</tr>
<tr>
<td>Host Node</td>
</tr>
<tr>
<td>Install Location</td>
</tr>
<tr>
<td>Execution Mode</td>
</tr>
<tr>
<td>Number of Compute Nodes</td>
</tr>
<tr>
<td>Number of Threads per Node</td>
</tr>
</tbody>
</table>

Because the two models being estimated are identical, it is reasonable to expect that Output 21.1.2 and Output 21.1.4 would show the same results except for the performance. However, in certain circumstances, you might observe slight numerical differences in the results (depending on the number of nodes and threads) because the order in which partial results are accumulated, the limits of numerical precision, and the propagation of error in numerical computations can make a difference in the final result.

**Output 21.1.4**  Censored Model on Ten Nodes with Eight Threads Each: Summary

<table>
<thead>
<tr>
<th>Model Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Source</td>
</tr>
<tr>
<td>Response Variable</td>
</tr>
<tr>
<td>Optimization Technique</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Observations Read</td>
</tr>
<tr>
<td>Number of Observations Used</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Summary Statistics of Continuous Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>y</td>
</tr>
</tbody>
</table>

Convergence criterion (FCONV=2.220446E-16) satisfied.
Chapter 21: The HPQLIM Procedure

Output 21.1.4 continued

<table>
<thead>
<tr>
<th>Model Fit Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Endogenous Variables</td>
</tr>
<tr>
<td>Endogenous Variable</td>
</tr>
<tr>
<td>Number of Observations</td>
</tr>
<tr>
<td>Log Likelihood</td>
</tr>
<tr>
<td>Maximum Absolute Gradient</td>
</tr>
<tr>
<td>Number of Iterations</td>
</tr>
<tr>
<td>Optimization Method</td>
</tr>
<tr>
<td>AIC</td>
</tr>
<tr>
<td>Schwarz Criterion</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>x1</td>
</tr>
<tr>
<td>x2</td>
</tr>
<tr>
<td>x3</td>
</tr>
<tr>
<td>x4</td>
</tr>
<tr>
<td>x5</td>
</tr>
<tr>
<td>x6</td>
</tr>
<tr>
<td>x7</td>
</tr>
<tr>
<td>Sigma</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Procedure Task Timing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task</td>
</tr>
<tr>
<td>Reading and Levelizing Data</td>
</tr>
<tr>
<td>Communication to Client</td>
</tr>
<tr>
<td>Optimization</td>
</tr>
<tr>
<td>Post-optimization</td>
</tr>
</tbody>
</table>

As this example suggests, increasing the number of nodes and the number of threads per node improves performance significantly. When you use the parallelism that a high-performance distributed environment affords, you can see an even more dramatic reduction in the time required for the optimization as the number of observations in the data set increases. When the data set is extremely large, the computations might not even be possible with the typical memory resources and computational constraints of a desktop computer. Under such circumstances the high-performance distributed environment becomes a necessity.

Example 21.2: Bayesian High-Performance Model with Censoring

This example shows the use of the Bayesian analysis available in the HPQLIM procedure with an emphasis on processing a large data set and on the performance improvements that are achieved by executing in a high-performance distributed environment.
The model and the data set are the same as in Example 21.1, and the priors are set to the defaults.

The model is executed in the distributed computing environment with 10 nodes, where each node spawns eight threads. To run the following statements successfully, you need to set the macro variables GRIDHOST and GRIDINSTALLLOC to resolve to appropriate values, or you can replace the references to the macro variables in the example with the appropriate values:

```plaintext
option set=GRIDHOST="&GRIDHOST";
option set=GRIDINSTALLLOC="&GRIDINSTALLLOC";

proc hpqlim data=simulate ;
bayes nbi=10000 nmc=30000;
    performance nthreads=8 nodes=10 details
        host="&GRIDHOST" install="&GRIDINSTALLLOC";
    model y=x1-x7 /censored(lb=0 ub=400);
    %*; ods output PerformanceInfo=perfInfo;
    %*; ods output Timing=time;
run;
```

Output 21.2.1 shows a summary of the posterior distribution that is associated with the censored model when you use diffuse prior distributions.

### Output 21.2.1 Posterior Summary for Bayesian Censored Model

#### Estimating a Tobit model

The HPQLIM Procedure

<table>
<thead>
<tr>
<th>Parameter</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>30000</td>
<td>2.2316</td>
<td>0.2219</td>
<td>2.0820</td>
<td>2.2342</td>
<td>2.3819</td>
</tr>
<tr>
<td>x1</td>
<td>30000</td>
<td>3.0580</td>
<td>0.2051</td>
<td>2.9209</td>
<td>3.0649</td>
<td>3.1964</td>
</tr>
<tr>
<td>x2</td>
<td>30000</td>
<td>4.0005</td>
<td>0.2049</td>
<td>3.8613</td>
<td>3.9979</td>
<td>4.1387</td>
</tr>
<tr>
<td>x3</td>
<td>30000</td>
<td>1.8445</td>
<td>0.2026</td>
<td>1.7150</td>
<td>1.8429</td>
<td>1.9782</td>
</tr>
<tr>
<td>x4</td>
<td>30000</td>
<td>4.1724</td>
<td>0.1980</td>
<td>4.0385</td>
<td>4.1717</td>
<td>4.3045</td>
</tr>
<tr>
<td>x5</td>
<td>30000</td>
<td>-3.0041</td>
<td>0.2034</td>
<td>-3.1420</td>
<td>-2.9997</td>
<td>-2.8658</td>
</tr>
<tr>
<td>x6</td>
<td>30000</td>
<td>-5.1667</td>
<td>0.2030</td>
<td>-5.3060</td>
<td>-5.1655</td>
<td>-5.0319</td>
</tr>
<tr>
<td>x7</td>
<td>30000</td>
<td>-2.6943</td>
<td>0.1997</td>
<td>-2.8248</td>
<td>-2.6970</td>
<td>-2.5630</td>
</tr>
<tr>
<td>_Sigma</td>
<td>30000</td>
<td>400.00</td>
<td>0.2573</td>
<td>399.8</td>
<td>400.0</td>
<td>400.2</td>
</tr>
</tbody>
</table>

Output 21.2.2 show a summary of the performance when you use a distributed computing environment with 10 nodes, where each node spawns eight threads.
Chapter 21: The HPQLIM Procedure

Output 21.2.2  Performance Analysis for Bayesian Censored Model on Ten Nodes with Eight Threads Each

Estimating a Tobit model

Performance Information

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Host Node</td>
<td>&lt;&lt; your grid host &gt;&gt;</td>
</tr>
<tr>
<td>Install Location</td>
<td>&lt;&lt; your grid install location &gt;&gt;</td>
</tr>
<tr>
<td>Execution Mode</td>
<td>Distributed</td>
</tr>
<tr>
<td>Number of Compute Nodes</td>
<td>10</td>
</tr>
<tr>
<td>Number of Threads per Node</td>
<td>8</td>
</tr>
</tbody>
</table>

Estimating a Tobit model

<table>
<thead>
<tr>
<th>Task</th>
<th>Seconds</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading and Levelizing Data</td>
<td>0.22</td>
<td>0.02%</td>
</tr>
<tr>
<td>Communication to Client</td>
<td>0.13</td>
<td>0.01%</td>
</tr>
<tr>
<td>Bayesian Analysis: Likelihood for MCMC</td>
<td>1328.41</td>
<td>99.71%</td>
</tr>
<tr>
<td>Bayesian Analysis: MCMC</td>
<td>0.63</td>
<td>0.05%</td>
</tr>
<tr>
<td>Optimization</td>
<td>2.92</td>
<td>0.22%</td>
</tr>
<tr>
<td>Post-optimization</td>
<td>0.00</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

Finally, Output 21.2.3 shows the diagnostic and summary plots that are associated with X1.

Output 21.2.3  Bayesian Diagnostic and Summary Plots for x1
The implementation took only 14.7 minutes to sample from the posterior distribution. The same implementation in a hypothetical environment resembling a desktop workstation with a dual-core CPU would have taken approximately 12 hours.

References


Subject Index

bounds on parameter estimates, 1146
BOUNDS statement, 1146
censored regression models
   HPQLIM procedure, 1158
covariates
   heteroscedasticity models, 1149
gamma distribution
   definition of (HPQLIM), 1164
   HPQLIM procedure, 1164
Gaussian distribution
   definition of (HPQLIM), 1165
   HPQLIM procedure, 1165
heteroscedasticity models
   covariates, 1149
   HPQLIM procedure, 1132
   BY groups, 1146
   censored regression models, 1158
   frontier, 1159
   gamma distribution, 1164
   Gaussian distribution, 1165
   heteroscedasticity, 1160
   inverse gamma distribution, 1164
   limited dependent variable models, 1158
   logit, 1132
   multithreading, 1154
   normal distribution, 1165
   ordinal discrete choice modeling, 1157
   output, 1166
   output ODS Graphics table names, 1171
   output table names, 1171
   probit, 1132
   selection, 1132
   standard distributions, 1163
   syntax, 1135
   t distribution, 1165
   tests on parameters, 1161
   Tobit, 1132
   truncated regression models, 1158
   uniform distribution, 1165
inverse gamma distribution
   definition of (HPQLIM), 1164
   HPQLIM procedure, 1164
Lagrange multiplier test
   nonlinear hypotheses, 1161
limited dependent variable models
   HPQLIM procedure, 1158
   logit
      HPQLIM procedure, 1132
   multithreading
      HPQLIM procedure, 1154
   Newton-Raphson
      optimization methods, 1141
   Newton-Raphson method, 1141
   nonlinear hypotheses
      Lagrange multiplier test, 1161
   normal distribution
      definition of (HPQLIM), 1165
      HPQLIM procedure, 1165
   optimization methods
      Newton-Raphson, 1141
      trust region, 1141
   ordinal discrete choice modeling
      HPQLIM procedure, 1157
   output ODS Graphics table names
      HPQLIM procedure, 1171
   output table names
      HPQLIM procedure, 1171
   prior distribution
      distribution specification (HPQLIM), 1154
   probit
      HPQLIM procedure, 1132
   quasi-Newton method, 1141
   selection
      HPQLIM procedure, 1132
   standard distributions
      HPQLIM procedure, 1163
   t distribution
      definition of (HPQLIM), 1165
      HPQLIM procedure, 1165
   Tobit
      HPQLIM procedure, 1132
   truncated regression models
      HPQLIM procedure, 1158
   trust region
      optimization methods, 1141
   trust region method, 1141
uniform distribution
  definition of (HPQLIM), 1165
HPQLIM procedure, 1165
Syntax Index

ALL option
  TEST statement (HPQLIM), 1156

BAYES statement
  HPQLIM procedure, 1142

BETA
  PRIOR statement (HPQLIM), 1154

BOUNDS statement
  HPQLIM procedure, 1146

BY statement
  HPQLIM procedure, 1146

CENSORED option
  ENDOGENOUS statement (HPQLIM), 1147, 1151

CONDITIONAL
  OUTPUT statement (HPQLIM), 1153

COPYVAR= option
  OUTPUT statement (HPQLIM), 1153

CORRB option
  HPQLIM procedure, 1139

CORROUT option
  PROC HPQLIM statement, 1138

COST option
  ENDOGENOUS statement (HPQLIM), 1148, 1151

COVB option
  HPQLIM procedure, 1139

COVEST= option
  HPQLIM procedure, 1139

COVOUT option
  PROC HPQLIM statement, 1138

DATA= option
  PROC HPQLIM statement, 1138

DETAILS option
  PERFORMANCE statement (HPQLIM), 1154

DIAGNOSTICS= option
  BAYES statement (HPQLIM), 1142

DISCRETE option
  ENDOGENOUS statement (HPQLIM), 1147, 1151

DISTRIBUTION= option
  ENDOGENOUS statement (HPQLIM), 1147, 1151

ERRSTD
  OUTPUT statement (HPQLIM), 1153

EXPECTED
  OUTPUT statement (HPQLIM), 1153

FRONTIER option
  ENDOGENOUS statement (HPQLIM), 1148, 1152

GAMMA
  PRIOR statement (HPQLIM), 1154
  HPQLIM procedure, 1135
  PERFORMANCE statement, 1154
  PRIOR statement, 1154
  syntax, 1135
  HPQLIM procedure, FREQ statement, 1149
  HPQLIM procedure, PERFORMANCE statement, 1154
  HPQLIM procedure, TEST statement, 1155
  HPQLIM procedure, WEIGHT statement, 1156

IGAMMA
  PRIOR statement (HPQLIM), 1154

INIT statement
  HPQLIM procedure, 1154

LM option
  TEST statement (HPQLIM), 1156

LOWERBOUND= option
  ENDOGENOUS statement (HPQLIM), 1148, 1151, 1152

LR option
  TEST statement (HPQLIM), 1156

MARGINAL
  OUTPUT statement (HPQLIM), 1153

MAXTUNE= option
  BAYES statement (HPQLIM), 1144

METHOD= option
  PROC HPQLIM statement, 1141

MILLS
  OUTPUT statement (HPQLIM), 1153

MINTUNE= option
  BAYES statement (HPQLIM), 1143

MODEL statement
  HPQLIM procedure, 1150

NBI= option
  BAYES statement (HPQLIM), 1144

NMC= option
  BAYES statement (HPQLIM), 1144

NODES option