

Preliminary Capabilities for Bayesian Analysis in SAS/STAT® Software

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2006. *Preliminary Capabilities for Bayesian Analysis in SAS/STAT® Software*. Cary, NC: SAS Institute Inc.

Preliminary Capabilities for Bayesian Analysis in SAS/STAT® Software

Copyright © 2006, SAS Institute Inc., Cary, NC, USA

All rights reserved. Produced in the United States of America.

For a hard-copy book: No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

For a Web download or e-book: Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

U.S. Government Restricted Rights Notice: Use, duplication, or disclosure of this software and related documentation by the U.S. government is subject to the Agreement with SAS Institute and the restrictions set forth in FAR 52.227-19, Commercial Computer Software-Restricted Rights (June 1987).

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, January 2007

SAS Publishing provides a complete selection of books and electronic products to help customers use SAS software to its fullest potential. For more information about our e-books, e-learning products, CDs, and hard-copy books, visit the SAS Publishing Web site at support.sas.com/pubs or call 1-800-727-3228.

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

Contents

Chapter 1. Introduction to Bayesian Analysis Procedures	1
Chapter 2. The BGENMOD Procedure	41
Chapter 3. The BLIFEREG Procedure	81
Chapter 4. The BPHREG Procedure	113
Subject Index	161
Syntax Index	163

Chapter 1

Introduction to Bayesian Analysis Procedures

Chapter Contents

OVERVIEW	3
INTRODUCTION	3
DETAILS	5
Prior Distributions	5
Bayesian Inference	9
Bayesian Analysis: Advantages and Disadvantages	11
Markov Chain Monte Carlo Method	12
Assessing Markov Chain Convergence	17
Summary Statistics	31
A BAYESIAN READING LIST	34
Textbooks	34
Tutorial and Review Papers on MCMC	35
ACKNOWLEDGMENTS	36
REFERENCES	36

Chapter 1

Introduction to Bayesian Analysis Procedures

Overview

SAS/STAT software now provides Bayesian analysis in downloadable, experimental versions of three procedures: GENMOD, LIFEREG, and PHREG. These procedures provide Bayesian modeling and inference capability in generalized linear models, accelerated life failure models, Cox regression models, and piecewise constant baseline hazard models (also known as piecewise exponential models). These procedures are available for the Windows 32-bit platform via a Web download and work with SAS 9.1.3. These versions are named BGENMOD, BLIFEREG, and BPHREG, respectively, and they otherwise contain the full functionality of the original procedures. (Note that the BPHREG procedure also contains the experimental CLASS statement documented for the TPHREG procedure.) The Bayesian capabilities will be included in the GENMOD, LIFEREG, and PHREG procedures in the next release of SAS/STAT software.

This chapter provides an overview of Bayesian statistics, describes specific sampling algorithms used in these three procedures, and discusses posterior inference and convergence diagnostics computations. Sources that provide in-depth treatment of Bayesian statistics can be found at the end of this chapter, in the section “[A Bayesian Reading List](#)” on page 34.

Additional chapters contain syntax, details, and examples for the individual procedures BGENMOD, BLIFEREG, and BPHREG. These chapters do not repeat information that is included in the SAS/STAT documentation for SAS 9.1.3. Only the syntax and details specific to the Bayesian capabilities are described.

Introduction

The most frequently used statistical methods are known as *frequentist* (or *classical*) methods. These methods assume that unknown parameters are fixed constants, and they define probability by using limiting relative frequencies. It follows from these assumptions that probabilities are objective and that you cannot make probability statements about parameters because they are fixed. Bayesian methods offer an alternative approach; they treat parameters as random variables and define probability as “degrees of belief”—that is, the probability of an event is the degree to which you believe the event is true. It follows from these postulates that probabilities are subjective and that you can make probability statements about parameters. The term “Bayesian” comes from the prevalent usage of Bayes’ theorem in this area.

Suppose you are interested in estimating θ from data $\mathbf{y} = \{y_1, \dots, y_n\}$ by using a statistical model described by a density $p(\mathbf{y}|\theta)$. Bayesian philosophy states that θ can-

not be determined exactly, and uncertainty about the parameter is expressed through probability statements and distributions. You can say that θ follows a normal distribution with mean 0 and variance 1, if it is believed that this distribution best describes the uncertainty associated with the parameter.

The following steps describe the essential elements of Bayesian inference:

1. A probability distribution for θ is formulated as $\pi(\theta)$, which is known as the *prior* distribution, or just the prior. The prior distribution expresses your beliefs, for example, on the mean, the spread, the skewness, and so forth, about the parameter before you examine the data.
2. Given the observed data \mathbf{y} , you choose a statistical model $p(\mathbf{y}|\theta)$ to describe the distribution of \mathbf{y} given θ .
3. You update your beliefs about θ by combining information from the prior distribution and the data through the calculation of the *posterior* distribution, $p(\theta|\mathbf{y})$.

The third step is carried out by using Bayes' theorem, which enables you to combine the prior distribution and the model in the following way:

$$p(\theta|\mathbf{y}) = \frac{p(\theta, \mathbf{y})}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\theta)\pi(\theta)}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\theta)\pi(\theta)}{\int p(\mathbf{y}|\theta)\pi(\theta)d\theta}$$

The quantity

$$p(\mathbf{y}) = \int p(\mathbf{y}|\theta)\pi(\theta)d\theta$$

is the normalizing constant of the posterior distribution. This quantity $p(\mathbf{y})$ is also the marginal distribution of \mathbf{y} , and it is sometimes called the marginal distribution of the data.

The likelihood function of θ is any function proportional to $p(\mathbf{y}|\theta)$ —that is, $L(\theta) \propto p(\mathbf{y}|\theta)$. Another way of writing Bayes' theorem is

$$p(\theta|\mathbf{y}) = \frac{L(\theta)\pi(\theta)}{\int L(\theta)\pi(\theta)d\theta}$$

The marginal distribution $p(\mathbf{y})$ is an integral; therefore, as long as it is finite, the particular value of the integral does not provide any additional information about the posterior distribution. Hence, $p(\theta|\mathbf{y})$ can be written up to an arbitrary constant, presented here in proportional form as

$$p(\theta|\mathbf{y}) \propto L(\theta)\pi(\theta)$$

Simply put, Bayes' theorem tells you how to update existing knowledge with new information. You begin with a prior belief $\pi(\theta)$, and, after learning information from

data \mathbf{y} , you change or update the belief on θ and obtain $p(\theta|\mathbf{y})$. These are the essential elements of the Bayesian approach to data analysis.

In theory, Bayesian methods offer a very simple alternative to statistical inference—all inferences follow from the posterior distribution $p(\theta|\mathbf{y})$. However, in practice, only the most rudimentary problems enable you to obtain the posterior distribution with straightforward analytical solutions. Most Bayesian analysis require sophisticated computations, including the use of simulation methods. You generate samples from the posterior distribution and use these samples to estimate the quantities of interest. The BGENMOD, BLIFEREG, and BPHREG procedures use the Gibbs sampler, one of the many algorithms that MCMC methods comprise. However, using these methods means that an important aspect of any analysis is assessing the convergence of the Markov chains. Inferences based on nonconverged Markov chains can be both inaccurate and misleading.

For more information, see the section “[Metropolis and Metropolis-Hastings Algorithms](#)” on page 13 and the section “[Gibbs Sampler](#)” on page 15.

Both Bayesian and classical analysis methods have their advantages and disadvantages. From a practical point of view, your choice of method depends on what you want to accomplish with your data analysis. If you have prior information, either expert opinion or historical knowledge, that you want to incorporate into the analysis, then you might consider Bayesian methods. In addition, if you want to communicate your findings in terms of probability notions that can be more easily understood by nonstatisticians, Bayesian methods might be appropriate. The Bayesian paradigm can often provide a framework for answering specific scientific questions that a single point estimate cannot sufficiently address. On the other hand, if you are interested only in estimating parameters based on the likelihood, then numerical optimization methods, such as the Newton-Raphson method, can give you very precise estimates and there is no need to use Bayesian analysis.

For further discussions of the relative advantages and disadvantages of Bayesian analysis, see the section “[Bayesian Analysis: Advantages and Disadvantages](#)” on page 11.

Details

Prior Distributions

A prior distribution of a parameter is the probability distribution that represents your uncertainty of the parameter before the current data are examined. Multiplying the prior distribution and the likelihood function together leads to the posterior distribution of the parameter. You use the posterior distribution to carry out all inferences. You cannot carry out any Bayesian inference or perform any modeling without using a prior distribution.

Objective Priors versus Subjective Priors

Bayesian probability measures the degree of belief that you have in a random event. By this definition, probability is highly subjective. It follows that all priors are *subjective priors*.

Not everyone agrees with this notion of subjectivity when it comes to specifying prior distributions. There has long been a desire to obtain results that are objectively valid. Within the Bayesian paradigm, this can be somewhat achieved by using prior distributions that are “objective”—that is, have a minimal impact on the posterior distribution. Such distributions are called *objective* or *noninformative* priors (see the next section). However, while noninformative priors are very popular in some applications, they are not always easy to construct.

See DeGroot and Schervish (2002, Section 1.2) and Press (2003, Section 2.2) for more information about interpretations of probability. See Berger (2006) and Goldstein (2006) for discussions about objective Bayesian versus subjective Bayesian analysis.

Noninformative Priors

Roughly speaking, a prior distribution is noninformative if the prior is “flat” relative to the likelihood function. Thus, a prior $\pi(\theta)$ is noninformative if it has minimal impact on the posterior distribution of θ . Other names for the noninformative prior are *vague* and *flat* prior. Many statisticians favor noninformative priors because they appear to be more objective. However, it is unrealistic to expect that noninformative priors represent total ignorance about the parameter of interest.

In some cases, noninformative priors can lead to *improper posteriors* (nonintegrable posterior density). You cannot make inferences with improper posterior distributions. In addition, noninformative priors are often not invariant under transformation; that is, a prior might be noninformative in one parameterization but not necessarily noninformative if a transformation is applied.

A common choice for a noninformative prior is the flat prior, which is a prior distribution that assigns equal likelihood on all possible values of the parameter. Intuitively this makes sense, and in some cases, such as linear regression, flat priors on the regression parameter are noninformative. However, this is not necessarily true in all cases.

For example, suppose there is a binomial experiment with n Bernoulli trials where y 1s are observed. You want to make inferences about the unknown success probability p . A uniform prior on p ,

$$\pi(p) \propto 1$$

might appear to be noninformative. But in fact, using the uniform prior is equivalent to adding two observations to the data, one 1 and one 0. With small n and y , the added observations can be very influential to the parameter estimate of p .

To see this, note that the likelihood is

$$p^y(1-p)^{n-y}$$

The MLE of p is y/n . The uniform prior can be written as a beta distribution with both the shape (α) and scale (β) parameters being 1:

$$\pi(p) \propto p^{\alpha-1}(1-p)^{\beta-1}$$

The posterior distribution of p is proportional to

$$p^{\alpha+y-1}(1-p)^{\beta+n-y-1}$$

which is $\text{beta}(\alpha + y, \beta + n - y)$.

The posterior mean is therefore

$$\frac{\alpha + y}{\alpha + \beta + n} = \frac{1 + y}{2 + n}$$

and it can be quite different from the MLE if both n and y are small.

See [Box and Tiao \(1973\)](#) for a more formal development of noninformative priors. See [Kass and Wasserman \(1996\)](#) for techniques for deriving noninformative priors.

Improper Priors

A prior $\pi(\theta)$ is said to be improper if

$$\int \pi(\theta) d\theta = \infty$$

For example, a uniform prior distribution on the real line, $\pi(\theta) \propto 1$ for $-\infty < \theta < \infty$, is an improper prior. Improper priors are often used in Bayesian inference since they usually yield noninformative priors and proper posterior distributions.

Improper prior distributions can lead to posterior impropriety (improper posterior distribution). To determine whether a posterior distribution is proper, you need to make sure that the normalizing constant $\int p(\mathbf{y}|\theta)p(\theta)d\theta$ is finite for all \mathbf{y} . If an improper prior distribution leads to an improper posterior distribution, inference based on the improper posterior distribution is invalid.

The BGENMOD, BLIFEREG, and BPHREG procedures allow the use of improper priors—that is, the flat prior on the real line—for regression coefficients. These improper priors do not lead to any improper posterior distributions in the models that these procedures fit.

Informative Priors

An informative prior is a prior that is not dominated by the likelihood and that has an impact on the posterior distribution. If a prior distribution dominates the likelihood, it is clearly an informative prior. These types of distributions must be specified with care in actual practice. On the other hand, the proper use of prior distributions illustrates the power of the Bayesian method; information gathered from the previous study, past experience or expert opinion can be combined with current information in a natural way. See the “Examples” sections of the BGENMOD and BPHREG procedure chapters for instructions about constructing informative prior distributions.

Conjugate Priors

A prior is said to be a conjugate prior for a family of distributions if the prior and posterior distributions are from the same family, meaning that the form of the posterior has the same distributional form as the prior distribution. For example, if the likelihood is binomial, $y \sim \text{Bin}(n, \theta)$, a conjugate prior on θ is the beta distribution; it follows that the posterior distribution of θ is also a beta distribution. Other commonly used conjugate prior/likelihood combinations include the normal/normal, gamma/Poisson, gamma/gamma, and gamma/beta cases. The development of conjugate priors was partially driven by a desire for computational convenience—conjugacy provides a practical way to obtain the posterior distributions. The Bayesian procedures do not use conjugacy in posterior sampling.

Jeffreys' Prior

A very useful prior is Jeffreys' prior. It satisfies the local uniformity property—a prior that does not change much over the region in which the likelihood is significant and does not assume large values outside that range. It is based on the Fisher information matrix.

Jeffreys' prior is defined as

$$\pi(\theta) \propto |I(\theta)|^{1/2}$$

where $|\cdot|$ denotes the determinant and $I(\theta)$ is the Fisher information matrix based on the likelihood function $p(\mathbf{y}|\theta)$:

$$I(\theta) = -E \left[\frac{\partial^2 \log p(\mathbf{y}|\theta)}{\partial \theta^2} \right]$$

Jeffreys' prior is locally uniform and hence noninformative. It provides an automated scheme for finding a noninformative prior for any parametric model $p(\mathbf{y}|\theta)$. It is important to recognize that Jeffreys' prior is not in violation of Bayesian philosophy—it is the form of the likelihood function that determines the prior but not the observed data, since the Fisher information is an expectation over all \mathbf{y} and not just the observed \mathbf{y} .

Another appealing property of Jeffreys' prior is that it is invariant with respect to one-to-one transformations. The invariance property means that, if you have a locally

uniform prior on θ and $\phi(\theta)$ is a one-to-one function of θ , then $p(\phi(\theta)) = \pi(\theta) \cdot |\phi'(\theta)|^{-1}$ is a locally uniform prior for $\phi(\theta)$. This invariance principle carries through to multidimensional parameters as well.

While Jeffreys' prior provides a general recipe for obtaining noninformative priors, it has some shortcomings: the prior is improper for many models and it can lead to improper posterior in some cases, and the prior can be cumbersome to use in high dimensions.

The BGENMOD procedure calculates Jeffreys' prior automatically for any generalized linear model. You can set it as your prior density for the coefficient parameters and it does not lead to improper posteriors.

Bayesian Inference

Bayesian inference about θ is primarily based on the posterior distribution of θ . There are various ways in which you can summarize this distribution. For example, you can report your findings through point estimates. You can also use the posterior distribution to construct hypothesis tests or probability statements.

Point Estimation and Estimation Error

In contrast with reporting the maximum likelihood estimator (MLE) or the method of moments estimator (MOME) of a parameter in classical analysis, Bayesian approaches often use the posterior mean. The definition of the posterior mean is given by

$$E(\theta|\mathbf{y}) = \int \theta p(\theta|\mathbf{y}) d\theta$$

Other commonly used posterior estimators include the posterior median, defined as

$$\theta: P(\theta \geq \text{median}|\mathbf{y}) = P(\text{median} \leq \theta|\mathbf{y}) = \frac{1}{2}$$

and the posterior mode, defined as the value of θ that maximizes $p(\theta|\mathbf{y})$.

The variance of the posterior density (simply referred to as the *posterior variance*) describes the uncertainty in the parameter, which is a random variable in the Bayesian paradigm. A Bayesian analysis typically uses the posterior variance, or the posterior standard deviation, to characterize the dispersion of the parameter. In multidimensional models, covariance or correlation matrices are used.

If you know the distributional form of the posterior density of interest, you can report the exact posterior point estimates. When models become too difficult to analyze analytically, you have to resort to using simulation algorithms, such as the Markov chain Monte Carlo (MCMC) method (see the section “[Markov Chain Monte Carlo Method](#)” on page 12) to obtain posterior estimates. The BGENMOD, BLIFEREG, and BPHREG procedures rely on MCMC to obtain all posterior estimates. Using only a finite number of samples, simulations introduce an additional level of uncertainty

to the accuracy of the estimates. *Monte Carlo standard error (MCSE)*, which is the standard error of the posterior mean estimate, measures the simulation accuracy. See the section “[Standard Error of the Mean Estimate](#)” on page 32 for more information.

The posterior standard deviation and the MCSE are two completely different concepts: the posterior standard deviation describes the uncertainty in the parameter, while the MCSE describes only the uncertainty in the parameter estimate as a result of MCMC simulation. The posterior standard deviation is a function of the sample size in the data set, and the MCSE is a function of the number of iterations in the simulation.

Hypothesis Testing

Suppose you have the following null and alternative hypotheses: H_0 is $\theta \in \Theta_0$ and H_1 is $\theta \in \Theta_0^c$, where Θ_0 is a subset of the parameter space and Θ_0^c is its complement. Using the posterior distribution $\pi(\theta|\mathbf{y})$, you can compute the posterior probabilities $P(\theta \in \Theta_0|\mathbf{y})$ and $P(\theta \in \Theta_0^c|\mathbf{y})$, or the probabilities that H_0 and H_1 are true, respectively. One way to perform a Bayesian hypothesis test is to accept the null hypothesis if $P(\theta \in \Theta_0|\mathbf{y}) \geq P(\theta \in \Theta_0^c|\mathbf{y})$ and vice versa, or to accept the null hypothesis if $P(\theta \in \Theta_0|\mathbf{y})$ is greater than a predefined threshold, such as 0.75, to guard against falsely accepted null distribution.

It is more difficult to carry out a point null hypothesis test in a Bayesian analysis. A point null hypothesis is a test of $H_0: \theta = \theta_0$ versus $H_1: \theta \neq \theta_0$. If the prior distribution $\pi(\theta)$ is a continuous density, then the posterior probability of the null hypothesis being true is 0, and there is no point in carrying out the test. One alternative is to restate the null to be a small interval hypothesis: $\theta \in \Theta_0 = (\theta_0 - a, \theta_0 + a)$, where a is a very small constant. The Bayesian paradigm can deal with an interval hypothesis more easily. Another approach is to give a mixture prior distribution to θ with a positive probability of p_0 on θ_0 and the density $(1 - p_0)\pi(\theta)$ on $\theta \neq \theta_0$. This prior ensures a nonzero posterior probability on θ_0 , and you can then make realistic probabilistic comparisons. For more detailed treatment of Bayesian hypothesis testing, see [Berger \(1985\)](#).

Interval Estimation

The analogous Bayesian concept to the confidence set is the *credible interval*, also known as the *credible set*. Given a posterior distribution $p(\theta|\mathbf{y})$, A is a credible set for θ if

$$P(\theta \in A|\mathbf{y}) = \int_A p(\theta|\mathbf{y})d\theta$$

For example, you can construct a 95% credible set for θ by finding an interval, A , over which $\int_A p(\theta|\mathbf{y}) = 0.95$.

You can construct credible sets that have equal tails: a $100(1 - \alpha)\%$ equal-tail interval corresponds to the $100(\alpha/2)$ th and $100(1 - \alpha/2)$ th percentiles of the posterior distribution. Some statisticians prefer this interval because it is invariant to transformation. Another frequently used Bayesian credible set is called the *highest posterior density* (HPD) interval.

A $100(1 - \alpha)\%$ HPD interval is a region that satisfies the following two conditions:

1. The posterior probability of that region is $100(1 - \alpha)\%$.
2. The minimum density of any point within that region is equal to or larger than the density of any point outside that region.

The HPD is an interval in which most of the distribution lies. Some statisticians prefer this interval because it is the smallest interval.

One major distinction between Bayesian and classical sets is their interpretation. The Bayesian probability reflects a person's subjective beliefs. Following this approach, a statistician can make the claim that θ is inside a credible interval with measurable probability. This property is appealing because it enables you to make a direct probability statement about parameters. Many people find this concept a more natural way of understanding a probability interval, which is also easier to explain to non-statisticians. A confidence interval, on the other hand, enables you to make a claim that the interval covers the true parameter. The interpretation reflects the uncertainty in the sampling procedure—a confidence interval of $100(1 - \alpha)\%$ asserts that, in the long run, $100(1 - \alpha)\%$ of the realized confidence intervals cover the true parameter.

Bayesian Analysis: Advantages and Disadvantages

Generally speaking, when the sample size is large, Bayesian inference often provides results for parametric models that are very similar to the results produced by frequentist methods. There are general advantages and disadvantages to Bayesian inference. The advantages to using Bayesian analysis include the following:

- Bayesian analysis provides a natural and principled way of combining prior information with data, within a solid decision-theoretical framework. You can incorporate past information about a parameter and form a prior distribution for future analysis. When new observations become available, the previous posterior distribution can be used as a prior. All inferences logically follow from Bayes' theorem.
- It provides inferences that are conditional on the data and are exact, without reliance on either asymptotic approximation or the “plug-in” principle. Small sample inference proceeds in the same manner as if one had a large sample.
- It obeys the likelihood principle: if two distinct sampling designs yield proportional likelihood functions for θ , then all inferences about θ should be identical from these two designs. Classical inference does not obey the likelihood principle, in general.
- It provides interpretable answers, such as “the true parameter θ has a probability of 0.95 of falling in a 95% credible interval.”
- In addition, it provides a convenient setting for a wide range of models, such as hierarchical models and missing data problems. MCMC, along with other numerical methods, makes computations tractable for virtually all parametric models.

There are also disadvantages to using Bayesian analysis:

- It does not tell you how to select a prior. There is no correct way to choose a prior. Bayesian inferences require skills to translate subjective prior beliefs into a mathematically formulated prior. If you do not proceed with caution, you can generate misleading results.
- It can produce posterior distributions that are heavily influenced by the priors. From a practical point of view, it might sometimes be difficult to convince subject matter experts who do not agree with the validity of the chosen prior.
- It often comes with a high computational cost, especially in models with a large number of parameters. In addition, simulations provide slightly different answers unless the same random seed is used. Note that slight variations in simulation results do not contradict the early claim that Bayesian inferences are exact: the posterior distribution of a parameter is exact, given the likelihood function and the priors, while simulation-based estimates of posterior quantities can vary due to the random number generator used in the procedures.

For more in-depth treatments of the pros and cons of Bayesian analysis, see [Berger \(1985, Sections 4.1 and 4.12\)](#), [Berger and Wolpert \(1988\)](#), [Bernardo and Smith \(1994, with a new edition coming out in 2007\)](#), [Carlin and Louis \(2000, Section 1.4\)](#), [Robert \(2001, Chapter 11\)](#), and [Wasserman \(2004, Section 11.9\)](#).

The following sections provide more detailed information about the computations of Bayesian analysis and the statistics provided by these procedures.

Markov Chain Monte Carlo Method

The Markov chain Monte Carlo (MCMC) method is a general simulation method for sampling from posterior distributions and computing posterior quantities of interest.

MCMC method samples successively from a target distribution, with each sample drawn depending on the previous one; hence the notion of the Markov chain. A Markov chain is a sequence of random variables, $\theta^1, \theta^2, \dots$, for which the random variable θ^t depends on all previous θ s only through its immediate predecessor θ^{t-1} . You can think of a Markov chain applied to sampling as a mechanism that traverses randomly through a target distribution without having any memory of where it has been. Where it moves next is entirely dependent on where it is now.

Monte Carlo, as in Monte Carlo integration, is mainly used to approximate an expectation by using the Markov chain samples. In the simplest version

$$\int_S g(\theta)p(\theta)d\theta \cong \frac{1}{n} \sum_{t=1}^n g(\theta^t)$$

where $g(\cdot)$ is a function of interest and θ^t are samples from $p(\theta)$ on its support S . This approximates the expected value of $g(\theta)$.

The earliest reference to MCMC simulation occurs in the physics literature. [Metropolis and Ulam \(1949\)](#) and [Metropolis et al. \(1953\)](#) describe what is known as the Metropolis algorithm (see the section “[Metropolis and Metropolis-Hastings Algorithms](#)” on page 13). The algorithm can be used to generate sequences of samples from the joint distribution of multiple variables, and it is the foundation of MCMC. [Hastings \(1970\)](#) generalized their work, resulting in the Metropolis-Hastings algorithm. [Geman and Geman \(1984\)](#) analyzed image data by using what is now called Gibbs sampling (see the section “[Gibbs Sampler](#)” on page 15). These MCMC methods first appeared in the mainstream statistical literature in [Tanner and Wong \(1987\)](#).

The Markov chain method has been quite successful in modern Bayesian computing. Only in the simplest Bayesian models can you recognize the analytical forms of the posterior distributions and summarize inferences directly. In moderately complex models, posterior densities can often be too difficult to work with directly. With the MCMC method, it is possible to generate samples from an arbitrary posterior density $p(\theta|\mathbf{y})$ and to use these samples to approximate expectations of quantities of interest.

Several other aspects of the Markov chain method also contributed to its success. Most important, if the simulation algorithm is implemented correctly, the Markov chain is guaranteed to converge to the target distribution $p(\theta|\mathbf{y})$, under rather broad conditions, regardless of where the chain was initialized. In other words, a Markov chain is able to improve its approximation to the true distribution at each step in the simulation. And, if the chain is allowed to run for a very long time (often required), you can recover $p(\theta|\mathbf{y})$ to any precision. Also, the simulation algorithm is easily extendable to models with a large number of parameters or high complexity, although the “curse of dimensionality” often causes problems in practice.

Properties of Markov chains are discussed in [Feller \(1968\)](#), [Breiman \(1968\)](#), and [Meyn and Tweedie \(1993\)](#). [Ross \(1997\)](#) and [Karlin and Taylor \(1975\)](#) give a non-measure-theoretic treatment of stochastic processes, including Markov chains.

For conditions governing Markov chain convergence and rates of convergence, see [Amit \(1991\)](#), [Applegate, Kannan, and Polson \(1990\)](#), [Chan \(1993\)](#), [Geman and Geman \(1984\)](#), [Liu, Wong, and Kong \(1991a, b\)](#), [Rosenthal \(1991a, b\)](#), [Tierney \(1994\)](#), and [Schervish and Carlin \(1992\)](#). [Besag \(1974\)](#) describes conditions under which a set of conditional distributions gives a unique joint distribution.

[Tanner \(1993\)](#), [Gilks, Richardson, and Spiegelhalter \(1996\)](#), [Chen, Shao, and Ibrahim \(2000\)](#), [Liu \(2001\)](#), [Gelman et al. \(2004\)](#), [Robert and Casella \(2004\)](#), and [Congdon \(2001, 2003, 2005\)](#) provide both theoretical and applied treatments of MCMC methods. You can also refer to the section “[A Bayesian Reading List](#)” on page 34 for a list of books with various levels of difficulty of treatment of the subject, and its application to Bayesian statistics.

Metropolis and Metropolis-Hastings Algorithms

The Metropolis algorithm was first proposed by [Metropolis et al. \(1953\)](#). It is simple but practical, and it can be used to obtain random samples from any arbitrarily complicated target distribution of any dimension that is known up to a normalizing

constant. The Bayesian procedures use a special case of the Metropolis algorithm called the Gibbs sampler to obtain posterior samplers.

Suppose you want to obtain T samples from a univariate distribution with probability density function $f(\theta|\mathbf{y})$. Let θ^t be the t th sample from f . To implement the Metropolis algorithm, you need to have an initial value θ^0 and a symmetric *proposal* density $q(\theta^{t+1}|\theta^t)$. For the $(t + 1)$ th iteration, the algorithm generates a sample from $q(\cdot|\cdot)$ based on the current sample θ^t , and makes a decision to either accept or reject the new sample. If the new sample is accepted, the algorithm repeats itself by starting at the new sample; if the new sample is rejected, the algorithm starts at the current point and repeats. The algorithm is self-repeating, so it can be carried out as long as required. In practice, you have to decide in advance the total number of samples needed and stop the sampler after that many iterations have been completed.

More specifically, let $q(\theta_{\text{new}}|\theta^t)$ be a symmetric distribution. The proposal distribution should be an easy distribution to sample from, and it must be such that $q(\theta_{\text{new}}|\theta^t) = q(\theta^t|\theta_{\text{new}})$, meaning that the likelihood of jumping to θ_{new} from θ^t is the same as the likelihood of jumping back to θ^t from θ_{new} . The most common choice of the proposal distribution is the normal distribution $N(\theta^t, \sigma)$ with a fixed σ . The Metropolis algorithm can be summarized as follows:

1. Set $t = 0$. Choose a starting point θ^0 . This can be an arbitrary point as long as $f(\theta^0|\mathbf{y}) > 0$.
2. Generate a new sample, θ_{new} , by using the proposal distribution $q(\cdot|\theta^t)$.
3. Calculate the following quantity:

$$r = \min \left\{ \frac{f(\theta_{\text{new}}|\mathbf{y})}{f(\theta^t|\mathbf{y})}, 1 \right\}$$

4. Sample u from the uniform distribution $U(0, 1)$.
5. Set $\theta^{t+1} = \theta_{\text{new}}$ if $u < r$; otherwise set $\theta^{t+1} = \theta^t$.
6. Set $t = t + 1$. If $t < T$, the number of desired samples, return to step 2. Otherwise, stop.

Note: The number of iteration keeps increasing regardless of whether a proposed sample is accepted.

This algorithm defines a chain of random variates whose distribution will converge to the desired distribution $f(\theta|\mathbf{y})$, and so from some point forward, the chain of samples is a sample from the distribution of interest. In Markov chain terminology, this distribution is called the *stationary distribution* of the chain, and in Bayesian statistics, it is the posterior distribution of the model parameters. The reason that the Metropolis algorithm works is beyond the scope of this documentation, but you can find more detailed descriptions and proofs in many standard textbooks, including [Roberts \(1996\)](#) and [Liu \(2001\)](#).

You are not limited to a symmetric random walk proposal distribution in establishing a valid sampling algorithm. A more general form, the Metropolis-Hastings (MH)

algorithm, was proposed by [Hastings \(1970\)](#). The MH algorithm uses an asymmetric proposal distribution: $q(\theta_{\text{new}}|\theta^t) \neq q(\theta^t|\theta_{\text{new}})$.

The difference in its implementation comes in calculating the ratio of densities:

$$r = \min \left\{ \frac{f(\theta_{\text{new}}|\mathbf{y})q(\theta^t|\theta_{\text{new}})}{f(\theta^t|\mathbf{y})q(\theta_{\text{new}}|\theta^t)}, 1 \right\}$$

Other steps remain the same.

Extension of the Metropolis algorithm to higher-dimensional $\boldsymbol{\theta}$ is straightforward. Suppose $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)'$ is the parameter vector. To start the Metropolis algorithm, select an initial value for each θ_k , and use a multivariate version of proposal distribution $q(\cdot|\cdot)$, such as a multivariate normal distribution, to select a k -dimensional new parameter. Other steps remain the same as those previously described, and this Markov chain eventually converges to the target distribution of $f(\boldsymbol{\theta}|\mathbf{y})$.

You might find [Chib and Greenberg \(1995\)](#) to be a useful tutorial on the algorithm.

Gibbs Sampler

The Gibbs sampler is a special case of the Metropolis-Hastings sampler in which the proposal distributions exactly match the posterior conditional distributions, and proposals are accepted 100% of the time. Gibbs sampling requires you to decompose the joint posterior distribution into full conditional distributions for each parameter in the model and to be able to sample from them. The sampler can be efficient when the parameters are not highly dependent on each other and the full conditional distributions are easy to sample from. Some people favor this algorithm because it does not require an instrumental proposal distribution as Metropolis methods do. However, while deriving the conditional distributions can be relatively easy, it is not always possible to find an efficient way to sample from these conditional distributions.

Suppose $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)'$ is the parameter vector, $p(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood, and $\pi(\boldsymbol{\theta})$ is the prior distribution. The full posterior conditional distribution of $\pi(\theta_i|\theta_j, i \neq j, \mathbf{y})$ is proportional to the joint posterior density; that is,

$$\pi(\theta_i|\theta_j, i \neq j, \mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

For instance, the one-dimensional conditional distribution of θ_1 given $\theta_j = \theta_j^*, 2 \leq j \leq k$, is computed as

$$\pi(\theta_1|\theta_j = \theta_j^*, 2 \leq j \leq k, \mathbf{y}) = p(\mathbf{y}|(\boldsymbol{\theta} = (\theta_1, \theta_2^*, \dots, \theta_k^*)'))\pi(\boldsymbol{\theta} = (\theta_1, \theta_2^*, \dots, \theta_k^*)')$$

The Gibbs sampler works as follows:

1. Set $t = 0$, and choose an arbitrary initial value of $\boldsymbol{\theta}^{(0)} = \{\theta_1^{(0)}, \dots, \theta_k^{(0)}\}$.
2. Generate each component of $\boldsymbol{\theta}$ as follows:

- draw $\theta_1^{(t+1)}$ from $\pi(\theta_1|\theta_2^{(t)}, \dots, \theta_k^{(t)}, \mathbf{y})$
 - draw $\theta_2^{(t+1)}$ from $\pi(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, \dots, \theta_k^{(t)}, \mathbf{y})$
 - ...
 - draw $\theta_k^{(t+1)}$ from $\pi(\theta_k|\theta_1^{(t+1)}, \dots, \theta_{k-1}^{(t+1)}, \mathbf{y})$
3. Set $t = t + 1$. If $t < T$, the number of desired samples, return to step 2. Otherwise, stop.

The name “Gibbs” was introduced by [Geman and Geman \(1984\)](#). [Gelfand et al. \(1990\)](#) first used Gibbs sampling to solve problems in Bayesian inference. See [Casella and George \(1992\)](#) for a tutorial on the sampler.

Adaptive Rejection Sampling Algorithm

The BGENMOD, BLIFEREG, and BPHREG procedures use the adaptive rejection sampling (ARS) algorithm to sample parameters sequentially from their univariate full conditional distributions. The ARS algorithm is a rejection algorithm that was originally proposed by [Gilks and Wild \(1992\)](#). Given a log-concave density (the log of the density is concave), you can construct an envelope to the density by using linear segments. You then use the linear segment envelope as a proposal density (it becomes a piecewise exponential density on the original scale and is easy to generate samplers from) in the rejection sampling. The log-concavity condition is met in some of the models fit by the procedures. For example, the posterior densities for the regression parameters in the generalized linear models are log-concave under flat priors. When this condition fails, the ARS algorithm calls for an additional Metropolis-Hasting step ([Gilks, Best, and Tan 1995](#)), and the modified algorithm becomes the adaptive rejection metropolis sampling (ARMS) algorithm. The Bayesian procedures can recognize whether a model is log-concave and select the appropriate sampler for the problem at hand.

The BGENMOD, BLIFEREG, and BPHREG procedures implement the ARMS algorithm based on code kindly provided by [Gilks \(2003\)](#) to obtain posterior samples.

For a detailed description and explanation of the algorithm, see [Gilks and Wild \(1992\)](#), [Gilks, Best, and Tan \(1995\)](#), and [Robert and Casella \(2004\)](#).

Burn-in, Thinning, and Markov Chain Samples

Burn-in refers to the practice of discarding an initial portion of a Markov chain sample so that the effect of initial values on the posterior inference is minimized. For example, suppose the target distribution is $N(0, 1)$, and the Markov chain was started at the value 10^6 . The chain might quickly travel to regions around 0 in a few iterations. However, including samples around the value 10^6 in the posterior mean calculation can produce substantial bias in the mean estimate. In theory, if the Markov chain is allowed to run for an infinite amount of time, the effect of the initial values decreases to zero. In practice, you do not have the luxury of infinite samples. The underlying assumption is that, after t iterations, the chain would have reached its target distribution and you can throw away the early portion and use the good samples for posterior inference. The value of t is the burn-in number.

With some models you might experience poor mixing, or slow convergence, of the Markov chain. This can happen, for example, when parameters are highly correlated with each other. Poor mixing means that the Markov chain slowly traverses the parameter space (see the section “[Visual Analysis via Trace Plots](#)” on page 18 for examples of poorly mixed chains) and the chain has high dependence. High sample autocorrelation can result in biased MCSE. A common strategy is to *thin* the Markov chain in order to reduce sample autocorrelations. To thin a chain with rate k is to keep every k th simulation draw from each sequence. You can safely use a thinned Markov chain for posterior inference as long as the chain converges.

It is important to note that thinning a Markov chain can be wasteful because you are throwing away a $\frac{k-1}{k}$ fraction of all the posterior samples generated. [MacEachern and Berliner \(1994\)](#) show that you always get more precise posterior estimates if the entire Markov chain is used. However, other factors, such as computer storage or plotting time, might prevent you from keeping all samples.

To use these Bayesian procedures, you need to determine the total number of samples to keep ahead of time. This number is not obvious and often depends on the type of inference you want to make. Mean estimates do not require nearly as many samples as small-tail percentile estimates. In most applications, you might find that keeping a few thousand iterations is sufficient for reasonably accurate posterior inference.

In all these procedures, the relationship between the number of iterations requested, the number of iterations kept, and the amount of thinning is as follows:

$$\text{kept} = \left[\frac{\text{requested}}{\text{thinning}} \right]$$

where $[]$ is the rounding operator.

Assessing Markov Chain Convergence

Simulation-based Bayesian inference requires using simulated draws to summarize the posterior distribution or calculate any relevant quantities of interest. You need to treat the simulation draws with care. There are usually two issues. First, you have to decide whether the Markov chain has reached its stationary, or the desired posterior, distribution. Second, you have to determine the number of iterations to keep after the Markov chain has reached stationarity. Convergence diagnostics help to resolve these issues.

Note that many diagnostic tools are designed to verify a necessary but not sufficient condition for convergence. There are no conclusive tests that can tell you when the Markov chain has converged to its stationary distribution. You should proceed with caution.

Also, note that you should check the convergence of **ALL** parameters, and not just those of interest, before proceeding to make any inference. With some models, certain parameters can appear to have very good convergence behavior, but that could be

misleading due to the slow convergence of other parameters. If some of the parameters have bad mixing, you cannot get accurate posterior inference for parameters that appear to have good mixing.

See [Cowles and Carlin \(1996\)](#) and [Brooks and Roberts \(1998\)](#) for discussions about convergence diagnostics.

Visual Analysis via Trace Plots

Trace plots of samples versus the simulation index can be very useful in assessing convergence. The trace tells you if the chain has not yet converged to its stationary distribution. That is, if it needs a longer burn-in period. A trace can also tell you whether the chain is mixing well. A chain might have reached stationarity if the distribution of points is not changing as the chain progresses. The aspects of stationarity that are most recognizable from a trace plot are a relatively constant mean and variance. A chain that mixes well traverses its posterior space rapidly, and it can jump from one remote region of the posterior to another in relatively few steps. [Figure 1.1](#) through [Figure 1.4](#) display some typical features that you might see in trace plots. The traces are for a parameter called γ .

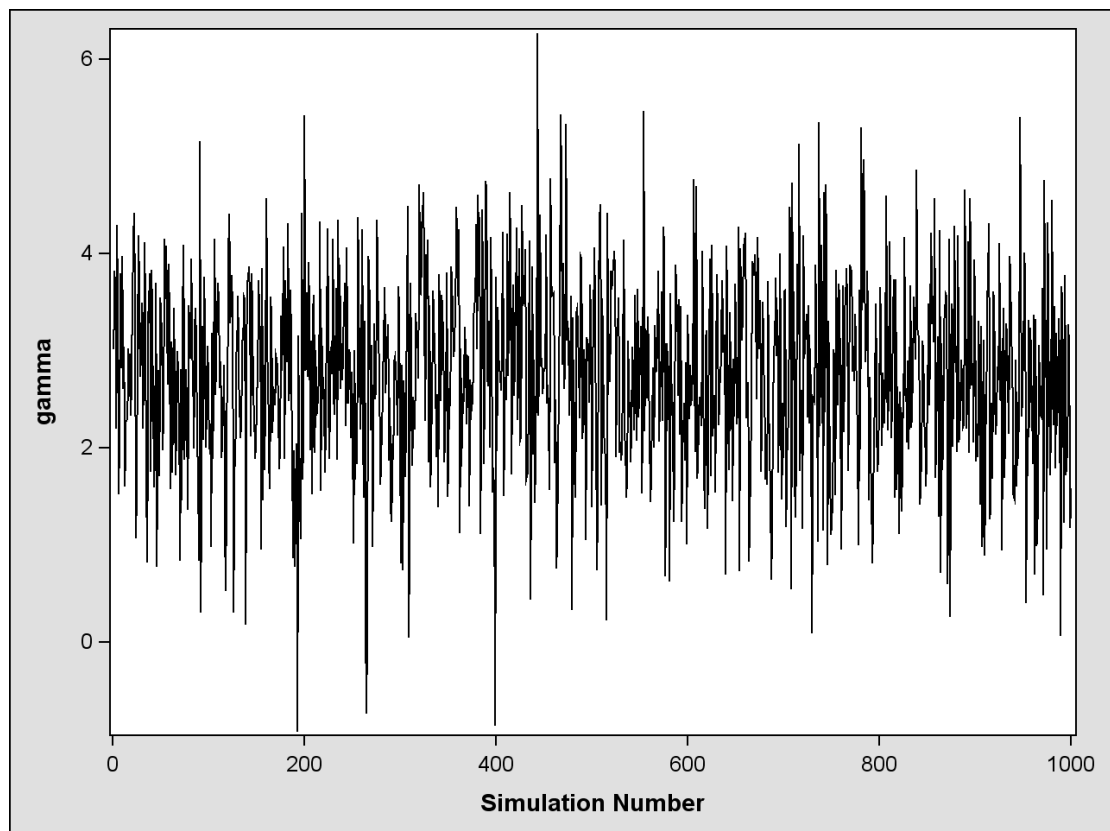


Figure 1.1. Essentially Perfect Trace for γ

Figure 1.1 displays a “perfect” trace plot. Note that the center of the chain appears to be around the value 3, with very small fluctuations. This indicates that the chain could have reached the right distribution. The chain is mixing well; it is exploring the distribution by traversing to areas where its density is very low. You can conclude that the mixing is quite good here.

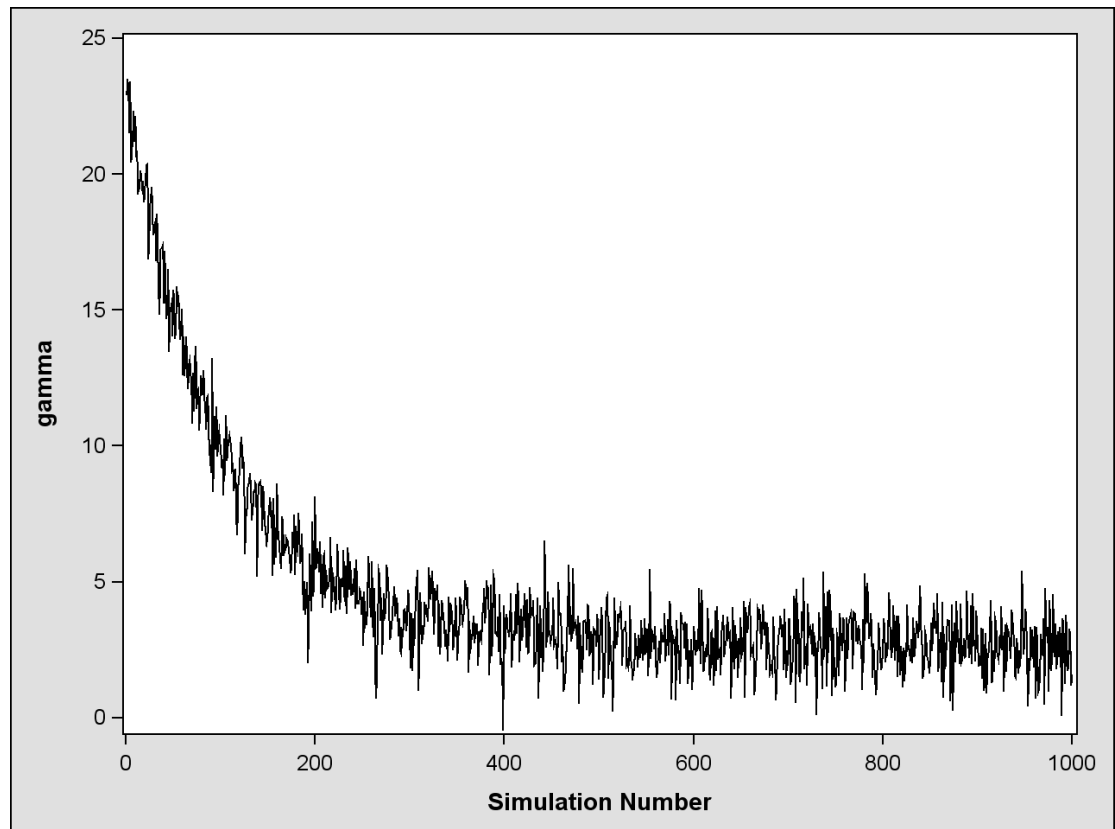


Figure 1.2. Nonconvergence of γ

Figure 1.2 displays a trace plot for a chain that starts at a very remote initial value and makes its way to the targeting distribution. The first few hundred observations should be discarded. This chain appears to be mixing very well locally. It travels relatively quickly to the target distribution, reaching it in a few hundred iterations. If you have a chain that looks like this, you will want to increase the burn-in sample size. If you need to use this sample to make inferences, you will want to use only the samples toward the end of the chain.

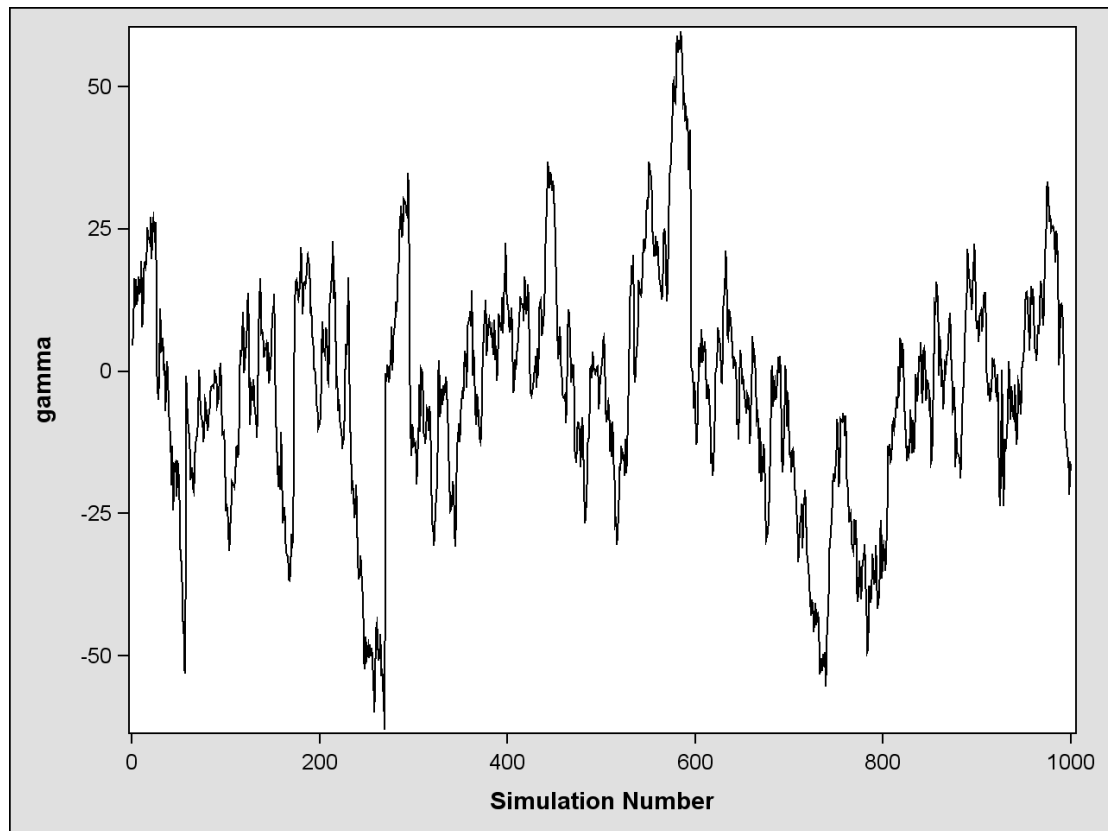


Figure 1.3. Marginal Mixing for γ

Figure 1.3 demonstrates marginal mixing. The chain is taking only small steps and does not traverse its distribution quickly. This type of trace plot is typically associated with high autocorrelation among the samples. To obtain a few thousand independent samples, you need to run the chain a lot longer.

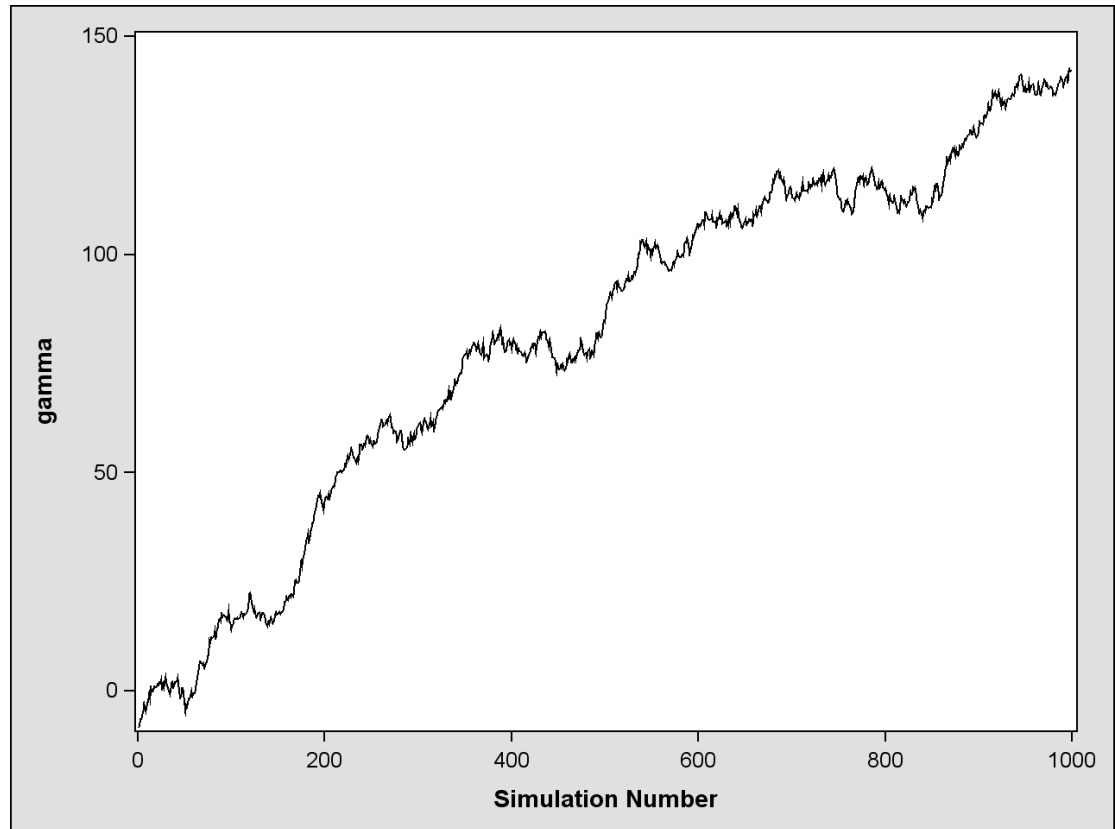


Figure 1.4. Bad Mixing, Nonconvergence of γ

The trace plot shown in [Figure 1.4](#) depicts a chain with serious problems. It is mixing very slowly, and it offers no evidence of convergence. You would want to try to improve the mixing of this chain. For example, you might consider reparameterizing your model on the log scale. Run the Markov chain for a long time to see where it goes. This type of chain is entirely unsuitable for making parameter inferences.

Statistical Diagnostic Tests

The BGENMOD, BLIFEREG, and BPHREG procedures include several statistical diagnostic tests that can help you assess Markov chain convergence. For a detailed description of each of the diagnostic tests, see the subsections. [Table 1.1](#) provides a summary of the diagnostic tests and their interpretations.

Table 1.1. Convergence Diagnostic Tests Available in the Bayesian Procedures

Name	Description	Interpretation of the Test
Gelman-Rubin	Uses parallel chains with dispersed initial values to test whether they all converge to the same target distribution. Failure could indicate the presence of a multi-mode posterior distribution (different chains converge to different local modes) or the need to run a longer chain (burn-in is yet to be completed).	One-sided test based on a variance ratio test statistic. Large \widehat{R}_c values indicate rejection.
Geweke	Tests whether the mean estimates have converged by comparing means from the early and latter part of the Markov chain.	Two-sided test based on a z -score statistic. Large absolute z values indicate rejection.
Heidelberger-Welch (stationarity test)	Tests whether the Markov chain is a covariance (or weakly) stationary process. Failure could indicate that a longer Markov chain is needed.	One-sided test based on a Cramer-von Mises statistic. Small p -values indicate rejection.
Heidelberger-Welch (halfwidth test)	Reports whether the sample size is adequate to meet the required accuracy for the mean estimate. Failure could indicate that a longer Markov chain is needed.	If a relative halfwidth statistic is greater than a predetermined accuracy measure, this indicates rejection.
Raftery-Lewis	Evaluates the accuracy of the estimated (desired) percentiles by reporting the number of samples needed to reach the desired accuracy of the percentiles. Failure could indicate that a longer Markov chain is needed.	If the total samples needed are less than the Markov chain sample, this indicates rejection.
autocorrelation	Measures dependency among Markov chain samples.	High correlations between long lags indicate poor mixing.
effective sample size	Relates to autocorrelation; measures mixing of the Markov chain.	Large discrepancy between the effective sample size and the actual simulation iteration indicates poor mixing.

Gelman and Rubin Diagnostics

Gelman and Rubin diagnostics (Gelman and Rubin 1992; Brooks and Gelman 1997) are based on analyzing multiple simulated MCMC chains by comparing the variances within each chain and the variance between chains. Large deviation between these two variances indicates nonconvergence.

Define $\{\theta^t\}$, where $t = 1, \dots, n$, to be the collection of a single Markov chain output. The parameter θ^t is the t th sample of the Markov chain. For notational simplicity, θ

is assumed to be single dimensional in this section.

Suppose you have M parallel MCMC chains that were initialized from various parts of the target distribution. Each chain is of length n (after discarding the burn-in). For each θ^t , the simulations are labeled as θ_m^t , where $t = 1, \dots, n$, and $m = 1, \dots, M$. The between-chain variance B , and the within-chain variance W are calculated as

$$B = \frac{n}{M-1} \sum_{m=1}^M (\bar{\theta}_m - \bar{\theta})^2, \quad \text{where } \bar{\theta}_m = \frac{1}{n} \sum_{t=1}^n \theta_m^t, \quad \bar{\theta} = \frac{1}{M} \sum_{m=1}^M \bar{\theta}_m$$

$$W = \frac{1}{M} \sum_{m=1}^M s_m^2, \quad \text{where } s_m^2 = \frac{1}{n-1} \sum_{t=1}^n (\theta_m^t - \bar{\theta}_m)^2$$

The posterior marginal variance, $\text{var}(\theta|\mathbf{y})$, is a weighted average of W and B . The estimate of the variance is

$$\hat{V} = \frac{n-1}{n}W + \frac{M+1}{nM}B$$

The idea is that if all M chains have reached the target distribution, this posterior variance estimate should be very close to the within-chain variance W . Therefore, you would expect to see the ratio \hat{V}/W be close to 1. The square root of this ratio is referred to as the *potential scale reduction factor* (PSRF). A large PSRF indicates that the between-chain variance is substantially greater than the within-chain variance, so that longer simulation is needed. If the PSRF is close to 1, you can conclude that each of the M chains has stabilized, and they are likely to have reached the target distribution.

A refined version of PSRF is calculated, as suggested by [Brooks and Gelman \(1997\)](#), as

$$\hat{R}_c = \sqrt{\frac{\hat{d}+3}{\hat{d}+1} \cdot \frac{\hat{V}}{W}} = \sqrt{\frac{\hat{d}+3}{\hat{d}+1} \left(\frac{n-1}{n} + \frac{M+1}{nM} \frac{B}{W} \right)}$$

where

$$\hat{d} = \frac{2\hat{V}^2}{\widehat{\text{Var}}(\hat{V})}$$

and

$$\widehat{\text{Var}}(\hat{V}) = \left(\frac{n-1}{n} \right)^2 \frac{1}{M} \widehat{\text{Var}}(s_m^2) + \left(\frac{M+1}{nM} \right)^2 \frac{2}{M-1} B^2$$

$$+ 2 \frac{(M+1)(n-1)}{n^2 M} \frac{n}{M} (\widehat{\text{cov}}(s_m^2, (\bar{\theta}_m)^2) - 2\bar{\theta} \cdot \widehat{\text{cov}}(s_m^2, \bar{\theta}_m))$$

All the Bayesian procedures also produce an upper $100(1-\alpha/2)\%$ confidence limit of \widehat{R}_c . Gelman and Rubin (1992) showed that the ratio B/W in \widehat{R}_c has an F distribution with degrees of freedom $M - 1$ and $2W^2M/\widehat{\text{Var}}(s_m^2)$. Because you are concerned only if the scale is large, not small, only the upper $100(1 - \alpha/2)\%$ confidence limit is reported. This is written as

$$\sqrt{\left(\frac{n-1}{n} + \frac{M+1}{nM} \cdot F_{1-\alpha/2}\left(M-1, \frac{2W^2}{\widehat{\text{Var}}(s_m^2)/M}\right)\right) \cdot \frac{\hat{d}+3}{\hat{d}+1}}$$

In the Bayesian procedures, you can specify the number of chains that you want to run. Typically three chains are sufficient. The first chain is used for posterior inference, such as mean and standard deviation; the other $M - 1$ chains are used for computing the diagnostics and discarded afterward. This test can be computationally costly, because it prolongs the simulation by M -fold.

It is best to choose different initial values for all M chains. The initial values should be as dispersed from each other as possible so that the Markov chains can fully explore different parts of the distribution before they converge to the target. Similar initial values can be risky because all of the chains can get stuck in a local maximum; that is something this convergence test cannot detect. If you do not supply initial values for all the different chains, the procedures generate them for you.

Geweke Diagnostics

The Geweke test (Geweke 1992) compares values in the early part of the Markov chain to those in the latter part of the chain in order to detect failure of convergence.

The statistic is constructed in the following way. Two subsequences of the Markov chain $\{\theta^t\}$ are taken out, with $\{\theta_1^t : t = 1, \dots, n_1\}$ and $\{\theta_2^t : t = n_a, \dots, n\}$, where $1 < n_1 < n_a < n$. Let $n_2 = n - n_a + 1$, and define

$$\bar{\theta}_1 = \frac{1}{n_1} \sum_{t=1}^{n_1} \theta^t \quad \text{and} \quad \bar{\theta}_2 = \frac{1}{n_2} \sum_{t=n_a}^n \theta^t$$

Let $\hat{s}_1(0)$ and $\hat{s}_2(0)$ denote consistent spectral density estimates at zero frequency (see the subsection “Spectral Density Estimate at Zero Frequency” for estimation details) for the two MCMC chains, respectively. If the ratios n_1/n and n_2/n are fixed, $(n_1 + n_2)/n < 1$, and the chain is stationary, then the statistic

$$Z_n = \frac{\bar{\theta}_1 - \bar{\theta}_2}{\sqrt{\frac{\hat{s}_1(0)}{n_1} + \frac{\hat{s}_2(0)}{n_2}}}$$

converges to a standard normal distribution as $n \rightarrow \infty$. This is a two-sided test, and large absolute z -scores indicate rejection.

Spectral Density Estimate at Zero Frequency

For one sequence of the Markov chain $\{\theta_t\}$, the relationship between the h -lag covariance sequence of a time series and the spectral density, f , is

$$s_h = \frac{1}{2\pi} \int_{-\pi}^{\pi} \exp(i\omega h) f(\omega) d\omega$$

where i indicates that ωh is the complex argument. Inverting this Fourier integral,

$$f(\omega) = \sum_{h=-\infty}^{\infty} s_h \exp(-i\omega h) = s_0 \left(1 + 2 \sum_{h=1}^{\infty} \rho_h \cos(\omega h) \right)$$

It follows that

$$f(0) = \sigma^2 \left(1 + 2 \sum_{h=1}^{\infty} \rho_h \right)$$

which gives an autocorrelation adjusted estimate of the variance. In this equation, σ^2 is the naive variance estimate of the sequence $\{\theta_t\}$ and ρ_h is the lag h autocorrelation. Due to obvious computational difficulties, such as calculation of autocorrelation at infinity, you cannot effectively estimate $f(0)$ by using the preceding formula.

The usual route is to first obtain the *periodogram* $p(\omega)$ of the sequence, and then estimate $f(0)$ by smoothing the estimated periodogram. The periodogram is defined to be

$$p(\omega) = \frac{1}{n} \left[\left(\sum_{t=1}^n \theta_t \sin(\omega t) \right)^2 + \left(\sum_{t=1}^n \theta_t \cos(\omega t) \right)^2 \right]$$

The procedures use the following way to estimate $\hat{f}(0)$ from p (Heidelberger and Welch 1981). In $p(\omega)$, let $\omega = \omega_k = 2\pi k/n$ and $k = 1, \dots, \lfloor n/2 \rfloor$.* A smooth spectral density in the domain of $(0, \pi]$ is obtained by fitting a gamma model with the log link function, using $p(\omega_k)$ as response and $x_1(\omega_k) = \sqrt{3}(4\omega_k/(2\pi) - 1)$ as the only regressor. The predicted value $\hat{f}(0)$ is given by

$$\hat{f}(0) = e^{\hat{\beta}_0 - \sqrt{3}\hat{\beta}_1}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates of the intercept and slope parameters, respectively.

*This is equivalent to the fast Fourier transformation of the original time series θ_t .

Heidelberger and Welch Diagnostics

The Heidelberger and Welch test (Heidelberger and Welch 1981, 1983) consists of two parts: a stationary portion test and a halfwidth test. The stationarity test assesses the stationarity of a Markov chain by testing a hypothesis that the chain comes from a covariance stationary process. The halfwidth test checks whether the Markov chain sample size is adequate to estimate the mean values accurately.

Given $\{\theta^t\}$, set $S_0 = 0$, $S_n = \sum_{t=1}^n \theta^t$, and $\bar{\theta} = (1/n) \sum_{t=1}^n \theta^t$. You can construct the following sequence with s coordinates on values from $\frac{1}{n}, \frac{2}{n}, \dots, 1$:

$$B_n(s) = (S_{[ns]} - [ns]\bar{\theta}) / (n\hat{p}(0))^{1/2}$$

where $[a]$ is the integer part of number a , and $\hat{p}(0)$ is an estimate of the spectral density at zero frequency (see the section “Spectral Density Estimate at Zero Frequency” on page 25 for estimation details) using the second half of the sequence. For large n , B_n converges in distribution to a Brownian bridge (Billingsley 1986). So you can construct a test statistic by using B_n . The statistic used in these procedures is the Cramer–von Mises statistic;* that is $\int_0^1 B_n(s)^2 ds = CVM(B_n)$. As $n \rightarrow \infty$, the statistic converges in distribution to a standard Cramer–von Mises distribution. The integral $\int_0^1 B_n(s)^2 ds$ is numerically approximated using Simpson’s rule.

Let $y_i = B_n(s)^2$, where $s = 0, \frac{1}{n}, \dots, \frac{n-1}{n}, 1$, and $i = ns = 0, 1, \dots, n$. If n is even, let $m = n/2$; otherwise, let $m = (n - 1)/2$. The Simpson’s approximation to the integral is

$$\int_0^1 B_n(s)^2 ds \approx \frac{1}{3n} [y_0 + 4(y_1 + \dots + y_{2m-1}) + 2(y_2 + \dots + y_{2m-2}) + y_{2m}]$$

Note that Simpson’s rule requires an even number of intervals. When n is odd, y_n is set to be 0, and the value does not contribute to the approximation.

This test can be performed repeatedly on the same chain and helps you identify a time t when the chain has reached stationarity. The whole chain, $\{\theta^t\}$, is first used to construct the Cramer–von Mises statistic. If it passes the test, you can conclude that the entire chain is stationary. If it fails the test, you drop the initial 10% of the chain and redo the test by using the remaining 90%. This process is repeated until either a time t is selected or it reaches a point where there are not enough data remaining to construct a confidence interval (the cutoff proportion is set to be 50%).

The part of the chain that is deemed stationary is put through a halfwidth test, which reports whether the sample size is adequate to meet certain accuracy requirements for the mean estimates. Running the simulation less than this length of time would not

*The von Mises distribution was first introduced by von Mises (1918). The density function is $p(\theta|\mu\kappa) \sim M(\mu, \kappa) = [2\pi I_0(\kappa)]^{-1} \exp(\kappa \cos(\theta - \mu))$ ($0 \leq \theta \leq 2\pi$), where the function $I_0(\kappa)$ is the modified Bessel function of the first kind and order zero, defined by $I_0(\kappa) = (2\pi)^{-1} \int_0^{2\pi} \exp(\kappa \cos(\theta - \mu)) d\theta$.

meet the requirement, while running it longer would not provide any additional information that is needed. The statistic calculated here is the *relative halfwidth* (RHW) of the confidence interval. The RHW for a confidence interval of level $1 - \alpha$ is

$$\text{RHW} = \frac{z_{(1-\alpha/2)} \cdot (\hat{s}_n/n)^{1/2}}{\hat{\theta}}$$

where $z_{(1-\alpha/2)}$ is the z -score of the $100(1 - \alpha/2)$ th percentile (for example, $z_{(1-\alpha/2)} = 1.96$ if $\alpha = 0.05$), \hat{s}_n is the variance of the chain estimated using the spectral density method (see explanation in the section “[Spectral Density Estimate at Zero Frequency](#)” on page 25), n is the length, and $\hat{\theta}$ is the estimated mean. The RHW quantifies accuracy of the $1 - \alpha$ level confidence interval of the mean estimate by measuring the ratio between the sample standard error of the mean and the mean itself. In other words, you can stop the Markov chain if the variability of the mean stabilizes with respect to the mean. An implicit assumption is that large means are often accompanied by large variances. If this assumption is not met, then this test can produce false rejections, such as a small mean around 0 and large standard deviation; or false acceptances, such as a very large mean with relative small variance. As with any other convergence diagnostics, you might want to exercise caution in interpreting the results.

The stationarity test is one-sided; rejection occurs when the p -value is greater than $1 - \alpha$. To perform the halfwidth test, you need to select an α level, the default of which is 0.05; and a predetermined tolerance value ϵ , the default of which is 0.1. If the calculated RHW is greater than ϵ , you conclude that there are not enough data to accurately estimate the mean with $1 - \alpha$ confidence under tolerance of ϵ .

Raftery and Lewis Diagnostics

If your interest lies in posterior percentiles, you want a diagnostic test that evaluates the accuracy of the estimated percentiles. The Raftery-Lewis test ([Raftery and Lewis 1992, 1995](#)) is designed for this purpose. Notation and deductions here closely resemble that in [Raftery and Lewis \(1995\)](#).

Suppose you are interested in a quantity θ_q such that $P(\theta \leq \theta_q | \mathbf{y}) = q$, where q can be an arbitrary cumulative probability, such as 0.025. This θ_q can be empirically estimated by finding the $[n \cdot 100 \cdot q]$ th number of the sorted $\{\theta^t\}$. Let $\hat{\theta}_q$ denote the estimand, which corresponds to an estimated probability $P(\theta \leq \hat{\theta}_q) = \hat{P}_q$. Because the simulated posterior distribution converges to the true distribution as the simulation sample size grows, $\hat{\theta}_q$ can achieve any degree of accuracy if the simulator is allowed to run for a very long time. However, running too long a simulation can be wasteful. Alternatively, you can use coverage probability to measure accuracy and stop the chain when a certain accuracy is reached.

A stopping criterion is reached when the estimated probability is within $\pm r$ of the true cumulative probability q , with probability s , such as $P(\hat{P}_q \in (q - r, q + r)) = s$. For example, suppose you want the coverage probability s to be 0.95 and the amount of tolerance r to be 0.005. This corresponds to requiring that the estimate of the cumulative distribution function of the 2.5th percentile be estimated to within ± 0.5 percentage points with probability 0.95.

The Raftery-Lewis diagnostics test finds the number of iterations, M , that need to be discarded (burn-ins), and the number of iterations needed, N , to achieve a desired precision.

Given a predefined cumulative probability q , these procedures first find $\hat{\theta}_q$, and then they construct a binary 0 – 1 process $\{Z_t\}$ by setting $Z_t = 1$ if $\theta^t \leq \hat{\theta}_q$ and 0 otherwise for all t . The sequence $\{Z_t\}$ is itself not a Markov chain, but you can construct a subsequence of $\{Z_t\}$ that is approximately Markovian if it is sufficiently k -thinned. When k becomes reasonably large, $\{Z_t^{(k)}\}$ starts to behave like a Markov chain.

Next, the procedures find this thinning parameter k . The number k is estimated by comparing the Bayesian information criterion (BIC) between two Markov models: a first-order and a second-order Markov model. A j th-order Markov model is one in which the current value of $\{Z_t^{(k)}\}$ depends on the previous j values. For example, in a second-order Markov model,

$$\begin{aligned} p\left(Z_t^{(k)} = z_t \mid Z_{t-1}^{(k)} = z_{t-1}, Z_{t-2}^{(k)} = z_{t-2}, \dots, Z_0^{(k)} = z_0\right) \\ = p\left(Z_t^{(k)} = z_t \mid Z_{t-1}^{(k)} = z_{t-1}, Z_{t-2}^{(k)} = z_{t-2}\right) \end{aligned}$$

where $z_i = \{0, 1\}$, $i = 0, \dots, t$. Given $\{Z_t^{(k)}\}$, you can construct two transition count matrices for a second-order Markov model,

$$\begin{array}{cc|cc} & & z_t = 0 & & z_t = 1 & \\ & & z_{t-1} = 0 & z_{t-1} = 1 & z_{t-1} = 0 & z_{t-1} = 1 & \\ z_{t-2} = 0 & & w_{000} & w_{010} & w_{001} & w_{011} & \\ z_{t-2} = 1 & & w_{100} & w_{110} & w_{101} & w_{111} & \end{array}$$

For each k , the procedures calculate the BIC that compares the two Markov models. The BIC is based on a likelihood ratio test statistic that is defined as

$$G_k^2 = 2 \sum_{i=0}^1 \sum_{j=0}^1 \sum_{l=0}^1 w_{ijl} \log \frac{w_{ijl}}{\hat{w}_{ijl}}$$

where \hat{w}_{ijl} is the expected cell count of w_{ijl} under the null model, the first-order Markov model, where the assumption $(Z_t^{(k)} \perp Z_{t-2}^{(k)}) \mid Z_{t-1}^{(k)}$ holds. The formula for the expected cell count is

$$\hat{w}_{ijl} = \frac{\sum_i w_{ijl} \cdot \sum_l w_{ijl}}{\sum_i \sum_l w_{ijl}}$$

The BIC is $G_k^2 - 2 \log(n_k - 2)$, where n_k is the k -thinned sample size (every k th sample starting with the first), with the last two data points discarded due to the construction of the second-order Markov model. The thinning parameter k is the smallest k for which the BIC is negative.

When k is found, you can estimate a transition probability matrix between state 0 and state 1 for $\{Z_t^{(k)}\}$:

$$Q = \begin{pmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{pmatrix}$$

Because $\{Z_t^{(k)}\}$ is a Markov chain, its equilibrium distribution exists and is estimated by

$$\pi = (\pi_0, \pi_1) = \frac{(\beta, \alpha)}{\alpha + \beta}$$

where $\pi_0 = P(\theta \leq \theta_q | \mathbf{y})$ and $\pi_1 = 1 - \pi_0$. The goal is to find an iteration number m such that after m steps, the estimated transition probability $P(Z_m^{(k)} = i | Z_0^{(k)} = j)$ is within ϵ of equilibrium π_i for $i, j = 0, 1$. Let $e_0 = (1, 0)$ and $e_1 = 1 - e_0$. The estimated transition probability after step m is

$$P(Z_m^{(k)} = i | Z_0^{(k)} = j) = e_j \left[\begin{pmatrix} \pi_0 & \pi_1 \\ \pi_0 & \pi_1 \end{pmatrix} + \frac{(1 - \alpha - \beta)^m}{\alpha + \beta} \begin{pmatrix} \alpha & -\alpha \\ -\beta & \beta \end{pmatrix} \right] e_j^\top$$

which holds when

$$m = \frac{\log \left(\frac{(\alpha + \beta)\epsilon}{\max(\alpha, \beta)} \right)}{\log(1 - \alpha - \beta)}$$

assuming $1 - \alpha - \beta > 0$.

Therefore, by time m , $\{Z_t^{(k)}\}$ is sufficiently close to its equilibrium distribution, and you know that a total size of $M = mk$ should be discarded as the burn-in.

Next, the procedures estimate N , the number of simulations needed to achieve desired accuracy on percentile estimation. The estimate of $P(\theta \leq \theta_q | \mathbf{y})$ is $\bar{Z}_n^{(k)} = \frac{1}{n} \sum_{t=1}^n Z_t^{(k)}$. For large n , $\bar{Z}_n^{(k)}$ is normally distributed with mean q , the true cumulative probability, and variance

$$\frac{1}{n} \frac{(2 - \alpha - \beta)\alpha\beta}{(\alpha + \beta)^3}$$

$P(q - r \leq \bar{Z}_n^{(k)} \leq q + r) = s$ is satisfied if

$$n = \frac{(2 - \alpha - \beta)\alpha\beta}{(\alpha + \beta)^3} \left\{ \frac{\Phi^{-1}\left(\frac{s+1}{2}\right)}{r} \right\}^2$$

Therefore, $N = nk$.

By using similar reasoning, the procedures first calculate the minimal number of iterations needed to achieve the desired accuracy, assuming the samples are independent:

$$N_{min} = \left\{ \Phi^{-1}\left(\frac{s+1}{2}\right) \right\}^2 \frac{q(1-q)}{r^2}$$

If $\{\theta^t\}$ does not have that required sample size, the Raftery-Lewis test is not carried out. If you still want to carry out the test, increase the number of Markov chain iterations.

The ratio N/N_{min} is sometimes referred to as the *dependence factor*. It measures deviation from posterior sample independence: the closer it is to 1, the less correlated are the samples.

There are a few things to keep in mind when you use this test. This diagnostic tool is specifically designed for the percentile of interest and does not provide information about convergence of the chain as a whole (Brooks and Roberts 1999). In addition, the test can be very sensitive to small changes. Both N and N_{min} are inversely proportional to r^2 , so you can expect to see large variations in these numbers with small changes to input variables, such as the desired coverage probability, or the cumulative probability of interest. Last, the time until convergence for a parameter can differ substantially for different cumulative probabilities.

Autocorrelations

The sample autocorrelation of lag h is defined in terms of the sample autocovariance function:

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \quad |h| < n$$

The sample autocovariance function of lag h (of $\{\theta_i^t\}$) is defined by

$$\hat{\gamma}(h) = \frac{1}{n-h} \sum_{t=1}^{n-h} \left(\theta_i^{t+h} - \bar{\theta}_i \right) \left(\theta_i^t - \bar{\theta}_i \right), \quad 0 \leq h < n$$

Effective Sample Size

You can use autocorrelation and trace plots to examine the mixing of a Markov chain. A closely related measure of mixing is the effective sample size (ESS) (Kass et al. 1998).

ESS is defined as follows:

$$\text{ESS} = \frac{n}{\tau} = \frac{n}{1 + 2 \sum_{k=1}^{\infty} \rho_k(\theta)}$$

where n is the total sample size, and $\rho_k(\theta)$ is the autocorrelation of lag k for θ . The quantity τ is referred to as the autocorrelation time. To estimate τ , the BGENMOD, BLIFEREG, and BPHREG procedures first find a cutoff point k after which the autocorrelations are very close to zero, and then sum all the ρ_k up to that point. The cutoff point k is such that $\rho_k < 0.05$ or $\rho_k < 2s_k$, where s_k is the estimated standard deviation:

$$s_k = 2 \sqrt{\left(\frac{1}{n} \left(1 + 2 \sum_{j=1}^{k-1} \rho_j^2(\theta) \right) \right)}$$

ESS and τ are inversely proportional to each other, and low ESS or high τ indicates bad mixing of the Markov chain.

Summary Statistics

Let θ be a p -dimensional parameter vector of interest: $\theta = \{\theta_1, \dots, \theta_p\}$. For each $i \in \{1, \dots, p\}$, there are n observations: $\theta_i = \{\theta_i^t, t = 1, \dots, n\}$.

Mean

The posterior mean is calculated using the following formula:

$$E(\theta_i | \mathbf{y}) \approx \bar{\theta}_i = \frac{1}{n} \sum_{t=1}^n \theta_i^t, \text{ for } i = 1, \dots, n$$

Standard Deviation

Sample standard deviation (expressed in variance term) is calculated using the following formula:

$$\text{Var}(\theta_i | \mathbf{y}) \approx s_i^2 = \frac{1}{n-1} \sum_{t=1}^n (\theta_i^t - \bar{\theta}_i)^2$$

Standard Error of the Mean Estimate

Suppose you have n iid samples, the mean estimate is $\bar{\theta}_i$, and the sample standard deviation is s_i . The standard error of the estimate is $\hat{\sigma}_i/\sqrt{n}$. However, positive autocorrelation (see the section “Autocorrelations” on page 30 for a definition) in the MCMC samples makes this an underestimate. To take account of the autocorrelation, the Bayesian procedures correct the standard error by using effective sample size (see the section “Effective Sample Size” on page 31).

Given an effective sample size of m , the standard error for $\bar{\theta}_i$ is $\hat{\sigma}_i/\sqrt{m}$. The procedures use the following formula (expressed in variance term):

$$\widehat{\text{Var}}(\bar{\theta}_i) = \frac{1 + 2 \sum_{k=1}^{\infty} \rho_k(\theta_i)}{n} \cdot \frac{\sum_{t=1}^n (\theta_i^t - \bar{\theta}_i)^2}{(n-1)}$$

The standard error of the mean is also known as the Monte Carlo standard error (MCSE). The MCSE provides a measurement of the accuracy of the posterior estimates, and small values do not necessarily indicate that you have recovered the true posterior mean.

Percentiles

Sample percentiles are calculated using Definition 5 (see the “Calculating Percentiles” section on page 274 in Chapter 3, “The UNIVARIATE Procedure” (SAS Base Statistical Procedures)).

Correlation

Correlation between θ_i and θ_j is calculated as

$$r_{ij} = \frac{\sum_{t=1}^n (\theta_i^t - \bar{\theta}_i) (\theta_j^t - \bar{\theta}_j)}{\sqrt{\sum_t (\theta_i^t - \bar{\theta}_i)^2 \sum_t (\theta_j^t - \bar{\theta}_j)^2}}$$

Covariance

Covariance θ_i and θ_j is calculated as

$$s_{ij} = \sum_{t=1}^n (\theta_i^t - \bar{\theta}_i) (\theta_j^t - \bar{\theta}_j)$$

Equal-Tail Credible Interval

Let $\pi(\theta_i|\mathbf{y})$ denote the marginal posterior cumulative distribution function of θ_i . A $100(1-\alpha)\%$ Bayesian equal-tail credible interval for θ_i is $(\theta_i^{\alpha/2}, \theta_i^{1-\alpha/2})$, where $\pi(\theta_i^{\alpha/2}|\mathbf{y}) = \frac{\alpha}{2}$, and $\pi(\theta_i^{1-\alpha/2}|\mathbf{y}) = 1 - \frac{\alpha}{2}$. The interval is obtained using the empirical $\frac{\alpha}{2}$ th and $(1 - \frac{\alpha}{2})$ th percentiles of $\{\theta_i^t\}$.

Highest Probability Density Interval (HPD)

For a definition of an HPD interval, see the section “Interval Estimation” on page 10. The procedures use the Chen-Shao algorithm (Chen and Shao 1999; Chen, Shao, and Ibrahim 2000) to estimate an empirical HPD interval of θ_i :

- Sort $\{\theta_i^t\}$ to obtain the ordered values:

$$\theta_{i(1)} \leq \theta_{i(2)} \leq \cdots \leq \theta_{i(n)}$$

- Compute the $100(1 - \alpha)\%$ credible intervals:

$$R_j(n) = (\theta_{i(j)}, \theta_{i(j + [(1 - \alpha)n])})$$

for $j = 1, 2, \dots, n - [(1 - \alpha)n]$.

- The $100(1 - \alpha)\%$ HPD interval, denoted by $R_{j^*}(n)$, is the one with the smallest interval width among all credible intervals.

Deviance Information Criterion (DIC)

The deviance information criterion (DIC) (Spiegelhalter et al. 2002) is a model assessment tool, and it is a Bayesian alternative to Akaike’s information criterion (AIC) and the Bayesian information criterion (BIC, also known as the Schwarz criterion). The DIC uses the posterior densities, which means that it takes the prior information into account. The criterion can be applied to non-nested models and models that have non-iid data. Calculation of the DIC in MCMC is trivial—it does not require maximization over the parameter space, like the AIC and BIC. A smaller DIC indicates a better fit to the data set.

Letting $\boldsymbol{\theta}$ be the parameters of the model, the deviance information formula is

$$\text{DIC} = \overline{D(\boldsymbol{\theta})} + p_D = D(\bar{\boldsymbol{\theta}}) + 2p_D$$

where

$$D(\boldsymbol{\theta}) = 2(\log(f(\mathbf{y})) - \log(p(\mathbf{y}|\boldsymbol{\theta}))) : \text{deviance}$$

where

$p(\mathbf{y}|\boldsymbol{\theta})$: likelihood function with the normalizing constants.

$f(\mathbf{y})$: a standardizing term that is a function of the data alone. This term is constant with respect to the parameter and is irrelevant when you compare different models that have the same likelihood function. Since the term cancels out in DIC comparisons, its calculation is often omitted.

Note: You can think of the deviance as the difference in twice the log likelihood between the saturated, $f(\mathbf{y})$, and fitted, $p(\mathbf{y}|\boldsymbol{\theta})$, models.

$\bar{\boldsymbol{\theta}}$: posterior mean, approximated by $\frac{1}{n} \sum_{t=1}^n \boldsymbol{\theta}^t$

$\overline{D(\boldsymbol{\theta})}$: posterior mean of the deviance, approximated by $\frac{1}{n} \sum_{t=1}^n D(\boldsymbol{\theta}^t)$. The expected deviation measures how well the model fits the data.

$D(\bar{\boldsymbol{\theta}})$: deviance evaluated at $\bar{\boldsymbol{\theta}}$, equal to $-2 \log(p(\mathbf{y}|\bar{\boldsymbol{\theta}}))$. It is the deviance evaluated at your “best” posterior estimate.

p_D : effective number of parameters. It is the difference between the measure of fit and the deviance at the estimates: $\overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}})$. This term describes the complexity of the model, and it serves as a penalization term that corrects deviance’s propensity toward models with more parameters.

A Bayesian Reading List

This section lists a number of Bayesian textbooks of various difficulty degrees and a few tutorial/review papers.

Textbooks

Introductory Books

- Berry, D. A. (1996), *Statistics: A Bayesian Perspective*, London: Duxbury Press.
- Bolstad, W. M. (2004), *Introduction to Bayesian Statistics*, New York: John Wiley & Sons.
- DeGroot, M. H. and Schervish, M. J. (2002) *Probability and Statistics*, Reading, MA: Addison Wesley.
- Gamerman, D. and Lopes, H. F. (2006), *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, 2nd ed. London: Chapman & Hall/CRC.
- Ghosh, J. K., Delampady, M. and Samanta, T. (2006), *An Introduction to Bayesian Analysis*, New York: Springer-Verlag.
- Lee, P. M. (2004), *Bayesian Statistics: An Introduction*, 3rd ed. London: Arnold.
- Sivia, D. S. (1996), *Data Analysis: A Bayesian Tutorial*, Oxford: Oxford University Press.

Intermediate-Level Books

- Box, G. E. P., and Tiao, G. C. (1992), *Bayesian Inference in Statistical Analysis*, New York: John Wiley & Sons.
- Chen, M. H., Shao Q. M., and Ibrahim, J. G. (2000), *Monte Carlo Methods in Bayesian Computation*, New York: Springer-Verlag.
- Harney, H. L. (2003), *Bayesian Inference: Parameter Estimation and Decisions*, New York: Springer-Verlag.
- Leonard, T. and Hsu, J. S. (1999), *Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers*, Cambridge: Cambridge University Press.

Liu, J. S. (2001), *Monte Carlo Strategies in Scientific Computing*, New York: Springer-Verlag.

Press, S. J. (2002), *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*, 2nd ed. New York: Wiley-Interscience.

Robert, C. P. (2001), *The Bayesian Choice*, 2nd ed. New York: Springer-Verlag.

Robert, C. P. and Casella, G. (2004), *Monte Carlo Statistical Methods*, 2nd ed. New York: Springer-Verlag.

Tanner, M. A. (1993), *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, New York: Springer-Verlag.

Advanced Titles

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, New York: Springer-Verlag.

Bernardo, J. M. and Smith, A. F. M. (2007), *Bayesian Theory*, 2nd ed. New York: John Wiley & Sons.

de Finetti, B. (1992), *Theory of Probability*, New York: John Wiley & Sons.

Jeffreys, H. (1998), *Theory of Probability*, Oxford: Oxford University Press.

O'Hagan, A. (1994), *Bayesian Inference*, volume 2B of *Kendall's Advanced Theory of Statistics*, London: Arnold.

Savage, L. J. (1954), *The Foundations of Statistics*, New York: John Wiley & Sons.

Books Motivated by Statistical Applications and Data Analysis

Carlin, B. and Louis, T. A. (2000), *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed. London: Chapman & Hall.

Congdon, P. (2001), *Bayesian Statistical Modeling*, New York: John Wiley & Sons.

Congdon, P. (2003), *Applied Bayesian Modeling*, New York: John Wiley & Sons.

Congdon, P. (2005), *Bayesian Models for Categorical Data*, New York: John Wiley & Sons.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis*, 3rd ed. London: Chapman & Hall.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996), *Markov Chain Monte Carlo in Practice*, London: Chapman & Hall.

Tutorial and Review Papers on MCMC

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995), "Bayesian Computation and Stochastic Systems," *Statistical Science*, 10(1), 3–66.

Cappè, O. and Robert, C. P. (2000), "Markov Chain Monte Carlo, 10 Years and Still Running!" *Journal of the American Statistical Association*, 95, 1282–1286.

Casella, G. and George, E. (1992), “Explaining the Gibbs Sampler,” *The American Statistician*, 46, 167–174.

Chib, S. and Greenberg, E. (1995), “Understanding the Metropolis-Hastings Algorithm,” *The American Statistician*, 49, 327–335.

Chib, S. and Greenberg, E. (1996), “Markov Chain Monte Carlo Simulation Methods in Econometrics,” *Econometric Theory*, 12, 409–431.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1998), “Markov Chain Monte Carlo in Practice: A Roundtable Discussion,” *Statistical Science*, 52(2), 93–100.

Acknowledgments

We are very grateful to Joseph G. Ibrahim, University of North Carolina at Chapel Hill, for the extensive time and effort that he generously contributed in providing input throughout the development of this software. His thoughtful insights and guidance on key issues of methodology and implementation were instrumental. We are also grateful to Christopher J. Paciorek, Harvard University, for his valuable feedback on the documentation.

References

Amit, Y. (1991), “On Rates of Convergence of Stochastic Relaxation for Gaussian and Non-Gaussian Distributions,” *Journal of Multivariate Analysis*, 38, 82–99.

Applegate, D., Kannan, R., and Polson, N. (1990), *Random Polynomial Time Algorithms for Sampling from Joint Distributions*, Technical report, School of Computer Science, Carnegie Mellon University.

Berger, J. (1985), *Statistical Decision Theory and Bayesian Analysis*, Second Edition, New York: Springer-Verlag.

Berger, J. and Wolpert, R. (1988), *The Likelihood Principle*, 9, Second Edition, Hayward, California: Institute of Mathematical Statistics, monograph series.

Berger, J. O. (2006), “The Case for Objective Bayesian Analysis,” *Bayesian Analysis*, 3, 385–402, <http://ba.stat.cmu.edu/journal/2006/vol01/issue03/berger.pdf>.

Bernardo, J. and Smith, A. (1994), *Bayesian Theory*, New York: John Wiley & Sons.

Besag, J. (1974), “Spatial Interaction and the Statistical Analysis of Lattice Systems,” *Journal of the Royal Statistical Society B*, 36, 192–326.

Billingsley, P. (1986), *Probability and Measure*, Second Edition, New York: John Wiley & Sons.

Box, G. E. and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, Wiley Classics Library Edition, published 1992, New York: John Wiley & Sons.

Breiman, L. (1968), *Probability*, Reading, MA: Addison-Wesley.

- Brooks, S. and Gelman, A. (1997), “General Methods for Monitoring Convergence of Iterative Simulations,” *Journal of Computational and Graphical Statistics*, 7, 434–455.
- Brooks, S. and Roberts, G. (1999), “On Quantile Estimation and Markov Chain Monte Carlo Convergence,” *Biometrika*, 86, 710–717.
- Brooks, S. P. and Roberts, G. O. (1998), “Assessing Convergence of Markov Chain Monte Carlo Algorithms,” *Statistics and Computing*, 8, 319–335.
- Carlin, B. P. and Louis, T. A. (2000), *Bayes and Empirical Bayes Methods for Data Analysis*, Second Edition, London: Chapman & Hall.
- Casella, G. and George, E. (1992), “Explaining the Gibbs Sampler,” *The American Statistician*, 46, 167–174.
- Chan, K. (1993), “Asymptotic Behavior of the Gibbs Sampler,” *Journal of the American Statistical Association*, 88, 320–326.
- Chen, M., Shao, Q., and Ibrahim, J. (2000), *Monte Carlo Methods in Bayesian Computation*, New York: Springer-Verlag.
- Chen, M. H. and Shao, Q. M. (1999), “Monte Carlo Estimation of Bayesian Credible and HPD Intervals,” *Journal of Computational and Graphical Statistics*, 8, 69–92.
- Chib, S. and Greenberg, E. (1995), “Understanding the Metropolis-Hastings Algorithm,” *The American Statistician*, 49, 327–335.
- Congdon, P. (2001), *Bayesian Statistical Modeling*, John Wiley & Sons.
- Congdon, P. (2003), *Applied Bayesian Modeling*, John Wiley & Sons.
- Congdon, P. (2005), *Bayesian Models for Categorical Data*, John Wiley & Sons.
- Cowles, M. K. and Carlin, B. P. (1996), “Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review,” *Journal of the American Statistical Association*, 883–904.
- DeGroot, M. and Schervish, M. (2002), *Probability and Statistics*, 34th Edition, Reading, MA: Addison-Wesley.
- Feller, W. (1968), *An Introduction to Probability Theory and Its Applications*, Third Edition, New York: John Wiley & Sons.
- Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. (1990), “Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling,” *Journal of the American Statistical Association*, 85, 972–985.
- Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004), *Bayesian Data Analysis*, Second Edition, London: Chapman & Hall.
- Gelman, A. and Rubin, D. B. (1992), “Inference from Iterative Simulation Using Multiple Sequences,” *Statistical Science*, 7, 457–472.

- Geman, S. and Geman, D. (1984), “Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 6, 721–741.
- Geweke, J. (1992), “Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments,” in J. Bernardo, J. Berger, A. Dawid, and A. Smith, eds., “Bayesian Statistics,” Vol. 4, Oxford, UK: Clarendon Press.
- Gilks, W. (2003), “Adaptive Metropolis Rejection Sampling (ARMS),” software from MRC Biostatistics Unit, Cambridge, UK, http://www.maths.leeds.ac.uk/~wally.gilks/adaptive.rejection/web_page/Welcome.html.
- Gilks, W., Best, N., and Tan, K. (1995), “Adaptive Rejection Metropolis Sampling with Gibbs Sampling,” *Applied Statistics*, 44, 455–472.
- Gilks, W., Richardson, S., and Spiegelhalter, D. (1996), *Markov Chain Monte Carlo in Practice*, London: Chapman & Hall.
- Gilks, W. and Wild, P. (1992), “Adaptive Rejection Sampling for Gibbs Sampling,” *Applied Statistics*, 41, 337–348.
- Goldstein, M. (2006), “Subjective Bayesian Analysis: Principles and Practice,” *Bayesian Analysis*, 3, 403–420, <http://ba.stat.cmu.edu/journal/2006/vol01/issue03/goldstein.pdf>.
- Hastings, W. (1970), “Monte Carlo Sampling Methods Using Markov Chains and their Applications,” *Biometrika*, 57, 97–109.
- Heidelberger, P. and Welch, P. (1981), “A Spectral Method for Confidence Interval Generation and Run Length Control in Simulations.” *Communication of the ACM*, 24, 233–245.
- Heidelberger, P. and Welch, P. (1983), “Simulation Run Length Control in the Presence of an Initial Transient,” *Operation Research*, 31, 1109–1144.
- Karlin, S. and Taylor, H. (1975), *A First Course in Stochastic Processes*, Second Edition, Orlando, FL: Academic Press.
- Kass, R., Carlin, B., Gelman, A., and Neal, R. (1998), “Markov Chain Monte Carlo in Practice: A Roundtable Discussion,” *The American Statistician*, 52, 93–100.
- Kass, R. E. and Wasserman, L. (1996), “Formal Rules of Selecting Prior Distributions: A Review and Annotated Bibliography,” *Journal of the American Statistical Association*, 91, 343–370.
- Liu, C., Wong, W., and Kong, A. (1991a), *Correlation Structure and Convergence Rate of the Gibbs Sampler (I): Application to the Comparison of Estimators and Augmentation Scheme*, Technical report, Department of Statistics, University of Chicago.
- Liu, C., Wong, W., and Kong, A. (1991b), *Correlation Structure and Convergence Rate of the Gibbs Sampler (II): Applications to Various Scans*, Technical report, Department of Statistics, University of Chicago.

- Liu, J. (2001), *Monte Carlo Strategies in Scientific Computing*, Springer-Verlag.
- MacEachern, S. and Berliner, L. (1994), “Subsampling the Gibbs Sampler,” *The American Statistician*, 48, 188–190.
- Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953), “Equation of State Calculations by Fast Computing Machines,” *Journal of Chemical Physics*, 21, 1087–1092.
- Metropolis, N. and Ulam, S. (1949), “The Monte Carlo Method,” *Journal of the American Statistical Association*, 44.
- Meyn, S. and Tweedie, R. (1993), *Markov Chains and Stochastic Stability*, Berlin: Springer-Verlag.
- Press, S. (2003), *Subjective and Objective Bayesian Statistics*, New York: Wiley.
- Raftery, A. and Lewis, S. (1992), “One Long Run with Diagnostics: Implementation Strategies for Markov Chain Monte Carlo,” *Statistical Science*, 7, 493–497.
- Raftery, A. and Lewis, S. (1995), “The Number of Iterations, Convergence Diagnostics and Generic Metropolis Algorithms,” in W. Gilks, D. Spiegelhalter, and S. Richardson, eds., “Practical Markov Chain Monte Carlo,” London, UK: Chapman & Hall.
- Robert, C. and Casella, G. (2004), *Monte Carlo Statistical Methods*, Second Edition, New York: Springer-Verlag.
- Robert, C. P. (2001), *The Bayesian Choice*, Second Edition, New York: Springer-Verlag.
- Roberts, G. (1996), “Markov Chain Concepts Related to Sampling Algorithms,” in W. Gilks, D. Spiegelhalter, and S. Richardson, eds., “Markov Chain Monte Carlo in Practice,” 45–58, London: Chapman & Hall.
- Rosenthal, J. (1991a), *Rates of Convergence for Data Augmentation on Finite Sample Spaces*, Technical report, Department of Mathematics, Harvard University.
- Rosenthal, J. (1991b), *Rates of Convergence for Gibbs Sampling for Variance Component Models*, Technical report, Department of Mathematics, Harvard University.
- Ross, S. (1997), *Simulation*, Second Edition, Orlando, FL: Academic Press.
- Schervish, M. and Carlin, B. (1992), “On the Convergence of Successive Substitution Sampling,” *Journal of Computational and Graphical Statistics*, 1, 111–127.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der Linde, A. (2002), “Bayesian Measures of Model Complexity and Fit (with Discussion),” *Journal of the Royal Statistical Society, Series B*, 64(4), 583–616.
- Tanner, M. (1993), *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, New York: Springer-Verlag.

Tanner, M. A. and Wong, W. H. (1987), “The Calculation of Posterior Distributions by Data Augmentation,” *Journal of the American Statistical Association*, 82, 528–550.

Tierney, L. (1994), “Markov Chains for Exploring Posterior Distributions,” *Annals of Statistics*, 22, 1701–1762.

von Mises, R. (1918), “Über die ‘Ganzzahligkeit’ der Atomgewicht und verwandte Fragen,” *Physikal. Z.*, 19, 490–500.

Wasserman, L. (2004), *All of Statistics: A Concise Course in Statistical Inference*, New York: Springer-Verlag.

Chapter 2

The BGENMOD Procedure (Experimental)

Chapter Contents

OVERVIEW	43
GETTING STARTED	44
Surgical Unit Data	44
SYNTAX	53
BAYES Statement	53
DETAILS	61
Gibbs Sampling	61
Priors for Model Parameters	63
Posterior Distribution	64
Starting Values of the Markov Chains	64
Displayed Output	65
ODS Table Names	67
ODS Graph Names	68
EXAMPLE	68
REFERENCES	79

Chapter 2

The BGENMOD Procedure

(Experimental)

Overview

The BGENMOD procedure adds Bayesian analysis by Gibbs sampling to the GENMOD procedure, which fits generalized linear models. Bayesian analysis of generalized linear models can be requested by using the BAYES statement in the BGENMOD procedure. These Bayesian capabilities will be included in PROC GENMOD in the next release of SAS/STAT software.

In Bayesian analysis, the model parameters are treated as random variables, and inference about parameters is based on the posterior distribution of the parameters, given the data. The posterior distribution is obtained using Bayes' theorem as the likelihood function of the data weighted with a prior distribution. The prior distribution enables you to incorporate knowledge or experience of the likely range of values of the parameters of interest into the analysis. If you have no prior knowledge of the parameter values, you can use a noninformative prior distribution, and the results of the Bayesian analysis will be very similar to a classical analysis based on maximum likelihood. A closed form of the posterior distribution is often not feasible, and a Markov chain Monte Carlo method by Gibbs sampling is used to simulate samples from the posterior distribution. See [Chapter 1, "Introduction to Bayesian Analysis Procedures,"](#) for an introduction to the basic concepts of Bayesian statistics. Also see the section ["Bayesian Analysis: Advantages and Disadvantages"](#) on page 11 for a discussion of the advantages and disadvantages of Bayesian analysis.

A Gibbs chain for the posterior distribution is generated for the model parameters. Summary statistics (mean, standard deviation, quartiles, HPD and credible intervals, correlation matrix) and convergence diagnostics (autocorrelations; Gelman-Rubin, Geweke, Raftery-Lewis, and Heidelberger and Welch tests; and the effective sample size) are computed for each parameter as well as the correlation matrix of the posterior sample. Trace plots, posterior density plots, and autocorrelation function plots that use ODS graphics are also provided for each parameter.

Note that the full functionality of the GENMOD procedure, as documented in the SAS/STAT 9.1 documentation, is included in the BGENMOD procedure.

We are eager for your feedback on this experimental procedure. Please send comments to bgenmod@sas.com.

Getting Started

Surgical Unit Data

If you are not familiar with Bayesian analysis, see [Chapter 1, “Introduction to Bayesian Analysis Procedures,”](#) for an introduction to setting up and interpreting the results of a Bayesian analysis.

Except for the new BAYES statement, PROC BGENMOD uses the same syntax as the GENMOD procedure. For many applications, only the PROC BGENMOD, MODEL, and BAYES statements are required.

[Neter and Wasserman \(1974\)](#) describe a study of 54 patients undergoing a certain kind of liver operation in a surgical unit. The data set `Surg`, shown in [Figure 2.1](#), contains the survival time and certain covariates. Consider the model

$$Y = \beta_0 + \beta_1 \text{Log}X_1 + \epsilon$$

where Y is the survival time, $\text{Log}X_1$ is $\log(\text{blood-clotting score})$, and ϵ is a $N(0, \sigma^2)$ error term.

The variables `Y`, `X1` in the data set correspond to Y and X_1 , and `LogX1` is $\log(X_1)$.

Obs	x1	x2	x3	x4	y	logy	id	logx1
1	6.7	62	81	2.59	199.986	2.3010	1	1.90211
2	5.1	59	66	1.70	100.995	2.0043	2	1.62924
3	7.4	57	83	2.16	203.986	2.3096	3	2.00148
4	6.5	73	41	2.01	100.995	2.0043	4	1.87180
5	7.8	65	115	4.30	508.979	2.7067	5	2.05412
6	5.8	38	72	1.42	80.002	1.9031	6	1.75786
7	5.7	46	63	1.91	80.002	1.9031	7	1.74047
8	3.7	68	81	2.57	126.999	2.1038	8	1.30833
9	6.0	67	93	2.50	202.023	2.3054	9	1.79176
10	3.7	76	94	2.40	203.002	2.3075	10	1.30833
11	6.3	84	83	4.13	329.003	2.5172	11	1.84055
12	6.7	51	43	1.86	64.998	1.8129	12	1.90211
13	5.8	96	114	3.95	830.042	2.9191	13	1.75786
14	5.8	83	88	3.95	329.989	2.5185	14	1.75786
15	7.7	62	67	3.40	167.996	2.2253	15	2.04122
16	7.4	74	68	2.40	217.020	2.3365	16	2.00148
17	6.0	85	28	2.98	86.996	1.9395	17	1.79176
18	3.7	51	41	1.55	34.002	1.5315	18	1.30833
19	7.3	68	74	3.56	214.981	2.3324	19	1.98787
20	5.6	57	87	3.02	171.989	2.2355	20	1.72277
21	5.2	52	76	2.85	108.993	2.0374	21	1.64866
22	3.4	83	53	1.12	135.988	2.1335	22	1.22378
23	6.7	26	68	2.10	70.000	1.8451	23	1.90211
24	5.8	67	86	3.40	219.989	2.3424	24	1.75786
25	6.3	59	100	2.95	275.994	2.4409	25	1.84055
26	5.8	61	73	3.50	144.012	2.1584	26	1.75786
27	5.2	52	86	2.45	181.009	2.2577	27	1.64866
28	11.2	76	90	5.59	573.984	2.7589	28	2.41591
29	5.2	54	56	2.71	71.995	1.8573	29	1.64866
30	5.8	76	59	2.58	177.992	2.2504	30	1.75786
31	3.2	64	65	0.74	71.007	1.8513	31	1.16315
32	8.7	45	23	2.52	57.996	1.7634	32	2.16332
33	5.0	59	73	3.50	116.011	2.0645	33	1.60944
34	5.8	72	93	3.30	294.985	2.4698	34	1.75786
35	5.4	58	70	2.64	115.001	2.0607	35	1.68640
36	5.3	51	99	2.60	183.992	2.2648	36	1.66771
37	2.6	74	86	2.05	118.005	2.0719	37	0.95551
38	4.3	8	119	2.85	120.005	2.0792	38	1.45862
39	4.8	61	76	2.45	151.008	2.1790	39	1.56862
40	5.4	52	88	1.81	148.013	2.1703	40	1.68640
41	5.2	49	72	1.84	94.995	1.9777	41	1.64866
42	3.6	28	99	1.30	75.007	1.8751	42	1.28093
43	8.8	86	88	6.40	483.059	2.6840	43	2.17475
44	6.5	56	77	2.85	153.003	2.1847	44	1.87180
45	3.4	77	93	1.48	190.985	2.2810	45	1.22378
46	6.5	40	84	3.00	122.999	2.0899	46	1.87180
47	4.5	73	106	3.05	311.028	2.4928	47	1.50408
48	4.8	86	101	4.10	398.016	2.5999	48	1.56862
49	5.1	67	77	2.86	158.016	2.1987	49	1.62924
50	3.9	82	103	4.55	310.027	2.4914	50	1.36098
51	6.6	77	46	1.95	123.994	2.0934	51	1.88707
52	6.4	85	40	1.21	124.997	2.0969	52	1.85630
53	6.4	59	85	2.33	198.016	2.2967	53	1.85630
54	8.8	78	72	3.20	312.968	2.4955	54	2.17475

Figure 2.1. Surgical Unit Data

A question of scientific interest is whether blood clotting score has a positive effect on survival time. Using PROC GENMOD, you can obtain a point estimate of the coefficient and construct a null point hypothesis to test whether β_1 is equal to 0. However, if you are interested in finding the probability that the coefficient is positive,

Bayesian analysis offers a convenient alternative. You can use Bayesian analysis to directly estimate the conditional probability, $\Pr(\beta_1 > 0|Y)$, using the posterior distribution samples, which are produced as part of the output by BGENMOD.

The example that follows shows how to use PROC BGENMOD to carry out a Bayesian analysis of the linear model with a normal error term. The SEED= option is specified to maintain reproducibility; no other options are specified in the BAYES statement. By default, a uniform prior distribution is assumed on the regression coefficients. The uniform prior is a flat prior on the real line with a distribution that reflects ignorance of the location of the parameter, placing equal likelihood on all possible values the regression coefficient can take. Using the uniform prior in the following example, you would expect the Bayesian estimates to resemble the classical results of maximizing the likelihood. If you can elicit an informative prior on the regression coefficients, you should use the COEFFPRIOR= option to specify it. A default noninformative gamma prior is used for the scale parameter σ .

You should make sure that the posterior distribution samples have achieved convergence before using them for Bayesian inference. PROC BGENMOD produces three convergence diagnostics by default. If ODS graphics are enabled as specified in the following SAS statements, diagnostic plots are also displayed.

Summary statistics of the posterior distribution samples are produced by default. However, these statistics might not be sufficient for carrying out your Bayesian inference. The BAYES statement in the following SAS statements invokes the Bayesian analysis, and the ODS OUTPUT statement saves the samples in the SAS data set Post for further processing:

```
ods graphics on;
proc bgenmod data=Surg;
  model y = logx1 / d=normal;
  bayes seed=1;
run;
ods output PosteriorSample=Post;
ods graphics off;
```

The results of this analysis are shown in the following figures.

The “Model Information” table in [Figure 2.2](#) summarizes information about the model you fit and the size of the simulation.

Model Information		
Data Set	WORK.SURG	
Burn-In Size	2000	
MC Sample Size	10000	
Thinning	1	
Distribution	Normal	
Link Function	Identity	
Dependent Variable	y	Survival Time

Figure 2.2. Model Information

The “Analysis of Maximum Likelihood Parameter Estimates” table in [Figure 2.3](#) summarizes maximum likelihood estimates of the model parameters. In the table, **Intercept** represents the MLE of β_0 in the model, **logx1** represents the MLE of β_1 , and **Scale** represents the MLE of σ .

Analysis Of Maximum Likelihood Parameter Estimates					
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	
Intercept	1	-94.9822	114.5279	-319.453	129.4884
logx1	1	170.1749	65.8373	41.1361	299.2137
Scale	1	135.7963	13.0670	112.4556	163.9815

NOTE: The scale parameter was estimated by maximum likelihood.

Figure 2.3. Maximum Likelihood Parameter Estimates

Since no prior distributions for the regression coefficients were specified, the default noninformative uniform distributions shown in the “Uniform Prior for Regression Coefficients” table in [Figure 2.4](#) are used. Noninformative priors are appropriate if you have no prior knowledge of the likely range of values of the parameters, and if you want to make probability statements about the parameters or functions of the parameters. See, for example, [Ibrahim, Chen, and Sinha \(2001\)](#) for more information about choosing prior distributions.

Uniform Prior for Regression Coefficients	
Parameter	Prior
Intercept	Constant
logx1	Constant

Figure 2.4. Regression Coefficient Priors

The default noninformative gamma prior distribution for the normal scale parameter is shown in the “Independent Prior Distributions for Model Parameters” table in [Figure 2.5](#).

Independent Prior Distributions for Model Parameters			
Parameter	Prior Distribution	Hyperparameters	
		Shape	Inverse Scale
Scale	Gamma	0.001	0.001

Figure 2.5. Scale Parameter Prior

By default, the maximum likelihood estimates of the regression parameters are used as the starting values for the simulation. These are listed in the “Initial Values and Seeds” table in [Figure 2.6](#).

Initial Values and Seeds				
Chain	_SEED_	Intercept	logx1	Scale
1	1	-94.9822	170.1749	135.7901

Figure 2.6. MCMC Initial Values and Seeds

Summary statistics for the posterior sample are displayed in the “Descriptive Statistics of the Posterior Samples” and “Interval Statistics of the Posterior Samples” tables in [Figure 2.7](#) and [Figure 2.8](#), respectively. Since noninformative prior distributions were used, these results are consistent with the maximum likelihood estimates shown in [Figure 2.3](#).

Descriptive Statistics of the Posterior Samples						
Parameter	N	Mean	Standard Deviation	25%	Quantiles 50%	75%
Intercept	10000	-78.5665	116.7	-153.9	-74.8423	-1.8472
logx1	10000	160.8	67.1044	116.7	159.3	203.7
Scale	10000	140.2	13.9387	130.3	139.1	148.8

Figure 2.7. Posterior Sample Summary Statistics

Interval Statistics of the Posterior Samples					
Parameter	Alpha	Credible Interval		HPD Interval	
Intercept	0.050	-314.2	143.9	-315.8	140.9
logx1	0.050	32.1162	296.2	32.6623	296.5
Scale	0.050	116.3	170.9	115.0	168.6

Figure 2.8. Posterior Sample Interval Statistics

By default, PROC BGENMOD computes three convergence diagnostics: the lag1, lag5, lag10, and lag50 autocorrelations ([Figure 2.9](#)); the Geweke diagnostic ([Figure 2.10](#)); and the effective sample size ([Figure 2.11](#)). There is no indication that the Markov chain has not converged. See the “[Assessing Markov Chain Convergence](#)” section on page 17 for more information about convergence diagnostics and their interpretation.

Autocorrelations of the Posterior Samples				
Parameter	Lag1	Lag5	Lag10	Lag50
Intercept	0.9728	0.8704	0.7522	0.2245
logx1	0.9726	0.8696	0.7516	0.2253
Scale	0.0334	0.0069	-0.0079	0.0019

Figure 2.9. Posterior Sample Autocorrelations

Geweke Diagnostics		
Parameter	z	Pr > z
Intercept	0.4912	0.6233
logx1	-0.4424	0.6582
Scale	-1.3765	0.1687

Figure 2.10. Geweke Diagnostic Statistics

Effective Sample Size			
Parameter	ESS	Correlation	
		Time	Efficiency
Intercept	133.3	74.9923	0.0133
logx1	133.4	74.9595	0.0133
Scale	8527.3	1.1727	0.8527

Figure 2.11. Effective Sample Sizes

Trace, autocorrelation, and density plots for the three model parameters shown in [Figure 2.12](#), [Figure 2.13](#), and [Figure 2.14](#) are useful in diagnosing whether the Markov chain of posterior samples has converged. These plots show no evidence that the chain has not converged.

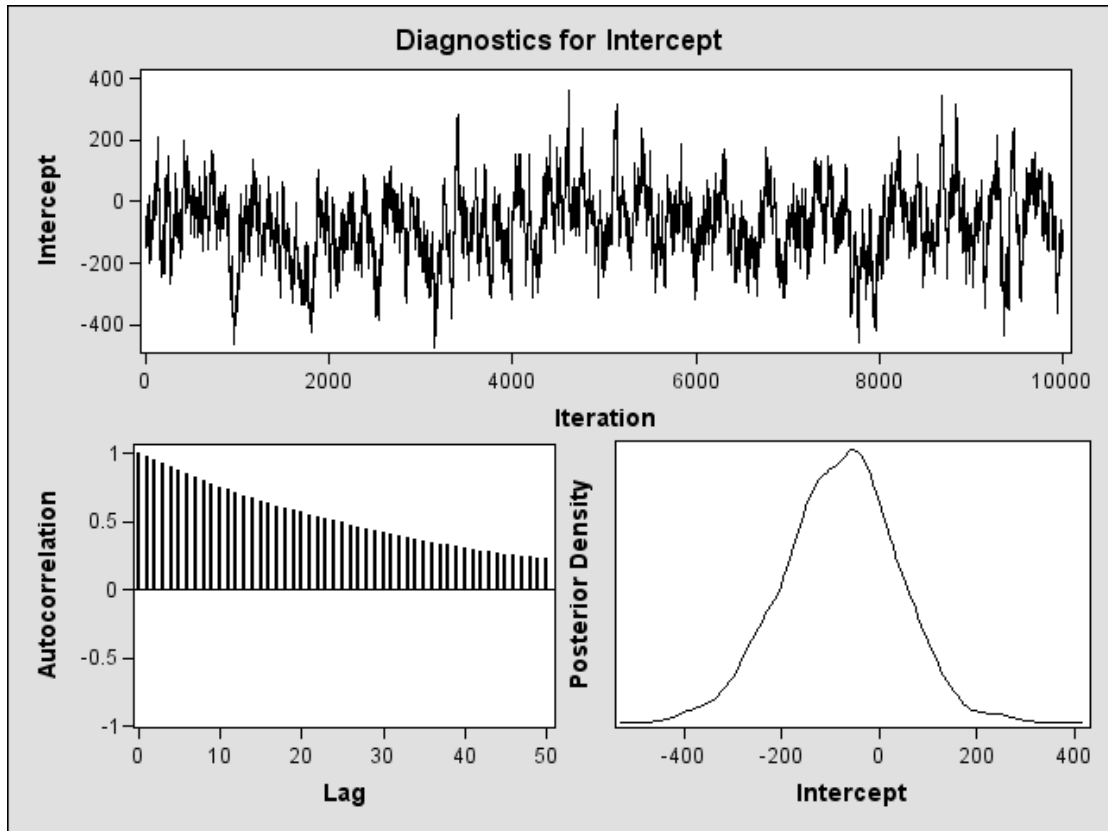


Figure 2.12. Diagnostic Plots for Intercept

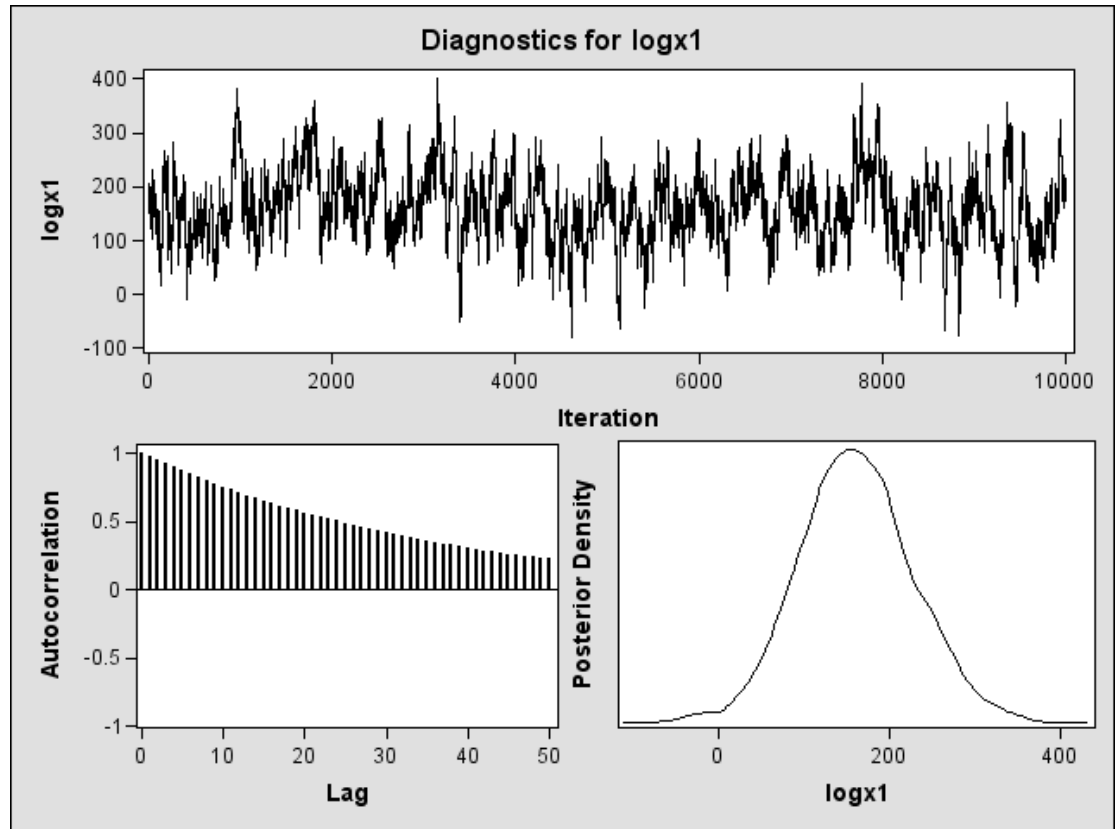


Figure 2.13. Diagnostic Plots for LogX1

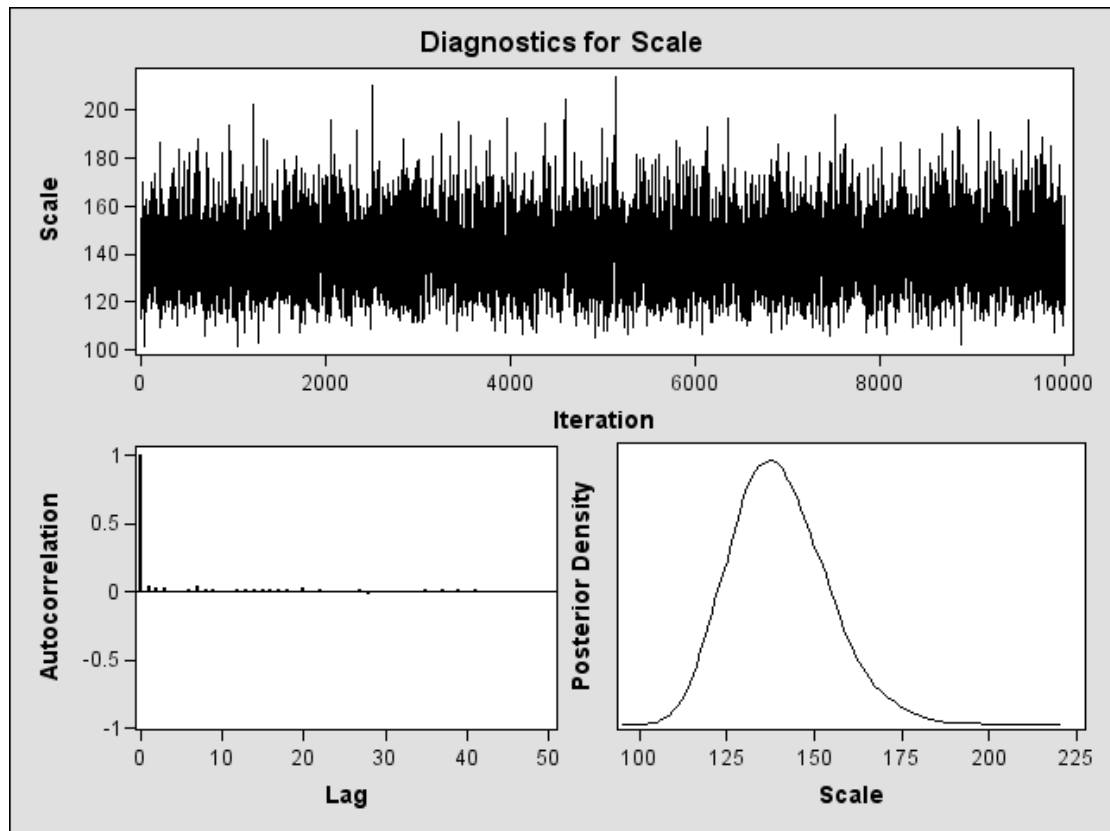


Figure 2.14. Diagnostic Plots for Scale

A question of scientific interest is whether blood clotting score has a positive effect on survival time. Since the model parameters are regarded as random quantities in a Bayesian analysis, you can answer this question by estimating the conditional probability of β_1 being positive, given the data, $\Pr(\beta_1 > 0 | \mathbf{Y})$, from the posterior distribution samples. The following SAS statements produce these estimates:

```
data Prob;
  set Post;
  Indicator = (logX1 > 0);
  label Indicator= 'log(Blood Clotting Score) > 0';
run;

proc means data = Prob(keep=Indicator) n mean;
run;
```

There is a 0.99 probability of a positive relationship between the logarithm of a blood clotting score and survival time.

The MEANS Procedure	
Analysis Variable : Indicator log(Blood Clotting Score) > 0	
N	Mean
10000	0.9896000

Figure 2.15. Probability That $\beta_1 > 0$

Syntax

The BGENMOD procedure has the same syntax as PROC GENMOD with the addition of the BAYES statement for requesting a Bayesian analysis.

To request a Bayesian analysis, you specify the new BAYES statement in addition to the PROC BGENMOD statement and the MODEL statement. You include a CLASS statement if you have effects that involve categorical variables. The FREQ or WEIGHT statement can be included if you have a frequency or weight variable in the input data.

The BY and REPEATED statements are not allowed when the BAYES statement is specified in PROC BGENMOD, and the multinomial distribution is not supported.

BAYES Statement

BAYES < options > ;

The BAYES statement requests a Bayesian analysis of the regression model by using Gibbs sampling. The Bayesian posterior samples (also known as the chain) for the regression parameters are not tabulated. In the following, “PosteriorSample” is the name of a nonprinting ODS table that is produced by default whenever the BAYES statement is used. You can create an ODS output data set (named *SAS-data-set*) of the chain by specifying the following:

ODS OUTPUT PosteriorSample = SAS-data-set ;

Table 2.1 summarizes the options available in the BAYES statement.

Table 2.1. BAYES Statement Options

Option	Description
Monte Carlo Options	
INITIAL=	specifies initial values of the chain
NBI=	specifies the number of burn-in iterations
NMC=	specifies the number of iterations after burn-in
SEED=	specifies the random number generator seed
THINNING=	controls the thinning of the Markov chain
Model and Prior Options	
COEFFPRIOR=	specifies the prior of the regression coefficients

Table 2.1. (continued)

Option	Description
DISPERSIONPRIOR=	specifies the prior of the dispersion parameter
PRECISIONPRIOR=	specifies the prior of the precision parameter
SCALEPRIOR=	specifies the prior of the scale parameter
Summary Statistics and Convergence Diagnostics	
DIAGNOSTICS=	displays convergence diagnostics
PLOTS=	displays diagnostic plots
SUMMARY=	displays summary statistics of the posterior samples

The following list describes these options and their suboptions.

COEFFPRIOR=JEFFREYS<(option)> | **NORMAL**<(options)> | **UNIFORM**
COEFF=JEFFREYS<(option)> | **NORMAL**<(options)> | **UNIFORM**

specifies the prior distribution for the regression coefficients. The default is COEFFPRIOR=UNIFORM, which specifies the noninformative and improper prior of a constant.

Jeffreys' prior is specified by COEFFPRIOR=JEFFREYS, which can be followed by the following option in parentheses. Jeffreys' prior is proportional to $|I(\beta)|^{\frac{1}{2}}$, where $I(\beta)$ is the Fisher information matrix. See the “Jeffreys' Prior” section on page 63 and Ibrahim and Laud (1991) for more details.

CONDITIONAL

specifies that the Jeffreys' prior, conditional on the current Markov chain value of the generalized linear model precision parameter τ , is proportional to $|\tau I(\beta)|^{\frac{1}{2}}$.

The normal prior is specified by COEFFPRIOR=NORMAL, which can be followed by one of the following options enclosed in parentheses. However, if you do not specify an option, the normal prior $N(\mathbf{0}, 10^6\mathbf{I})$, where \mathbf{I} is the identity matrix, is used. See the “Normal Prior” section on page 64 for more details.

CONDITIONAL

specifies that the normal prior, conditional on the current Markov chain value of the generalized linear model precision parameter τ , is $N(\boldsymbol{\mu}, \tau^{-1}\boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and covariance of the normal prior specified by other normal options.

INPUT=SAS-data-set

specifies a SAS data set containing the mean and covariance information of the normal prior. The data set must have a `_TYPE_` variable to represent the type of each observation and a variable for each regression coefficient. If the data set also contains a `_NAME_` variable, the values of this variable are used to identify the covariances for the `_TYPE_='COV'` observations; otherwise, the `_TYPE_='COV'` observations are assumed to be in the same order as the explanatory variables in the MODEL statement. PROC BGENMOD

reads the mean vector from the observation with `_TYPE_='MEAN'` and reads the covariance matrix from observations with `_TYPE_='COV'`. For an independent normal prior, the variances can be specified with `_TYPE_='VAR'`; alternatively, the precisions (inverse of the variances) can be specified with `_TYPE_='PRECISION'`.

`RELVAR<=c>`

specifies normal prior $N(\mathbf{0}, c\mathbf{J})$, where \mathbf{J} is a diagonal matrix with diagonal elements equal to the variances of the corresponding ML estimator. By default, $c=10^6$.

`VAR=c`

specifies the normal prior $N(\mathbf{0}, c\mathbf{I})$, where \mathbf{I} is the identity matrix.

DIAGNOSTICS=ALL | NONE | (*keyword-list*)

DIAG=ALL | NONE | (*keyword-list*)

controls the number of diagnostics produced. You can request all the following diagnostics by specifying `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 a subset of the following keywords. The default is `DIAGNOSTICS=(AUTOCORR ESS GEWEKE)`.

`AUTOCORR` computes the autocorrelations of lags 1, 5, 10, and 50 for each variable. See the “[Autocorrelations](#)” section on page 30 for details.

`ESS` computes Carlin’s estimate of the effective sample size, the correlation time, and the efficiency of the chain for each parameter. See the “[Effective Sample Size](#)” section on page 31 for details.

`HEIDELBERGER <(heidel-options)>`

computes the Heidelberg and Welch diagnostic for each variable, 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*:

`SALPHA=value`

specifies the α level ($0 < \alpha < 1$). for the stationarity test.

`HALPHA=value`

specifies the α level ($0 < \alpha < 1$). for the halfwidth test.

`EPS=value`

specifies a positive number ϵ such that if the halfwidth is less than ϵ times the sample mean of the retaining iterates, the halfwidth test is passed.

See the “[Heidelberg and Welch Diagnostics](#)” section on page 26 for details.

`GELMAN <(gelman-options)>`

computes the Gelman and Rubin convergence diagnostics. You can specify one or more of the following *gelman-options*:

NCHAIN | N=*number*

specifies the number of parallel chains used to compute the diagnostic, and has to be 2 or larger. The default is NCHAIN=3. If an INITIAL= data set is used, NCHAIN defaults to the number of rows in the INITIAL= data set. If any number other than this is specified with the NCHAIN= option, the NCHAIN= value is ignored.

ALPHA=*value*

specifies the significance level for the upper bound. The default is ALPHA=0.05, resulting in a 97.5% bound.

See the “Gelman and Rubin Diagnostics” section on page 22 for details.

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 default is $f_1 = 0.1$ 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.

See the “Geweke Diagnostics” section on page 24 for details.

RAFTERY<(raftery-options)>

computes the Raftery and Lewis diagnostics that evaluate the accuracy of the estimated quantile ($\hat{\theta}_Q$ for a given $Q \in (0, 1)$) of a chain. $\hat{\theta}_Q$ can achieve any degree of accuracy when the chain is allowed to run for a long time. A stopping criterion is when the estimated probability $\hat{P}_Q = \Pr(\theta \leq \hat{\theta}_Q)$ 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 *raftery-options* enable you to specify Q , R , S , and a precision level ϵ 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 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 tolerance level (a small positive number) for the stationary test. The default is 0.001.

See the “Raftery and Lewis Diagnostics” section on page 27 for details.

**DISPERSIONPRIOR=GAMMA<(options)> | IGAMMA<(options)> | IMPROPER
DPRIOR=GAMMA<(options)> | IGAMMA<(options)> | IMPROPER**

specifies that Gibbs sampling be performed on the generalized linear model dispersion parameter and the prior distribution for the dispersion parameter, if there is a dispersion parameter in the model. For models that do not have a dispersion parameter (the Poisson and binomial), this option is ignored. Note that you can specify Gibbs sampling on either the dispersion parameter ϕ , the scale parameter $\sigma = \phi^{\frac{1}{2}}$, or the precision parameter $\tau = \phi^{-1}$, with the DPRIOR=, SPRIOR=, and PPRIOR= options, respectively. These three parameters are transformations of one another, and you should specify Gibbs sampling for only one of them.

A gamma prior $G(a, b)$ with density $f(t) = \frac{b(bt)^{a-1}e^{-bt}}{\Gamma(a)}$ is specified by DISPERSIONPRIOR=GAMMA, which can be followed by one of the following *gamma-options* enclosed in parentheses. The hyperparameters a and b are the shape and inverse-scale parameters of the gamma distribution, respectively. See the “Gamma Prior” section on page 63. The default is $G(10^{-4}, 10^{-4})$.

RELSHAPE<=c>

specifies independent $G(c\hat{\phi}, c)$ distribution, where $\hat{\phi}$ is the MLE of the dispersion parameter. With this choice of hyperparameters, the mean of the prior distribution is $\hat{\phi}$ and the variance is $\frac{\hat{\phi}}{c}$. By default, $c=10^{-4}$.

SHAPE=a

and

ISCALE=b

specify the $G(a, b)$ prior.

SHAPE=c

specifies the $G(c, c)$ prior.

ISCALE=c

specifies the $G(c, c)$ prior.

An inverse gamma prior $IG(a, b)$ with density $f(t) = \frac{b^a}{\Gamma(a)} t^{-(a+1)} e^{-b/t}$ is specified by DISPERSIONPRIOR=IGAMMA, which can be followed by one of the following *gamma-options* enclosed in parentheses. The hyperparameters a and b are the shape and scale parameters of the gamma distribution, respectively. See the “Inverse Gamma Prior” section on page 63. The default is $IG(2.001, 0.001)$.

RELSHAPE<=c>

specifies independent $IG(\frac{c+\hat{\phi}}{\hat{\phi}}, c)$ distribution, where $\hat{\phi}$ is the MLE of the dispersion parameter. With this choice of hyperparameters, the mean of the prior distribution is $\hat{\phi}$. By default, $c=10^{-4}$.

SHAPE=a

and

ISCALE=b

specify the $IG(a, b)$ prior.

SHAPE= c
specifies the $IG(c, c)$ prior.

ISCALE= c
specifies the $IG(c, c)$ prior.

An improper prior with density $f(t)$ proportional to t^{-1} is specified with DISPERSIONPRIOR=IMPROPER.

INITIAL=SAS-data-set

specifies the SAS data set that contains the initial values of the Markov chains. The INITIAL= data set must contain all the variables of the model. You can specify multiple rows as the initial values of the parallel chains for the Gelman-Rubin statistics, but posterior summaries, diagnostics, and plots are computed only for the first chain. If the data set also contains the variable _SEED_, the value of the _SEED_ variable is used as the seed of the random number generator for the corresponding chain.

NBI=number

specifies the number of burn-in iterations before the chains are saved. The default is 2000.

NMC=number

specifies the number of iterations after the burn-in. The default is 10000.

PRECISIONPRIOR=GAMMA<(options)> | IMPROPER

PPRIOR=GAMMA<(options)> | IMPROPER

specifies that Gibbs sampling be performed on the generalized linear model precision parameter and the prior distribution for the precision parameter, if there is a precision parameter in the model. For models that do not have a precision parameter (the Poisson and binomial), this option is ignored. Note that you can specify Gibbs sampling on either the dispersion parameter ϕ , the scale parameter $\sigma = \phi^{\frac{1}{2}}$, or the precision parameter $\tau = \phi^{-1}$, with the DPRIOR=, SPRIOR=, and PPRIOR= options, respectively. These three parameters are transformations of one another, and you should specify Gibbs sampling for only one of them.

A gamma prior $G(a, b)$ with density $f(t) = \frac{b(bt)^{a-1}e^{-bt}}{\Gamma(a)}$ is specified by PRECISIONPRIOR=GAMMA, which can be followed by one of the following *gamma-options* enclosed in parentheses. The hyperparameters a and b are the shape and inverse-scale parameters of the gamma distribution, respectively. See the “Gamma Prior” section on page 63. The default is $G(10^{-4}, 10^{-4})$.

RELSHAPE<=c>

specifies independent $G(c\hat{\tau}, c)$ distribution, where $\hat{\tau}$ is the MLE of the dispersion parameter. With this choice of hyperparameters, the mean of the prior distribution is $\hat{\tau}$ and the variance is $\frac{\hat{\tau}}{c}$. By default, $c=10^{-4}$.

SHAPE= a
and

ISCALE= b
specify the $G(a, b)$ prior.

SHAPE= c
 specifies the $G(c, c)$ prior.

ISCALE= c
 specifies the $G(c, c)$ prior.

An improper prior with density $f(t)$ proportional to t^{-1} is specified with PRECISIONPRIOR=IMPROPER.

SCALEPRIOR=GAMMA<(options)> | **IMPROPER**

SPRIOR=GAMMA<(options)> | **IMPROPER**

specifies that Gibbs sampling be performed on the generalized linear model scale parameter and the prior distribution for the scale parameter, if there is a scale parameter in the model. For models that do not have a scale parameter (the Poisson and binomial), this option is ignored. Note that you can specify Gibbs sampling on either the dispersion parameter ϕ , the scale parameter $\sigma = \phi^{\frac{1}{2}}$, or the precision parameter $\tau = \phi^{-1}$, with the DPRIOR=, SPRIOR=, and PPRIOR= options, respectively. These three parameters are transformations of one another, and you should specify Gibbs sampling for only one of them.

A gamma prior $G(a, b)$ with density $f(t) = \frac{b(bt)^{a-1}e^{-bt}}{\Gamma(a)}$ is specified by SCALEPRIOR=GAMMA, which can be followed by one of the following *gamma-options* enclosed in parentheses. The hyperparameters a and b are the shape and inverse-scale parameters of the gamma distribution, respectively. See the “Gamma Prior” section on page 63. The default is $G(10^{-4}, 10^{-4})$.

RELSHAPE<= c >

specifies independent $G(c\hat{\sigma}, c)$ distribution, where $\hat{\sigma}$ is the MLE of the dispersion parameter. With this choice of hyperparameters, the mean of the prior distribution is $\hat{\sigma}$ and the variance is $\frac{\hat{\sigma}}{c}$. By default, $c=10^{-4}$.

SHAPE= a
 and

ISCALE= b
 specify the $G(a, b)$ prior.

SHAPE= c
 specifies the $G(c, c)$ prior.

ISCALE= c
 specifies the $G(c, c)$ prior.

An improper prior with density $f(t)$ proportional to t^{-1} is specified with SCALEPRIOR=IMPROPER.

PLOTS<(global-plot-options)>= *plot-request*

PLOTS<(global-plot-options)>= (*plot-request* <...*plot-request*>)

controls the display of diagnostic plots. Three types of plots can be requested: trace plots, autocorrelation function plots, and kernel density plots. By default, the plots are displayed in panels unless the global plot option UNPACK is specified. Also,

when specifying more than one type of plots, the plots are displayed by parameters unless the global plot option TYPE is specified. When you specify only one plot request, you can omit the parentheses around the plot request. For example:

```
plots=trace
plots=(trace density)
```

The global plot options are as follows:

TYPE

specifies how the plots are grouped when there is more than one type of plot. TYPE specifies that the plots are grouped by type. If you do not specify TYPE, then the plots are grouped by parameter.

UNPACKPANEL | UNPACK

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

The plot requests are as follows:

ALL	specifying all types of plots. PLOTS=ALL is equivalent to specifying PLOTS=(TRACE AUTOCORR DENSITY).
AUTOCORR	displays autocorrelation function plots.
DENSITY	displays kernel density plots.
NONE	suppresses the display of any plots.
TRACE	displays trace plots.

The default plots request is as follows:

```
PLOTS=ALL
```

See the “[Visual Analysis via Trace Plots](#)” section on page 18 for details about interpretation of diagnostic plots.

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 the SEED= option is not specified, or if you specify a nonpositive seed, a random seed is derived from the time of day.

SUMMARIES=ALL | NONE | (*keyword-list*)

SUMMARY=ALL | NONE | (*keywords-list*)

SUM=ALL | NONE | (*keywords-list*)

controls the number of posterior summaries produced. SUMMARIES=ALL produces all the summary statistics, which include the mean, standard deviation, quartiles, credible intervals, and HPD intervals for each parameter. If you do not want any posterior summaries, you specify SUMMARIES=NONE. You can use

the following keywords to request only the descriptive statistics or the credible and HPD intervals of a given level, or the correlation matrix. The default is `SUMMARIES=(DESCRIPTIVE INTERVAL)`.

DESCRIPTIVE

DESC

produces the means, standard deviations, and quartiles for the posterior sample.

INTERVAL<(ALPHA=*numeric-list*)>

produces the $100(1 - \alpha)\%$ credible interval and the $100(1 - \alpha)\%$ HPD interval for each parameter and for each α in the *numeric-list* specified in the ALPHA= option. The default is ALPHA=0.05.

CORR

produces the correlation matrix of the posterior samples.

See the “[Summary Statistics](#)” section on page 31 for details.

THINNING=*number*

THIN=*number*

controls the thinning of the Markov chain. Only one in every k samples is used when THINNING= k , and if NBI= n_0 and NMC= n , the number of samples kept is

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

where $[a]$ represents the integer part of the number a . Only the kept samples are used in computing posterior statistics. The default is THINNING=1.

Details

Gibbs Sampling

This section provides details for Gibbs sampling in generalized linear models. See the “[Gibbs Sampler](#)” section on page 15 for a general discussion of Gibbs sampling. In generalized linear models, the response has a probability distribution from a family of distributions of the exponential form. That is, the probability density of the response Y for continuous response variables, or the probability function for discrete responses, can be expressed as

$$f(y) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

for some functions a , b , and c that determine the specific distribution. The canonical parameters θ depend only on the means of the response μ_i , which are related to the regression parameters β through the link function $g(\mu_i) = \mathbf{x}_i' \beta$. The additional parameter ϕ is the dispersion parameter. The GENMOD procedure estimates the regression parameters and the scale parameter $\sigma = \phi^{\frac{1}{2}}$ by maximum likelihood.

However, the BGENMOD procedure can also provide Bayesian estimates of the regression parameters and either the scale σ , the dispersion ϕ , or the precision $\tau = \phi^{-1}$ by Gibbs sampling. Except where noted, the following discussion applies to either σ , ϕ , or τ , although ϕ is used to illustrate the formulas. Note that the Poisson and binomial distributions do not have a dispersion parameter, and the dispersion is considered to be fixed at $\phi = 1$.

Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)'$ be the parameter vector. For generalized linear models, the θ_i s are the regression coefficients β_i s and the dispersion parameter ϕ . Let $L(D|\boldsymbol{\theta})$ be the likelihood function, where D is the observed data. Let $\pi(\boldsymbol{\theta})$ be the prior distribution. The full conditional distribution of $[\theta_i|\theta_j, i \neq j]$ is proportional to the joint distribution; that is,

$$\pi(\theta_i|\theta_j, i \neq j, D) \propto L(D|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

For instance, the one-dimensional conditional distribution of θ_1 given $\theta_j = \theta_j^*, 2 \leq j \leq k$, is computed as

$$\pi(\theta_1|\theta_j = \theta_j^*, 2 \leq j \leq k, D) = L(D|(\boldsymbol{\theta} = (\theta_1, \theta_2^*, \dots, \theta_k^*)')p(\boldsymbol{\theta} = (\theta_1, \theta_2^*, \dots, \theta_k^*)')$$

Suppose you have a set of arbitrary starting values $\{\theta_1^{(0)}, \dots, \theta_k^{(0)}\}$. Using the ARMS (adaptive rejection Metropolis sampling) algorithm of [Gilks and Wild \(1992\)](#) and [Gilks, Best, and Tan \(1995\)](#), you can do the following:

```
draw  $\theta_1^{(1)}$  from  $[\theta_1|\theta_2^{(0)}, \dots, \theta_k^{(0)}]$ 
draw  $\theta_2^{(1)}$  from  $[\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \dots, \theta_k^{(0)}]$ 
...
draw  $\theta_k^{(1)}$  from  $[\theta_k|\theta_1^{(1)}, \dots, \theta_{k-1}^{(1)}]$ 
```

This completes one iteration of the Gibbs sampler. After one iteration, you have $\{\theta_1^{(1)}, \dots, \theta_k^{(1)}\}$. After n iterations, you have $\{\theta_1^{(n)}, \dots, \theta_k^{(n)}\}$. PROC BGENMOD implements the ARMS algorithm based on code provided by [Gilks \(2003\)](#) to draw a sample from a full conditional distribution.

You can output these posterior samples into a SAS data set through ODS. The following SAS statement outputs the posterior samples into the SAS data set `Post`:

```
ods output PosteriorSample=Post;
```

The data set also includes the variable `_LOGPOST_`, representing the log of the posterior log likelihood.

Priors for Model Parameters

The model parameters are the regression coefficients and the dispersion parameter (or the precision or scale), if the model has one. The priors for the dispersion parameter and the priors for the regression coefficients are assumed to be independent, while you can have a joint multivariate normal prior for the regression coefficients.

Dispersion, Precision, or Scale Parameter

Gamma Prior

The gamma distribution $G(a, b)$ has a pdf

$$f(u) = \frac{b(bu)^{a-1}e^{-bu}}{\Gamma(a)}, \quad u > 0$$

where a is the shape parameter and b is the inverse-scale parameter. The mean is $\frac{a}{b}$ and the variance is $\frac{a}{b^2}$.

Improper Prior

The joint prior density is given by

$$p(u) \propto u^{-1}, \quad u > 0$$

Inverse Gamma Prior

The inverse gamma distribution $IG(a, b)$ has a pdf

$$f(u) = \frac{b^a}{\Gamma(a)} u^{-(a+1)} e^{-b/u}, \quad u > 0$$

where a is the shape parameter and b is the scale parameter. The mean is $\frac{b}{a-1}$ if $a > 1$ and the variance is $\frac{b^2}{(a-1)^2(a-2)}$ if $a > 2$.

Regression Coefficients

Let β be the regression coefficients.

Jeffreys' Prior

The joint prior density is given by

$$p(\beta) \propto |I(\beta)|^{\frac{1}{2}}$$

where $I(\beta)$ is the Fisher information matrix for the model. See [Ibrahim and Laud \(1991\)](#) for a full discussion, with examples, of Jeffreys' prior for generalized linear models.

Normal Prior

Assume β has a multivariate normal prior with mean vector β_0 and covariance matrix Σ_0 . The joint prior density is given by

$$p(\beta) \propto e^{-\frac{1}{2}(\beta-\beta_0)'\Sigma_0^{-1}(\beta-\beta_0)}$$

Uniform Prior

The joint prior density is given by

$$p(\beta) \propto 1$$

Posterior Distribution

Denote the observed data by D .

The posterior distribution is

$$\pi(\beta|D) \propto L_P(D|\beta)p(\beta)$$

where $L_P(D|\beta)$ is the likelihood function with regression coefficients β as parameters.

Starting Values of the Markov Chains

When the BAYES statement is specified, PROC BGENMOD generates one Markov chain containing the approximate posterior samples of the model parameters. Additional chains are produced when the Gelman-Rubin diagnostics are requested. Starting values (or initial values) can be specified in the INITIAL= data set in the BAYES statement. If INITIAL= option is not specified, PROC BGENMOD picks its own initial values for the chains.

Denote $[x]$ as the integral value of x . Denote $\hat{s}(X)$ as the estimated standard error of the estimator X .

Regression Coefficients

For the first chain for which the summary statistics and regression diagnostics are based, the initial values are the maximum likelihood estimates; that is,

$$\beta_i^{(0)} = \hat{\beta}_i$$

Initial values for the r th chain ($2 \leq r$) are given by

$$\beta_i^{(0)} = \hat{\beta}_i \pm \left(2 + \left\lceil \frac{r}{2} \right\rceil\right) \hat{s}(\hat{\beta}_i)$$

with the plus sign for odd r and minus sign for even r .

Dispersion, Scale, or Precision Parameter λ

Let λ be the generalized linear model parameter you choose to sample, either the dispersion, scale, or precision parameter. Note that the Poisson and binomial distributions do not have this additional parameter.

For the first chain that the summary statistics and diagnostics are based on, the initial values are the maximum likelihood estimates; that is,

$$\lambda^{(0)} = \hat{\lambda}$$

The initial values of the r th chain ($2 \leq r$) are given by

$$\lambda^{(0)} = \hat{\lambda} e^{\pm \left(\left[\frac{r}{2} \right] + 2 \right) \hat{s}(\hat{\lambda})}$$

with the plus sign for odd r and minus sign for even r .

Displayed Output

The displayed output for a Bayesian analysis includes the following.

Model Information

The “Model Information” table displays the two-level name of the input data set, the number of burn-in iterations, the number of iterations after the burn-in, the number of thinning iterations, the distribution name, the link function, and the name and label of the dependent variable; if you use the OFFSET= option in the MODEL statement, it displays the name and label of the offset variable; if you specify the FREQ statement, it displays the name and label of the frequency variable.

Class Level Information

The “Class Level Information” table displays the levels of class variables if you specify a CLASS statement.

Criteria for Assessing Goodness of Fit

The “Criteria for Assessing Goodness of Fit” table displays the value, degrees of freedom, and value divided by degrees of freedom of the deviance, the scaled deviance, the Pearson chi-square, the scaled Pearson chi-square, the log likelihood, and the full log likelihood (including all constant terms).

Maximum Likelihood Estimates

The “Analysis of Maximum Likelihood Parameter Estimates” table displays the maximum likelihood estimate of each parameter, the estimated standard error of the parameter estimator, and confidence limits for each parameter.

Coefficient Prior

The “Coefficient Prior” table displays the prior distribution of the regression coefficients.

Independent Prior Distributions for Model Parameters

The “Independent Prior Distributions for Model Parameters” table displays the prior distribution of the additional model parameter (dispersion, scale, or precision).

Initial Values and Seeds

The “Initial Values and Seeds” table displays the initial values and random number generator seeds for the Gibbs chains.

Descriptive Statistics of the Posterior Samples

The “Descriptive Statistics of the Posterior Sample” table contains the size of the sample, the mean, the standard error, and the quartiles for each model parameter.

Interval Estimates for Posterior Sample

The “Interval Estimates for Posterior Sample” table contains the HPD intervals and the credible intervals for each model parameter.

Correlation Matrix of the Posterior Samples

The “Correlation Matrix of the Posterior Samples” table is produced if you include the CORR suboption in the SUMMARY= option in the BAYES statement. This table displays the sample correlation of the posterior samples.

Autocorrelations of the Posterior Samples

The “Autocorrelations of the Posterior Samples” table displays the lag1, lag5, lag10, and lag50 autocorrelations for each parameter.

Gelman and Rubin Diagnostics

The “Gelman and Rubin Diagnostics” table is produced if you include the GELMAN suboption in the DIAGNOSTIC= option in the BAYES statement. This table displays the estimate of the potential scale reduction factor and its 97.5% upper confidence limit for each parameter.

Geweke Diagnostics

The “Geweke Diagnostics” table displays the Geweke statistic and its p -value for each parameter.

Raftery and Lewis Diagnostics

The “Raftery Diagnostics” tables is produced if you include the RAFTERY suboption in the DIAGNOSTIC= option in the BAYES statement. This table displays the Raftery and Lewis diagnostics for each variable.

Heidelberger and Welch Diagnostics

The “Heidelberger and Welch Diagnostics” table is displayed if you include the HEIDELBERGER suboption in the DIAGNOSTIC= option in the BAYES statement. This table shows the results of a stationary test and a halfwidth test for each parameter.

Effective Sample Size

The “Effective Sample Size” table displays, for each parameter, the effective sample size, the correlation time, and the efficiency.

ODS Table Names

PROC BGENMOD assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in Table 2.2.

Table 2.2. ODS Tables Produced by PROC BGENMOD

ODS Table Name	Description	Statement	Option
AutoCorr	Autocorrelations of the posterior samples	BAYES	default
ChainStatistics	Simple descriptive statistics for the chain	BAYES	default
ClassLevels	Levels of class variables	CLASS	default
CoeffPrior	Prior distribution of the regression coefficients	BAYES	default
ConvergenceStatus	Convergence status of maximum likelihood estimation	MODEL	default
Corr	Correlation matrix of the posterior samples	BAYES	SUMMARY=(CORR)
ESS	Effective sample size	BAYES	default
Gelman	Gelman and Rubin convergence diagnostics	BAYES	DIAG=GELMAN
Geweke	Geweke convergence diagnostics	BAYES	default
Heidelberger	Heidelberger and Welch convergence diagnostics	BAYES	DIAG=HEIDELBERGER
InitialValues	Initial values of the Markov chains	BAYES	default
IntervalStatistics	HPD and credible intervals for the posterior samples	BAYES	default
ModelFit	Criteria for assessing MLE fit	MODEL	default
ModelInfo	Model information	PROC	default
NObs	Number of observations		default
ParameterEstimates	Maximum likelihood estimates of model parameters	MODEL	default

Table 2.2. (continued)

ODS Table Name	Description	Statement	Option
ParmPrior	Prior distribution for the dispersion, precision, or scale	BAYES	default
PosteriorSample	Posterior samples (for ODS output data set only)	BAYES	
Raftery	Raftery and Lewis convergence diagnostics	BAYES	DIAG=(RAFTERY)

ODS Graph Names

Each statistical graphic created by PROC BGENMOD has a name associated with it, and you can reference the graph by using ODS statements. These names are listed in Table 2.3.

Table 2.3. ODS Graphics Produced by PROC BGENMOD

ODS Graph Name	Description	Statement	Option
ADPanel	Autocorrelation function and density panel	BAYES	PLOTS=(AUTOCORR DENSITY)
AutocorrPanel	Autocorrelation function panel	BAYES	PLOTS= AUTOCORR
AutocorrPlot	Autocorrelation function plot	BAYES	PLOTS(UNPACK)=AUTOCORR
DensityPanel	Density panel	BAYES	PLOTS=DENSITY
DensityPlot	Density plot	BAYES	PLOTS(UNPACK)=DENSITY
TAPanel	Trace and autocorrelation function panel	BAYES	PLOTS=(TRACE AUTOCORR)
TADPanel	Trace, autocorrelation, and density function panel	BAYES	default
TDPanel	Trace and density panel	BAYES	PLOTS=(TRACE DENSITY)
TracePanel	Trace panel	BAYES	PLOTS=TRACE
TracePlot	Trace plot	BAYES	PLOTS(UNPACK)=TRACE

Example

Consider the data on patients from clinical trials in Figure 2.16. The data set is a subset of the data described in Ibrahim, Chen, and Lipsitz (1999).

The primary interest is in prediction of the number of cancerous liver nodes when a patient enters the trials, by using six other baseline characteristics. The number of nodes is modeled by a Poisson regression model with the six baseline characteristics as covariates. The response and regression variables are as follows:

y	Number of Cancerous Liver Nodes
x1	Body Mass Index
x2	Age, in Years
x3	Time Since Diagnosis of Disease, in Weeks
x4	Two Biochemical Markers (each classified as normal=1 or abnormal=0)
x5	Anti Hepatitis B Antigen
x6	Associated Jaundice (yes=1, no=0)

Two analyses are performed using PROC BGENMOD. The first analysis uses non-informative normal prior distributions, and the second analysis uses an informative normal prior for one of the regression parameters.

In the following BAYES statement, COEFFPRIOR=NORMAL specifies a noninformative independent normal prior distribution with zero mean and variance 10^6 for each parameter. DIAGNOSTICS=(AUTOCORR) specifies that the table of auto-correlations be produced for each parameter, and PLOTS=(TRACE) specifies that diagnostic trace plots be created for each parameter.

The initial analysis is performed using PROC BGENMOD to obtain Bayesian estimates of the regression coefficients by using the following SAS statements:

```
ods graphics on;
proc bgenmod data=Liver;
  model y = x1-x6 / dist=Poisson link=log;
  bayes diagnostics=(AutoCorr)
  plots=(Trace)
  coeffprior=normal;
run;
ods graphics off;
```

Obs	x1	x2	x3	x4	x5	x6	y
1	19.1358	50.0110	51.000	0	0	1	3
2	23.5970	18.4959	3.429	0	0	1	9
3	20.0474	56.7699	3.429	1	1	0	6
4	28.0277	59.7836	4.000	0	0	1	6
5	28.6851	74.1589	5.714	1	0	1	1
6	18.8092	31.0630	2.286	0	1	1	61
7	28.7201	52.9178	37.286	1	0	1	6
8	21.3669	61.6603	54.143	0	1	1	6
9	23.7332	42.2904	0.571	1	0	1	21
10	20.4783	22.1260	19.000	1	0	1	6
.							
.							
.							
127	25.9924	66.5151	2.857	1	0	1	6
128	31.0735	73.0493	8.714	1	0	0	2
129	20.9840	48.2027	4.857	1	0	0	2
130	21.4536	69.1808	2.571	0	0	0	1
131	26.2346	60.3425	2.571	1	0	1	1
132	24.1633	60.8329	11.000	1	0	1	1
133	26.8519	58.6877	3.429	1	0	1	2
134	17.0993	48.8384	3.000	0	0	0	9
135	19.1327	65.3425	2.571	1	0	0	1
136	17.3010	51.4493	4.429	1	0	0	6

Figure 2.16. Liver Cancer Data

Maximum likelihood estimates of the model parameters are computed by default. These are shown in [Figure 2.17](#).

Analysis Of Maximum Likelihood Parameter Estimates					
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits	
Intercept	1	2.4508	0.2284	2.0032	2.8984
x1	1	-0.0044	0.0080	-0.0201	0.0114
x2	1	-0.0135	0.0024	-0.0181	-0.0088
x3	1	-0.0029	0.0022	-0.0072	0.0014
x4	1	-0.2715	0.0795	-0.4272	-0.1157
x5	1	0.3215	0.0832	0.1585	0.4845
x6	1	0.2077	0.0827	0.0456	0.3698
Scale	0	1.0000	0.0000	1.0000	1.0000

NOTE: The scale parameter was held fixed.

Figure 2.17. Maximum Likelihood Parameter Estimates

Noninformative independent normal prior distributions with zero means and variances of 10^6 were used in the initial analysis. These are shown in the “Independent Normal Prior for Regression Coefficients” table in [Figure 2.18](#).

Initial values for the Markov chain are listed in the “Initial Values” table in [Figure 2.18](#). The random number seed is also listed so that you can reproduce the analysis. Since no seed was specified, the seed shown was derived from the time of day.

Independent Normal Prior for Regression Coefficients								
		Parameter	Mean	Precision				
		Intercept	0	1E-6				
		x1	0	1E-6				
		x2	0	1E-6				
		x3	0	1E-6				
		x4	0	1E-6				
		x5	0	1E-6				
		x6	0	1E-6				
Initial Values and Seeds								
Chain	_SEED_	Intercept	x1	x2	x3	x4	x5	x6
1	209058001	2.450813	-0.00435	-0.01347	-0.00291	-0.27149	0.321507	0.207713

Figure 2.18. Regression Coefficient Priors

Descriptive and interval statistics for the posterior sample are displayed in the tables in Figure 2.19. Since noninformative prior distributions for the regression coefficients were used, the mean and standard deviations of the posterior distributions for the model parameters are close to the maximum likelihood estimates and standard errors.

Descriptive Statistics of the Posterior Samples						
Parameter	N	Mean	Standard Deviation	25%	50%	75%
Intercept	10000	2.4727	0.2449	2.3168	2.4752	2.6331
x1	10000	-0.00565	0.00821	-0.0113	-0.00590	-0.00024
x2	10000	-0.0134	0.00238	-0.0150	-0.0134	-0.0118
x3	10000	-0.00307	0.00220	-0.00451	-0.00300	-0.00155
x4	10000	-0.2712	0.0793	-0.3257	-0.2710	-0.2180
x5	10000	0.3190	0.0828	0.2634	0.3187	0.3749
x6	10000	0.2064	0.0850	0.1481	0.2052	0.2629
Interval Statistics of the Posterior Samples						
Parameter	Alpha	Credible Interval		HPD Interval		
Intercept	0.050	1.9753	2.9564	1.9570	2.9343	
x1	0.050	-0.0211	0.0109	-0.0212	0.0108	
x2	0.050	-0.0180	-0.00858	-0.0181	-0.00870	
x3	0.050	-0.00763	0.00107	-0.00736	0.00122	
x4	0.050	-0.4243	-0.1155	-0.4245	-0.1160	
x5	0.050	0.1556	0.4803	0.1587	0.4828	
x6	0.050	0.0431	0.3754	0.0461	0.3778	

Figure 2.19. Posterior Sample Statistics

The requested autocorrelation table is shown in Figure 2.20. The parameters *Intercept*, *x1*, and *x3* have fairly high autocorrelations at lag 10, indicating that additional thinning of the Markov chain might be appropriate to ensure that the samples used in computing statistics for the posterior distribution are independent. However, this analysis will continue with no additional thinning of the Markov chain (thinning=1).

Autocorrelations of the Posterior Samples				
Parameter	Lag1	Lag5	Lag10	Lag50
Intercept	0.9758	0.8778	0.7581	0.1129
x1	0.9594	0.8068	0.6526	0.0985
x2	0.9176	0.6479	0.4265	-0.0417
x3	0.2128	-0.0170	-0.0128	0.0097
x4	0.6360	0.1400	0.0370	-0.0218
x5	0.3296	0.0108	0.0089	0.0306
x6	0.6917	0.2097	0.1375	0.0114

Figure 2.20. Posterior Sample Autocorrelations

Trace plots for the seven model parameters are shown in [Figure 2.21](#) through [Figure 2.23](#).

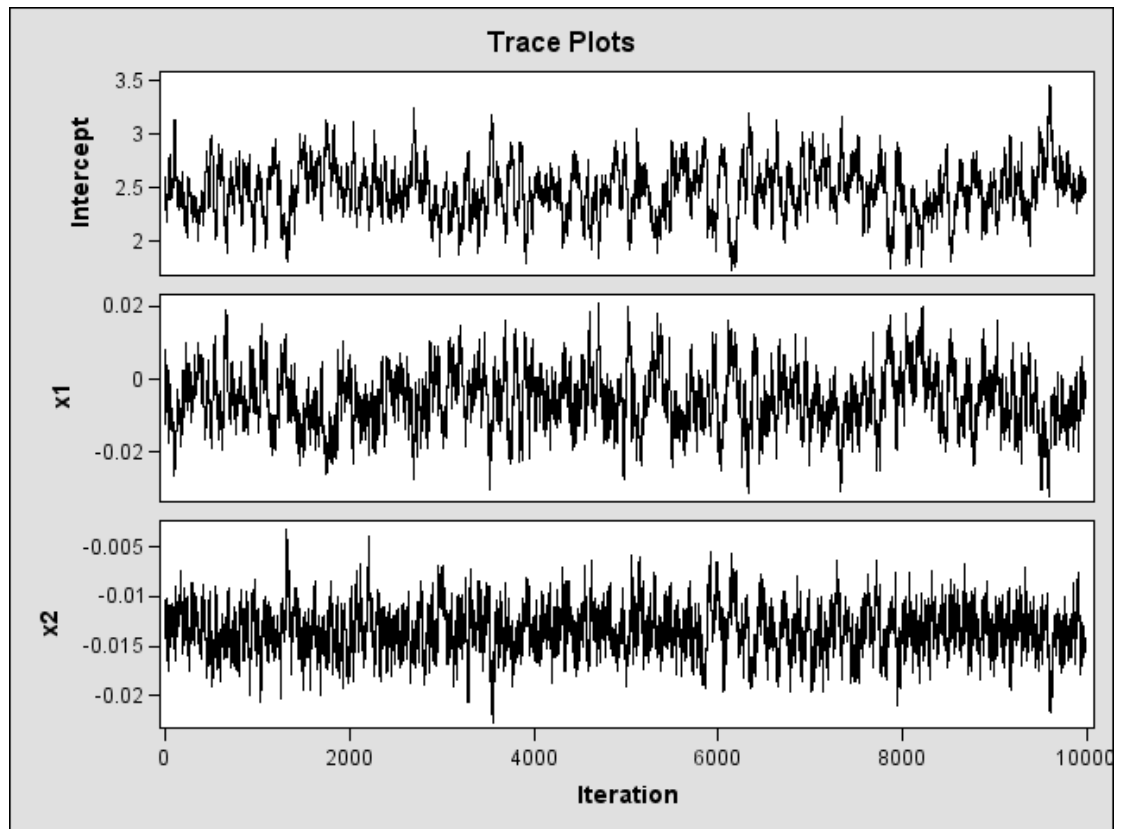


Figure 2.21. Trace Plots

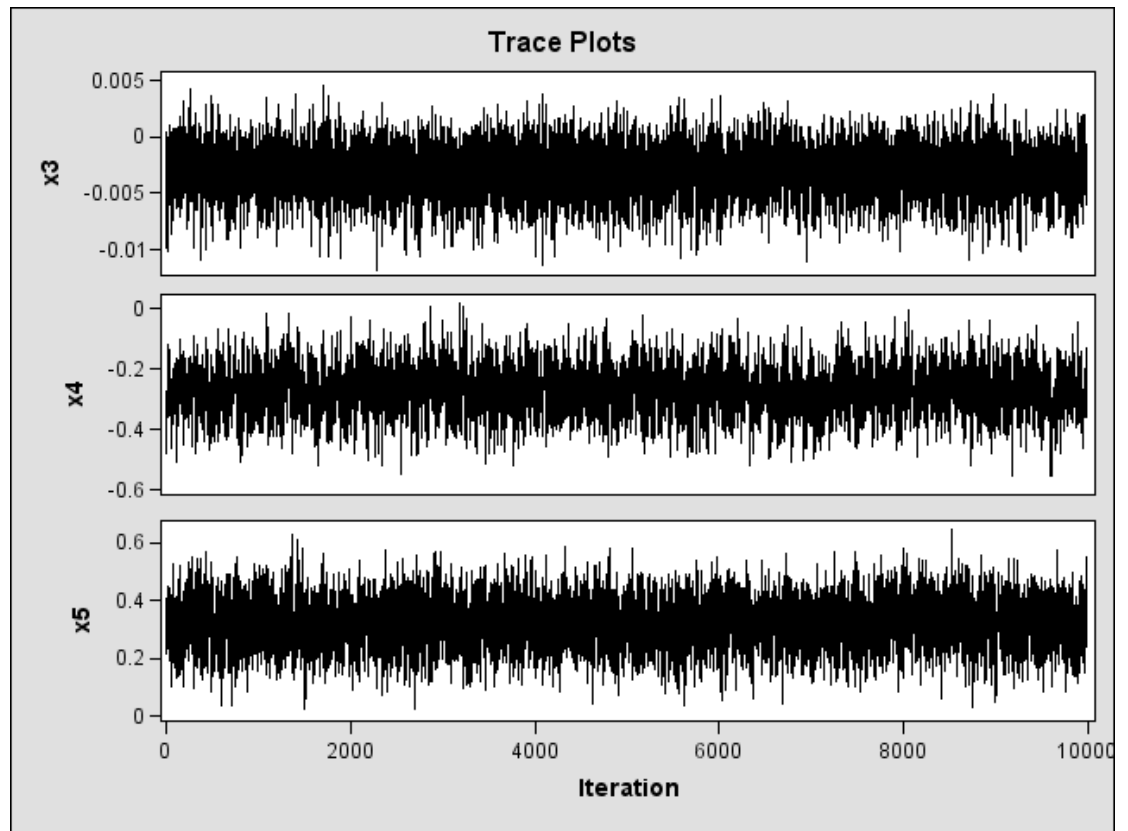


Figure 2.22. Trace Plots

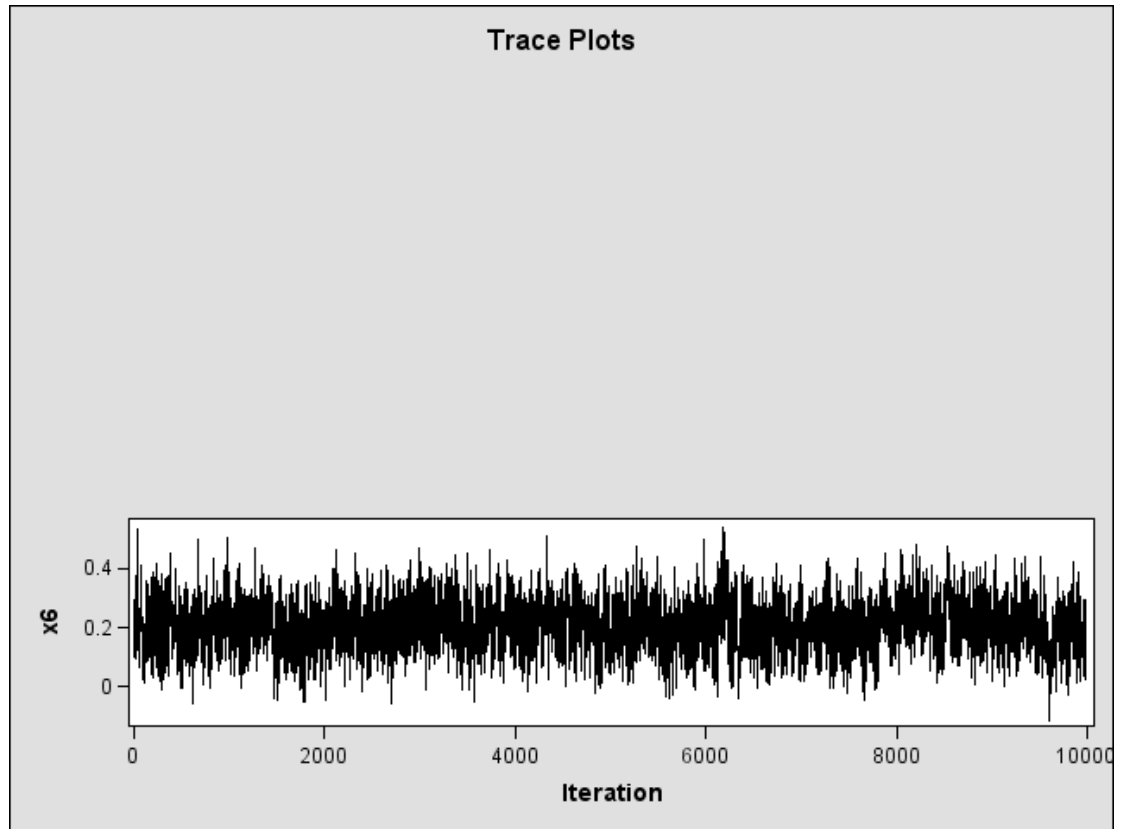


Figure 2.23. Trace Plots

In order to illustrate the use of an informative prior distribution, suppose that researchers expect that a unit increase in body mass index (x_1) will be associated with an increase in the mean number of nodes of between 10% and 20%, and they want to incorporate this prior knowledge in the Bayesian analysis. For log-linear models, the mean and linear predictor are related by $\log(\mu_i) = \mathbf{x}_i' \boldsymbol{\beta}$. If x_{1_1} and x_{1_2} are two values of body mass index, μ_1 and μ_2 are the two mean values, and all other covariates remain equal for the two values of x_1 , then

$$\frac{\mu_1}{\mu_2} = \exp(\beta(x_{1_1} - x_{1_2}))$$

so that for a unit change in x_1 ,

$$\frac{\mu_1}{\mu_2} = \exp(\beta)$$

If $1.1 \leq \frac{\mu_1}{\mu_2} \leq 1.2$, then $1.1 \leq \exp(\beta) \leq 1.2$, or $0.095 \leq \beta \leq 0.182$. This gives you guidance in specifying a prior distribution for the β for body mass index. Taking the mean of the prior normal distribution to be the midrange of the values of β , and taking $\mu \pm 2\sigma$ to be the extremes of the range, a $N(0.435, 0.005)$ is the resulting prior

distribution. The second analysis uses this informative normal prior distribution for the coefficient of x_1 and uses independent noninformative normal priors with zero means and variances equal to 10^6 for the remaining model regression parameters.

In the following BAYES statement, COEFFPRIOR=NORMAL(INPUT=NormalPrior) specifies the normal prior distribution for the regression coefficients with means and variances contained in the data set NormalPrior.

The analysis is performed using PROC BGENMOD to obtain Bayesian estimates of the regression coefficients by using the following SAS statements:

```
ods graphics on;
data NormalPrior;
  input _type_ $ Intercept x1-x6;
  datalines;
Var 1e6 0.0005 1e6 1e6 1e6 1e6 1e6
Mean 0.0 0.0435 0.0 0.0 0.0 0.0 0.0
;
run;

proc bgenmod data=Liver;
  model y = x1-x6 / dist=Poisson link=log;
  bayes diagnostics=(AutoCorr)
  plots=(Trace)
  coeffprior=normal(input=NormalPrior);
run;
ods graphics off;
```

The prior distributions are shown in the “Independent Normal Prior for Regression Coefficients” table in [Figure 2.24](#).

Independent Normal Prior for Regression Coefficients			
Parameter	Mean	Precision	
Intercept	0	1E-6	
x1	0.0435	2000	
x2	0	1E-6	
x3	0	1E-6	
x4	0	1E-6	
x5	0	1E-6	
x6	0	1E-6	

Figure 2.24. Regression Coefficient Priors

Summary and interval statistics are shown in [Figure 2.25](#). Except for x_1 , the statistics shown in [Figure 2.25](#) are very similar to the previous statistics for noninformative priors shown in [Figure 2.19](#). The point estimate for x_1 is now positive. This is expected because the prior distribution on x_1 is quite informative. The distribution reflects the belief that the coefficient is positive. The $N(0.0435, 0.005)$ distribution places the majority of its probability density on positive values. As a result, the posterior density of β_1 places more likelihood on positive values than in the noninformative case.

Descriptive Statistics of the Posterior Samples						
Parameter	N	Mean	Standard Deviation	25%	Quantiles 50%	75%
Intercept	10000	2.3394	0.2160	2.1985	2.3391	2.4837
x1	10000	0.000850	0.00729	-0.00377	0.000779	0.00550
x2	10000	-0.0138	0.00223	-0.0152	-0.0138	-0.0123
x3	10000	-0.00310	0.00218	-0.00451	-0.00303	-0.00161
x4	10000	-0.2741	0.0798	-0.3277	-0.2752	-0.2206
x5	10000	0.3252	0.0823	0.2698	0.3259	0.3813
x6	10000	0.2160	0.0833	0.1602	0.2168	0.2720

Interval Statistics of the Posterior Samples					
Parameter	Alpha	Credible Interval		HPD Interval	
Intercept	0.050	1.9064	2.7597	1.9147	2.7653
x1	0.050	-0.0133	0.0157	-0.0129	0.0161
x2	0.050	-0.0181	-0.00941	-0.0181	-0.00934
x3	0.050	-0.00753	0.000992	-0.00729	0.00116
x4	0.050	-0.4289	-0.1171	-0.4343	-0.1231
x5	0.050	0.1631	0.4845	0.1613	0.4822
x6	0.050	0.0530	0.3790	0.0559	0.3808

Figure 2.25. Posterior Sample Statistics

Trace plots for the second analysis are shown in [Figure 2.26](#) through [Figure 2.28](#). These plots indicate, as in the first analysis, satisfactory convergence of the Markov chains.

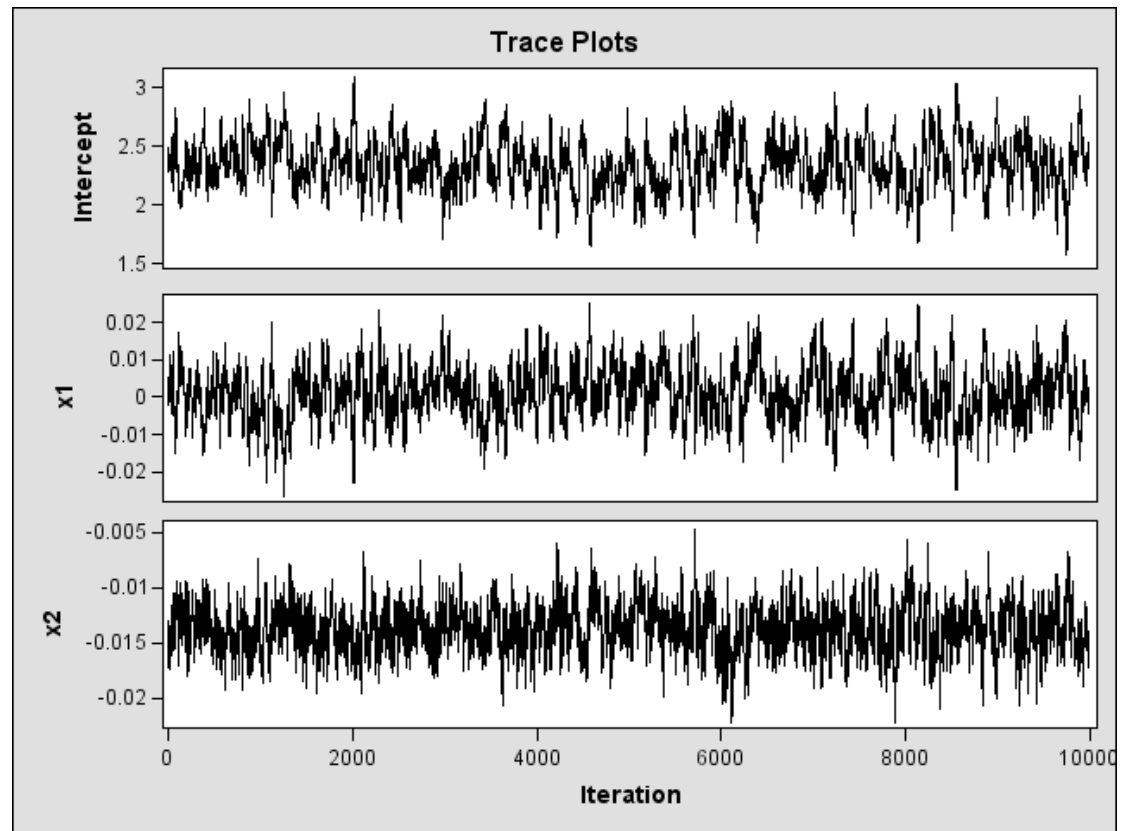


Figure 2.26. Trace Plots

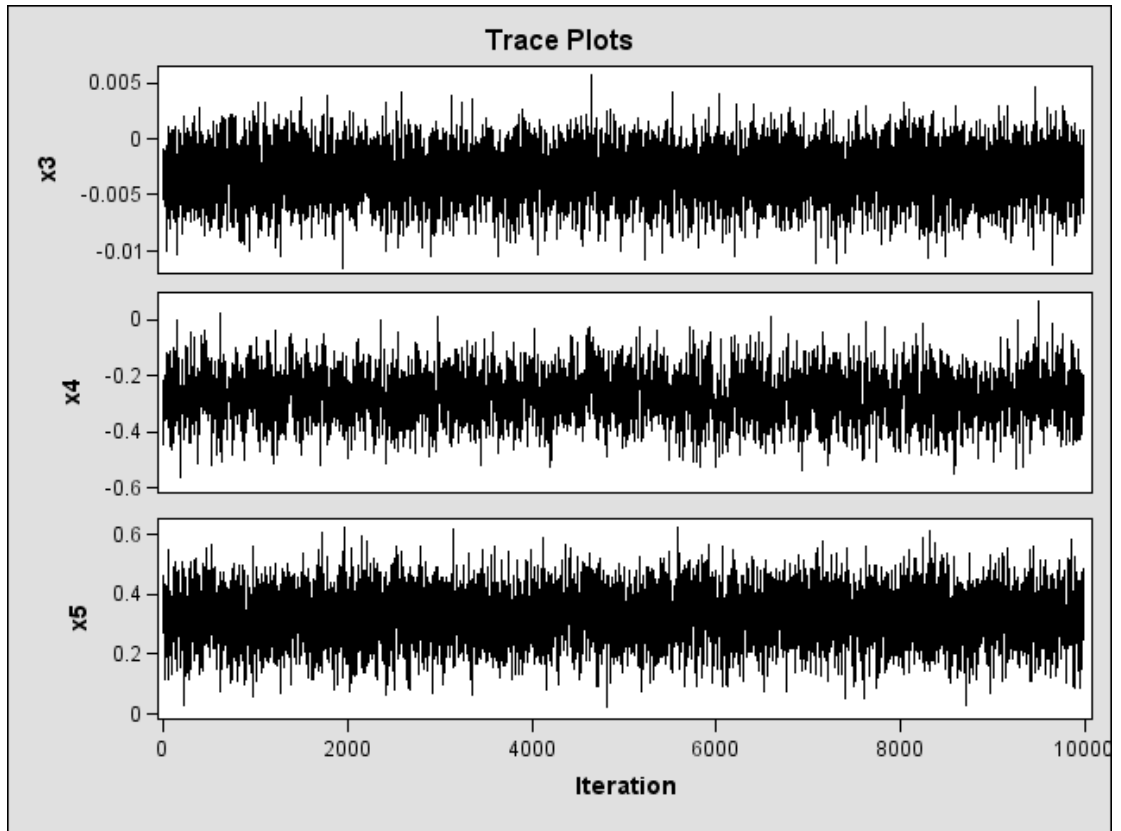


Figure 2.27. Trace Plots

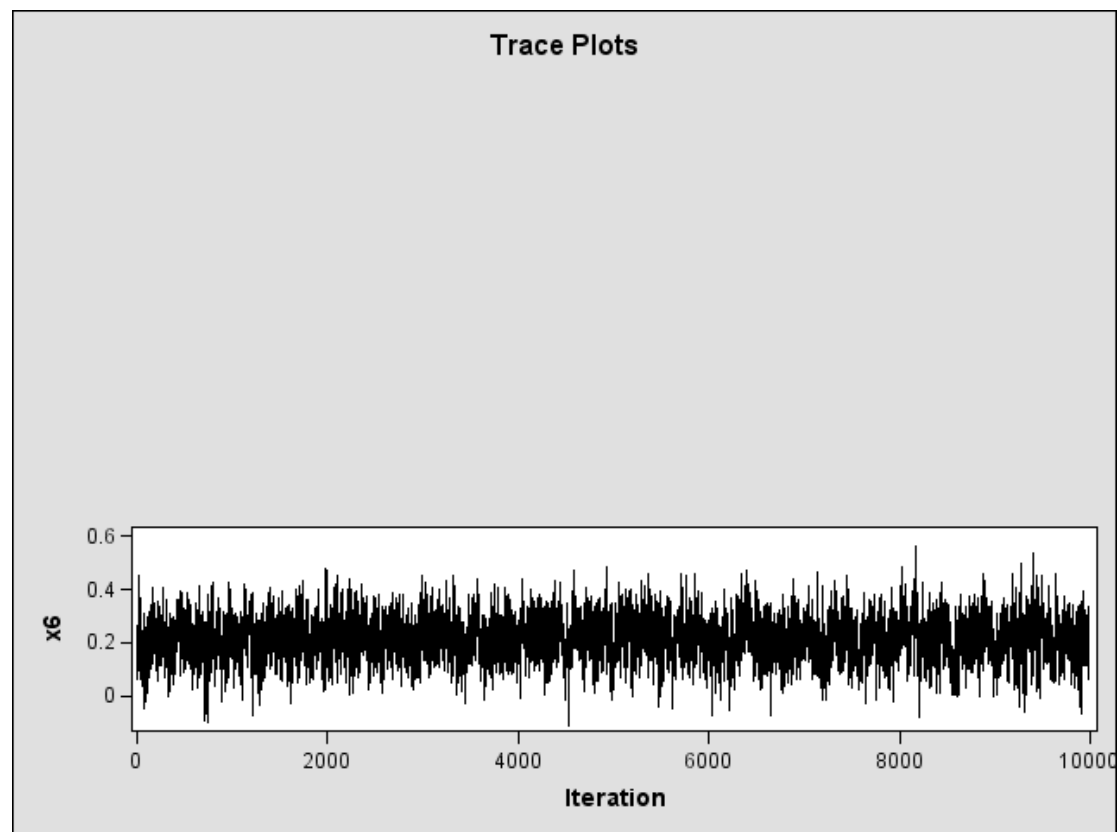


Figure 2.28. Trace Plots

References

- Gilks, W. (2003), "Adaptive Metropolis Rejection Sampling (ARMS)," software from MRC Biostatistics Unit, Cambridge, UK, http://www.maths.leeds.ac.uk/~wally.gilks/adaptive.rejection/web_page/Welcome.html.
- Gilks, W., Best, N., and Tan, K. (1995), "Adaptive Rejection Metropolis Sampling with Gibbs Sampling," *Applied Statistics*, 44, 455–472.
- Gilks, W. and Wild, P. (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Applied Statistics*, 41, 337–348.
- Ibrahim, J., Chen, M., and Sinha, D. (2001), *Bayesian Survival Analysis*, New York: Springer-Verlag.
- Ibrahim, J., Chen, M.-H., and Lipsitz, S. (1999), "Monte Carlo EM for Missing Covariates in Parametric Regression Models," *Biometrics*, 55, 591–596.
- Ibrahim, J. and Laud, P. (1991), "On Bayesian Analysis of Generalized Linear Models Using Jeffreys' Prior," *Journal of the American Statistical Association*, 86, 981–986.

Neter, J. and Wasserman, W. (1974), *Applied Linear Statistical Models*, Homewood, IL: Irwin.

Chapter 3

The BLIFEREG Procedure (Experimental)

Chapter Contents

OVERVIEW	83
GETTING STARTED	84
Engine Fan Data	84
SYNTAX	90
BAYES Statement	90
DETAILS	98
Gibbs Sampling	98
Priors for Model Parameters	100
Posterior Distribution	100
Starting Values of the Markov Chains	101
Displayed Output	102
ODS Table Names	103
ODS Graph Names	104
EXAMPLE	105
REFERENCES	112

Chapter 3

The BLIFEREG Procedure

(Experimental)

Overview

The BLIFEREG procedure adds Bayesian analysis by Gibbs sampling to the LIFEREG procedure, which fits parametric models for survival analysis. Bayesian analysis of parametric survival models can be requested using the BAYES statement in the experimental BLIFEREG procedure. These Bayesian capabilities will be included in PROC LIFEREG in the next release of SAS/STAT software.

In Bayesian analysis, the model parameters are treated as random variables, and inference about parameters is based on the posterior distribution of the parameters, given the data. The posterior distribution is obtained using Bayes' theorem as the likelihood function of the data weighted with a prior distribution. The prior distribution enables you to incorporate knowledge or experience of the likely range of values of the parameters of interest into the analysis. If you have no prior knowledge of the parameter values, you can use a noninformative prior distribution, and the results of the Bayesian analysis will be very similar to a classical analysis based on maximum likelihood. A closed form of the posterior distribution is often not feasible, and a Markov chain Monte Carlo method by Gibbs sampling is used to simulate samples from the posterior distribution. See [Chapter 1, "Introduction to Bayesian Analysis Procedures,"](#) for an introduction to the basic concepts in Bayesian statistics. Also see the section "[Bayesian Analysis: Advantages and Disadvantages](#)" on page 11 for a discussion of the advantages and disadvantages of Bayesian analysis.

A Gibbs chain for the posterior distribution is generated for the model parameters. Summary statistics (mean, standard deviation, quartiles, HPD and credible intervals, correlation matrix) and convergence diagnostics (autocorrelations; Gelman-Rubin, Geweke, Raftery-Lewis, Heidelberger and Welch tests; and the effective sample size) are computed for each parameter as well as the correlation matrix of the posterior sample. Trace plots, posterior density plots, and the autocorrelation function plots that use ODS graphics are also provided for each parameter.

Note that the full functionality of the LIFEREG procedure, as documented in the SAS/STAT 9.1 documentation, is included in the BLIFEREG procedure.

We are eager for your feedback on this experimental procedure. Please send comments to blifereg@sas.com.

Getting Started

Engine Fan Data

If you are not familiar with Bayesian analysis, see [Chapter 1, “Introduction to Bayesian Analysis Procedures,”](#) for an introduction to setting up and interpreting the results of a Bayesian analysis.

Except for the new BAYES statement, PROC BLIFEREG uses the same syntax as the LIFEREG procedure. For many applications, only the PROC BLIFEREG, MODEL, and BAYES statements are required.

[Nelson \(1982\)](#) describes a study of the lifetimes of locomotive engine fans. The data set `Fan`, shown in [Figure 3.1](#), contains the right-censored survival times and a censoring indicator variable. Consider a lognormal model for the data. There are no covariates, so the model is an intercept-only model.

Obs	Lifetime	Censor
1	450	0
2	460	1
3	1150	0
4	1150	0
5	1560	1
6	1600	0
7	1660	1
8	1850	1
9	1850	1
10	1850	1
.		
.		
.		
61	8500	1
62	8750	1
63	8750	0
64	8750	1
65	9400	1
66	9900	1
67	10100	1
68	10100	1
69	10100	1
70	11500	1

Figure 3.1. Locomotive Engine Fan Data

The example that follows shows how to use PROC BLIFEREG to carry out a Bayesian analysis of the lognormal model. The `SEED=` option is specified to maintain reproducibility; no other options are specified in the BAYES statement. By default, a uniform prior distribution is assumed on the intercept coefficient. The uniform prior is a flat prior on the real line with a distribution that reflects ignorance of the location of the parameter, placing equal probability on all possible values the regression coefficient can take. Using the uniform prior in the following example, you would expect the Bayesian estimates to resemble the classical results of maximizing the likelihood. If you can elicit an informative prior on the regression coefficients,

you should use the `COEFFPRIOR=` option to specify it. A default noninformative gamma prior is used for the lognormal scale parameter σ .

You should make sure that the posterior distribution samples have achieved convergence before using them for Bayesian inference. PROC BLIFEREG produces three convergence diagnostics by default. If ODS graphics are enabled as specified in the following code, diagnostics plots are also displayed.

Summary statistics of the posterior distribution samples are produced by default. However, these statistics might not be sufficient for carrying out your Bayesian inference. The BAYES statement in the following SAS statements invokes the Bayesian analysis, and the ODS OUTPUT statement saves the samples in the SAS data set Post for further processing:

```
ods graphics on;
proc blifereg data=Fan ;
  model Lifetime*Censor( 1 )= / dist=Lognormal;
  bayes seed=1;
  ods output PosteriorSample=Post;
run;
ods graphics off;
```

The results of this analysis are shown in the following figures.

The “Model Information” table in [Figure 3.2](#) summarizes information about the model you fit and the size of the simulation.

The BLIFEREG Procedure	
Bayesian Analysis	
Model Information	
Data Set	WORK.FAN
Dependent Variable	Log(lifetime)
Censoring Variable	censor
Censoring Value(s)	1
Number of Observations	70
Noncensored Values	12
Right Censored Values	58
Left Censored Values	0
Interval Censored Values	0
Burn-In Size	2000
MC Sample Size	10000
Thinning	1
Name of Distribution	Lognormal
Log Likelihood	-41.64492483

Figure 3.2. Model Information

The “Analysis of Maximum Likelihood Parameter Estimates” table in [Figure 3.3](#) summarizes maximum likelihood estimates of the lognormal intercept and scale parameters.

Analysis of Maximum Likelihood Parameter Estimates					
Parameter	DF	Estimate	Standard Error	95% Confidence Limits	
Intercept	1	10.1432	0.5211	9.1219	11.1646
Scale	1	1.6796	0.3893	1.0664	2.6453

Figure 3.3. Maximum Likelihood Parameter Estimates

Since no prior distribution for the intercept was specified, the default uniform improper distribution shown in the “Uniform Prior for Regression Coefficients” table in Figure 3.4 is used. Noninformative priors are appropriate if you have no prior knowledge of the likely range of values of the parameters, and if you want to make probability statements about the parameters or functions of the parameters. See, for example, [Ibrahim, Chen, and Sinha \(2001\)](#) for more information about choosing prior distributions.

The default noninformative gamma prior distribution for the lognormal scale parameter is shown in the “Independent Prior Distributions for Model Parameters” table in Figure 3.4.

Uniform Prior for Regression Coefficients					
Parameter	Prior				
Intercept	Constant				

Independent Prior Distributions for Model Parameters					
Parameter	Prior Distribution	Hyperparameters			
		Shape	Inverse Scale	Scale	Seed
Scale	Gamma	0.001	0.001	0.001	0.001

Figure 3.4. Model Parameter Priors

By default, the maximum likelihood estimates of the regression parameters are used as the starting value for the simulation. These are listed in the “Initial Values and Seeds” table in Figure 3.5. The specified value of the random number seed is also displayed.

Initial Values and Seeds				
Chain	_SEED_	Intercept	Scale	
1	1	10.14324	1.679545	

Figure 3.5. Markov Chain Initial Values

Summary statistics for the posterior sample are displayed in the “Descriptive

Statistics of the Posterior Samples” and “Interval Statistics of the Posterior Samples” tables in Figure 3.6. Since noninformative prior distributions were used, these results are consistent with the maximum likelihood estimates shown in Figure 3.3.

Descriptive Statistics of the Posterior Samples						
Parameter	N	Mean	Standard Deviation	25%	Quantiles 50%	75%
Intercept	10000	10.4208	0.6273	9.9675	10.3235	10.7694
Scale	10000	1.9216	0.4819	1.5741	1.8383	2.1925

Interval Statistics of the Posterior Samples			
Parameter	Alpha	Credible Interval	HPD Interval
Intercept	0.050	9.4597 11.9495	9.3454 11.7347
Scale	0.050	1.2067 3.0725	1.1093 2.8894

Figure 3.6. Posterior Sample Statistics

By default, PROC BLIFEREG computes three convergence diagnostics: the lag1, lag5, lag10, and lag50 autocorrelations; the Geweke diagnostic; and the effective sample size. These are displayed in Figure 3.7. There is no indication that the Markov chain has not converged. See the “Assessing Markov Chain Convergence” section on page 17 for more information about convergence diagnostics and their interpretation.

Autocorrelations of the Posterior Samples				
Parameter	Lag1	Lag5	Lag10	Lag50
Intercept	0.7085	0.1890	0.0358	-0.0331
Scale	0.7129	0.1919	0.0292	-0.0400

Geweke Diagnostics		
Parameter	z	Pr > z
Intercept	0.1404	0.8884
Scale	0.2753	0.7831

Effective Sample Size			
Parameter	ESS	Correlation Time	Efficiency
Intercept	1680.4	5.9511	0.1680
Scale	1678.7	5.9569	0.1679

Figure 3.7. Diagnostic Statistics

Trace, autocorrelation, and density plots for the three model parameters shown in Figure 3.8 and Figure 3.9 are useful in diagnosing whether the Markov chain of posterior samples has converged. These plots show no evidence that the chain has not converged.

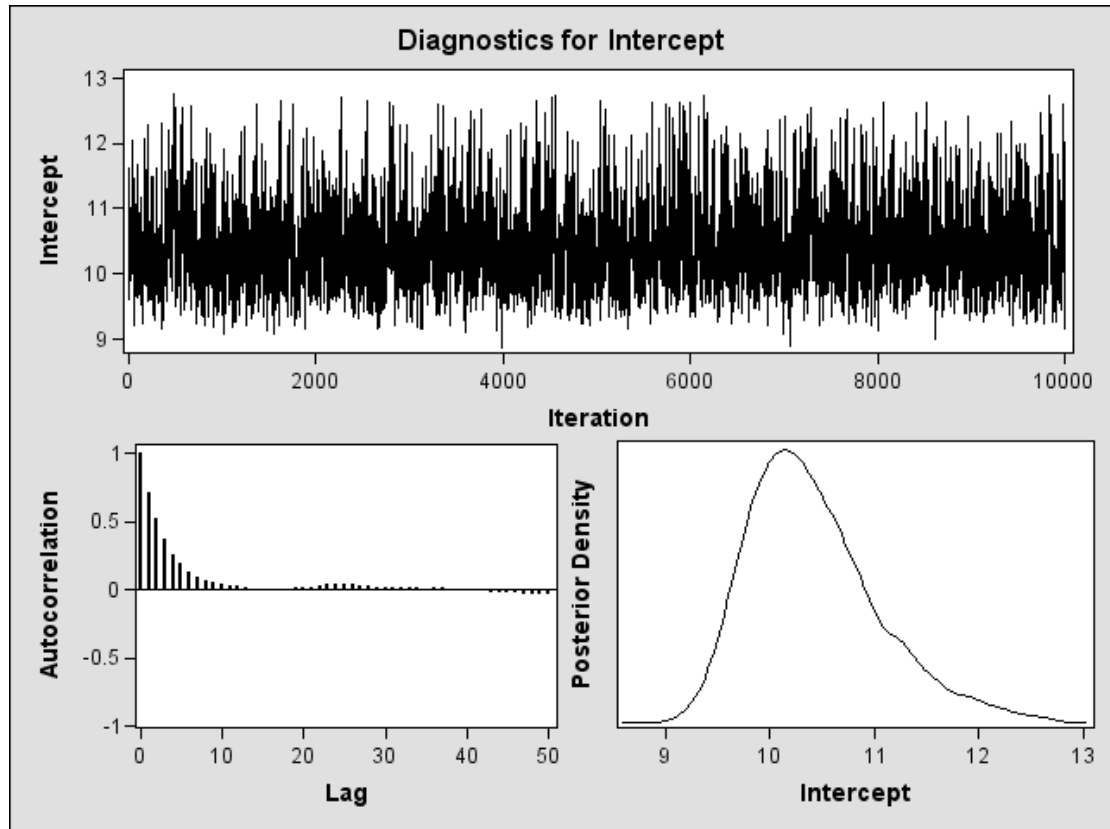


Figure 3.8. Diagnostic Plots for Intercept

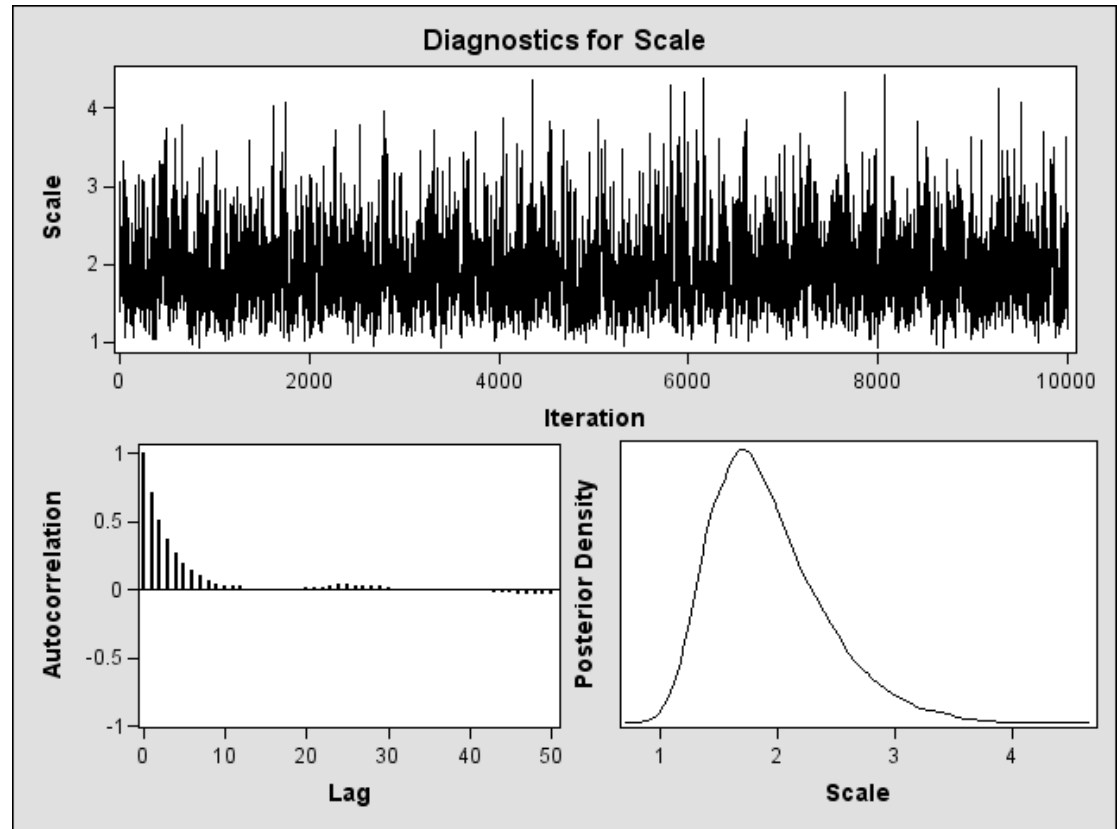


Figure 3.9. Diagnostic Plots for Scale

The fraction failing in the first 8000 hours of operation is a quantity of interest. This kind of information could be useful, for example, in determining whether to improve the reliability of the engine components due to warranty considerations. The following SAS statements compute the mean and percentiles of the distribution of the fraction failing in the first 8000 hours from the posterior sample data set:

```
data Prob;
  set Post;
  Frac = ProbNorm( ( log(8000) - Intercept ) / Scale );
  label Frac= 'Fraction Failing in 8000 Hours';
run;

proc means data = Prob(keep=Frac) n mean p10 p25 p50 p75 p90;
run;
```

The mean fraction of failures in the first 8000 hours is about 0.24, which could be used in further analysis of warranty costs. The 10th percentile is about 0.16 and the 90th percentile is about 0.32, which gives an assessment of the probable range of the fraction failing in the first 8000 hours.

Analysis Variable : Frac Fraction Failing in 8000 Hours						
N	Mean	10th Pctl	25th Pctl	50th Pctl	75th Pctl	90th Pctl
10000	0.2383963	0.1643652	0.1957282	0.2344985	0.2758637	0.3172884

Figure 3.10. Fraction Failing in 8000 Hours

Syntax

The following statements are available in PROC BLIFEREG:

```

PROC BLIFEREG < options > ;
  BAYES < options > ;
  CLASS variable <(options)> <variable <(options)>... >
    < / options >;
  FREQ variable ;
  MODEL response < *censor(list) > = effects < /options > ;

```

The BLIFEREG procedure has the same syntax as PROC LIFEREG, with the addition of the BAYES statement for requesting a Bayesian analysis.

To request a Bayesian analysis, you specify the new BAYES statement in addition to the PROC BLIFEREG statement and the MODEL statement. You include a CLASS statement if you have effects that involve categorical variables. The WEIGHT statement can be included if you have a weight variable in the input data. The binomial distribution is not supported for Bayesian analysis.

BAYES Statement

```
BAYES < options > ;
```

The BAYES statement requests a Bayesian analysis of the regression model by using Gibbs sampling. The Bayesian posterior samples (commonly known as the chain) for the regression parameters are not tabulated. In the following, “PosteriorSample” is the name of a non-printing ODS table that is produced by default whenever the BAYES statement is used. You can create an ODS output data set (named *SAS-data-set*) of the chain by specifying the following:

```
ODS OUTPUT PosteriorSample = SAS-data-set ;
```

Table 3.1 summarizes the options available in the BAYES statement.

Table 3.1. BAYES Statement Options

Option	Description
Monte Carlo Options	
INITIAL=	specifies initial values of the chain
NBI=	specifies the number of burn-in iterations samples

Table 3.1. (continued)

Option	Description
NMC=	specifies the number of iterations after burn-in
SEED=	specifies the random number generator seed
THINNING=	controls the thinning of the Markov chain
Model and Prior Options	
COEFFPRIOR=	specifies the prior of the regression coefficients
EXPONENTIALSCALEPRIOR=	specifies the prior of the exponential scale parameter
SCALEPRIOR=	specifies the prior of the scale parameter
WEIBULLSCALEPRIOR=	specifies the prior of the Weibull scale parameter
WEIBULLSHAPEPRIOR=	specifies the prior of the Weibull shape parameter
Summary Statistics and Convergence Diagnostics	
DIAGNOSTIC=	displays convergence diagnostics
PLOTS=	displays diagnostic plots
SUMMARY=	displays summary statistics of the posterior samples

The following list describes these options and their suboptions.

COEFFPRIOR=NORMAL | UNIFORM <(option)>

COEFF=NORMAL | UNIFORM <(option)>

specifies the prior distribution for the regression coefficients. The default is COEFFPRIOR=UNIFORM, which specifies the noninformative and improper prior of a constant.

The normal prior is specified by COEFFPRIOR=NORMAL, which can be followed by one of the following options enclosed in parentheses. However, if you do not specify an option, the normal prior $N(\mathbf{0}, 10^6\mathbf{I})$, where \mathbf{I} is the identity matrix, is used. See the “Normal Prior” section on page 100.

CONDITIONAL

specifies that the normal prior, conditional on the current Markov chain value of the generalized linear model precision parameter τ , is $N(\boldsymbol{\mu}, \tau^{-1}\boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and covariance of the normal prior specified by other normal options.

INPUT=SAS-data-set

specifies a SAS data set containing the mean and covariance information of the normal prior. The data set must have a `_TYPE_` variable to represent the type of each observation and a variable for each regression coefficient. If the data set also contains a `_NAME_` variable, the values of this variable are used to identify the covariances for the `_TYPE_='COV'` observations; otherwise, the `_TYPE_='COV'` observations are assumed to be in the same order as the explanatory variables in the MODEL statement. PROC BLIFEREG reads the mean vector from the observation with `_TYPE_='MEAN'` and reads the covariance matrix from observations with `_TYPE_='COV'`. For an independent normal prior, the variances can be specified with `_TYPE_='VAR'`; alternatively, the precisions (inverse of the variances) can be specified with `_TYPE_='PRECISION'`.

RELVAR<=*c*>

specifies normal prior $N(\mathbf{0}, c\mathbf{J})$, where \mathbf{J} is a diagonal matrix with diagonal elements equal to the variances of the corresponding ML estimator. By default, $c=10^6$.

VAR=*c*

specifies the normal prior $N(\mathbf{0}, c\mathbf{I})$, where \mathbf{I} is the identity matrix.

DIAGNOSTICS=ALL | NONE | (*keyword-list*)

DIAG=ALL | NONE | (*keyword-list*)

controls the number of diagnostics produced. You can request all the following diagnostics by specifying 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 a subset of the following keywords. The default is DIAGNOSTICS=(AUTOCORR ESS GEWEKE).

AUTOCORR computes the autocorrelations of lags 1, 5, 10, and 50 for each variable. See the “Autocorrelations” section on page 30 for details.

ESS computes Carlin’s estimate of the Effective Sample Size, the correlation time, and the efficiency of the chain for each parameter. See the “Effective Sample Size” section on page 31 for details.

HEIDELBERGER <(heidel-options)>

computes the Heidelberg and Welch diagnostic for each variable, which consists of a stationarity test of the null hypothesis that the sample values form a stationary process. If the stationarity test is passed, a halfwidth test is then carried out. Optionally, you can specify one or more of the following *heidel-options*:

SALPHA=*value*

specifies the α level ($0 < \alpha < 1$) for the stationarity test.

HALPHA=*value*

specifies the α level ($0 < \alpha < 1$) for the halfwidth test.

EPS=*value*

specifies the a small positive number ϵ that if the halfwidth is less than ϵ times the sample mean of the retaining iterates, the halfwidth test is passed.

See the “Heidelberg and Welch Diagnostics” section on page 26 for details.

GELMAN< (*gelman-options*)>

computes the Gelman and Rubin convergence diagnostics. You can specify one or more of the following *gelman-options*:

NCHAIN | **N=***number*

specifies the number of parallel chains used to compute the diagnostic, and has to be 2 or larger. The default is NCHAIN=3. If an INITIAL= data set is used, NCHAIN defaults to the number of rows in the INITIAL= data set. If any number other than this is specified with the NCHAIN= option, the NCHAIN= value is ignored.

ALPHA=*value*

specifies the significance level for the upper bound. The default is ALPHA=0.05, resulting in a 97.5% bound.

See the “[Gelman and Rubin Diagnostics](#)” section on page 22 for details.

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 default is $f_1 = 0.1$ 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.

See the “[Geweke Diagnostics](#)” section on page 24 for details.

RAFTERY<(raftery-options)>

computes the Raftery and Lewis diagnostics that evaluate the accuracy of the estimated quantile ($\hat{\theta}_Q$ for a given $Q \in (0, 1)$) of a chain. $\hat{\theta}_Q$ can achieve any degree of accuracy when the chain is allowed to run for a long time. A stopping criterion is when the estimated probability $\hat{P}_Q = \Pr(\theta \leq \hat{\theta}_Q)$ reaches within $\pm R$ of the value Q with probability S ; i.e., $\Pr(Q - R \leq \hat{P}_Q \leq Q + R) = S$. The *raftery-options* enable you to specify Q , R , S , and a precision level ϵ for a stationary 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 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 tolerance level (a small positive number) for the stationary test. The default is 0.001.

See the “[Raftery and Lewis Diagnostics](#)” section on page 27 for details.

EXPSCALEPRIOR=GAMMA | IMPROPER<(option)>

ESCALEPRIOR=GAMMA | IMPROPER<(option)>

ESCPRIOR=GAMMA | IMPROPER<(option)>

specifies that Gibbs sampling be performed on the exponential distribution scale parameter and the prior distribution for the scale parameter. This prior distribution applies only when the exponential distribution and no covariates are specified.

A gamma prior $G(a, b)$ with density $f(t) = \frac{b(bt)^{a-1}e^{-bt}}{\Gamma(a)}$ is specified by `EXPSCALEPRIOR=GAMMA`, which can be followed by one of the following *gamma-options* enclosed in parentheses. The hyperparameters a and b are the shape and inverse-scale parameters of the gamma distribution, respectively. See the “Gamma Prior” section on page 100 for more details. The default is $G(10^{-4}, 10^{-4})$.

`RELSHAPE<=c>`

specifies independent $G(c\hat{\alpha}, c)$ distribution, where $\hat{\alpha}$ is the MLE of the exponential scale parameter. With this choice of hyperparameters, the mean of the prior distribution is $\hat{\alpha}$ and the variance is $\frac{\hat{\alpha}}{c^2}$. By default, $c=10^{-4}$.

`ESHAPE=a`

and

`EISCALE=b`

specify the $G(a, b)$ prior.

`ESHAPE=c`

specifies the $G(c, c)$ prior.

`EISCALE=c`

specifies the $G(c, c)$ prior.

An improper prior with density $f(t)$ proportional to t^{-1} is specified with `EXPSCALEPRIOR=IMPROPER`.

`GAMMASHAPEPRIOR=NORMAL<(option)>`

`GAMASHAPEPRIOR=NORMAL<(option)>`

`SHAPE1PRIOR=NORMAL<(option)>`

specifies the prior distribution for the gamma distribution shape parameter. If you do not specify any options in a gamma model, the $N(0, 10^6)$ prior for the shape is used. You can specify `MEAN=` and `VAR=` or `RELVAR=` options, either alone or together, to specify the mean and variance of the normal prior for the gamma shape parameter.

`MEAN=a`

specifies a normal prior $N(a, 10^6)$. By default, $a=0$.

`RELVAR<=b>`

specifies the normal prior $N(0, bJ)$, where J is the variance of the MLE of the shape parameter. By default, $b=10^6$.

`VAR=c`

specifies the normal prior $N(0, c)$. By default, $c=10^6$.

`INITIAL=SAS-data-set`

specifies the SAS data set that contains the initial values of the Markov chains. The `INITIAL=` data set must contain all the independent variables of the model. You can specify multiple rows as the initial values of the parallel chains for the Gelman-Rubin statistics, but posterior summaries, diagnostics, and plots are computed only for the first chain. If the data set also contains the variable `_SEED_`, the value of

the `_SEED_` variable is used as the seed of the random number generator for the corresponding chain.

NBI=number

specifies the number of burn-in iterations before the chains are saved. The default is 2000.

NMC=number

specifies the number of iterations after the burn-in. The default is 10000.

SCALEPRIOR=GAMMA<(option)>

specifies that Gibbs sampling be performed on the location-scale model scale parameter and the prior distribution for the scale parameter.

A gamma prior $G(a, b)$ with density $f(t) = \frac{b(bt)^{a-1}e^{-bt}}{\Gamma(a)}$ is specified by `SCALEPRIOR=GAMMA`, which can be followed by one of the following *gamma-options* enclosed in parentheses. The hyperparameters a and b are the shape and inverse-scale parameters of the gamma distribution, respectively. See the “Gamma Prior” section on page 100. The default is $G(10^{-4}, 10^{-4})$.

RELSHAPE<=c>

specifies independent $G(c\hat{\sigma}, c)$ distribution, where $\hat{\sigma}$ is the MLE of the scale parameter. With this choice of hyperparameters, the mean of the prior distribution is $\hat{\sigma}$ and the variance is $\frac{\hat{\sigma}}{c}$. By default, $c=10^{-4}$.

SHAPE=a

and

ISCALE=b

specify the $G(a, b)$ prior.

SHAPE=c

specifies the $G(c, c)$ prior.

ISCALE=c

specifies the $G(c, c)$ prior.

PLOTS<(global-plot-options)>= plot-request

PLOTS<(global-plot-options)>= (plot-request <...plot-request>)

controls the display of diagnostic plots. Three types of plots can be requested: trace plots, autocorrelation function plots, and kernel density plots. By default, the plots are displayed in panels unless the global plot option `UNPACK` is specified. Also, when specifying more than one type of plots, the plots are displayed by parameters unless the global plot option `TYPE` is specified. When you specify only one plot request, you can omit the parentheses around the plot request. For example:

```
plots=trace
plots=(trace density)
```

The global plot options are as follows:

TYPE specifies how the plots are grouped when there is more than one type of plot. TYPE specifies that the plots are grouped by type. The default is that the plots are grouped by parameter.

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

The plot requests are as follows:

ALL specifies all types of plots. PLOTS=ALL is equivalent to specifying PLOTS=(TRACE AUTOCORR DENSITY).

AUTOCORR displays the autocorrelation function plots for the regression parameters.

DENSITY displays the kernel density plots for the regression parameters.

NONE suppresses the display of any plots.

TRACE displays the trace plots for the regression parameters.

See the “[Visual Analysis via Trace Plots](#)” section on page 18 for details.

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 to reproduce identical Markov chains for the same specification. If the SEED= option is not specified, or if you specify a nonpositive seed, a seed is derived from the time of day.

SUMMARIES=ALL | NONE | (keyword-list)

SUMMARY=ALL | NONE | (keywords-list)

SUM=ALL | NONE | (keywords-list)

control the number of posterior summaries produced. SUMMARIES=ALL produces all the summary statistics, which include the mean, standard deviation, quartiles, credible intervals, and HPD intervals for each parameter. If you do not want any posterior summaries, you specify SUMMARIES=NONE. You can use the following keywords to request only the descriptive statistics or the credible and HPD intervals of a given level, or the correlation matrix. The default is SUMMARIES=(DESCRIPTIVE INTERVAL).

DESCRIPTIVE

DESC

produces the means, standard deviations, and quartiles for the posterior sample.

INTERVAL<(ALPHA=numeric-list)>

produces the $100(1 - \alpha)\%$ credible interval and the $100(1 - \alpha)\%$ HPD interval for each parameter and for each α in the *numeric-list* specified in the ALPHA= option. The default is ALPHA=0.05.

CORR

produces the correlation matrix of the posterior samples.

See the “[Summary Statistics](#)” section on page 31 for details.

THINNING=number

THIN=number

controls the thinning of the Markov chain. Only one in every k samples is used when THINNING= k , and if NBI= n_0 and NMC= n , the number of samples kept is

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

where $[a]$ represents the integer part of the number a . The default is THINNING=1.

WEIBULLSCALEPRIOR=GAMMA<(option)>

WSCALEPRIOR=GAMMA<(option)>

WSCPRIOR=GAMMA<(option)>

specifies that Gibbs sampling be performed on the Weibull model scale parameter and the prior distribution for the scale parameter. This option applies only when a Weibull distribution and no covariates are specified. When this option is specified, PROC BLIFEREG performs Gibbs sampling on the Weibull scale parameter, which is defined as $\exp(\mu)$, where μ is the intercept term.

A gamma prior $G(a, b)$ with density $f(t) = \frac{b(bt)^{a-1}e^{-bt}}{\Gamma(a)}$ is specified by WEIBULLSCALEPRIOR=GAMMA, which can be followed by one of the following *gamma-options* enclosed in parentheses. The hyperparameters a and b are the shape and inverse-scale parameters of the gamma distribution, respectively. See the “[Gamma Prior](#)” section on page 100. The default is $G(10^{-4}, 10^{-4})$.

RELSHAPE<=c>

specifies independent $G(c\hat{\alpha}, c)$ distribution, where $\hat{\alpha}$ is the MLE of the Weibull scale parameter. With this choice of hyperparameters, the mean of the prior distribution is $\hat{\alpha}$ and the variance is $\frac{\hat{\alpha}}{c}$. By default, $c=10^{-4}$.

WCSHAPE=a

and

WSCISCALE=b

specify the $G(a, b)$ prior

WCSHAPE=c

specifies the $G(c, c)$ prior.

WSCISCALE=c

specifies the $G(c, c)$ prior.

WEIBULLSHAPEPRIOR=GAMMA<(option)>

WSHAPEPRIOR=GAMMA<(option)>

WSPRIOR=GAMMA<(option)>

specifies that Gibbs sampling be performed on the Weibull model shape parameter and the prior distribution for the shape parameter. When this option is specified, PROC BLIFEREG performs Gibbs sampling on the Weibull shape parameter, which is defined as σ^{-1} , where σ is the location-scale model scale parameter.

A gamma prior $G(a, b)$ with density $f(t) = \frac{b(bt)^{a-1}e^{-bt}}{\Gamma(a)}$ is specified by WEIBULLSHAPEPRIOR=GAMMA, which can be followed by one of the following *gamma-options* enclosed in parentheses. The hyperparameters a and b are the shape and inverse-scale parameters of the gamma distribution, respectively. See the “Gamma Prior” section on page 100. The default is $G(10^{-4}, 10^{-4})$.

RELSHAPE<= c >

specifies independent $G(c\hat{\beta}, c)$ distribution, where $\hat{\beta}$ is the MLE of the Weibull shape parameter. With this choice of hyperparameters, the mean of the prior distribution is $\hat{\beta}$ and the variance is $\frac{\hat{\beta}}{c}$. By default, $c=10^{-4}$.

WSHSHAPE= a

and

WSHISCALE= b

specify the $G(a, b)$ prior.

WSHSHAPE= c

specifies the $G(c, c)$ prior.

WSHISCALE= c

specifies the $G(c, c)$ prior.

Details

Gibbs Sampling

This section provides details about Gibbs sampling in the location-scale models for survival data available in PROC LIFEREG. See the “Gibbs Sampler” section on page 15 for a general discussion of Gibbs sampling. PROC LIFEREG fits parametric location-scale survival models. That is, the probability density of the response Y can be expressed in the general form

$$f(y) = g\left(\frac{y - \mu}{\sigma}\right)$$

where $Y = \log(T)$ for lifetimes T . The function g determines the specific distribution. The location parameter μ_i is modeled through regression parameters as $\mu_i = \mathbf{x}_i' \boldsymbol{\beta}$. The LIFEREG procedure jointly estimates the regression parameters and the scale parameter σ by maximum likelihood. The BLIFEREG procedure can provide Bayesian estimates of the regression parameters and σ .

For the Weibull distribution, you can specify that Gibbs sampling be performed on the Weibull shape parameter $\beta = \sigma^{-1}$ instead of the scale parameter σ by specifying a prior distribution for the shape parameter with the WEIBULLSHAPEPRIOR= option. In addition, if there are no covariates in the model, you can specify Gibbs sampling on the Weibull scale parameter $\alpha = \exp(\mu)$, where μ is the intercept term, with the WEIBULLSCALEPRIOR= option.

In the case of the exponential distribution with no covariates, you can specify Gibbs sampling on the exponential scale parameter $\alpha = \exp(\mu)$, where μ is the intercept term, with the `EXPSCALEPRIOR=` option.

Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)'$ be the parameter vector. For location-scale models, the θ_i 's are the regression coefficients β_i 's and the scale parameter σ . In the case of the three parameter gamma distribution, there is an additional gamma shape parameter τ . Let $L(D|\boldsymbol{\theta})$ be the likelihood function, where D is the observed data. Let $\pi(\boldsymbol{\theta})$ be the prior distribution. The full conditional distribution of $[\theta_i|\theta_j, i \neq j]$ is proportional to the joint distribution; that is,

$$\pi(\theta_i|\theta_j, i \neq j, D) \propto L(D|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

For instance, the one-dimensional conditional distribution of θ_1 given $\theta_j = \theta_j^*, 2 \leq j \leq k$, is computed as

$$\pi(\theta_1|\theta_j = \theta_j^*, 2 \leq j \leq k, D) = L(D|(\boldsymbol{\theta} = (\theta_1, \theta_2^*, \dots, \theta_k^*)'))p(\boldsymbol{\theta} = (\theta_1, \theta_2^*, \dots, \theta_k^*)')$$

Suppose you have a set of arbitrary starting values $\{\theta_1^{(0)}, \dots, \theta_k^{(0)}\}$. Using the ARMS (adaptive rejection Metropolis sampling) algorithm of [Gilks and Wild \(1992\)](#) and [Gilks, Best, and Tan \(1995\)](#), you can do the following:

```
draw  $\theta_1^{(1)}$  from  $[\theta_1|\theta_2^{(0)}, \dots, \theta_k^{(0)}]$ 
draw  $\theta_2^{(1)}$  from  $[\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \dots, \theta_k^{(0)}]$ 
...
draw  $\theta_k^{(1)}$  from  $[\theta_k|\theta_1^{(1)}, \dots, \theta_{k-1}^{(1)}]$ 
```

This completes one iteration of the Gibbs sampler. After one iteration, you have $\{\theta_1^{(1)}, \dots, \theta_k^{(1)}\}$. After n iterations, you have $\{\theta_1^{(n)}, \dots, \theta_k^{(n)}\}$. PROC BLIFEREG implements the ARMS algorithm based on code provided by [Gilks \(2003\)](#) to draw a sample from a full conditional distribution.

You can output these posterior samples into a SAS data set through ODS. The following SAS statement outputs the posterior samples into the SAS data set `Post`:

```
ods output PosteriorSample=Post;
```

The data set also includes the variable `_LOGPOST_`, representing the log of the posterior log likelihood.

Priors for Model Parameters

The model parameters are the regression coefficients and the dispersion parameter (or the precision or scale), if the model has one. The priors for the dispersion parameter and the priors for the regression coefficients are assumed to be independent, while you can have a joint multivariate normal prior for the regression coefficients.

Scale and Shape Parameters

Gamma Prior

The gamma distribution $G(a, b)$ has a pdf

$$f_{a,b}(u) = \frac{b(bu)^{a-1}e^{-bu}}{\Gamma(a)}, \quad u > 0$$

where a is the shape parameter and b is the inverse-scale parameter. The mean is $\frac{a}{b}$ and the variance is $\frac{a}{b^2}$.

Improper Prior

The joint prior density is given by

$$p(u) \propto u^{-1}, \quad u > 0$$

Regression Coefficients

Let β be the regression coefficients.

Normal Prior

Assume β has a multivariate normal prior with mean vector β_0 and covariance matrix Σ_0 . The joint prior density is given by

$$p(\beta) \propto e^{-\frac{1}{2}(\beta-\beta_0)'\Sigma_0^{-1}(\beta-\beta_0)}$$

Uniform Prior

The joint prior density is given by

$$p(\beta) \propto 1$$

Posterior Distribution

Denote the observed data by D .

The posterior distribution is

$$\pi(\beta|D) \propto L_P(D|\beta)p(\beta)$$

where $L_P(D|\beta)$ is the likelihood function with regression coefficients β as parameters.

Starting Values of the Markov Chains

When the BAYES statement is specified, PROC BLIFEREG generates one Markov chain containing the approximate posterior samples of the model parameters. Additional chains are produced when the Gelman-Rubin diagnostics are requested. Starting values (or initial values) can be specified in the INITIAL= data set in the BAYES statement. If INITIAL= option is not specified, PROC BLIFEREG picks its own initial values for the chains.

Denote $[x]$ as the integral value of x . Denote $\hat{s}(X)$ as the estimated standard error of the estimator X .

Regression Coefficients and Gamma Shape Parameter

For the first chain that the summary statistics and regression diagnostics are based on, the initial values are the maximum likelihood estimates; that is,

$$\beta_i^{(0)} = \hat{\beta}_i$$

Initial values for the r th chain ($2 \leq r$) are given by

$$\beta_i^{(0)} = \hat{\beta}_i \pm \left(2 + \left[\frac{r}{2}\right]\right) \hat{s}(\hat{\beta}_i)$$

with the plus sign for odd r and minus sign for even r .

Scale, Exponential Scale, Weibull Scale, or Weibull Shape Parameter λ

Let λ be the parameter sampled.

For the first chain that the summary statistics and diagnostics are based on, the initial values are the maximum likelihood estimates; that is,

$$\lambda^{(0)} = \hat{\lambda}$$

The initial values of the r th chain ($2 \leq r$) are given by

$$\lambda^{(0)} = \hat{\lambda} e^{\pm \left(\left[\frac{r}{2}\right] + 2\right) \hat{s}(\hat{\lambda})}$$

with the plus sign for odd r and minus sign for even r .

Displayed Output

The displayed output for a Bayesian analysis includes the following.

Model Information

The “Model Information” table displays the two-level name of the input data set, the number of burn-in iterations, the number of iterations after the burn-in, the number of thinning iterations, the distribution name, and the name and label of the dependent variable; the name and label of the censor indicator variable, for right-censored data; if you specify the WEIGHT statement, the name and label of the weight variable; and the maximum value of the log likelihood.

Class Level Information

The “Class Level Information” table displays the levels of class variables if you specify a CLASS statement.

Maximum Likelihood Estimates

The “Analysis of Maximum Likelihood Parameter Estimates” table displays the maximum likelihood estimate of each parameter, the estimated standard error of the parameter estimator, and confidence limits for each parameter.

Coefficient Prior

The “Coefficient Prior” table displays the prior distribution of the regression coefficients.

Independent Prior Distributions for Model Parameters

The “Independent Prior Distributions for Model Parameters” table displays the prior distributions of additional model parameters (scale, exponential scale, Weibull scale, Weibull shape, gamma shape).

Initial Values and Seeds

The “Initial Values and Seeds” table displays the initial values and random number generator seeds for the Gibbs chains.

Descriptive Statistics of the Posterior Samples

The “Descriptive Statistics of the Posterior Sample” table contains the size of the sample, the mean, the standard error, and the quartiles for each model parameter.

Interval Estimates for Posterior Sample

The “Interval Estimates for Posterior Sample” table contains the HPD intervals and the credible intervals for each model parameter.

Correlation Matrix of the Posterior Samples

The “Correlation Matrix of the Posterior Samples” table is produced if you include the CORR suboption in the SUMMARY= option in the BAYES statement. This table displays the sample correlation of the posterior samples.

Autocorrelations of the Posterior Samples

The “Autocorrelations of the Posterior Samples” table displays the lag1, lag5, lag10, and lag50 autocorrelations for each parameter.

Gelman and Rubin Diagnostics

The “Gelman and Rubin Diagnostics” table is produced if you include the GELMAN suboption in the DIAGNOSTIC= option in the BAYES statement. This table displays the estimate of the potential scale reduction factor and its 97.5% upper confidence limit for each parameter.

Geweke Diagnostics

The “Geweke Diagnostics” table displays the Geweke statistic and its p -value for each parameter.

Raftery and Lewis Diagnostics

The “Raftery Diagnostics” tables is produced if you include the RAFTERY suboption in the DIAGNOSTIC= option in the BAYES statement. This table displays the Raftery and Lewis diagnostics for each variable.

Heidelberger and Welch Diagnostics

The “Heidelberger and Welch Diagnostics” table is displayed if you include the HEIDELBERGER suboption in the DIAGNOSTIC= option in the BAYES statement. This table shows the results of a stationary test and a halfwidth test for each parameter.

Effective Sample Size

The “Effective Sample Size” table displays, for each parameter, the effective sample size, the correlation time, and the efficiency.

ODS Table Names

PROC BLIFEREG assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in Table 3.2.

Table 3.2. ODS Tables Produced by PROC BLIFEREG

ODS Table Name	Description	Statement	Option
AutoCorr	Autocorrelations of the posterior samples	BAYES	default
ChainStatistics	Simple descriptive statistics for the chain	BAYES	default

Table 3.2. (continued)

ODS Table Name	Description	Statement	Option
ClassLevels	Levels of class variables	CLASS	default
CoeffPrior	Prior distribution of the regression coefficients	BAYES	default
ConvergenceStatus	Convergence status of maximum likelihood estimation	MODEL	default
Corr	Correlation matrix of the posterior samples	BAYES	SUMMARY=CORR
ESS	Effective sample size	BAYES	default
Gelman	Gelman and Rubin convergence diagnostics	BAYES	DIAG=GELMAN
Geweke	Geweke convergence diagnostics	BAYES	default
Heidelberger	Heidelberger and Welch convergence diagnostics	BAYES	DIAG=HEIDELBERGER
InitialValues	Initial values of the Markov chains	BAYES	default
IntervalStatistics	HPD and credible intervals for the posterior samples	BAYES	default
ModelInfo	Model information	PROC	default
NObs	Number of observations		default
ParameterEstimates	Maximum likelihood estimates of model parameters	MODEL	default
ParmPrior	Prior distribution for scale and shape	BAYES	default
PosteriorSample	Posterior samples (for ODS output data set only)	BAYES	
Raftery	Raftery and Lewis convergence diagnostics	BAYES	DIAG=RAFTERY
Type3Analysis	Type III statistics for model effects	MODEL	default

ODS Graph Names

Each statistical graphic created by PROC BLIFEREG has a name associated with it, and you can reference the graph by using ODS statements. These names are listed in Table 3.3.

Table 3.3. ODS Graphics Produced by PROC BLIFEREG

ODS Graph Name	Description	Statement	Option
ADPanel	Autocorrelation function and density panel	BAYES	PLOTS=(AUTOCORR DENSITY)
AutocorrPanel	Autocorrelation function panel	BAYES	PLOTS= AUTOCORR
AutocorrPlot	Autocorrelation function plot	BAYES	PLOTS(UNPACK)=AUTOCORR
DensityPanel	Density panel	BAYES	PLOTS=DENSITY
DensityPlot	Density plot	BAYES	PLOTS(UNPACK)=DENSITY

Table 3.3. (continued)

ODS Graph Name	Description	Statement	Option
TAPanel	Trace and autocorrelation function panel	BAYES	PLOTS=(TRACE AUTOCORR)
TADPanel	Trace, density, and autocorrelation function panel	BAYES	default
TDPanel	Trace and density panel	BAYES	PLOTS=(TRACE DENSITY)
TracePanel	Trace panel	BAYES	PLOTS=TRACE
TracePlot	Trace plot	BAYES	PLOTS(UNPACK)=TRACE

Example

Consider the data on melanoma patients from a clinical trial in [Figure 3.11](#). These data are described in [Ibrahim, Chen, and Sinha \(2001\)](#).

The survival time is modeled by a Weibull regression model with three covariates. An analysis of the right-censored survival data is performed using PROC BLIFEREG to obtain Bayesian estimates of the regression coefficients by using the following SAS statements:

```
ods graphics on;
proc blifereg data=e1684;
  class Sex;
  model Survtime*Surv cens(1)=Age Sex Perform / dist=Weibull;
  bayes WeibullShapePrior=gamma diagnostics=(AutoCorr);
run;
ods graphics off;
```

Obs	Survtime	Surv cens	Age	Sex	Perform
1	1.57808	2	35.9945	1	0
2	1.48219	2	41.9014	1	0
3	7.33425	1	70.2164	2	0
4	0.65479	2	58.1753	2	1
5	2.23288	2	33.7096	1	0
6	9.38356	1	47.9726	1	0
7	3.27671	2	31.8219	2	0
8	0.00000	1	72.3644	2	0
9	0.80274	2	40.7151	2	0
10	9.64384	1	32.9479	1	0
.					
.					
.					
277	4.81370	1	57.3726	2	0
278	4.50137	2	29.7726	2	0
279	3.92329	2	51.8822	2	0
280	4.86027	1	65.3123	2	0
281	0.52603	2	52.0658	2	0
282	2.10685	2	60.9534	2	0
283	4.24384	1	32.6055	2	0
284	3.39178	2	51.5123	2	1
285	4.36164	1	48.6548	1	0
286	4.81918	1	43.8438	2	0

Figure 3.11. Melanoma Data

Type III statistics and maximum likelihood estimates of the model parameters shown in [Figure 3.12](#) are displayed by default.

Type III Analysis of Effects					
Effect	DF	Wald			
		Chi-Square	Pr > ChiSq		
Age	1	2.6683	0.1024		
Sex	1	0.3496	0.5544		
Perform	1	0.8127	0.3673		
Analysis of Maximum Likelihood Parameter Estimates					
Parameter	DF	Estimate	Standard Error	95% Confidence Limits	
Intercept	1	2.4402	0.3716	1.7119	3.1685
Age	1	-0.0115	0.0070	-0.0253	0.0023
Sex	1	-0.1170	0.1978	-0.5046	0.2707
Sex	2	0.0000	.	.	.
Perform	1	0.2905	0.3222	-0.3411	0.9220
Scale	1	1.2537	0.0824	1.1021	1.4260
Weibull Shape	1	0.7977	0.0524	0.7012	0.9073

Figure 3.12. Maximum Likelihood Parameter Estimates

Since no prior distributions for the regression coefficients were specified, the default uniform improper distributions shown in the “Uniform Prior for Regression

Coefficients” table in [Figure 3.13](#) are used. The specified gamma prior for the Weibull shape parameter is also shown in [Figure 3.13](#).

Uniform Prior for Regression Coefficients					
Parameter	Prior				
Intercept	Constant				
Age	Constant				
Sex1	Constant				
Perform	Constant				

Independent Prior Distributions for Model Parameters					
Parameter	Prior Distribution	Hyperparameters			
Weibull Shape	Gamma	Shape	0.001	Inverse Scale	0.001

Figure 3.13. Model Parameter Priors

Descriptive and interval statistics for the posterior sample are displayed in the tables in [Figure 3.14](#). Since noninformative prior distributions for the regression coefficients were used, the mean and standard deviations of the posterior distributions for the model parameters are close to the maximum likelihood estimates and standard errors.

Descriptive Statistics of the Posterior Samples						
Parameter	N	Mean	Standard Deviation	25%	50%	75%
Intercept	10000	2.4710	0.3816	2.2072	2.4553	2.7184
Age	10000	-0.0116	0.00730	-0.0165	-0.0113	-0.00661
Sex1	10000	-0.1261	0.2029	-0.2653	-0.1242	0.0128
Perform	10000	0.3314	0.3307	0.1036	0.3181	0.5464
WeibShape	10000	0.7839	0.0521	0.7482	0.7823	0.8190

Interval Statistics of the Posterior Samples					
Parameter	Alpha	Credible Interval		HPD Interval	
Intercept	0.050	1.7621	3.2529	1.7611	3.2514
Age	0.050	-0.0264	0.00222	-0.0268	0.00174
Sex1	0.050	-0.5295	0.2619	-0.5184	0.2703
Perform	0.050	-0.2786	1.0170	-0.3068	0.9784
WeibShape	0.050	0.6849	0.8887	0.6853	0.8889

Figure 3.14. Posterior Sample Statistics

The requested autocorrelation table is shown in [Figure 3.15](#).

Autocorrelations of the Posterior Samples				
Parameter	Lag1	Lag5	Lag10	Lag50
Intercept	0.9344	0.7079	0.4989	0.0996
Age	0.9278	0.6886	0.4810	0.0999
Sex1	0.6324	0.1165	0.0093	0.0205
Perform	0.0930	0.0076	0.0031	-0.0084
WeibShape	0.0864	0.0122	0.0191	0.0093

Figure 3.15. Posterior Sample Autocorrelations

Trace, autocorrelation, and density plots for the seven model parameters are shown in [Figure 3.16](#) through [Figure 3.20](#). These plots show no indication that the Markov chains have not converged.

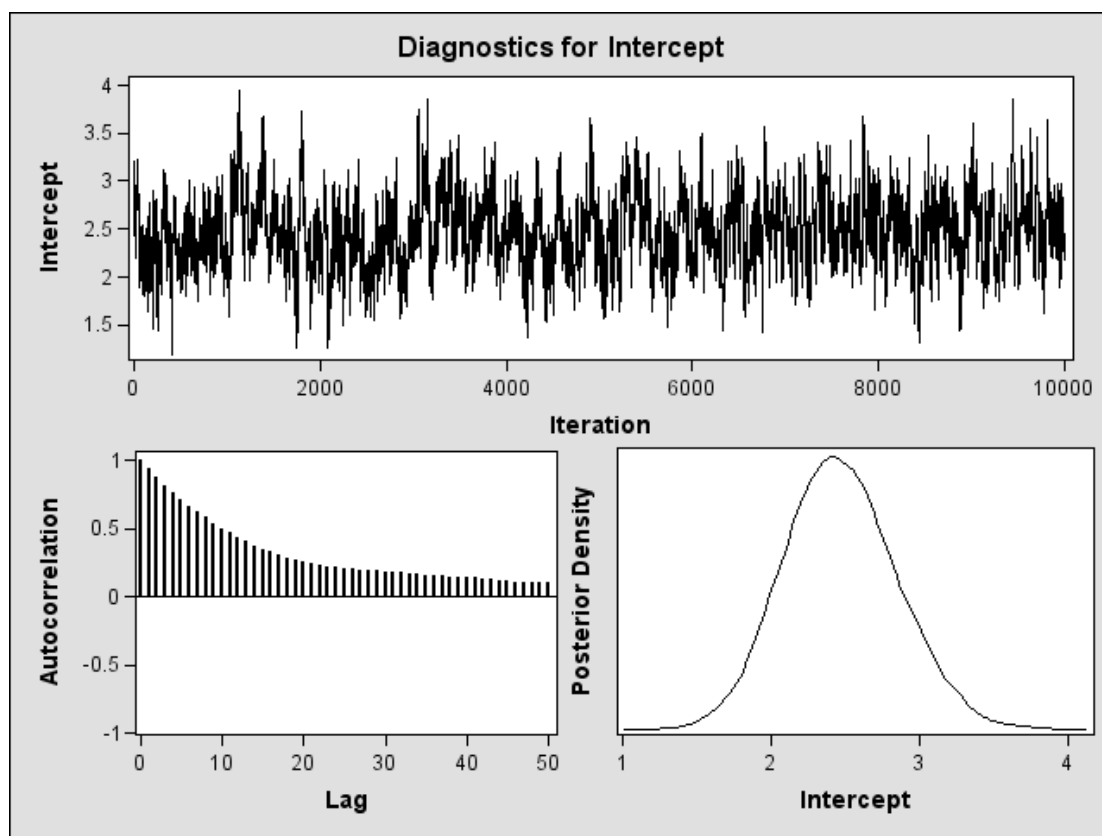


Figure 3.16. Diagnostic Plots for Intercept

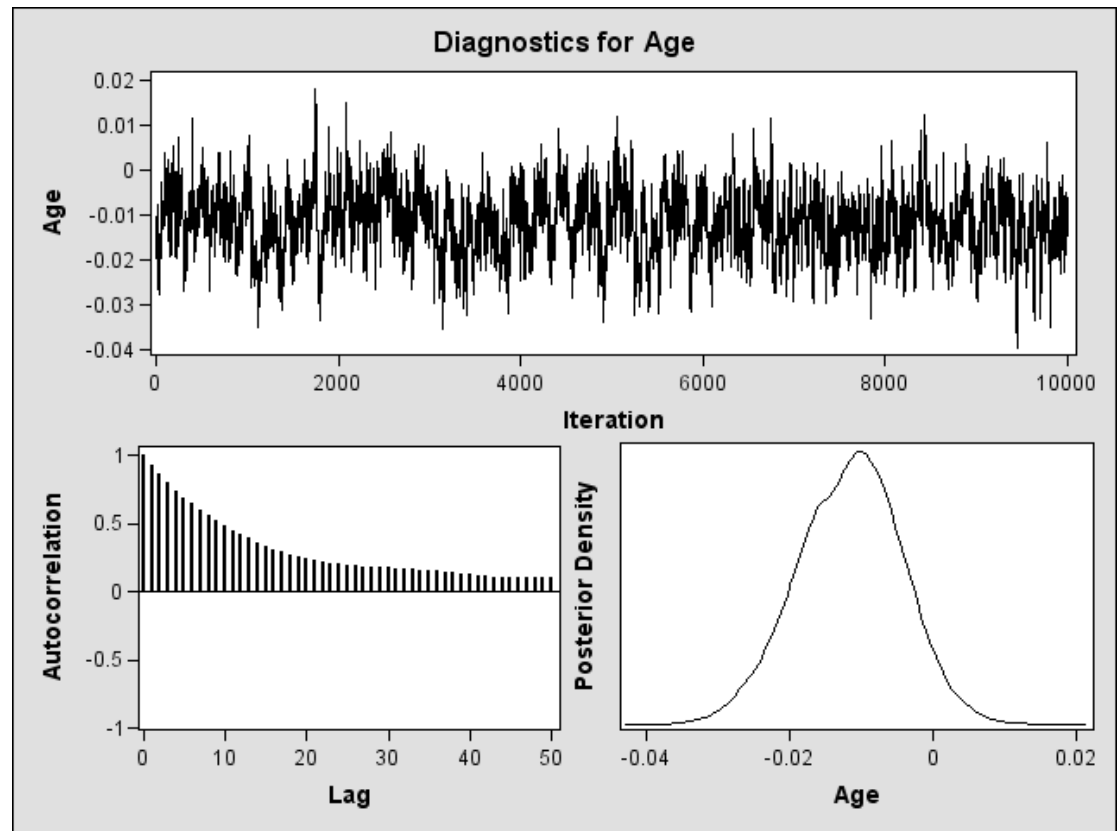


Figure 3.17. Diagnostic Plots for Age

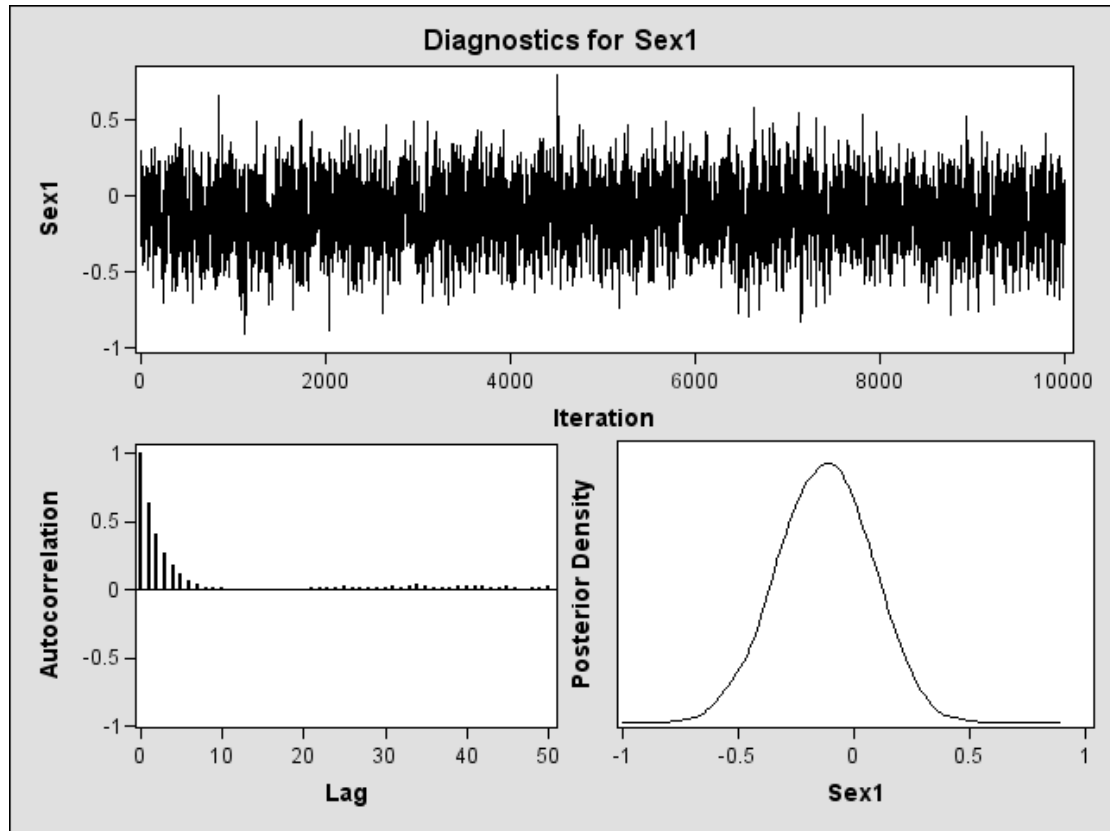


Figure 3.18. Diagnostic Plots for Sex

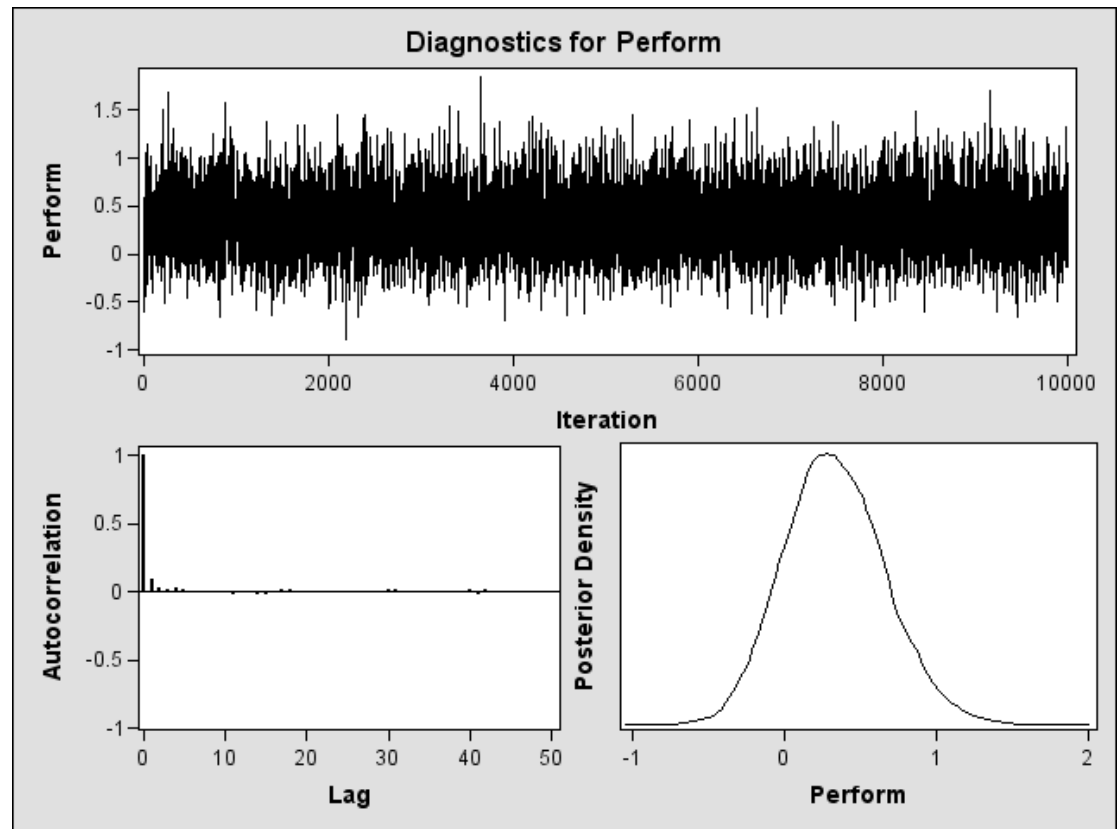


Figure 3.19. Diagnostic Plots for Perform

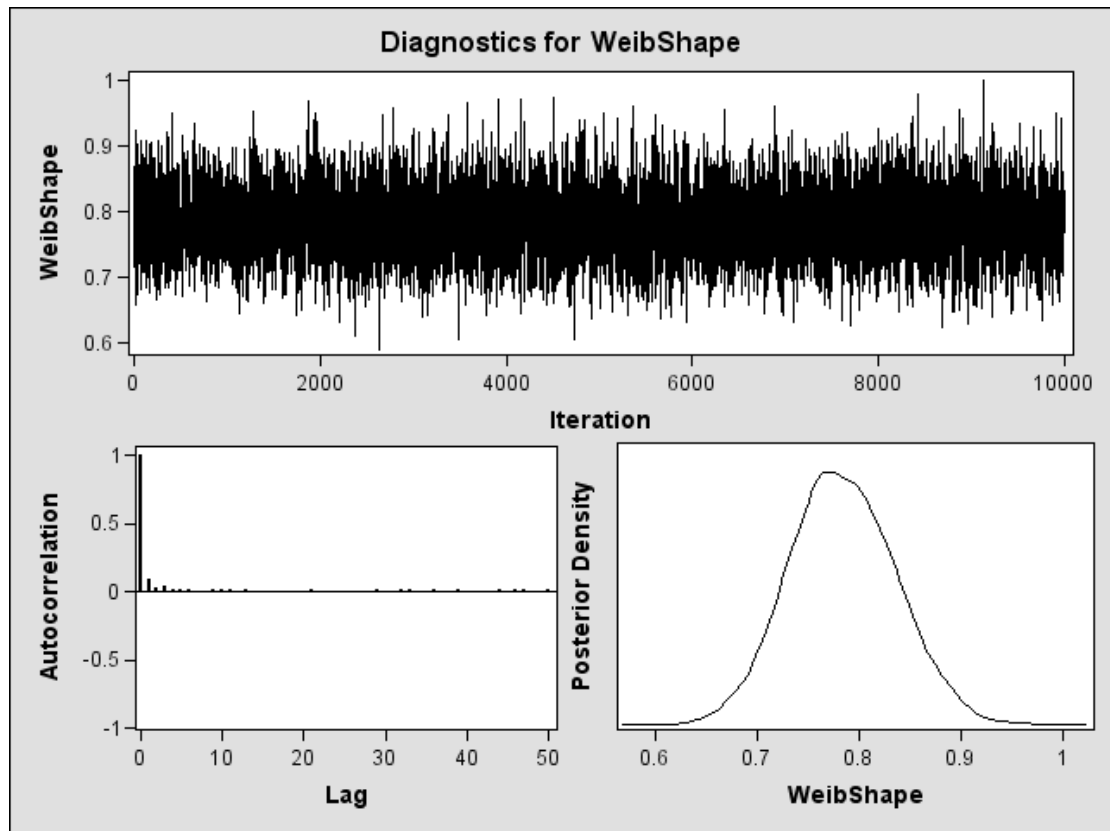


Figure 3.20. Diagnostic Plots for Weibull Shape

References

- Gilks, W. (2003), “Adaptive Metropolis Rejection Sampling (ARMS),” software from MRC Biostatistics Unit, Cambridge, UK, http://www.maths.leeds.ac.uk/~wally.gilks/adaptive.rejection/web_page/Welcome.html.
- Gilks, W., Best, N., and Tan, K. (1995), “Adaptive Rejection Metropolis Sampling with Gibbs Sampling,” *Applied Statistics*, 44, 455–472.
- Gilks, W. and Wild, P. (1992), “Adaptive Rejection Sampling for Gibbs Sampling,” *Applied Statistics*, 41, 337–348.
- Ibrahim, J., Chen, M., and Sinha, D. (2001), *Bayesian Survival Analysis*, New York: Springer-Verlag.
- Nelson, W. (1982), *Applied Life Data Analysis*, New York: John Wiley & Sons.

Chapter 4

The BPHREG Procedure (Experimental)

Chapter Contents

OVERVIEW	115
GETTING STARTED	116
Two-Sample Carcinogenesis Data	116
SYNTAX	121
BAYES Statement	122
HAZARDRATIO Statement	130
DETAILS	132
Piecewise Constant Baseline Hazard Model	132
Priors for Model Parameters	135
Posterior Distribution	137
Sampling from the Posterior Distribution	138
Starting Values of the Markov Chains	139
Fit Statistics	140
Displayed Output	140
ODS Table Names	143
ODS Graph Names	144
EXAMPLES	145
Example 4.1. Informative Prior and Hazard Ratio Analysis	145
Example 4.2. Piecewise Exponential Model	154
REFERENCES	158

Chapter 4

The BPHREG Procedure

(Experimental)

Overview

The experimental BPHREG procedure adds Bayesian analysis to the PHREG procedure. In essence, the Bayesian paradigm treats parameters as random variables, and inference (measurement of uncertainty) about parameters is based on the posterior distribution of the parameters. A posterior distribution is a weighted likelihood function of the data with a prior distribution using the Bayes theorem. Without any past experience or knowledge of what prior distribution to use, you can always start with a noninformative prior. Knowledge of the prior is accumulated over time, and the Bayesian approach can be viewed as a process of learning from experience. A closed form of the posterior distribution is hard to come by, and a Markov chain Monte Carlo method is used to simulate samples from the distribution.

See [Chapter 1, “Introduction to Bayesian Analysis Procedures,”](#) for an introduction to the basic concepts in Bayesian statistics. You can also refer to the section [“Bayesian Analysis: Advantages and Disadvantages”](#) on page 11 for a discussion of the advantages and disadvantages of Bayesian analysis.

Bayesian analysis of the Cox model and the piecewise constant baseline hazard model (also known as the piecewise exponential model) can be requested using the BAYES statement in the BPHREG procedure. For the Cox model, the partial likelihood is used as the likelihood, which is justified by [Sinha, Ibrahim, and Chen \(2003\)](#). For a Bayesian analysis, PROC BPHREG generates a chain of posterior distribution samples by the Gibbs sampler, using the adaptive rejection sampling algorithm ([Gilks and Wild 1992](#); [Gilks, Best, and Tan 1995](#)) to sample each parameter value from its full conditional distribution. Summary statistics (mean, standard deviation, quartiles, HPD intervals, and credible intervals) and convergence diagnostics (autocorrelations; Gelman-Rubin, Geweke, Raftery-Lewis, and Heidelberger-Welch tests; and the effective sample size) are computed for each parameter, as well as the correlation matrix of the posterior samples. Trace plots, posterior density plots, and autocorrelation function plots are also provided using the experimental ODS graphics.

In addition to the BAYES statement, a new HAZARDRATIO statement is available in PROC BPHREG. This statement enables you to compute the hazard ratios at customized settings for both Bayesian and non-Bayesian analyses.

The BPHREG procedure includes the full functionality of the PHREG procedure as documented in the SAS/STAT 9.1 documentation. It also includes the CLASS statement as documented for the experimental TPHREG procedure. The Bayesian capabilities will be included in the PHREG procedure in the next release of SAS/STAT software.

We are eager for your feedback on this experimental procedure. Please send comments to bphreg@sas.com.

Getting Started

Two-Sample Carcinogenesis Data

Consider the vaginal cancer mortality data from [Kalbfleisch and Prentice \(1980, p. 2\)](#). Two groups of rats received different pretreatment regimes and then were exposed to a carcinogen. Survival times of the rats were recorded, from exposure to death from vaginal cancer. Four rats died of other causes, so their survival times are censored. Investigators are interested in knowing whether the different pretreatment regimes affect the survival of the rats; in particular, they would like to compute the probability that the hazard rate of one group is greater than that of the other group. Bayesian analysis provides the framework to answer such specific questions.

In the following DATA step, the data set `Rats` contains the variable `Days` (the survival time in days), the variable `Status` (the censoring indicator variable: 0 if censored and 1 if not censored), and the variable `Group` (the pretreatment group indicator with values 0 and 1).

```
data Rats;
  label Days = 'Days from Exposure to Death';
  input Days Status Group @@;
  datalines;
143 1 0   164 1 0   188 1 0   188 1 0
190 1 0   192 1 0   206 1 0   209 1 0
213 1 0   216 1 0   220 1 0   227 1 0
230 1 0   234 1 0   246 1 0   265 1 0
304 1 0   216 0 0   244 0 0   142 1 1
156 1 1   163 1 1   198 1 1   205 1 1
232 1 1   232 1 1   233 1 1   233 1 1
233 1 1   233 1 1   239 1 1   240 1 1
261 1 1   280 1 1   280 1 1   296 1 1
296 1 1   323 1 1   204 0 1   344 0 1
;
run;
```

The Cox model is used with `Group` as the only explanatory variable. The `BAYES` statement invokes the Bayesian analysis. The `SEED=` option is specified to maintain reproducibility; no other options are specified in the `BAYES` statement. By default, a uniform prior distribution is assumed on the regression coefficient `Group`. The uniform prior is a flat prior on the real line with a distribution that reflects ignorance of the location of the parameter, placing equal probability on all possible values the regression coefficient can take. Using the uniform prior in the following example, you would expect the Bayesian estimates to resemble the classical results of maximizing the likelihood. If you can elicit an informative prior on the regression coefficients, you should use the `COEFFPRIOR=` option to specify it.

You should make sure that the posterior distribution samples have achieved convergence before using them for Bayesian inference. PROC BPHREG produces three convergence diagnostics by default. If you enable ODS graphics before calling PROC BPHREG as in the following code, diagnostics plots are also displayed.

Summary statistics of the posterior distribution samples are produced by default. However, these statistics might not provide enough information for you to carry out your Bayesian inference, and therefore you need to have access to the posterior distribution samples. To save these samples for postprocessing, you use the ODS OUTPUT statement as follows to save the samples into the SAS data set `Post`:

```
ods graphics on;
proc bphreg data=Rats;
  model Days*Status(0)=Group;
  bayes seed=1;
  ods output PosteriorSample=Post;
run;
ods graphics off;
```

The results of this analysis are shown in the following figures.

The “Model Information” table in [Figure 4.1](#) summarizes information about the model you fit and the size of the simulation.

The BPHREG Procedure		
Bayesian Analysis		
Model Information		
Data Set	WORK.RATS	
Dependent Variable	Days	Days from Exposure to Death
Censoring Variable	Status	
Censoring Value(s)	0	
Model	Cox	
Ties Handling	BRESLOW	
Burn-In Size	2000	
MC Sample Size	10000	
Thinning	1	

Figure 4.1. Model Information

PROC BPHREG first fits the Cox model by maximizing the partial likelihood. The only parameter in the model is the regression coefficient of `Group`. The maximum likelihood estimate (MLE) of the parameter and its 95% confidence interval are shown in [Figure 4.2](#).

Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	95% Confidence Limits	
Group	1	-0.5959	0.3484	-1.2788	0.0870

Figure 4.2. Parameter Estimates

Since no prior is specified for the regression coefficient, the default uniform prior is used. This information is displayed in the “Uniform Prior for Regression Coefficients” table in Figure 4.3.

Uniform Prior for Regression Coefficients	
Parameter	Prior
Group	Constant

Figure 4.3. Coefficient Prior

The “Fit Statistics” table in Figure 4.4 lists information about the fitted model. The AIC and BIC statistics are the same as those produced by PROC PHREG, while the DIC (deviance information criterion) and pD are computed from the posterior samples. See the “Fit Statistics” section on page 140 for details.

Fit Statistics	
AIC (smaller is better)	203.438
BIC (smaller is better)	205.022
DIC (smaller is better)	203.444
pD (Effective Number of Parameters)	1.003

Figure 4.4. Fit Statistics

Summaries of the posterior samples are displayed in the “Descriptive Statistics of the Posterior Samples” table and “Interval Statistics of the Posterior Samples” table as shown in Figure 4.5. Note that the mean and standard deviation of the posterior samples are comparable to the MLE and its standard error, respectively, due to the use of the uniform prior.

Descriptive Statistics of the Posterior Samples						
Parameter	N	Mean	Standard Deviation	25%	Quantiles 50%	75%
Group	10000	-0.5998	0.3511	-0.8326	-0.5957	-0.3670

Interval Statistics of the Posterior Samples					
Parameter	Alpha	Credible Interval		HPD Interval	
Group	0.050	-1.3042	0.0721	-1.2984	0.0756

Figure 4.5. Summary Statistics

PROC BPHREG provides diagnostics to assess the convergence of the generated Markov chain. [Figure 4.6](#) shows three of these diagnostics: the lag1, lag5, lag10, and lag50 autocorrelations; the Geweke diagnostic; and the effective sample size. There is no indication that the Markov chain has not reached convergence. Refer to the “[Statistical Diagnostic Tests](#)” section on page 21 for information on interpreting these diagnostics.

Autocorrelations of the Posterior Samples				
Parameter	Lag1	Lag5	Lag10	Lag50
Group	-0.0079	0.0091	-0.0161	0.0101

Geweke Diagnostics		
Parameter	z	Pr > z
Group	0.0149	0.9881

Effective Sample Size			
Parameter	ESS	Correlation Time	Efficiency
Group	10000.0	1.0000	1.0000

Figure 4.6. Convergence Diagnostics

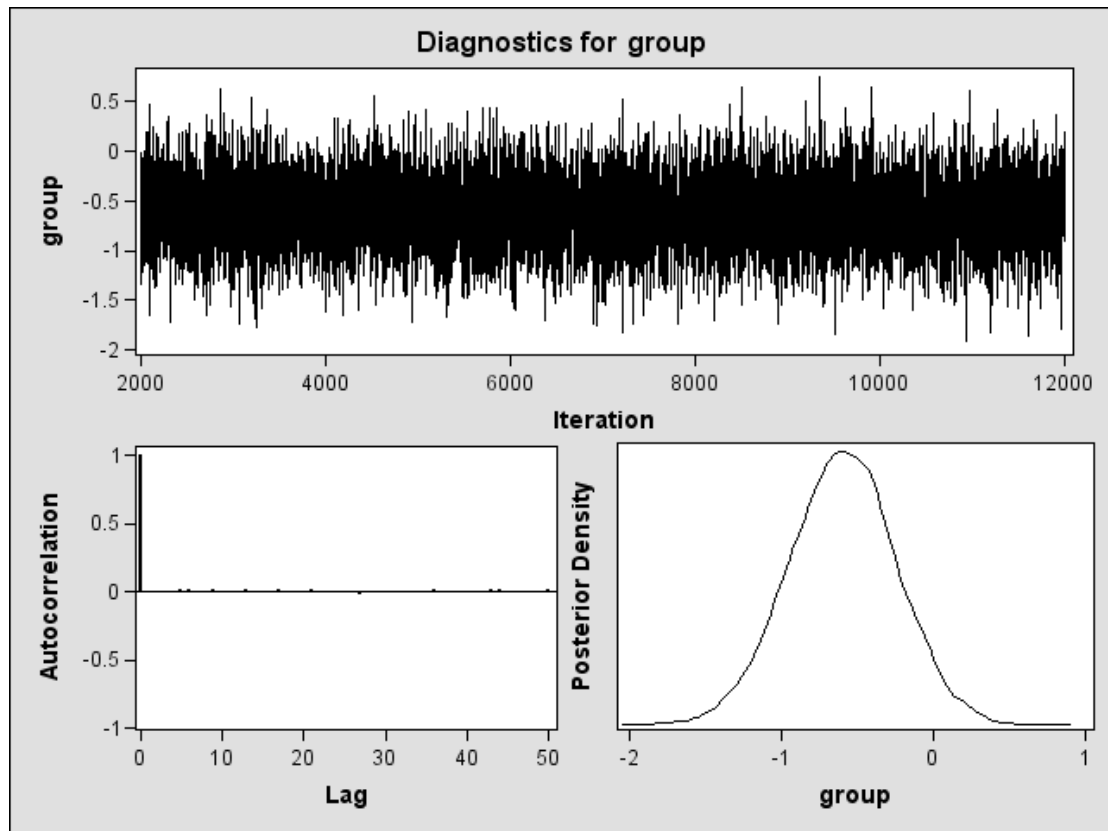


Figure 4.7. Diagnostic Plots

You can also assess the convergence of the generated Markov chain by examining the trace plot, the autocorrelation function plot, and the posterior density plot. [Figure 4.7](#) displays a panel of these three plots for the parameter `Group`. This graphical display is automatically produced when the experimental ODS graphics is enabled. Note that the trace of the samples centers on -0.6 with only small fluctuations, the autocorrelations are quite small, and the posterior density appears bell-shaped—all exemplifying the behavior of a converged Markov chain.

The proportional hazards model for comparing the two pretreatment groups is:

$$\lambda(t) = \begin{cases} \lambda_0(t) & \text{if Group}=0 \\ \lambda_0(t)e^\beta & \text{if Group}=1 \end{cases}$$

The probability that the hazard of `Group=0` is greater than that of `Group=1` is:

$$\Pr(\lambda_0(t) > \lambda_0(t)e^\beta) = \Pr(\beta < 0)$$

This probability can be enumerated from the posterior distribution samples by computing the fraction of samples with a coefficient less than 0. The following `DATA` step and `PROC MEANS` perform this calculation.

```

data New;
  set Post;
  Indicator=(Group < 0);
  label Indicator='Group < 0';
run;
proc means data=New(keep=Indicator) n mean;
run;

```

The MEANS Procedure	
Analysis Variable : Indicator (Group < 0)	
N	Mean
10000	0.9581000

Figure 4.8. Prob(Hazard(Group=0) > Hazard(Group=1))

The PROC MEANS results are displayed in [Figure 4.8](#). There is a 95.8% chance that the hazard rate of Group=0 is greater than that of Group=1. The result is consistent with the fact that the average survival time of Group=0 is less than that of Group=1.

Syntax

PROC BPHREG is a superset of PROC PHREG and the experimental PROC TPHREG, both available in the SAS/STAT 9.1 release. You can find the syntax for PROC PHREG and PROC TPHREG in the SAS/STAT 9.1 documentation.

To request a Bayesian analysis, you specify the new BAYES statement in addition to the PROC BPHREG statement and the MODEL statement. You include a CLASS statement if you have effects that involve categorical variables. The FREQ statement can be included if you have a frequency variable in the input data. The STRATA statement can be used to carry out a stratified analysis for the Cox model, but it is not allowed in the piecewise constant baseline hazard model. Programming statements can be used to create time-dependent covariates for the Cox model, but they are not allowed in the piecewise constant baseline hazard model. However, you can use the counting process style of input to accommodate time-dependent covariates that are not continuously changing with time for the piecewise constant baseline hazard model and the Cox model as well.

The new HAZARDRATIO statement enables you to request a hazard ratio analysis for the Bayesian and non-Bayesian analyses. All other statements (ASSESS, BASELINE, BY, CONTRAST, OUTPUT, TEST) are ignored when the BAYES statement is specified in PROC BPHREG.

BAYES Statement

BAYES < options > ;

The BAYES statement requests a Bayesian analysis of the regression model by using Gibbs sampling. The Bayesian posterior samples (also known as the chain) for the regression parameters are not tabulated. You can create an ODS output data set of the chain by specifying the following:

ODS OUTPUT PosteriorSample = SAS-data-set ;

Table 4.1 summarizes the options available in the BAYES statement.

Table 4.1. BAYES Statement Options

Option	Description
Monte Carlo Options	
INITIAL=	specifies initial values of the chain
NBI=	specifies the number of burn-in iterations
NMC=	specifies the number of iterations after burn-in
SEED=	specifies the random number generator seed
THINNING=	controls the thinning of the Markov chain
Model and Prior Options	
COEFFPRIOR=	specifies the prior of the regression coefficients
PIECEWISE=	specifies details of the piecewise exponential model
Summaries and Diagnostics of the Posterior Samples	
DIAGNOSTIC=	displays convergence diagnostics
PLOTS=	displays diagnostic plots
SUMMARY=	displays summary statistics

The following list describes these options and their suboptions.

COEFFPRIOR=UNIFORM | NORMAL<(option)>

COEFF=UNIFORM | NORMAL<(option)>

specifies the prior distribution for the regression coefficients. The default is COEFFPRIOR=UNIFORM, which assumes that the prior is proportional to a constant ($p(\beta_1, \dots, \beta_k) \propto 1$ for all $-\infty < \beta_i < \infty$).

Normal prior is specified by COEFFPRIOR=NORMAL, which can be followed by one of the following options enclosed in parentheses. However, if you do not specify a suboption, the normal prior $N(\mathbf{0}, 10^6\mathbf{I})$, where \mathbf{I} is the identity matrix, is used. See the “Normal Prior” section on page 137.

INPUT=SAS-data-set

specifies a SAS data set containing the mean and covariance information of the normal prior. The data set must contain the `_TYPE_` variable to identify the observation type, and it must contain a variable to represent each regression coefficient. If the data set also contains the `_NAME_` variable, values of this variable are used to identify the covariances for the `_TYPE_='COV'` observations; otherwise, the `_TYPE_='COV'` observations are assumed to be in

the same order as the explanatory variables in the MODEL statement. PROC BPHREG reads the mean vector from the observation with `_TYPE_='MEAN'` and the covariance matrix from observations with `_TYPE_='COV'`. For an independent normal prior, the variances can be specified with `_TYPE_='VAR'`; alternatively, the precisions (inverse of the variances) can be specified with `_TYPE_='PRECISION'`.

`RELVAR<=c>`

specifies a normal prior $N(\mathbf{0}, c\mathbf{J})$, where \mathbf{J} is a diagonal matrix with diagonal elements equal to the variances of the corresponding ML estimator. By default, $c=10^6$.

`VAR=c`

specifies the normal prior $N(\mathbf{0}, c\mathbf{I})$, where \mathbf{I} is the identity matrix.

DIAGNOSTIC=ALL | NONE | (*keyword-list*)

DIAG=ALL | NONE | (*keyword-list*)

controls the number of diagnostics produced. You can request all the diagnostics listed below by specifying `DIAGNOSTICS=ALL`. If you don't want any of these diagnostics, you specify `DIAGNOSTICS=NONE`. If you want some but not all of the diagnostics, or if you want to change certain settings of these diagnostics, you specify a subset of the following keywords. The default is `DIAGNOSTICS=(AUTOCORR ESS GEWEKE)`.

AUTOCORR computes the autocorrelations of lags 1, 5, 10, and 50 for each variable. See the “[Autocorrelations](#)” section on page 30 for details.

ESS computes the effective sample size of [Kass et al. \(1998\)](#), the correlation time, and the efficiency of the chain for each parameter. See the “[Effective Sample Size](#)” section on page 31 for details.

HEIDELBERGER <(heidel-options)>

computes the Heidelberg and Welch tests for each parameter. The tests consist of a stationary test and a halfwidth test. The former tests the null hypothesis that the sample values form a stationary process. If the stationarity test is passed, a halfwidth test is then carried out. Optionally, you can specify one or more of the following *heidel-options*:

`SALPHA=value`

specifies the α level ($0 < \alpha < 1$) for the stationarity test.

`HALPHA=value`

specifies the α level ($0 < \alpha < 1$) for the halfwidth test.

`EPS=value`

specifies a small positive number ϵ such that if the halfwidth is less than ϵ times the sample mean of the retaining samples, the halfwidth test is passed.

See the “[Heidelberg and Welch Diagnostics](#)” section on page 26 for details.

GELMAN <(gelman-options)>

computes the Gelman and Rubin convergence diagnostics. You can specify one or more of the following *gelman-options*:

NCHAIN | N=*number*

specifies the number of parallel chains used to compute the diagnostic and has to be 2 or larger. The default is NCHAIN=3. The NCHAIN= option is ignored when the INITIAL= option is specified in the BAYES statement, and in such a case, the number of parallel chains is determined by the number of valid observations in the INITIAL= data set.

ALPHA=*value*

specifies the significance level for the upper bound. The default is ALPHA=0.05, resulting in a 97.5% bound.

See the “[Gelman and Rubin Diagnostics](#)” section on page 22 for details.

GEWEKE<(geweke-options)>

computes the Geweke diagnostic, which is essentially a two-sample t -test between the first f_1 portion and the last f_2 portion of the chain. The default is $f_1=0.1$ and $f_2=0.5$, but you can choose other fractions by using the following *geweke-options*:

FRAC1=*value*

specifies the early f_1 fraction of the Markov chain.

FRAC2=*value*

specifies the latter f_2 fraction of the Markov chain.

See the “[Geweke Diagnostics](#)” section on page 24 for details.

RAFTERY<(raftery-options)>

computes the Raftery and Lewis diagnostics that evaluate the accuracy of the estimated quantile ($\hat{\theta}_Q$ for a given $Q \in (0, 1)$) of a chain. $\hat{\theta}_Q$ can achieve any degree of accuracy when the chain is allowed to run for a long time. A stopping criterion is when the estimated probability $\hat{P}_Q = \Pr(\theta \leq \hat{\theta}_Q)$ 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 ϵ for a stationary 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 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 tolerance level (a small positive number) for the test. The default is 0.001.

See the “[Raftery and Lewis Diagnostics](#)” section on page 27 for details.

INITIAL=SAS-data-set

specifies the SAS data set that contains the initial values of the Markov chains. The INITIAL= data set must contain a variable for each parameter in the model. You can specify multiple rows as the initial values of the parallel chains for the Gelman-Rubin statistics, but posterior summaries, diagnostics, and plots are computed only for the first chain.

NBI=number

specifies the number of burn-in iterations before the chains are saved. The default is 2000.

NMC=number

specifies the number of iterations after the burn-in. The default is 10000.

PIECEWISE=<keyword>(<N=number> <INTERVAL=(numeric-list)> <PRIOR=option>)>

specifies that the piecewise constant baseline hazard model be used in the Bayesian analysis. The keyword is either HAZARD or LOGHAZARD. Use PIECEWISE=HAZARD< (...) > if you want to model the baseline hazards in the original scale and the hazard parameters are named Lambda1, Lambda2, ... and so on. Alternatively, if you are modeling the baseline hazards in the log scale, use PIECEWISE=LOGHAZARD< (...) >, and the log-hazard parameters are named Alpha1, Alpha2, ..., and so on. Specifying PIECEWISE by itself is the same as specifying PIECEWISE=LOGHAZARD(N=8 PRIOR=UNIFORM).

Specify one of the following suboptions to partition the time axis into intervals of constant baseline hazards:

N=number**NINTERVAL=number**

specifies the number of intervals with constant baseline hazard rates. PROC BPHREG partitions the time axis into the given number of intervals with approximately equal number of events in each interval.

INTERVAL=(numeric-list) specifies the list of numbers that partition the time axis into disjoint intervals with constant baseline hazard in each interval. For example, INTERVAL=(100, 150, 200, 250, 300) specifies a model with a constant hazard in the intervals [0,100), [100,150), [150,200), [200,250), [250,300), and [300,∞). Each interval must contain at least one event; otherwise, the posterior distribution can be improper, and inferences cannot be derived from an improper posterior distribution.

If neither N= nor INTERVAL= is specified, the default is N=8.

To specify the prior for the baseline hazards $(\lambda_1, \dots, \lambda_J)$ in the original scale, choose one of the following PIECEWISE=HAZARD(PRIOR=) suboptions. The default is PRIOR=IMPROPER.

IMPROPER

specifies the noninformative and improper prior $p(\lambda_1, \dots, \lambda_J) \propto \prod_i \lambda_i^{-1}$ for all $\lambda_i > 0$.

UNIFORM specifies a uniform prior on the real line; that is, $p(\lambda_i) \propto 1$ for all $\lambda_i > 0$.

GAMMA<(gamma-option)> specifies an independent gamma prior $G(a, b)$ with density $f(t) = \frac{b(bt)^{a-1}e^{-bt}}{\Gamma(a)}$, which can be followed by one of the following *gamma-options* enclosed in parentheses. The hyperparameters a and b are the shape and inverse-scale parameters of the gamma distribution, respectively. See the “Independent Gamma Prior” section on page 136 for details. The default is $G(10^{-4}, 10^{-4})$ for each λ_j , setting the prior mean to 1 with variance 10^4 . This prior is proper and reasonably noninformative.

INPUT=SAS-data-set

specifies a data set containing the hyperparameters of the independent gamma prior. The data set must contain the `_TYPE_` variable to identify the observation type, and it must contain the variables named `Lambda1`, `Lambda2`, . . . , and so forth, to represent the hazard parameters. The observation with `_TYPE_='SHAPE'` identifies the shape parameters, and the observation with `_TYPE_='ISCALE'` identifies the inverse-scale parameters.

RELSHAPE=<c>

specifies independent $G(c\hat{\lambda}_j, c)$ distribution, where $\hat{\lambda}_j$'s are the MLEs of the hazard rates. This prior has mean $\hat{\lambda}_j$ and variance $\frac{\hat{\lambda}_j}{c}$. By default, $c=10^{-4}$.

SHAPE=a and SCALE=b

together specify the $\text{Gamma}(a, b)$ prior.

SHAPE=c

specifies the $G(c, c)$ prior.

ISCALE=c

specifies the $G(c, c)$ prior.

ARGAMMA<(argamma-option)> specifies an autoregressive gamma prior of order 1, which can be followed by one of the following *argamma-options*. See the “AR1 Prior” section on page 136 for details.

INPUT=SAS-data-set

specifies a data set containing the hyperparameters of the correlated gamma prior. The data set must contain the `_TYPE_` variable to identify the observation type, and it must contain the variables named `Lambda1`, `Lambda2`, . . . , and so forth, to represent the hazard parameters. The observation with `_TYPE_='SHAPE'` identifies the shape parameters, and the observation with `_TYPE_='ISCALE'` identifies the *relative* inverse-scale parameters; that is, if a_j and b_j are, respectively, the SHAPE and ISCALE values for λ_j , $1 \leq j \leq J$, then $\lambda_1 \sim G(a_1, b_1)$, and $\lambda_j \sim G(a_j, b_j/\lambda_{j-1})$ for $2 \leq j \leq J$.

SHAPE=a and SCALE=b

together specify that $\lambda_1 \sim G(a, b)$ and $\lambda_j \sim G(a, b/\lambda_{j-1})$ for $2 \leq j \leq J$.

SHAPE=c

specifies that $\lambda_1 \sim G(c, c)$ and $\lambda_j \sim G(c, c/\lambda_{j-1})$ for $2 \leq j \leq J$.

ISCALE=*c*

specifies that $\lambda_1 \sim G(c, c)$ and $\lambda_j \sim G(c, c/\lambda_{j-1})$ for $2 \leq j \leq J$.

To specify the prior for the baseline hazards $(\alpha_1, \dots, \alpha_J)$ in the log scale, you use the following suboptions:

PRIOR=UNIFORM | NORMAL<(option)>

The default is PRIOR=UNIFORM, which specifies the uniform prior on the real line; that is, $\alpha_i \propto 1$ for all $-\infty < \alpha_i < \infty$.

Normal prior is specified by PRIOR=NORMAL, which can be followed by one of the following options enclosed in parentheses. However, if you do not specify an option, the normal prior $N(\mathbf{0}, 10^6\mathbf{I})$, where \mathbf{I} is the identity matrix, is used.

INPUT=*SAS-data-set*

specifies a SAS data set containing the mean and covariance information of the normal prior. The data set must contain the `_TYPE_` variable to identify the observation type, and it must contain variables named `Alpha1`, `Alpha2`, `...`, and so forth, to represent the log-hazard parameters. If the data set also contains the `_NAME_` variable, the value of this variable will be used to identify the covariances for the `_TYPE_='COV'` observations; otherwise, the `_TYPE_='COV'` observations are assumed to be in the same order as the explanatory variables in the MODEL statement. PROC BPHREG reads the mean vector from the observation with `_TYPE_='MEAN'` and the covariance matrix from observations with `_TYPE_='COV'`. See the “[Normal Prior](#)” section on page 136 for details. For an independent normal prior, the variances can be specified with `_TYPE_='VAR'`; alternatively, the precisions (inverse of the variances) can be specified with `_TYPE_='PRECISION'`.

If you have a joint normal prior for the log-hazard parameters and the regression coefficients, specify the same data set containing the mean and covariance information of the multivariate normal distribution in both the `COEFFPRIOR=NORMAL(INPUT=)` and the `PIECEWISE=LOGHAZARD(PRIOR=NORMAL(INPUT=))` options. See the “[Joint Multivariate Normal Prior for Log-Hazards and Regression Coefficients](#)” section on page 137 for details.

RELVAR<=*c*>

specifies the normal prior $N(\mathbf{0}, c\mathbf{J})$, where \mathbf{J} is a diagonal matrix with diagonal elements equal to the variances of the corresponding ML estimator. By default, $c=10^6$.

VAR=*c*

specifies the normal prior $N(\mathbf{0}, c\mathbf{I})$, where \mathbf{I} is the identity matrix.

PLOTS<(global-plot-options)>=*plot-request*

PLOTS<(global-plot-options)>=(*plot-request* <... *plot-request*>)

controls the diagnostic plots produced through the experimental ODS graphics. Three types of plots can be requested: trace plots, autocorrelation function plots, and kernel density plots. By default, the plots are displayed in panels unless the global plot

option UNPACK is specified. If you specify more than one type of plot, the plots are displayed by parameters unless the global plot option GROUPBY=TYPE is specified. When you specify only one plot request, you can omit the parentheses around the plot request. For example:

```
PLOTS=NONE
PLOTS (UNPACK) =TRACE
PLOTS= (TRACE AUTOCORR)
```

You must enable ODS graphics before requesting plots, for example, like this:

```
ods graphics on;

proc bphreg;
  model y=x;
  bayes plots=trace;
run;
end;

ods graphics off;
```

If you have enabled ODS graphics but do not specify the PLOTS= option in the BAYES statement, then PROC BPHREG produces, for each parameter, a panel containing the trace plot, the autocorrelation function plot, and the density plot. This is equivalent to specifying PLOTS=(TRACE AUTOCORR DENSITY).

The global plot options include the following:

FRINGE creates a fringe plot on the X axis of the density plot.

GROUPBY = PARAMETER

GROUPBY = TYPE

specifies how the plots are to be grouped when there is more than one type of plot. GROUPBY=TYPE specifies that the plots are to be grouped by type, and GROUPBY=PARAMETER, which is the default, specifies that the plots are to be grouped by parameter.

UNPACKPANEL

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

The plot requests include the following:

ALL specifies all types of plots. PLOTS=ALL is equivalent to specifying PLOTS=(TRACE AUTOCORR DENSITY).

AUTOCORR displays the autocorrelation function plots for the parameters.

DENSITY displays the kernel density plots for the parameters.

NONE suppresses the display of any plots.
 TRACE displays the trace plots for the parameters. See the “[Visual Analysis via Trace Plots](#)” section on page 18 for details.

Consider a model with four parameters, X1–X4. Displays for various specification are depicted as follows:

1. PLOTS=(TRACE AUTOCORR) displays the trace and autocorrelation plots for each parameter side by side with two parameters per panel.

Display 1	Trace(X1)	Autocorr(X1)
	Trace(X2)	Autocorr(X2)
Display 2	Trace(X3)	Autocorr(X3)
	Trace(X4)	Autocorr(X4)

2. PLOTS(GROUPBY=TYPE)=(TRACE AUTOCORR) displays all the paneled trace plots, followed by panels of autocorrelation plots.

Display 1	Trace(X1)	
	Trace(X2)	
Display 2	Trace(X3)	
	Trace(X4)	
Display 3	Autocorr(X1)	Autocorr(X2)
	Autocorr(X3)	Autocorr(X4)

3. PLOTS(UNPACK)=(TRACE AUTOCORR) displays a separate trace plot and a separate correlation plot, parameter by parameter.

Display 1	Trace(X1)
Display 2	Autocorr(X1)
Display 3	Trace(X2)
Display 4	Autocorr(X2)
Display 5	Trace(X3)
Display 6	Autocorr(X3)
Display 7	Trace(X4)
Display 8	Autocorr(X4)

4. PLOTS(UNPACK GROUPBY=TYPE) = (TRACE AUTOCORR) displays all the separate trace plots followed by the separate autocorrelation plots.

Display 1	Trace(X1)
Display 2	Trace(X2)
Display 3	Trace(X3)
Display 4	Trace(X4)
Display 5	Autocorr(X1)
Display 6	Autocorr(X2)
Display 7	Autocorr(X3)
Display 8	Autocorr(X4)

SEED=number

specifies an integer seed ranging from 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 the SEED= option is not specified, or if you specify a nonpositive seed, a random seed is derived from the time of day.

SUMMARY=ALL | NONE | (keyword-list)**SUM=ALL | NONE | (keywords-list)**

controls the number of posterior summaries produced. SUMMARY=ALL produces all the summary statistics, which include the mean, standard deviation, quartiles, credible intervals, and HPD intervals for each parameter, and the sample correlation matrix. If you don't want any posterior summaries, specify SUMMARY=NONE. You can use the following keywords to request only the descriptive statistics or the interval statistics of a given level, or the correlation matrix. The default is SUMMARY=(DESCRIPTIVE INTERVAL).

DESCRIPTIVE

DESC

produces the means, standard deviations, and quartiles for the posterior samples.

INTERVAL<(ALPHA=numeric-list)>

produces a $100(1 - \alpha)\%$ credible interval and a $100(1 - \alpha)\%$ HPD interval for each parameter and for each α in the *numeric-list* specified in the ALPHA= option. The default is ALPHA=0.05.

CORR

produces the correlation matrix of the posterior samples.

See the “[Summary Statistics](#)” section on page 31 for details.

THINNING=number**THIN=number**

controls the thinning of the Markov chain. Only one in every k samples is used when THINNING= k , and if NBI= n_0 and NMC= n , the number of samples kept is

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

where $[a]$ represents the integer part of the number a . The default is THINNING=1.

HAZARDRATIO Statement

The HAZARDRATIO statement enables you to request hazard ratios for any variable in the model at customized settings.

For example, if the model contains the interaction of a CLASS variables A and a continuous variable X, the following specification displays a table of hazard ratios comparing the hazards of each pair of levels of A at X=3:

```
hazardratio A / at (X=3) diff=ALL;
```

The syntax of the HAZARDRATIO statement is as follows:

```
HAZARDRATIO <'label'> variable < / options > ;
```

The HAZARDRATIO statement identifies the variable whose hazard ratios are to be evaluated. If the variable is a continuous variable, the hazard ratio compares the hazards for a given change (by default, a increase of 1 unit) in the variable. For a CLASS variable, a hazard ratio compares the hazards of two levels of the variable. More than one HAZARDRATIO statement can be specified, and an optional label (specified as a quoted string) helps identify the output.

Options for the HAZARDRATIO statement are as follows.

ALPHA=number

specifies the alpha level of the credible and HPD intervals for the hazard ratios. The value must be between 0 and 1. The default is ALPHA= 0.05.

AT (variable=ALL | REF | list <... variable=ALL | REF| list>)

specifies the variables that interact with the variable of interest and the corresponding values of the interacting variables. If the interacting variable is continuous and a numeric list is specified after the equal sign, hazard ratios are computed for each value in the list. If the interacting variable is a CLASS variable, you can specify, after the equal sign, a list of quoted strings corresponding to various levels of the CLASS variable, or you can specify the keyword ALL or REF. Hazard ratios are computed at each value of the list if the list is specified, or at each level of the interacting variable if ALL is specified, or at the reference level of the interacting variable if REF is specified.

If this option is not specified, PROC BPHREG finds all the variables that interact with the variable of interest. If an interacting *variable* is a CLASS variable, *variable=*ALL is the default; if the interacting *variable* is continuous, *variable=m* is the default, where *m* is the average of all the sampled values of the continuous *variable*.

Suppose the model contains two interactions: an interaction A*B of CLASS variables A and B, and another interaction A*X of A with a continuous variable X. If 3.5 is the average of the sampled values of X, the following two HAZARDRATIO statements are equivalent:

```
hazardratio A;  
hazardratio A / at (B=ALL X=3.5);
```

DIFF=ALL | REF

specifies which differences to consider for the level comparisons of a CLASS variable. The default is DIFF=ALL. This option is ignored in the estimation of hazard ratios for a continuous variable. DIFF=ALL requests all differences, and DIFF=REF requests comparisons between the reference level and all other levels of the CLASS variable.

UNITS=value

specifies the units of change in the continuous explanatory variable for which the

customized hazard ratio is estimated. The default is UNITS=1. This option is ignored in the computation of the hazard ratios for a CLASS variable.

E

displays the \mathbf{x} vector such that the hazard ratio is given by $\exp(\mathbf{x}'\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the vector of parameter estimates.

Details

Piecewise Constant Baseline Hazard Model

Single Failure Time Variable

Let $\{(t_i, \mathbf{x}_i, \delta_i), i = 1, 2, \dots, n\}$ be the observed data. Let $a_0 = 0 < a_1 < \dots < a_{J-1} < a_J = \infty$ be a partition of the time axis.

Hazards in Original Scale

The hazard function for subject i is

$$h(t|\mathbf{x}_i; \boldsymbol{\theta}) = h_0(t) \exp(\boldsymbol{\beta}'\mathbf{x}_i)$$

where

$$h_0(t) = \lambda_j \quad a_{j-1} \leq t < a_j \quad (j = 1, \dots, J)$$

The baseline cumulative hazard function is

$$H_0(t) = \sum_{j=1}^J \lambda_j \Delta_j(t)$$

where

$$\Delta_j(t) = \begin{cases} 0 & t < a_{j-1} \\ t - a_{j-1} & a_{j-1} \leq t < a_j \\ a_j - a_{j-1} & t \geq a_j \end{cases}$$

The log likelihood is given by

$$\begin{aligned} l(\boldsymbol{\lambda}, \boldsymbol{\beta}) &= \sum_{i=1}^n \delta_i \left[\sum_{j=1}^J I(a_{j-1} \leq t_i < a_j) \log \lambda_j + \boldsymbol{\beta}'\mathbf{x}_i \right] - \sum_{i=1}^n \left[\sum_{j=1}^J \Delta_j(t_i) \lambda_j \right] \exp(\boldsymbol{\beta}'\mathbf{x}_i) \\ &= \sum_{j=1}^J d_j \log \lambda_j + \sum_{i=1}^n \delta_i \boldsymbol{\beta}'\mathbf{x}_i - \sum_{j=1}^J \lambda_j \left[\sum_{i=1}^n \Delta_j(t_i) \exp(\boldsymbol{\beta}'\mathbf{x}_i) \right] \end{aligned}$$

where $d_j = \sum_{i=1}^n \delta_i I(a_{j-1} \leq t_i < a_j)$.

Note that for $1 \leq j \leq J$, the full conditional for λ_j is log-concave only when $d_j > 0$, but the full conditionals for the β 's are always log-concave.

For a given β , $\frac{\partial l}{\partial \lambda} = 0$ gives

$$\tilde{\lambda}_j(\beta) = \frac{d_j}{\sum_{i=1}^n \Delta_j(t_i) \exp(\beta' \mathbf{x}_i)} \quad (j = 1, \dots, J)$$

Substituting these values into $l(\lambda, \beta)$ gives the profile log likelihood for β

$$l_p(\beta) = \sum_{i=1}^n \delta_i \beta' \mathbf{x}_i - \sum_{j=1}^J d_j \log \left[\sum_{l=1}^n \Delta_j(t_l) \exp(\beta' \mathbf{x}_l) \right] + c$$

where $c = \sum_j (d_j \log d_j - d_j)$. Since the constant c does not depend on β , it can be discarded from $l_p(\beta)$ in the optimization.

The MLE $\hat{\beta}$ of β is obtained by maximizing

$$l_p(\beta) = \sum_{i=1}^n \delta_i \beta' \mathbf{x}_i - \sum_{j=1}^J d_j \log \left[\sum_{l=1}^n \Delta_j(t_l) \exp(\beta' \mathbf{x}_l) \right]$$

with respect to β , and the MLE $\hat{\lambda}$ of λ is given by

$$\hat{\lambda} = \tilde{\lambda}(\hat{\beta})$$

Let

$$\begin{aligned} \mathbf{S}_j^{(r)}(\beta) &= \sum_{l=1}^n \Delta_j(t_l) e^{\beta' \mathbf{x}_l} \mathbf{x}_l^{\otimes r} \quad r = 0, 1, 2 \quad (j = 1, \dots, J) \\ \mathbf{E}_j(\beta) &= \frac{\mathbf{S}_j^{(1)}(\beta)}{S_j^{(0)}(\beta)} \end{aligned}$$

The partial derivatives of $l_p(\beta)$ are

$$\begin{aligned} \frac{\partial l_p(\beta)}{\partial \beta} &= \sum_{i=1}^n \delta_i \mathbf{x}_i - \sum_{j=1}^J d_j \mathbf{E}_j(\beta) \\ -\frac{\partial^2 l_p(\beta)}{\partial \beta^2} &= \sum_{j=1}^J d_j \left\{ \frac{\mathbf{S}_j^{(2)}(\beta)}{S_j^{(0)}(\beta)} - \left[\mathbf{E}_j(\beta) \right] \left[\mathbf{E}_j(\beta) \right]' \right\} \end{aligned}$$

The asymptotic covariance matrix for $(\hat{\lambda}, \hat{\beta})$ is obtained as the inverse of the information matrix given by

$$\begin{aligned} -\frac{\partial^2 l(\hat{\lambda}, \hat{\beta})}{\partial \lambda^2} &= \mathcal{D}\left(\frac{d_1}{\hat{\lambda}_1^2}, \dots, \frac{d_J}{\hat{\lambda}_J^2}\right) \\ -\frac{\partial^2 l(\hat{\lambda}, \hat{\beta})}{\partial \beta^2} &= \sum_{j=1}^J \hat{\lambda}_j \mathbf{S}_j^{(2)}(\hat{\beta}) \\ -\frac{\partial^2 l(\hat{\lambda}, \hat{\beta})}{\partial \lambda \partial \beta} &= (\mathbf{S}_1^{(1)}(\hat{\beta}), \dots, \mathbf{S}_J^{(1)}(\hat{\beta})) \end{aligned}$$

See Example 6.5.1 in Lawless (2003) for details.

Hazards in Log Scale

By letting

$$\alpha_j = \log(\lambda_j) \quad j = 1, \dots, J$$

you can build a prior correlation among the λ_j 's by using a correlated prior $\alpha \sim N(\alpha_0, \Sigma_\alpha)$, where $\alpha = (\alpha_1, \dots, \alpha_J)'$.

The log likelihood is given by

$$l(\alpha, \beta) = \sum_{j=1}^J d_j \alpha_j + \sum_{i=1}^n \delta_i \beta' \mathbf{x}_i - \sum_{j=1}^J e^{\alpha_j} S_j^{(0)}(\beta)$$

Then the MLE of λ_j is given by

$$e^{\hat{\alpha}_j} = \hat{\lambda}_j = \frac{d_j}{S_j^0(\hat{\beta})}$$

Note that the full conditionals for α 's and β 's are always log-concave.

The asymptotic covariance matrix for $(\hat{\alpha}, \hat{\beta})$ is obtained as the inverse of the information matrix formed by

$$\begin{aligned} -\frac{\partial^2 l(\hat{\alpha}, \hat{\beta})}{\partial \alpha^2} &= \mathcal{D}\left(e^{\hat{\alpha}_1} S_1^0(\hat{\beta}), \dots, e^{\hat{\alpha}_J} S_J^0(\hat{\beta})\right) \\ -\frac{\partial^2 l(\hat{\alpha}, \hat{\beta})}{\partial \beta^2} &= \sum_{j=1}^J e^{\hat{\alpha}_j} \mathbf{S}_j^{(2)}(\hat{\beta}) \\ -\frac{\partial^2 l(\hat{\alpha}, \hat{\beta})}{\partial \alpha \partial \beta} &= (e^{\hat{\alpha}_1} \mathbf{S}_1^{(1)}(\hat{\beta}), \dots, e^{\hat{\alpha}_J} \mathbf{S}_J^{(1)}(\hat{\beta})) \end{aligned}$$

Counting Process Style of Input

Let $\{(s_j, t_i], \mathbf{x}_i, \delta_i), i = 1, 2, \dots, n\}$ be the observed data. Let $a_0 = 0 < a_1 < \dots < a_k$ be a partition of the time axis, where $a_k > t_i$ for all $i = 1, 2, \dots, n$.

Replacing $\Delta_j(t_i)$ with

$$\Delta_j((s_i, t_i]) = \begin{cases} 0 & t_i < a_{j-1} \vee s_i > a_j \\ t_i - \max(s_i, a_{j-1}) & a_{j-1} \leq t_i < a_j \\ a_j - \max(s_i, a_{j-1}) & t_i \geq a_j \end{cases}$$

the formulation for the single failure time variable applies.

Priors for Model Parameters

For a Cox model, the model parameters are the regression coefficients. For a piecewise exponential model, the model parameters consist of the regression coefficients and the hazards or log-hazards. The priors for the hazards and the priors for the regression coefficients are assumed to be independent, while you can have a joint multivariate normal prior for the log-hazards and the regression coefficients.

Hazard Parameters

Let $\lambda_1, \dots, \lambda_J$ be the constant baseline hazards.

Improper Prior

The joint prior density is given by

$$p(\lambda_1, \dots, \lambda_J) = \prod_{j=1}^J \frac{1}{\lambda_j}, \quad \forall \lambda_j > 0$$

This prior is improper (nonintegrable), but the posterior distribution is proper as long as there is at least one event time in each of the constant hazard intervals.

Uniform Prior

The joint prior density is given by

$$p(\lambda_1, \dots, \lambda_J) \propto 1, \quad \forall \lambda_j > 0$$

This prior is improper (nonintegrable), but the posteriors are proper as long as there is at least one event time in each of the constant hazard intervals.

Gamma Prior

The gamma distribution $G(a, b)$ has a pdf

$$f_{a,b}(t) = \frac{b(bt)^{a-1}e^{-bt}}{\Gamma(a)}, \quad t > 0$$

where a is the shape parameter and b^{-1} is the scale parameter. The mean is $\frac{a}{b}$ and the variance is $\frac{a}{b^2}$.

Independent Gamma Prior

Suppose for $j = 1, \dots, J$, λ_j has an independent $G(a_j, b_j)$ prior. The joint prior density is given by

$$p(\lambda_1, \dots, \lambda_J) \propto \prod_{j=1}^J \left\{ \lambda_j^{a_j-1} e^{-b_j \lambda_j} \right\}, \quad \forall \lambda_j > 0$$

AR1 Prior

$\lambda_1, \dots, \lambda_J$ are correlated as follows:

$$\begin{aligned} \lambda_1 &\sim G(a_1, b_1) \\ \lambda_2 &\sim G\left(a_2, \frac{b_2}{\lambda_1}\right) \\ \dots &\dots \\ \lambda_J &\sim G\left(a_J, \frac{b_J}{\lambda_{J-1}}\right) \end{aligned}$$

The joint prior density is given by

$$p(\lambda_1, \dots, \lambda_J) \propto \lambda_1^{a_1-1} e^{-b_1 \lambda_1} \prod_{j=2}^J \left(\frac{b_j}{\lambda_{j-1}} \right)^{a_j} \lambda_j^{a_j-1} e^{-\frac{b_j}{\lambda_{j-1}} \lambda_j}$$

Log-Hazard Parameters

Write $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_J)' \equiv (\log \lambda_1, \dots, \log \lambda_J)'$.

Uniform Prior

The joint prior density is given by

$$p(\alpha_1 \dots \alpha_J) \propto 1, \quad \forall -\infty < \alpha_i < \infty$$

Note that the uniform prior for the log-hazards is the same as the improper prior for the hazards.

Normal Prior

Assume $\boldsymbol{\alpha}$ has a multivariate normal prior with mean vector $\boldsymbol{\alpha}_0$ and covariance matrix $\boldsymbol{\Psi}_0$. The joint prior density is given by

$$p(\boldsymbol{\alpha}) \propto e^{-\frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)'\boldsymbol{\Psi}_0^{-1}(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)}$$

Regression Coefficients

Let $\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)'$ be the vector of regression coefficients.

Uniform Prior

The joint prior density is given by

$$p(\beta_1, \dots, \beta_k) \propto 1, \forall -\infty < \beta_i < \infty$$

This prior is improper, but the posterior distributions for β are proper.

Normal Prior

Assume β has a multivariate normal prior with mean vector β_0 and covariance matrix Σ_0 . The joint prior density is given by

$$p(\beta) \propto e^{-\frac{1}{2}(\beta-\beta_0)'\Sigma_0^{-1}(\beta-\beta_0)}$$

Joint Multivariate Normal Prior for Log-Hazards and Regression Coefficients

Assume $(\alpha', \beta)'$ has a multivariate normal prior with mean vector $(\alpha_0', \beta_0)'$ and covariance matrix Φ_0 . The joint prior density is given by

$$p(\alpha, \beta) \propto e^{-\frac{1}{2}[(\alpha-\alpha_0)', (\beta-\beta_0)']\Phi_0^{-1}[(\alpha-\alpha_0)', (\beta-\beta_0)]'}$$

Posterior Distribution

Denote the observed data by D .

Cox Model

$$\pi(\beta|D) \propto L_P(D|\beta)p(\beta)$$

where $L_P(D|\beta)$ is the partial likelihood function with regression coefficients β as parameters.

Piecewise Exponential Model**Hazard Parameters**

$$\pi(\lambda, \beta|D) \propto L_H(D|\lambda, \beta)p(\lambda)p(\beta)$$

where $L_H(D|\lambda, \beta)$ is the likelihood function with hazards λ and regression coefficients β as parameters.

Log-Hazard Parameters

$$\pi(\alpha, \beta|D) \propto \begin{cases} L_{LH}(D|\alpha, \beta)p(\alpha, \beta) & \text{if } (\alpha', \beta)' \sim \text{MVN} \\ L_{LH}(D|\alpha, \beta)p(\alpha)p(\beta) & \text{otherwise} \end{cases}$$

where $L_{LH}(D|\alpha, \beta)$ is the likelihood function with log-hazards α and regression coefficients β as parameters.

Sampling from the Posterior Distribution

PROC BPHREG uses a Gibbs sampler to generate the posterior samples. See the “Gibbs Sampler” section on page 15 for a general discussion.

Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)'$ be the parameter vector. For the Cox model, the θ_i 's are the regression coefficients β_i 's, and for the piecewise constant baseline hazard model, the θ_i 's consist of the baseline hazards λ_i 's (or log baseline hazards α_i 's) and the regression coefficients β_j 's. Let $L(D|\boldsymbol{\theta})$ be the likelihood function, where D is the observed data. Note that for the Cox model, the likelihood contains the infinite-dimensional baseline hazard function and the Gamma process is perhaps the most commonly used prior process (Ibrahim, Chen, and Sinha 2001); however, Sinha, Ibrahim, and Chen (2003) justify using the partial likelihood as the likelihood function for the Bayesian analysis. Let $p(\boldsymbol{\theta})$ be the prior distribution. The full conditional distribution of θ_i is proportional to the joint distribution; that is,

$$\pi(\theta_i|\theta_j, i \neq j, D) \propto L(D|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

For instance, the one-dimensional conditional distribution of θ_1 given $\theta_j = \theta_j^*$, $2 \leq j \leq k$, is computed as

$$\pi(\theta_1|\theta_j = \theta_j^*, 2 \leq j \leq k, D) = L(D|(\boldsymbol{\theta} = (\theta_1, \theta_2^*, \dots, \theta_k^*)')p(\boldsymbol{\theta} = (\theta_1, \theta_2^*, \dots, \theta_k^*)')$$

Suppose you have a set of arbitrary starting values $\{\theta_1^{(0)}, \dots, \theta_k^{(0)}\}$. Using the ARMS (adaptive rejection Metropolis sampling) algorithm of Gilks, Best, and Tan (1995), you do the following:

```
draw  $\theta_1^{(1)}$  from  $\pi(\theta_1|\theta_2^{(0)}, \dots, \theta_k^{(0)}, D)$ 
draw  $\theta_2^{(1)}$  from  $\pi(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \dots, \theta_k^{(0)}, D)$ 
...
draw  $\theta_k^{(1)}$  from  $\pi(\theta_k|\theta_1^{(1)}, \dots, \theta_{k-1}^{(1)}, D)$ 
```

This completes one iteration of the Gibbs sampler. After one iteration, you have $\{\theta_1^{(1)}, \dots, \theta_k^{(1)}\}$. After n iterations, you have $\{\theta_1^{(n)}, \dots, \theta_k^{(n)}\}$. PROC BPHREG implements the ARMS algorithm based on code provided by Gilks (2003) to draw a sample from a full conditional distribution.

You can output these posterior samples into a SAS data set through ODS. The following SAS statement outputs the posterior samples into the SAS data set `Post`:

```
ods output PosteriorSample=Post;
```

The data set also includes the variable `_LOGPOST_`, representing the log of the posterior log likelihood.

Starting Values of the Markov Chains

When the BAYES statement is specified, PROC BPHREG generates one Markov chain containing the approximate posterior samples of the model parameters. Additional chains are produced when the Gelman-Rubin diagnostics are requested. Starting values (or initial values) can be specified in the INITIAL= data set in the BAYES statement. If INITIAL= option is not specified, PROC BPHREG picks its own initial values for the chains. If the prior distribution of the parameter ω is proper, the starting values for ω are based on the estimated mean ($\hat{\omega}$) and standard deviation ($\hat{s}(\hat{\omega})$) of the posterior distribution given the MLE. If the prior distribution of ω is improper, the starting values for ω are based on the MLE and its standard error estimate; that is, $\hat{\omega}$ is the MLE of ω and $\hat{s}(\hat{\omega})$ is the standard error of maximum likelihood estimator.

Denote $[x]$ as the integral value of x .

Constant Baseline Hazards λ_i 's

For the first chain that the summary statistics and diagnostics are based on, the initial values are

$$\lambda_i^{(0)} = \hat{\lambda}_i$$

For subsequent chains, the starting values are picked in two different ways according to the total number of chains specified. If the total number of chains specified is less than or equal to 10, initial values of the r th chain ($2 \leq r \leq 10$) are given by

$$\lambda_i^{(0)} = \hat{\lambda}_i e^{\pm \left(\left[\frac{r}{2} \right] + 2 \right) \hat{s}(\hat{\lambda}_i)}$$

with the plus sign for odd r and minus sign for even r . If the total number of chains is greater than 10, initial values are picked at random over a wide range of values. Let u_i be a uniform random number between 0 and 1; the initial value for λ_i is given by

$$\lambda_i^{(0)} = \hat{\lambda}_i e^{16(u_i - 0.5)\hat{s}(\hat{\lambda}_i)}$$

Regression Coefficients and Log-Hazard Parameters θ_i 's

The θ_i 's are the regression coefficients β_i 's, and in the piecewise exponential model, include the log-hazard parameters α_i 's. For the first chain that the summary statistics and regression diagnostics are based on, the initial values are

$$\theta_i^{(0)} = \hat{\theta}_i$$

If the number of chains requested is less than or equal to 10, initial values for the r th chain ($2 \leq r \leq 10$) are given by

$$\theta_i^{(0)} = \hat{\theta}_i \pm \left(2 + \left[\frac{r}{2} \right] \right) \hat{s}(\hat{\theta}_i)$$

with the plus sign for odd r and minus sign for even r . When there are more than 10 chains, the initial value for the θ_i is picked at random over the range $(\hat{\theta}_i - 8\hat{s}(\hat{\theta}_i), \hat{\theta}_i + 8\hat{s}(\hat{\theta}_i))$; that is,

$$\theta_i^{(0)} = \hat{\theta}_i + 16(u_i - 0.5)\hat{s}(\hat{\theta}_i)$$

where u_i is a uniform random number between 0 and 1.

Fit Statistics

Denote the observed data by D . Let θ be the vector of parameters of length k . Let $L(D|\theta)$ be the likelihood. Let $\hat{\theta}$ be the MLE of θ .

The two most commonly used penalized model selection criteria, Akaike's information criterion (AIC) and the Bayesian information criterion (BIC), are computed as follows:

$$\begin{aligned} \text{AIC} &= -2\log(L(D|\hat{\theta})) + 2k \\ \text{BIC} &= -2\log(L(D|\hat{\theta})) + k\log(N_e) \end{aligned}$$

where N_e is the number of uncensored observations.

The deviance information criterion (DIC) proposed in Spiegelhalter et al. (2002) is a Bayesian model assessment tool. Let $\text{Dev}(\theta) = -2\log L(D|\theta)$. Let $\overline{\text{Dev}(\theta)}$ and $\bar{\theta}$ be the corresponding posterior means of $\text{Dev}(\theta)$ and θ , respectively. The deviance information criterion is computed as

$$\text{DIC} = 2\overline{\text{Dev}(\theta)} - \text{Dev}(\bar{\theta})$$

Also computed is

$$pD = \overline{\text{Dev}(\theta)} - \text{Dev}(\bar{\theta})$$

where pD is interpreted as the effective number of parameters.

Note that $\text{Dev}(\theta)$ defined here does not have the standardizing term as in the “Deviance Information Criterion (DIC)” section on page 33. Nevertheless, DIC calculated here is still useful for variable selection.

Displayed Output

If you use the NOPRINT option in the PROC BPHREG statement, the procedure does not display any output. Otherwise, the displayed output for the Bayesian analysis includes the following.

Model Information

The “Model Information” table displays the two-level name of the input data set, the name and label of the failure time variable, the name and label of the censoring variable and the values indicating censored times, the model (either the Cox model or the piecewise constant baseline hazard model), the name and label of the OFFSET variable, the name and label of the FREQ variable, the name and label of the WEIGHT variable, the method of handling ties in the failure time for the Cox model, the number of burn-in iterations, the number of iterations after the burn-in, and the number of thinning iterations. For ODS purposes, the name of the “Model Information” table is “ModelInfo.”

Summary of the Number of Event and Censored Values

The “Summary of the Number of Event and Censored Values” table displays, for each stratum, the breakdown of the number of events and censored values. For ODS purposes, the name of the “Summary of the Number of Event and Censored Values” table is “CensoredSummary.”

Class Level Information

The “Class Level Information” table lists the levels of every variable in the CLASS statement that is used in the model and the corresponding design variable values. For ODS purposes, the name of the “Class Level Information” table is “ClassLevelInfo.”

Regression Parameter Information

The “Regression Parameter Information” table displays the names of the parameters and the corresponding level information of effects containing the CLASS variables. For ODS purposes, the name of the “Regression Parameter Information” table is “ParmInfo.”

Constant Baseline Hazard Time Intervals

The “Constant Baseline Hazard Time Intervals” table displays the intervals of constant baseline hazard and the corresponding numbers of failure times and event times. This table is produced only if you specify the PIECEWISE option in the BAYES statement. For ODS purposes, the name of the “Constant Baseline Hazard Time Intervals” table is “Interval”.

Maximum Likelihood Estimates

The “Maximum Likelihood Estimates” table displays, for each parameter, the maximum likelihood estimate, the estimated standard error, and the 95% confidence limits. For ODS purposes, the name of the “Maximum Likelihood Estimates” table is “ParameterEstimates.”

Hazard Prior

The “Hazard Prior” table is displayed if you specify the PIECEWISE=HAZARD option in the BAYES statement. It describes the prior distribution of the hazard parameters. For ODS purposes, the name of the “Hazard Prior” table is “HazardPrior.”

Log-Hazard Prior

The “Log-hazard Prior” table is displayed if you specify the `PIECEWISE=LOGHAZARD` option in the `BAYES` statement. It describes the prior distribution of the log-hazard parameters. For ODS purposes, the name of the “Log-hazard Prior” table is “HazardPrior.”

Coefficient Prior

The “Coefficient Prior” table displays the prior distribution of the regression coefficients. For ODS purposes, the name of the “Coefficient Prior” table is “CoeffPrior.”

Initial Values

The “Initial Values” table displays the initial values of the parameters for the Gibbs sampling. For ODS purposes, the name of the “Initial Values” table is “InitialValues.”

Fit Statistics

The “Fit Statistics” table displays the AIC, BIC, DIC, and pD statistics for each parameter. For ODS purposes, the name of the “Fit Statistics” table is “FitStatistics.”

Descriptive Statistics of the Posterior Samples

The “Descriptive Statistics of the Posterior Sample” table displays the size of the posterior samples, the mean, the standard error, and the quartiles for each model parameter. For ODS purposes, the name of the “Descriptive Statistics of the Posterior Sample” table is “ChainStatistics.”

Interval Estimates for Posterior Samples

The “Interval Estimates for Posterior Sample” table displays the credible interval and the HPD interval for each model parameter. For ODS purposes, the name of the “Interval Estimates for Posterior Sample” table is “IntervalStatistics.”

Correlation Matrix of the Posterior Samples

The “Correlation Matrix of the Posterior Samples” table is produced if you include the `CORR` suboption in the `SUMMARY=` option in the `BAYES` statement. This table displays the sample correlation of the posterior samples. For ODS purposes, the name of the “Correlation Matrix of the Posterior Samples” table is “Corr.”

Autocorrelations of the Posterior Samples

The “Autocorrelations of the Posterior Samples” table displays the lag1, lag5, lag10, and lag50 autocorrelations for each parameter. For ODS purposes, the name of the “Autocorrelations of the Posterior Samples” table is “AutoCorr.”

Gelman and Rubin Diagnostics

The “Gelman and Rubin Diagnostics” table is produced if you include the `GELMAN` suboption in the `DIAGNOSTIC=` option in the `BAYES` statement. This table displays the estimate of the potential scale reduction factor and its 97.5% upper confidence limit for each parameter. For ODS purposes, the name of the “Gelman and Rubin Diagnostics” table is “Gelman.”

Geweke Diagnostics

The “Geweke Diagnostics” table displays the Geweke statistic and its p -value for each parameter. For ODS purposes, the name of the “Geweke Diagnostics” table is “Geweke.”

Raftery and Lewis Diagnostics

The “Raftery Diagnostics” table is produced if you include the RAFTERY suboption in the DIAGNOSTIC= option in the BAYES statement. This table displays the Raftery and Lewis diagnostics for each variable. For ODS purposes, the name of the “Raftery and Lewis Diagnostics” table is “Raftery.”

Heidelberger and Welch Diagnostics

The “Heidelberger and Welch Diagnostics” table is displayed if you include the HEIDELBERGER suboption in the DIAGNOSTIC= option in the BAYES statement. This table describes the results of a stationary test and a halfwidth test for each parameter. For ODS purposes, the name of the “Heidelberger and Welch Diagnostics” table is “Heidelberger.”

Effective Sample Size

The “Effective Sample Size” table displays, for each parameter, the effective sample size, the correlation time, and the efficiency. For ODS purposes, the name of the “Effective Sample Size” table is “ESS.”

ODS Table Names

PROC BPHREG assigns a name to each table it creates. You can use these names to reference the table when using ODS statements to select tables and create output data sets. For a Bayesian analysis, the ODS table names are listed in the Table 4.2.

Table 4.2. ODS Tables Produced by PROC BPHREG

ODS Table Name	Description	Statement	Option
AutoCorr	Autocorrelations of the posterior samples	BAYES	default
CensoredSummary	Numbers of the event and censored observations	PROC	default
ChainStatistics	Descriptive statistics of the posterior samples	BAYES	default
ClassLevelInfo	Levels of class variables and design variables	CLASS	default
CoeffPrior	Prior distribution of the regression coefficients	BAYES	default
Corr	Correlation matrix of the posterior samples	BAYES	SUMMARY=CORR
ESS	Effective sample size	BAYES	default
HazardPrior	Prior distribution of the baseline hazards	BAYES	PIECEWISE

Table 4.2. (continued)

ODS Table Name	Description	Statement	Option
Heidelberger	Heidelberger and Welch convergence diagnostics	BAYES	DIAGNOSTIC=HEIDELBERGER
Gelman	Gelman and Rubin convergence diagnostics	BAYES	DIAGNOSTIC=GELMAN
Geweke	Geweke convergence diagnostics	BAYES	default
InitialValues	Initial values of the Markov chains	BAYES	default
IntervalStatistics	HPD and credible intervals for the model parameters	BAYES	default
ModelInfo	Model information	PROC	default
NObs	Number of observations	PROC	default
ParameterEstimates	Maximum likelihood estimates of model parameters	PROC	default
ParmInfo	Names of regression coefficients	CLASS+BAYES	
Partition	Partition of constant baseline hazard intervals	BAYES	PIECEWISE
PosteriorSample	Posterior samples	BAYES	(for ODS output data set only)
Raftery	Raftery and Lewis convergence diagnostics	BAYES	DIAGNOSTIC=RAFTERY

ODS Graph Names

Each statistical graphic created by PROC BPHREG has a name associated with it, and you can reference the graph by using ODS statements. These names are listed in Table 4.3.

Table 4.3. ODS Graphics Produced by PROC BPHREG

ODS Graph Name	Description	Statement	Option
ADPanel	Autocorrelation function and density panel	BAYES	PLOTS=(AUTOCORR DENSITY)
AutocorrPanel	Autocorrelation function panel	BAYES	PLOTS= AUTOCORR
AutocorrPlot	Autocorrelation function plot	BAYES	PLOTS(UNPACK)=AUTOCORR
DensityPanel	Density panel	BAYES	PLOTS=DENSITY
DensityPlot	Density plot	BAYES	PLOTS(UNPACK)=DENSITY
TAPanel	Trace and autocorrelation function panel	BAYES	PLOTS=(TRACE AUTOCORR)
TADPanel	Trace, density, and autocorrelation function panel	BAYES	default
TDPanel	Trace and density panel	BAYES	PLOTS=(TRACE DENSITY)
TracePanel	Trace panel	BAYES	PLOTS=TRACE
TracePlot	Trace plot	BAYES	PLOTS(UNPACK)=TRACE

Examples

Example 4.1. Informative Prior and Hazard Ratio Analysis

This example illustrates the use of an informative prior. Hazard ratios, which are transformations of the regression parameters, are useful for interpreting survival models. This example also demonstrates the use of the HAZARDRATIO statement to obtain customized hazard ratios.

Consider the data presented in Appendix I of [Kalbfleisch and Prentice \(1980\)](#). The response variable, `Time`, is the survival time in days of a lung cancer patient. Negative values of `Time` are censored values. The explanatory variables are `Cell` (type of cancer cell), `Therapy` (type of therapy: standard or test), `Prior` (prior therapy: 0=no, 10=yes), `Age` (age in years), `Duration` (time in months from diagnosis to entry into the trial), and `Kps` (Karnofsky performance scale). The following DATA step saves the data in the data set `VALung`. An indicator variable `Status` is also created, with the value 1 indicating an uncensored time and the value 0 indicating a censored time.

```
proc format;
  value yesno 0='no' 10='yes';
run;

data VALung;
  drop check m;
  retain Therapy Cell;
  infile cards column=column;
  length Check $ 1;
  label Time='time to death in days'
        Kps='Karnofsky performance scale'
        Duration='months from diagnosis to randomization'
        Age='age in years'
        Prior='prior therapy'
        Cell='cell type'
        Therapy='type of treatment';
  format Prior yesno.;
  M=Column;
  input Check $@@;
  if M>Column then M=1;
  if Check='s'|Check='t' then do;
    input @M Therapy $ Cell $;
    delete;
  end;
  else do;
    input @M Time Kps Duration Age Prior @@;
    Status=(Time>0);
    Time=abs(Time);
  end;
  datalines;
standard squamous
  72 60 7 69 0 411 70 5 64 10 228 60 3 38 0 126 60 9 63 10
118 70 11 65 10 10 20 5 49 0 82 40 10 69 10 110 80 29 68 0
314 50 18 43 0 -100 70 6 70 0 42 60 4 81 0 8 40 58 63 10
144 30 4 63 0 -25 80 9 52 10 11 70 11 48 10
standard small
```

```

30 60 3 61 0 384 60 9 42 0 4 40 2 35 0 54 80 4 63 10
13 60 4 56 0 -123 40 3 55 0 -97 60 5 67 0 153 60 14 63 10
59 30 2 65 0 117 80 3 46 0 16 30 4 53 10 151 50 12 69 0
22 60 4 68 0 56 80 12 43 10 21 40 2 55 10 18 20 15 42 0
139 80 2 64 0 20 30 5 65 0 31 75 3 65 0 52 70 2 55 0
287 60 25 66 10 18 30 4 60 0 51 60 1 67 0 122 80 28 53 0
27 60 8 62 0 54 70 1 67 0 7 50 7 72 0 63 50 11 48 0
392 40 4 68 0 10 40 23 67 10
standard adeno
8 20 19 61 10 92 70 10 60 0 35 40 6 62 0 117 80 2 38 0
132 80 5 50 0 12 50 4 63 10 162 80 5 64 0 3 30 3 43 0
95 80 4 34 0
standard large
177 50 16 66 10 162 80 5 62 0 216 50 15 52 0 553 70 2 47 0
278 60 12 63 0 12 40 12 68 10 260 80 5 45 0 200 80 12 41 10
156 70 2 66 0 -182 90 2 62 0 143 90 8 60 0 105 80 11 66 0
103 80 5 38 0 250 70 8 53 10 100 60 13 37 10
test squamous
999 90 12 54 10 112 80 6 60 0 -87 80 3 48 0 -231 50 8 52 10
242 50 1 70 0 991 70 7 50 10 111 70 3 62 0 1 20 21 65 10
587 60 3 58 0 389 90 2 62 0 33 30 6 64 0 25 20 36 63 0
357 70 13 58 0 467 90 2 64 0 201 80 28 52 10 1 50 7 35 0
30 70 11 63 0 44 60 13 70 10 283 90 2 51 0 15 50 13 40 10
test small
25 30 2 69 0 -103 70 22 36 10 21 20 4 71 0 13 30 2 62 0
87 60 2 60 0 2 40 36 44 10 20 30 9 54 10 7 20 11 66 0
24 60 8 49 0 99 70 3 72 0 8 80 2 68 0 99 85 4 62 0
61 70 2 71 0 25 70 2 70 0 95 70 1 61 0 80 50 17 71 0
51 30 87 59 10 29 40 8 67 0
test adeno
24 40 2 60 0 18 40 5 69 10 -83 99 3 57 0 31 80 3 39 0
51 60 5 62 0 90 60 22 50 10 52 60 3 43 0 73 60 3 70 0
8 50 5 66 0 36 70 8 61 0 48 10 4 81 0 7 40 4 58 0
140 70 3 63 0 186 90 3 60 0 84 80 4 62 10 19 50 10 42 0
45 40 3 69 0 80 40 4 63 0
test large
52 60 4 45 0 164 70 15 68 10 19 30 4 39 10 53 60 12 66 0
15 30 5 63 0 43 60 11 49 10 340 80 10 64 10 133 75 1 65 0
111 60 5 64 0 231 70 18 67 10 378 80 4 65 0 49 30 3 37 0
;

```

In this example, the Cox model is used for the Bayesian analysis. The parameters are the coefficients of the continuous explanatory variables (*Kps*, *Duration*, and *Age*) and the coefficients of the design variables for the categorical explanatory variables (*Prior*, *Cell*, and *Therapy*). You use the *CLASS* statement in *PROC BPHREG* to specify the categorical variables and their reference levels. Using the default reference parameterization, the design variables for the categorical variables are *Prioryes* (for *Prior* with *Prior*=*'no'* as reference), *Celladeno*, *Cellsmall*, *Cellsquamous* (for *Cell* with *Cell*=*'large'* as reference), and *Therapytest* (for *Therapy*=*'standard'* as reference).

Consider the explanatory variable *Kps*. The Karnofsky performance scale index enables patients to be classified according to their functional impairment. The scale can range from 0 to 100—0 for dead, and 100 for a normal, healthy person with no evidence of disease. Recall that a flat prior was used for the regression coefficient

in the example in the “Getting Started” section on page 116. A flat prior on the Kps coefficient implies that the coefficient is as likely to be 0.1 as it is to be -100000 . A coefficient of -5 means that a decrease of 20 points in the scale increases the hazard by $e^{-20 \times -5} (=2.68 \times 10^{43})$ -fold, which is a rather unreasonable and unrealistic expectation for the effect of the Karnofsky index, much less than the value of -100000 . Suppose you have a more realistic expectation: the effect is somewhat small and is more likely to be negative than positive, and a decrease of 20 points in the Karnofsky index will change the hazard from 0.9-fold (some minor positive effect) to 4-fold (a large negative effect). You can convert this opinion to a more informative prior on the Kps coefficient β_1 . Mathematically,

$$0.9 < e^{-20\beta_1} < 4$$

which is equivalent to

$$-0.0693 < \beta_1 < 0.0053$$

This becomes the plausible range that you believe the Kps coefficient can take. Now you can find a normal distribution that best approximates this belief by placing the majority of the prior distribution mass within this range. Assuming this interval is $\mu \pm 2\sigma$, where μ and σ are the mean and standard deviation of the normal prior, respectively, the hyperparameters μ and σ are computed as follows:

$$\begin{aligned}\mu &= \frac{-0.0693 + 0.0053}{2} = -0.032 \\ \sigma &= \frac{0.0053 - (-0.0693)}{4} = 0.0186\end{aligned}$$

Note that a normal prior distribution with mean -0.0320 and standard deviation 0.0186 indicates that you believe, before looking at the data, that a decrease of 20 points in the Karnofsky index will probably change the hazard rate by 0.9-fold to 4-fold. This does not rule out the possibility that the Kps coefficient can take a more extreme value such as -5 , but the probability of having such extreme values is very small.

Assume the prior distributions are independent for all the parameters. For the coefficient of Kps, you use a normal prior distribution with mean -0.0320 and variance $0.0186^2 (=0.00035)$. For other parameters, you resort to using a normal prior distribution with mean 0 and variance 10^6 , which is fairly noninformative. Means and variances of these independent normal distributions are saved in the data set `Prior` as follows:

```
data Prior;
  input _TYPE_ $ Kps Duration Age Priories Celladeno Cellsmall
          Cellsquamous Therapytest;
  datalines;
  Mean -0.0320 0 0 0 0 0 0 0
  Var 0.00035 1e6 1e6 1e6 1e6 1e6 1e6 1e6
  ;
run;
```

In the following BAYES statement, COEFFPRIOR=NORMAL(INPUT=Prior) specifies the normal prior distribution for the regression coefficients with details contained in the data set Prior. Summary statistics of the posterior distribution are produced by default. Autocorrelations and effective sample size are requested as convergence diagnostics as well as the trace plots for visual analysis. For comparisons of hazards, three HAZARDRATIO statements are specified—one for the variable Therapy, one for the variable Age, and one for the variable Cell.

```
ods graphics on;
proc bphreg data=VALung;
  class Prior(ref='no') Cell(ref='large') Therapy(ref='standard');
  model Time*Status(0) = Kps Duration Age Prior Cell Therapy;
  bayes seed=1 coeffprior=normal diagnostic=(autocorr ess) plots=trace;
  hazardratio 'Hazard Ratio Statement 1' Therapy;
  hazardratio 'Hazard Ratio Statement 2' Age / unit=10;
  hazardratio 'Hazard Ratio Statement 3' Cell;
run;
ods graphics off;
```

This analysis generates a posterior chain of 10,000 iterations after 2,000 iterations of burn-in, as depicted in [Output 4.1.1](#).

Output 4.1.1. Model Information

Model Information		
Data Set	WORK.VALUNG	
Dependent Variable	Time	time to death in days
Censoring Variable	Status	
Censoring Value(s)	0	
Model	Cox	
Ties Handling	BRESLOW	
Burn-In Size	2000	
MC Sample Size	10000	
Thinning	1	

[Output 4.1.2](#) displays the names of the parameters and their corresponding effects and categories.

Output 4.1.2. Parameter Names

Regression Parameter Information				
Parameter	Effect	Prior	Cell	Therapy
Kps	Kps			
Duration	Duration			
Age	Age			
Priories	Prior	yes		
Celladeno	Cell		adeno	
Cellsmall	Cell		small	
Cellsquamous	Cell		squamous	
Therapytest	Therapy			test

PROC BPHREG computes the maximum likelihood estimates of regression parameters (Output 4.1.3). These estimates are used as the starting values for the simulation of posterior samples.

Output 4.1.3. Parameter Estimates

Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	95% Confidence Limits	
Kps	1	-0.0326	0.00551	-0.0434	-0.0218
Duration	1	-0.00009	0.00913	-0.0180	0.0178
Age	1	-0.00855	0.00930	-0.0268	0.00969
Prioryes	1	0.0723	0.2321	-0.3826	0.5273
Celladeno	1	0.7887	0.3027	0.1955	1.3819
Cellsmall	1	0.4569	0.2663	-0.0650	0.9787
Cellsquamous	1	-0.3996	0.2827	-0.9536	0.1544
Therapytest	1	0.2899	0.2072	-0.1162	0.6961

Output 4.1.4 displays the independent normal prior for the analysis.

Output 4.1.4. Coefficient Prior

Independent Normal Prior for Regression Coefficients		
Parameter	Mean	Precision
Kps	-0.032	2857.143
Duration	0	1E-6
Age	0	1E-6
Prioryes	0	1E-6
Celladeno	0	1E-6
Cellsmall	0	1E-6
Cellsquamous	0	1E-6
Therapytest	0	1E-6

Fit statistics are displayed in Output 4.1.5. These statistics are useful for variable selection.

Output 4.1.5. Fit Statistics

The PHREG Procedure	
Bayesian Analysis	
Fit Statistics	
AIC (smaller is better)	966.359
BIC (smaller is better)	989.175
DIC (smaller is better)	966.418
pD (Effective Number of Parameters)	8.012

Summary statistics of the posterior samples are shown in Output 4.1.6 and Output

4.1.7. These results are quite comparable to the classical results based on maximizing the likelihood as shown in [Output 4.1.3](#), since the prior distribution for the regression coefficients is relatively flat.

Output 4.1.6. Descriptive Statistics

Descriptive Statistics of the Posterior Samples						
Parameter	N	Mean	Standard Deviation	25%	Quantiles 50%	75%
Kps	10000	-0.0326	0.00523	-0.0362	-0.0326	-0.0291
Duration	10000	-0.00159	0.00954	-0.00756	-0.00093	0.00504
Age	10000	-0.00844	0.00928	-0.0147	-0.00839	-0.00220
Prioryes	10000	0.0742	0.2348	-0.0812	0.0737	0.2337
Celladeno	10000	0.7881	0.3065	0.5839	0.7876	0.9933
Cellsmall	10000	0.4639	0.2709	0.2817	0.4581	0.6417
Cellsquamous	10000	-0.4024	0.2862	-0.5927	-0.4025	-0.2106
Therapytest	10000	0.2892	0.2038	0.1528	0.2893	0.4240

Output 4.1.7. Interval Statistics

Interval Statistics of the Posterior Samples					
Parameter	Alpha	Credible Interval		HPD Interval	
Kps	0.050	-0.0429	-0.0222	-0.0433	-0.0226
Duration	0.050	-0.0220	0.0156	-0.0210	0.0164
Age	0.050	-0.0263	0.00963	-0.0265	0.00941
Prioryes	0.050	-0.3936	0.5308	-0.3832	0.5384
Celladeno	0.050	0.1879	1.3920	0.1764	1.3755
Cellsmall	0.050	-0.0571	1.0167	-0.0888	0.9806
Cellsquamous	0.050	-0.9687	0.1635	-0.9641	0.1667
Therapytest	0.050	-0.1083	0.6930	-0.1284	0.6710

With autocorrelations retreating quickly to 0 ([Output 4.1.8](#)) and large effective sample sizes ([Output 4.1.9](#)), both diagnostics indicate a reasonably good mixing of the Markov chain. The trace plots in [Output 4.1.10](#) also confirm the convergence of the Markov chain.

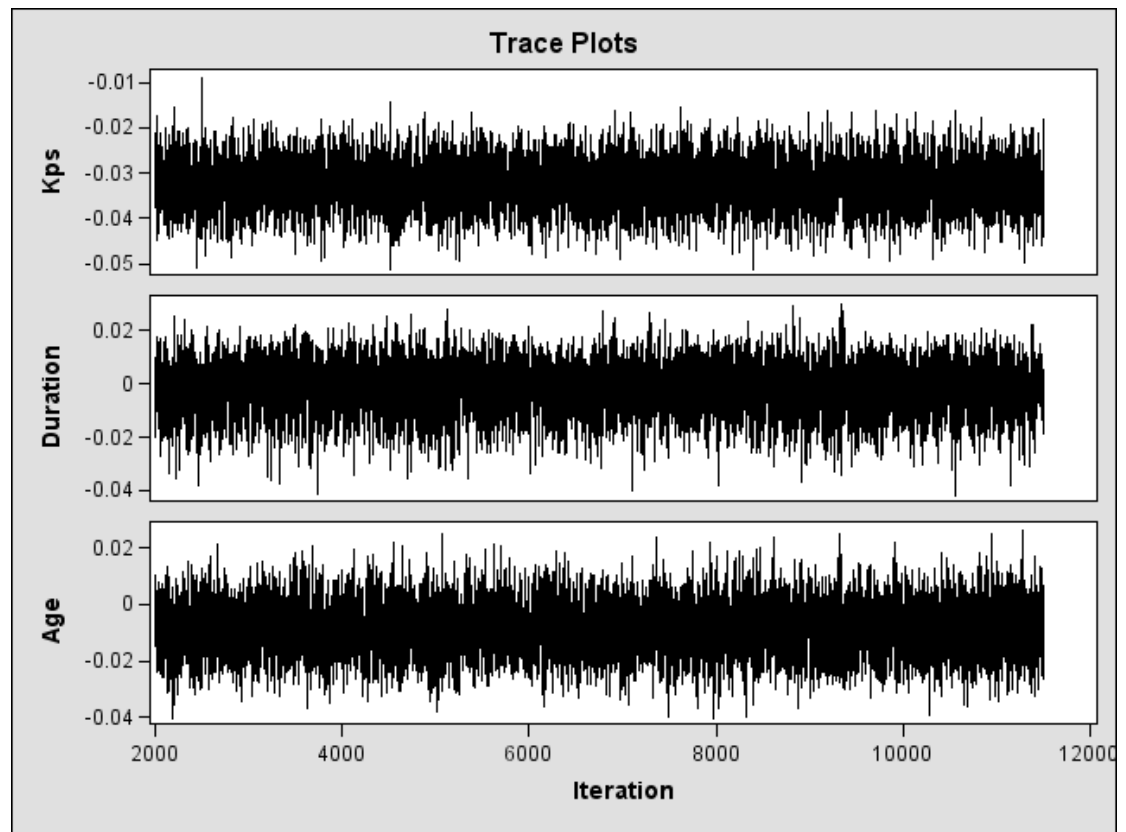
Output 4.1.8. Autocorrelation Diagnostics

Autocorrelations of the Posterior Samples				
Parameter	Lag1	Lag5	Lag10	Lag50
Kps	0.1442	-0.0016	0.0096	-0.0013
Duration	0.2672	-0.0054	-0.0004	-0.0011
Age	0.1374	-0.0044	0.0129	0.0084
Prioryes	0.2507	-0.0271	-0.0012	0.0004
Celladeno	0.4160	0.0265	-0.0062	0.0190
Cellsmall	0.5055	0.0277	-0.0011	0.0271
Cellsquamous	0.3586	0.0252	-0.0044	0.0107
Therapytest	0.2063	0.0199	-0.0047	-0.0166

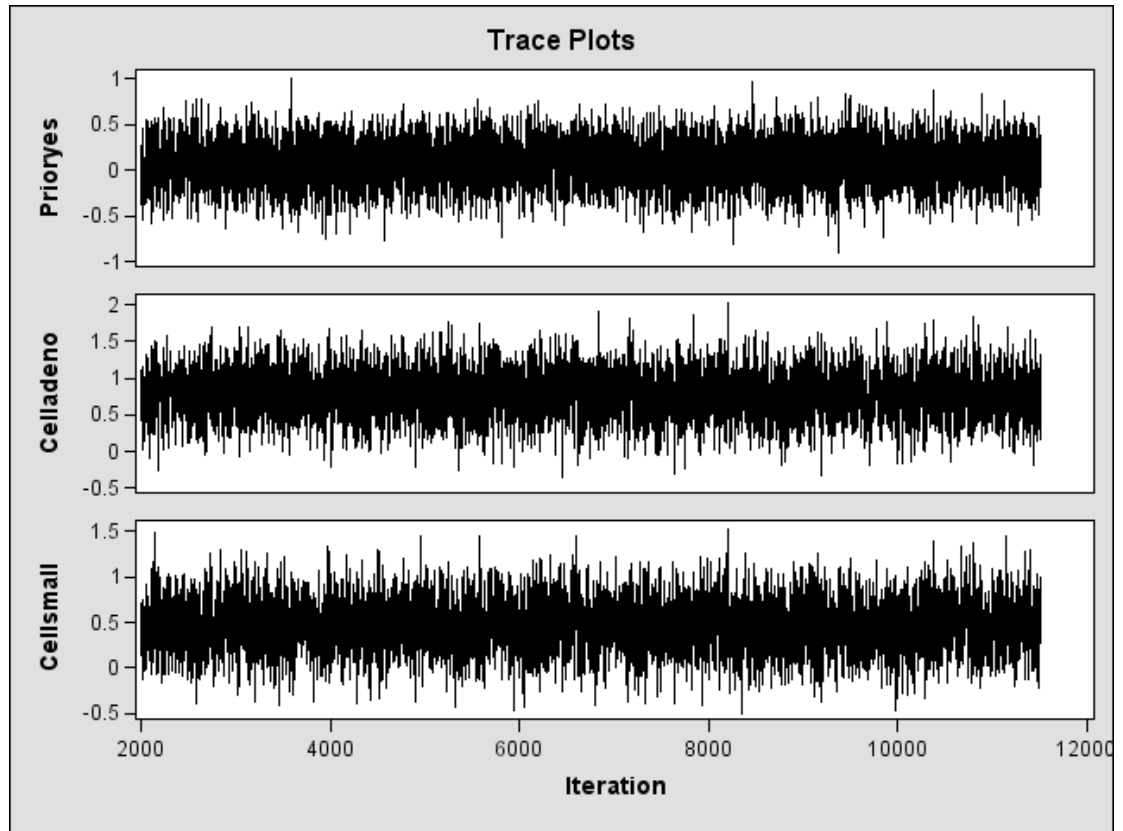
Output 4.1.9. Effective Sample Size Diagnostics

Effective Sample Size			
Parameter	ESS	Correlation Time	Efficiency
Kps	7046.7	1.4191	0.7047
Duration	5790.0	1.7271	0.5790
Age	7426.1	1.3466	0.7426
Prioryes	6102.2	1.6388	0.6102
Celladeno	3673.4	2.7223	0.3673
Cellsmall	3346.4	2.9883	0.3346
Cellsquamous	4052.8	2.4674	0.4053
Therapytest	6870.8	1.4554	0.6871

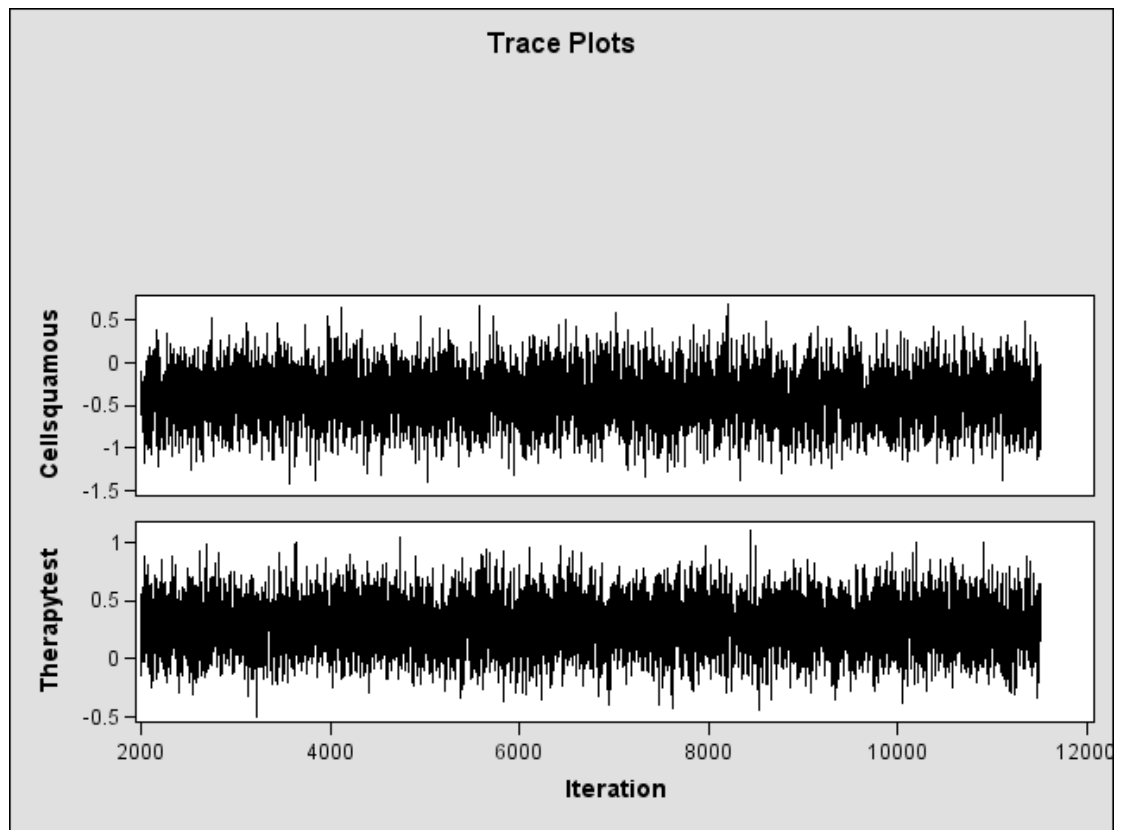
Output 4.1.10. Trace Plots



Output 4.1.10. (continued)



Output 4.1.10. (continued)



The first HAZARDRATIO statement compares the hazards between the standard therapy and the test therapy. Summaries of the posterior distribution of the corresponding hazard ratio are shown in [Output 4.1.11](#). There is a 95% chance that the hazard ratio of standard therapy versus test therapy lies between 0.5 and 1.1.

Output 4.1.11. Hazard Ratio for Treatment

Hazard Ratios for Therapy						
Description	N		Mean	Standard Deviation		
Therapy standard vs test	10000		0.7645	0.1573		
Hazard Ratios for Therapy						
25%	Quantiles		75%	95% Credible Interval		95% HPD Interval
	50%					
0.6544	0.7488	0.8583	0.5001	1.1143	0.4788	1.0805

The second HAZARDRATIO statement assesses the change of hazards for an increase in Age of 10 years. Summaries of the posterior distribution of the corresponding hazard ratio are shown in [Output 4.1.12](#).

Output 4.1.12. Hazard Ratio for Age

Hazard Ratios for Age						
Description	N	Mean	Standard Deviation	25%	Quantiles 50%	75%
Age Unit=10	10000	0.9230	0.0859	0.8635	0.9195	0.9782
Hazard Ratios for Age						
		95% Credible Interval		95% HPD Interval		
		0.7685	1.1011	0.7650	1.0960	

The third HAZARDRATIO statement compares the changes of hazards between two types of cells. For four types of cells, there are six different pairs of cell comparisons. The results are shown in [Output 4.1.13](#).

Output 4.1.13. Hazard Ratios for Cell

Hazard Ratios for Cell						
Description	N	Mean	Standard Deviation			
Cell adeno vs large	10000	2.3048	0.7224			
Cell adeno vs small	10000	1.4377	0.4078			
Cell adeno vs squamous	10000	3.4449	1.0745			
Cell large vs small	10000	0.6521	0.1780			
Cell large vs squamous	10000	1.5579	0.4548			
Cell small vs squamous	10000	2.4728	0.7081			

Hazard Ratios for Cell						
25%	Quantiles		95% Credible		95% HPD Interval	
	50%	75%	Interval			
1.7929	2.1982	2.7000	1.2067	4.0227	1.0053	3.7057
1.1522	1.3841	1.6704	0.7930	2.3999	0.7309	2.2662
2.6789	3.2941	4.0397	1.8067	5.9727	1.6303	5.5946
0.5264	0.6325	0.7545	0.3618	1.0588	0.3331	1.0041
1.2344	1.4955	1.8089	0.8492	2.6346	0.7542	2.4575
1.9620	2.3663	2.8684	1.3789	4.1561	1.2787	3.9263

Example 4.2. Piecewise Exponential Model

This example illustrates using a piecewise exponential model in a Bayesian analysis. Consider the *Rats* data set in the “Getting Started” section on page 116. In the following statements, PROC BPHREG is used to carry out a Bayesian analysis for the piecewise exponential model. In the BAYES statement, the option `PIECEWISE` stipulates a piecewise exponential model, and `PIECEWISE=HAZARD` requests that the constant hazards be modeled in the original scale. By default, eight intervals of constant hazards are used, and the intervals are chosen such that each has roughly the same number of events.

```
proc bphreg data=Rats;
  model Days*Status(0)=Group;
  bayes seed=1 piecewise=hazard;
run;
```

The “Model Information” table in [Output 4.2.1](#) shows that the piecewise exponential model is being used.

Output 4.2.1. Model Information

Model Information		
Data Set	WORK.RATS	
Dependent Variable	days	Days from Exposure to Death
Censoring Variable	status	
Censoring Value(s)	0	
Model	Piecewise Exponential	
Burn-In Size	2000	
MC Sample Size	10000	
Thinning	1	

By default the time axis is partitioned into eight intervals of constant hazard. [Output 4.2.2](#) details the number of events and observations in each interval. Note that the constant hazard parameters are named `Lambda1 ... Lambda8`. You can supply your own partition by using the `INTERVALS=` suboption within the `PIECEWISE=HAZARD` option.

Output 4.2.2. Interval Partition

Bayesian Analysis				
Constant Hazard Time Intervals				
Interval		N	Event	Hazard Parameter
[Lower,	Upper)			
0	176	5	5	Lambda1
176	201.5	5	5	Lambda2
201.5	218	7	5	Lambda3
218	232.5	5	5	Lambda4
232.5	233.5	4	4	Lambda5
233.5	253.5	5	4	Lambda6
253.5	288	4	4	Lambda7
288	Infty	5	4	Lambda8

The model parameters consist of the eight hazard parameters `Lambda1, ..., Lambda8`, and the regression coefficient `Group`. The maximum likelihood estimates are displayed in [Output 4.2.3](#). Again, these estimates are used as the starting values for simulation of the posterior distribution.

Output 4.2.3. Maximum Likelihood Estimates

Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	95% Confidence Limits	
Lambda1	1	0.000953	0.000443	0.000084	0.00182
Lambda2	1	0.00794	0.00371	0.000672	0.0152
Lambda3	1	0.0156	0.00734	0.00120	0.0300
Lambda4	1	0.0236	0.0115	0.00112	0.0461
Lambda5	1	0.3669	0.1959	-0.0172	0.7509
Lambda6	1	0.0276	0.0148	-0.00143	0.0566
Lambda7	1	0.0262	0.0146	-0.00233	0.0548
Lambda8	1	0.0545	0.0310	-0.00626	0.1152
group	1	-0.6223	0.3468	-1.3020	0.0573

Without using the PRIOR= suboption within the PIECEWISE=HAZARD option to specify the prior of the hazard parameters, the default is to use the noninformative and improper prior displayed in [Output 4.2.4](#).

Output 4.2.4. Hazard Prior

Improper Prior for Hazards	
Parameter	Prior
Lambda1	1 / Lambda1
Lambda2	1 / Lambda2
Lambda3	1 / Lambda3
Lambda4	1 / Lambda4
Lambda5	1 / Lambda5
Lambda6	1 / Lambda6
Lambda7	1 / Lambda7
Lambda8	1 / Lambda8

The noninformative uniform prior is used for the regression coefficient Group ([Output 4.2.5](#)), as in the “Getting Started” section on page 116.

Output 4.2.5. Coefficient Prior

Uniform Prior for Regression Coefficients	
Parameter	Prior
group	Constant

Summary statistics for all model parameters are shown in [Output 4.2.6](#) and [Output 4.2.7](#).

Output 4.2.6. Descriptive Statistics

Descriptive Statistics of the Posterior Samples						
Parameter	N	Mean	Standard Deviation	Quantiles		
				25%	50%	75%
Lambda1	10000	0.000945	0.000444	0.000624	0.000876	0.00118
Lambda2	10000	0.00782	0.00363	0.00519	0.00724	0.00979
Lambda3	10000	0.0155	0.00735	0.0102	0.0144	0.0195
Lambda4	10000	0.0236	0.0116	0.0152	0.0217	0.0297
Lambda5	10000	0.3634	0.1965	0.2186	0.3266	0.4685
Lambda6	10000	0.0278	0.0153	0.0166	0.0249	0.0356
Lambda7	10000	0.0265	0.0151	0.0157	0.0236	0.0338
Lambda8	10000	0.0558	0.0323	0.0322	0.0488	0.0721
group	10000	-0.6154	0.3570	-0.8569	-0.6186	-0.3788

Output 4.2.7. Interval Statistics

Interval Statistics of the Posterior Samples					
Parameter	Alpha	Credible Interval		HPD Interval	
		Lower	Upper	Lower	Upper
Lambda1	0.050	0.000289	0.00199	0.000208	0.00182
Lambda2	0.050	0.00247	0.0165	0.00194	0.0152
Lambda3	0.050	0.00484	0.0331	0.00341	0.0301
Lambda4	0.050	0.00699	0.0515	0.00478	0.0462
Lambda5	0.050	0.0906	0.8325	0.0541	0.7469
Lambda6	0.050	0.00676	0.0654	0.00409	0.0580
Lambda7	0.050	0.00614	0.0648	0.00421	0.0569
Lambda8	0.050	0.0132	0.1368	0.00637	0.1207
group	0.050	-1.3190	0.0893	-1.3379	0.0652

The default diagnostics—namely, lag1, lag5, lag10, lag50 autocorrelations (Output 4.2.8, the Geweke diagnostics (Output 4.2.9), and the effective sample size diagnostics (Output 4.2.10)—show a good mixing of the Markov chain.

Output 4.2.8. Autocorrelations

Autocorrelations of the Posterior Samples				
Parameter	Lag1	Lag5	Lag10	Lag50
Lambda1	0.0705	0.0015	0.0017	-0.0076
Lambda2	0.0909	0.0206	-0.0013	-0.0039
Lambda3	0.0861	-0.0072	0.0011	0.0002
Lambda4	0.1447	-0.0023	0.0081	0.0082
Lambda5	0.1086	0.0072	-0.0038	-0.0028
Lambda6	0.1281	0.0049	-0.0036	0.0048
Lambda7	0.1925	-0.0011	0.0094	-0.0011
Lambda8	0.2128	0.0322	-0.0042	-0.0045
group	0.5638	0.0410	-0.0003	-0.0071

Output 4.2.9. Geweke Diagnostics

Geweke Diagnostics		
Parameter	z	Pr > z
Lambda1	-0.0705	0.9438
Lambda2	-0.4936	0.6216
Lambda3	0.5751	0.5652
Lambda4	1.0514	0.2931
Lambda5	0.8910	0.3729
Lambda6	0.2976	0.7660
Lambda7	1.6543	0.0981
Lambda8	0.6686	0.5038
group	-1.2621	0.2069

Output 4.2.10. Effective Sample Size

Effective Sample Size			
Parameter	ESS	Correlation	
		Time	Efficiency
Lambda1	7775.3	1.2861	0.7775
Lambda2	6874.8	1.4546	0.6875
Lambda3	7655.7	1.3062	0.7656
Lambda4	6337.1	1.5780	0.6337
Lambda5	6563.3	1.5236	0.6563
Lambda6	6720.8	1.4879	0.6721
Lambda7	5968.7	1.6754	0.5969
Lambda8	5137.2	1.9466	0.5137
group	2980.4	3.3553	0.2980

References

- Gilks, W. (2003), "Adaptive Metropolis Rejection Sampling (ARMS)," software from MRC Biostatistics Unit, Cambridge, UK, http://www.maths.leeds.ac.uk/~wally.gilks/adaptive.rejection/web_page/Welcome.html.
- Gilks, W., Best, N., and Tan, K. (1995), "Adaptive Rejection Metropolis Sampling with Gibbs Sampling," *Applied Statistics*, 44, 455–472.
- Gilks, W. and Wild, P. (1992), "Adaptive Rejection Sampling for Gibbs Sampling," *Applied Statistics*, 41, 337–348.
- Ibrahim, J., Chen, M., and Sinha, D. (2001), *Bayesian Survival Analysis*, New York: Springer-Verlag.
- Kalbfleisch, J. D. and Prentice, R. L. (1980), *The Statistical Analysis of Failure Time Data*, New York: John Wiley & Sons.
- Kass, R., Carlin, B., Gelman, A., and Neal, R. (1998), "Markov Chain Monte Carlo in Practice: A Roundtable Discussion," *The American Statistician*, 52, 93–100.

Lawless, J. (2003), *Statistical Model and Methods for Lifetime Data*, Second Edition, New York: John Wiley & Sons.

Sinha, D., Ibrahim, J., and Chen, M. (2003), “A Bayesian Justification of Cox’s Partial Likelihood,” *Biometrika*, 90, 629–641.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der Linde, A. (2002), “Bayesian Measures of Model Complexity and Fit (with Discussion),” *Journal of the Royal Statistical Society, Series B*, 64(4), 583–616.

Subject Index

A

- adaptive algorithm
 - adaptive rejection Metropolis sampling (ARMS), 16
 - adaptive rejection sampling (ARS), 16
- advantages and disadvantages of Bayesian analysis, 11
- assessing MCMC convergence, 17
 - autocorrelation, 30
 - effective sample size (ESS), 31
 - Gelman and Rubin diagnostics, 22
 - Geweke diagnostics, 24
 - Heidelberger and Welch diagnostics, 26
 - Raftery and Lewis diagnostics, 27
 - visual inspection, 18
- autocorrelation
 - BPHREG procedure, 142

B

- Bayes' theorem, 4
- BGENMOD procedure
 - displayed output, 65
 - ODS table names, 67
 - parameter estimates, 65
 - standard error, 65
- BLIFEREG procedure
 - displayed output, 102
 - ODS table names, 103
 - parameter estimates, 102
 - standard error, 102
- BPHREG procedure
 - autocorrelation, 142
 - censored values summary, 141
 - coefficient prior, 142
 - correlation matrix, 142
 - credible interval, 142
 - descriptive statistics, 142
 - displayed output, 140
 - effective sample size, 143
 - event values summary, 141
 - fit statistics, 142
 - Gelman and Rubin Diagnostics, 142
 - Geweke Diagnostics, 143
 - initial values, 142
 - interval estimates, 142
 - maximum likelihood estimates, 141
 - model information, 141
 - ODS graph names, 144
 - ODS table names, 143
 - parameter information, 141

- piecewise constant baseline hazard model, 115, 125, 132
- time intervals, 141
- burn-in of MCMC sequences, 16

C

- censored values summary
 - BPHREG procedure, 141
- coefficient prior
 - BPHREG procedure, 142
- correlation matrix
 - BPHREG procedure, 142
- credible interval, 10
 - BPHREG procedure, 142
- equal tail, 10
- highest posterior density (HPD), 10, 33

D

- descriptive statistics
 - BPHREG procedure, 142
- deviance information criterion (DIC), 33
- displayed output
 - BPHREG procedure, 140

E

- effective sample size
 - BPHREG procedure, 143
- effective sample size (ESS), 31
- event values summary
 - BPHREG procedure, 141

F

- fit statistics
 - BPHREG procedure, 142

G

- Gelman and Rubin Diagnostics
 - BPHREG procedure, 142
- Geweke Diagnostics
 - BPHREG procedure, 143
- Gibbs sampler, 15

H

- Heidelberger and Welch diagnostics
 - T, 143
- hypothesis testing, 10

I

initial values
 BPHREG procedure, 142

interval estimates
 BPHREG procedure, 142

interval estimation, 10

L

likelihood, 4

likelihood principle, 11

M

marginal distribution, 4

Markov chain Monte Carlo (MCMC), 12

- adaptive rejection Metropolis sampling (ARMS), 16
- adaptive rejection sampling (ARS), 16
- assessing convergence, 17
- burn-in, 16
- Gibbs sampler, 13, 15
- Metropolis algorithm, 13
- Metropolis-Hastings algorithm, 13
- sample size, 16
- summary statistics, 31
- thinning, 16

maximum likelihood estimates
 BPHREG procedure, 141

Metropolis algorithm, 13

Metropolis-Hastings algorithm, 13, 15

model information
 BPHREG procedure, 141

Monte Carlo standard error (MCSE), 10, 32

N

normalizing constant, 4

O

ODS graph names
 BPHREG procedure, 144

ODS table names
 BPHREG procedure, 143

P

parameter estimates
 BGENMOD procedure, 65
 BLIFEREG procedure, 102

parameter information
 BPHREG procedure, 141

PHREG procedure, 115

piecewise constant baseline hazard model
 BPHREG procedure, 115, 125, 132

piecewise exponential model,
 See piecewise constant baseline hazard model

point estimation, 9

posterior distribution, 4
 improper, 6, 7

posterior summary statistics, 31
 correlation, 32
 covariance, 32

equal-tail interval, 32

highest posterior density (HPD), 33

mean, 31

median, 9

mode, 9

Monte Carlo standard error (MCSE), 10, 32

percentile, 32

standard deviation, 31

standard error of the mean estimate, 32

prior distribution, 4, 5

- conjugate, 8
- flat, 6
- improper, 7
- informative, 8
- Jeffreys', 8
- noninformative, 6, 8
- objective, 5
- subjective, 5
- vague, 6

probability

- Bayesian, 3
- frequency, 3
- interpretation, 11

R

Raftery and Lewis diagnostics
 T, 143

S

spectral density estimate at zero frequency, 25

standard error
 BGENMOD procedure, 65
 BLIFEREG procedure, 102

T

T
 Heidelberg and Welch diagnostics, 143
 Raftery and Lewis diagnostics, 143

thinning of MCMC sequences, 16

time intervals
 BPHREG procedure, 141

TPHREG procedure, 115

Syntax Index

A

ALPHA= option
 HAZARDRATIO statement(BPHREG), 131
AT= option
 HAZARDRATIO statement(BPHREG), 131

B

BAYES statement
 BPHREG procedure, 122
BPHREG PROCEDURE, BAYES statement, 122
BPHREG procedure, BAYES statement
 COEFFPRIOR= option, 122
 DIAGNOSTIC= option, 123
 INITIAL= option, 125
 NBI= option, 125
 NMC= option, 125
 PIECEWISE= option, 125
 PLOTS= option, 127
 SEED= option, 130
 SUMMARY= option, 130
 THINNING= option, 130
BPHREG procedure, HAZARDRATIO statement,
 130
 ALPHA= option, 131
 AT= option, 131
 DIFF= option, 131
 E= option, 132
 UNITS= option, 131

C

COEFFPRIOR= option
 BAYES statement(BPHREG), 122

D

DIAGNOSTIC= option
 BAYES statement(BPHREG), 123
DIFF= option
 HAZARDRATIO statement(BPHREG), 131

E

E= option
 HAZARDRATIO statement(BPHREG), 132

H

HAZARDRATIO statement
 BPHREG procedure, 130

I

INITIAL= option
 BAYES statement(BPHREG), 125

N

NBI= option
 BAYES statement(BPHREG), 125
NMC= option
 BAYES statement(BPHREG), 125

P

PIECEWISE= option
 BAYES statement(BPHREG), 125
PLOTS= option
 BAYES statement(BPHREG), 127

S

SEED= option
 BAYES statement(BPHREG), 130
SUMMARY= option
 BAYES statement(BPHREG), 130

T

THINNING= option
 BAYES statement(BPHREG), 130

U

UNITS= option
 HAZARDRATIO statement(BPHREG), 131

