Bayesian Analysis in SAS/STAT® 12.1 Software

Growing Bayesian Capabilities

Overview

SAS/STAT software now provides Bayesian analysis in two ways: procedures with built-in Bayesian capabilities and a procedure for general Bayesian modeling.

Why Bayesian?

Bayesian methods have become increasingly popular in recent years. Bayesian analysis provides:

- a natural and principled way of combining prior information with data within a solid decision-theoretical framework
- inferences that are conditional on the data and are exact, without relying on asymptotic approximation
- straightforward interpretation because inference is based on the distribution of a parameter
- a convenient framework for answering specific scientific questions
- convenient formulation of hierarchical models and missing data problems

Built-In Bayesian Capabilities

You have easy access to Bayesian analysis in several standard SAS/STAT procedures. You can perform Bayesian analysis for:

- generalized linear models, including logistic regression and Poisson regression models
- Cox regression models
- frailty models
- accelerated life failure models
- piecewise exponential models
- finite mixture models

These capabilities are available via the BAYES statement in the GENMOD, LIFEREG, PHREG, and FMM procedures of SAS/STAT software. These procedures provide:

- posterior point and interval statistics
- several standard prior distributions, such as Jeffreys’ prior in the GENMOD procedure and the Zellner g-prior for the regression coefficients in the PHREG procedure
- a SAS data set that includes posterior samples for use in additional analyses

Various algorithms for obtaining posterior samples are available:

- The Gamerman algorithm is now the default sampling mechanism in the GENMOD procedure except when the linear models have conjugacy, in which case the GENMOD procedure takes advantage of the conjugacy. The adaptive rejection Metropolis sampling (ARMS) algorithm and the independent Metropolis algorithm are also provided.
- The ARMS algorithm is the default sampling algorithm for the PHREG and LIFEREG procedures. In addition, the random walk Metropolis algorithm is available in PROC PHREG.
- The FMM procedure offers the conjugate sampler and the Gamerman algorithm.

Assessing convergence of the Markov chain is an important aspect of Bayesian analysis, and these procedures provide diagnostics such as the Gelman-Rubin, Geweke, Heidelberger-Welch, and Raftery-Lewis tests. Diagnostic plots are also produced.

The following statements illustrate a basic invocation for a Bayesian regression analysis.

```
proc genmod data=Surg;
    model y = Logx1 x2 x3 x4 / dist=normal;
bayes;
run;
```

A trace plot, an autocorrelation plot, and a posterior density plot are provided for each parameter that is estimated.
General Bayesian Modeling Tool

The MCMC procedure is a general-purpose simulation procedure that uses Markov chain Monte Carlo (MCMC) techniques to fit Bayesian models. You can use the MCMC procedure to analyze data that have any likelihood, prior, or hyperprior, as long as you can program them with SAS DATA step functions. It then produces summary and diagnostic statistics, and saves the posterior samples to an output data set for further analysis.

By default, PROC MCMC takes advantage of certain forms of conjugacy in the model in order to sample directly from the target conditional distributions. PROC MCMC uses an adaptive blocked random walk Metropolis algorithm that uses a normal proposal distribution. A slice sampler is also available.

You specify prior distributions for the parameters with the PRIOR statement, specify the likelihood function with the MODEL statement, and specify the random effects with the RANDOM statement.

You can fit models such as random-effects models, change point models, Weibull models, and normal regression models with interval censoring. The parameters can enter the model linearly or in any nonlinear functional form.

The RANDOM statement simplifies the construction of hierarchical random-effects models by enabling you to define random effects that can enter the model in a linear or nonlinear fashion. Some features of the RANDOM statement are:

- It significantly reduces simulation time and improves model convergence, especially in models with a large number of subjects or clusters.
- It supports both standard and general (nonstandard) prior distributions, which can be either univariate or multivariate.
- It supports arbitrary hierarchy, meaning that a random effect can be a hyperparameter to another random effect.

PROC MCMC provides support for standard distributions including the Dirichlet, inverse Wishart, multivariate normal, and multinomial distributions. In addition, the multivariate normal distribution with autocorrelation covariance structure is available for the PRIOR, RANDOM, and MODEL statements.

PROC MCMC now models missing data for the response variable. In addition, you can also model missing covariates.

Example of Using PROC MCMC

This example considers $y$ deaths of $n$ beetles that have been exposed to a contaminant. Logistic regression is a suitable technique to assess the contaminant’s effect on beetle deaths. The following are the data:

```
data beetles;
  input n y x @@;
datalines;
  6 0 25.7 8 2 35.9
  5 2 32.9 7 7 50.4
  6 0 28.3 7 2 32.3
  5 1 33.2 8 3 40.9
  6 0 36.5 6 1 36.5
  6 6 49.6 6 3 39.8
  6 4 43.6 6 1 34.1
  7 1 37.4 8 2 35.2
  6 6 51.3 5 3 42.5
  7 0 31.3 3 2 40.6
;
```

The following PROC MCMC statements fit a logistic regression model to these data. Diffuse, normal priors are specified for both the intercept and slope parameters.

```
proc mcmc data=beetles nmc=20000
   outpost=beetleout seed=246810;
   parms alpha beta;
   prior alpha beta ~normal(0, var=10000);
   p = logistic(alpha + beta*x);
   model y ~ binomial(n,p);
run;
```

The NMC option requests an MCMC sample of 20,000. In the PRIOR statement, diffuse normal priors are specified for $\alpha$ and $\beta$. The probability $p$ is expressed as a logistic transformation, and the MODEL statement specifies the likelihood. The following table displays summary measures for the posterior samples.

<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>alpha</td>
<td>10000</td>
<td>-11.7202</td>
<td>2.0529</td>
<td>-13.0094</td>
<td>-11.6026</td>
<td>-10.3474</td>
</tr>
<tr>
<td>beta</td>
<td>10000</td>
<td>0.2907</td>
<td>0.0529</td>
<td>0.2550</td>
<td>0.2876</td>
<td>0.3247</td>
</tr>
</tbody>
</table>

For More Information

SAS/STAT 12.1 is scheduled for release during the third quarter of 2012. See the details at support.sas.com/statistics